

AUTOMATIC DECEIT DETECTION THROUGH MULTIMODAL ANALYSIS OF SPEECH VIDEOS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Berat Biçer
September 2022

Automatic Deceit Detection Through Multimodal Analysis of Speech
Videos

By Berat Biçer

September 2022

We certify that we have read this thesis and that in our opinion it is fully adequate,
in scope and in quality, as a thesis for the degree of Master of Science.

Hamdi Dibeklioglu(Advisor)

Ayşegül Dünder

Albert Ali Salah

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

AUTOMATIC DECEIT DETECTION THROUGH MULTIMODAL ANALYSIS OF SPEECH VIDEOS

Berat Biçer

M.S. in Computer Engineering

Advisor: Hamdi Dibeklioglu

September 2022

In this study we propose the use of self-attention for spatial representation learning, while explore transformers as the backbone of our speech model for inferring apparent deceptive intent based on multimodal analysis of speech videos. The proposed model applies separate modality-specific representation learning from visual, vocal, and speech modality representations and applies fusion afterwards to merge information channels. We test our method on the popular, high-stake Real-Life Trial (RLT) dataset. We also introduce a novel, low-stake, in-the-wild dataset named *PoliDB* for deceit detection; and report the first results on this dataset as well. Experiments suggest the proposed design surpasses previous studies performed on RLT dataset, while it achieves significant classification performance on the proposed PoliDB dataset. Following our analysis, we report (1) convolutional self-attention successfully achieves joint representation learning and attention computation with up to three times less number of parameters than its competitors, (2) apparent deceptive intent is a continuous function of time that can fluctuate throughout the videos, and (3) studying particular abnormal behaviors out of context can be an unreliable way to predict deceptive intent.

Keywords: Automatic deceit detection, behavioral analysis, affective computing, multimodal data analysis, deep learning.

ÖZET

KONUŞMA VİDEOLARININ ÇOK-KİPLİ ANALİZİYLE OTOMATİK ALDATMA TESPİTİ

Berat Biçer
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Hamdi Dibeklioglu
September 2022

Bu çalışmada, evrimsel özdeğer birimlerinin uzamsal gösterim öğrenmede kullanılmasını ve transformatör ağlarının konuşma modellerinin omurgasını oluşturmasını öneriyoruz. Tasarladığımız derin ağ görsel, ses, ve konuşma kipi-bazlı derin ağ eğitimi ve gösterim öğrenimi yaptıktan sonra elde ettiği kip gösterimlerini füzyona uğratarak bu kanallardan gelen bilgiyi birleştirmektedir. Deneysel çalışmaların sonucunda, tasarlanan derin ağ ile literatürde popüler *Real-Life Trial (RLT)* veri tabanı üzerinde yapılan tüm çalışmaları geçen bir performans elde edilmiştir. Dahası, önerdiğimiz model, yeni topladığımız çok-kipli, kontrolsüz şartlarda düşük riskli senaryolar içeren ve doğruluk araştırmalarıyla etikenlenmiş *PoliDB* isimli veri tabanı üzerinde de yüksek sınıflandırma performansı göstermiştir. Yaptığımız analiz sonunda (1) evrimsel özdeğer birimlerinin aldatma tespiti için anlamlı gösterimleri, diğer alternatiflere göre üç kata kadar daha az parametre kullanarak elde ettiğini gösteriyor; (2) izlenimsel aldatma niyetinin sürekli bir fonksiyon olduğunu ve videolar süresince her zaman biriminde aynı değerde yada miktarda aldatıcı davranış görülmediğini gösteriyor; (3) elde ettiğimiz bulgular, gözlenen davranışların birbirinden bağımsız olarak incelenmesinin aldatıcı niyeti tespit etmek için güvenilir bir bilgi kaynağı olmadığını işaret ediyor.

Anahtar sözcükler: Otomatik aldatma tespiti, davranışsal analiz, duyuşsal analiz, çok-kipli veri analizi, derin öğrenme.

Acknowledgement

The author is supported by TUBITAK BİDEB 2211 Scholarship Program, from 2019 to 2021. We thank Bahadır Durmaz and Serhat Aras for their contributions in data collection for PoliDB.



Contents

1	Introduction	1
2	Literature Review	4
2.1	Theory of Deception	4
2.2	Deceit Detection in Criminal Psychology	5
2.2.1	Deceptive Cues	6
2.2.2	Why Lies Fail?	7
2.2.3	Practical Deceit Detection for Law Enforcement	9
2.3	Political Misinformation and Deception	10
2.3.1	Roots of Political Misinformation	10
2.3.2	Fact-Checking Journalism	11
2.4	Multimodal Deceit Detection	12
2.5	Attention, Self-Attention and Transformers	13
2.6	An Alternate Strategy for Feature Importance Assessment	15

3	Methodology	17
3.1	Face Tracking and Normalization	19
3.2	Voice Representation	22
3.3	Speech Representation	23
3.4	Visual Model	23
3.4.1	Convolutional Self-Attention	25
3.4.2	Design Details	28
3.5	Voice Model	30
3.5.1	Speech Model	30
3.6	Modality Fusion	32
4	Datasets	33
4.1	Real-Life Trial Dataset	33
4.2	PoliDB	34
4.2.1	Politifact and Its Fact-Check Procedure	35
4.2.2	Statement Acquisition From Politifact	36
4.2.3	Data Acquisition Procedure	38
4.2.4	Database Details	40
5	Experiments & Results	45

5.1	Experimental Setup	45
5.1.1	Experimental Setup for RLT Dataset	45
5.1.2	Experimental Setup for PoliDB	46
5.2	Results	47
5.2.1	Results on RLT Dataset	47
5.2.2	Results on PoliDB	57
5.3	Error Analysis	58
5.3.1	Sources of Visual Error	58
5.3.2	The Human Factor	62
5.3.3	Curse of Dimensionality and Generalizability	62
5.3.4	Interpretability	63
5.3.5	Representation	63
5.3.6	Explainability and Real-Life Usage	64
6	Conclusion and Future Work	67

List of Figures

3.1	Overview of the proposed method.	18
3.2	Example case of a crowded scene.	19
3.3	Example histogram equalization output.	20
3.4	Example spectrograms before and after vocal separation.	22
3.5	Overview of visual model.	24
3.6	Visualization of convolutional self-attention.	25
3.7	Visualization of the speech model.	31
4.1	Sample screenshot of a Politifact article.	37
4.2	Sample screenshots from PoliDB.	40
4.3	Sample screenshots from PoliDB, continued.	42
4.4	Sample screenshots from PoliDB, continued.	44
5.1	Example self-attention response overlays for RLT dataset.	51
5.2	Example self-attention response overlays for RLT dataset, continued.	52

5.4	Deceit level predictions throughout selected videos from RLT dataset.	55
5.5	Deceit plots, continued.	56
5.6	Example extreme head orientation.	59
5.7	Example blur caused by rapid camera motion.	59
5.8	Example problematic viewpoint.	60
5.9	Example complex background.	60
5.10	Example visual artifacts and glare effect.	61
5.11	Example negative effect of visual quality on visual representation learning.	61
5.3	Attention overlays and action unit (AU) correspondence.	66

List of Tables

3.1	Comparison of convolutional self-attention and convolutions with respect to kernel spatial extend.	27
4.1	Statistics of PoliDB.	43
5.1	Optimization algorithms and hyperparameters considered for cross validation.	46
5.2	Comparison of attention proposal performance.	47
5.3	Effect of employed language backbone on speech model performance	48
5.4	Correct classification rates (CCR) calculated for each fusion experiment performed on RLT dataset.	49
5.5	Comparison to other methods on RLT.	49
5.6	Correct classification rates (CCR) calculated for each fusion experiment performed on PoliDB dataset.	58

Chapter 1

Introduction

Deception occurs commonly throughout our daily lives and is an integral part of social life. Whether for personal gain or for avoiding punishment, deception is used as a social tool so frequently that deceptive behaviors arise even in daily conversations between relational partners [1] and as much as a quarter of all conversations [2]. Naturally, this engenders some negative consequences as well. For example, patients report that they feel dissatisfied when their doctors are unclear or vague in answering their healthcare related questions [3]. Findings suggest romantic relationships are negatively affected by partners showing dishonest behavior such as indirect refusal or vague responses [4]. Despite deception's significant presence and role in social interactions, ordinary individuals can hardly identify deceptive intent since deceptive cues aren't universal and has no obvious, singular tell-tale signs. For this reason, statistics show that an ordinary individual is barely as good as random chance at correctly identifying deceptive intent occurring during a social interaction [5]. Research suggests that only extensive training at capturing deceptive indicators, which vary significantly due to interpersonal and circumstantial reasons, reportedly improves the rate of capturing deceptive behavior for law enforcement and military personnel, resulting in a significant surpass over this margin [6]. Moreover, theories from popular culture and literature fuel the misconception that deception has obvious, tell-tale signs around deception by their many false claims and statements [5], which tends to

attract many people. Studies which attempted to test the validity of these claims obtained from the literature and pop-culture show that people display various behaviors as they attempt a deceitful act, which supports the argument that no singular tell-tale sign of deceit exists. Even more interestingly, researchers show that these stereotypical signs can also be observed in truthful individuals in high stress circumstances [6]. Therefore, it's no surprising that most scientists believe deceit detection is an inexact science [5].

These difficulties in regards to deception's nature motivates research into automatically inferring deceptive intent, which is based on capturing deceptive behavioral cues. Polygraphs, or as incorrectly known as *lie detectors*, are amongst the first attempts towards this goal. These devices rely on localizing abnormal physiological measurements to infer deception, however, findings suggest individuals can easily confuse these readings which renders polygraphs unreliable [5]. Based on polygraphs, one can infer that changes in an individual's physiology is an unreliable source of information; however, there's also evidence to suggest some nonverbal behaviors occur outside of subject's conscious control and cannot be inhibited. Vocal characteristics, facial muscle movements are examples of this phenomenon. Based on these findings, Ekman suggests that capturing such occurrences form the basis of nonverbal detection of deceit [7], which has been the norm in nonverbal deceit detection literature since.

Inspired by Ekman's findings, nonverbal deceit detection research attempts to localize involuntary and instantaneous behavioral patterns outside of conscious control for inferring deceptive intent. However, doing so usually requires comparing the candidate behavioral pattern with the subject's truthful baseline profile [8], which requires advanced computer vision techniques capable of modelling temporal dependencies for multimodal data. Despite this, our observations suggest that focus on comparison of a subject's truthful behavioral baseline to their deceptive state is severely lacking in affective computing literature. This problem is partially engendered by a lack of high-quality, easily accessible, rich databases; forcing researchers to employ legacy tools and traditional machine learning techniques and limiting their ability to learn adequate representations. Moreover,

contemporary research shows that state of the art deep learning techniques require lots of data for successful training; which limits their applicability to deceit detection task due to sparsely-occurring nature of high-quality videos capturing deceitful behavior, despite how common deception is in our daily lives. Lastly, while there’s evidence speech signals show deceptive characteristics and can be analysed similar to visual data utilizing computer vision-based approaches, this has not yet been demonstrated clearly for deception task.

In summary, we propose the use of convolutional self-attention layers that computes attention on a filter-basis, producing attention responses directly as image features instead of applying feature extraction and attention in sequence. We also propose utilizing a transformer as a transfer-learning backbone and vectorization algorithm which allows the language model to exploit characteristics of a wider language task for successful inference of deceitful intent, reducing the negative effects of data sparsity. The proposed architecture is a multimodal data analysis framework that employs these design choices and is shown to surpass the state-of-the-art on the popular high-stake Real-Life Trial (RLT) dataset. Moreover, we propose a novel, low-stake, in-the-wild dataset annotated through fact-checks named *PoliDB* to address the data sparsity and annotation bias issues in deceit detection; and report the performance of the proposed model on this dataset as well. Following experimental studies, we show that convolutional self-attention responses captures facial dynamics that, according to psychological studies, correlate with deceptive intent. We also show our language model captures characteristics of deceptive and truthful language patterns similarly validated by psychological research. Our analysis suggests that deceptive intent is a continuous, heterogeneous function of time: A subject may not display deceptive behaviors at all times during a clip and deceptive patterns can be localized to specific time intervals. We also conclude that deceptive cues are only meaningful within the context of the conversation: Singular instances of abnormal behavior, not unlike how the pop-culture claims the existence of false tell-tale signs of deception, can be unreliable indicators of deceit.

Chapter 2

Literature Review

2.1 Theory of Deception

Deception, despite having been studied extensively, lack a unified theory to explain its related phenomena. One of these major theories is *Interpersonal Deception Theory* (IDT). IDT attempts to explain deception in the context of an interpersonal relationship between a sender and a receiver. IDT proposes 18 individually verifiable propositions for its deception model explaining roles of the context, the sender, and the listener in conversations. While IDT provides a comprehensive description of deceptive interactions and how they can be detected, its propositions lack the interconnectivity and predictive power of a complete, unifying theory [9].

Next, *Information Manipulation Theory* (IMT) explains how a sender constructs a message that will deceive the receiver. IMT is more concerned with the content of the messages, the context of the relationship and the message itself, and the social consequences precipitated by the outcome of the interaction between the sender and the receiver. IMT argues that the degree of deceptiveness of a message can be determined based on conversational maxims of Paul Grice: quantity, quality, relevance, and equivocality. IMT provides explanations for how

deception can occur, yet it does not provide predictions regarding which maxims will be violated [9]. Also, even though the maxims themselves mostly have objective definitions, their violations are subjectively quantified which is problematic for describing the violations [9].

Lastly, *Truth-Default Theory* (TDT) suggests that by default, people are inclined to assume others are truthful because either they don't consider deception a possibility or there is insufficient evidence pointing towards deception. This is named the truth bias. Truth bias commonly occurs between people who are members of the same community which would imply they hold strong presumptions that members are truthful towards each other and makes them susceptible to deception [10, 11]. As a consequence of truth bias and the diversity of interpersonal signals of deceit, it's argued that people can detect truths better than lies, which is referred to as the veracity effect [10, 12, 13].

2.2 Deceit Detection in Criminal Psychology

It is a common misconception that there are specific tell-tale signs to detecting deceptive behavior, and this is fueled by many false statements commonly seen in the literature [14, 15, 6, 5]. For example, in spite of claims made by proponents of neuro-linguistic programming and contrary to the widespread belief, eye movements are shown to be unreliable indicators of deceit [16]. People's responses in similar studies performed to test the validity of these so-called tell-tale signs of deceit vary based on the kind of lie, preparation, interviewer's strategy, and the subject's confidence [17, 6]. Moreover, some signs appear more commonly amongst liars than truth tellers despite no such universal identifiers are detected [18, 6]. This is complicated even further by observations suggesting the same individual may display different behavior depending on the context [17, 6] and individuals differ from one another in expressing said behavioral signs [17]. Interestingly, it is also true that innocent individuals can also demonstrate many of the stereotypical signs of deceptive behavior due to high stress induced by the interviews/interrogations [19, 6]. Hence, some scientists believe deceit detection

is an inexact science [5] and, it's hardly surprising that average person is no better than random chance at detecting deceitful behavior. In fact, a study shows that only U.S. Secret Service agents exceeded this margin amongst the participants [20, 6]. Others support this observation and even suggest that law enforcement officials become progressively more confident in their ability to detect deception over the course of the experiment despite no such progress is observed [21, 6].

2.2.1 Deceptive Cues

Despite these difficulties, many studies establish that there are two major types of deceptive indicators: verbal and nonverbal ones. The prominent actors in nonverbal indicators are facial expressions of emotion known as microexpressions which are fleeting expressions of concealed emotion that are involuntary and has strong correlation with deceptive behavior [7, 8]. Due to their near-instantaneous occurrence, specialized training is required to accurately detect them in real time. For example, FBI trainees can increase their recognition rate of such behaviors to as high as 90 percent, 70 percent on average. Other populations show similar results of detection accuracy reaching as high as 80 percent. These studies report the officers in question doubled their ability to read microexpressions in real world and in real time. It's further noted that these officers retained the ability to recognize microexpressions even weeks after the training [22, 8]. It is also noteworthy that emotional facial expressions including macro-, micro-, and subtle expressions are universal, immediate, automatic, and unconscious reactions.

Second major deceptive indicator is the broadly categorized verbal indicators. Verbal indicators come from statement analysis that applies grammatical rules of the language to the information provided. Statement analysis consists of including but not limited to: detecting verbs describing communication and uncompleted action; changes in verb tense; minimizing, intensifying, and editing adverbs; extraneous information; unique sensory details; and statement structure, which identifies the person's focus—on the incident or somewhere else. Previous research shows that these categories can differentiate deceitful individuals from

truth-tellers at statistically significant levels [23, 8, 5] and some are shown to be reliable even across cultures and ethnicities [24, 25].

2.2.2 Why Lies Fail?

Experts believe there are three major reasons why lies fail. First one is the emotional approach, which suggests the lies fail because of the difficulty of concealing or falsifying emotions in social circumstances [18, 6]. Emotional approach argues there are two broad categories of emotional cues, namely lying about the feeling and the feelings about lying [17, 6]. While people lie about their feelings in the first instance, in the second instance the act of lying generates strong negative emotions such as guilt, anxiety, and fear independent from the individual's innocence. It is argued that not all lies involve emotions, yet those that do are distinctly susceptible to social interactions as people don't choose which emotions to feel and their intensities [26, 6]. This phenomenon can be observed more clearly under high-stakes cases such as interrogations: when the stakes are low, the subjects are less likely to show emotional activation and corresponding behavioral cues [18, 6, 27]. It is argued that fear, guilt, and delight are three emotions which are closely associated with deceit [18]. The level of fear directly correlates to one's confidence in deceiving others as some are more experienced deceivers, and these individuals are rarely detected [28, 6]. Guilt refers to one's feeling regarding deception and can be about the content of the lie or the act of lying itself. Studies suggest people with strict beliefs about dishonesty and those perceiving it as a sin are more susceptible to deception guilt [18, 6]. Delight, or 'duping delight' is the positive feeling a liar experiences when anticipating the challenge of a lie or during the act when its success isn't assured [18, 6]. These people can display contempt, enjoyment, or excitement and are known to confess their crimes to close associates, family members, and even investigators for approval and acknowledgement of their intellect [6]. However, emotional expressions don't reveal their source; hence, law enforcement experts seek situations where the perceived emotional response doesn't match the context of the conversation when probing for deceptive behavior [6, 29, 30].

In cognitive approach, it's argued that lying persuasively is difficult and requires great mental effort for formulating a plausible story, monitoring body language and emotional expressions, anticipating further questions, keeping track of the interviewer's responses, remembering the content of the construct, etc. There is evidence that this mental effort is subjective, yet liars are more likely to exhibit certain behaviors such as speech errors, pauses and waits between answers, increased pupil size [17, 31, 6, 5], indicating major cognitive overload. Alas, these cues are weak indicators of deception: a study shows that officers who relied on verbal cues such as vague responses distinguished deceptive behavior better than those who relied on visual signs [17, 32, 33]. Thus, interviewers attempt to gather as much information as possible through conversations instead, and studies suggest information gathering and accusatory interviewing styles are suitable for this purpose [34]. In accusatory approach, which is heavily used in Reid Technique [35, 30], allegations are utilized to construct a stressful environment for the suspect and disrupt the cognitive process for deception. Minimization, maximization, and other techniques are heavily utilized in this kind of interrogation to steer the suspect into a confession [30]; however, there's evidence that Reid Technique can lead to false confessions caused by three consecutive errors: first, the investigator misclassifies the innocent person as guilty, the suspect then is lead to a guilt-trip which frequently lies about the evidence, and then the suspect is forced a narrative that's jointly shaped by the police that often supplies facts about the crime [36, 30]. This is a major shortcoming of the Reid Method, and the evidence points out that it rarely succeeds in obtaining correct and accurate confessions [37]. On the contrary, the information gathering is heavily utilized in the PEACE Model, the Cognitive approach, and the Kinesic Method. These interrogation techniques seek to gather relevant facts as opposed to seeking confessions, though they follow different approaches [30]. Despite some promising improvements, each has their own shortcomings that impacts their practical use [38, 39].

Attempted control approach argues that most liars are aware of observers paying close attention to behavioral cues, so they spend extra effort attempting to keep their responses and body language in check, going to great lengths as to not

display any stereotypical deceptive behaviors such as avoiding eye contact, postural shifts, etc [40, 6]. One strategy criminals prefer to limit exposing information is to hide the deceptive content amongst a set of truths, which greatly diminishes the mental load required to sustain the lie [17]. However, it's impossible to control or disguise every reaction as some physiological changes accompanied by strong emotional reactions are beyond conscious control [18, 6]; which, frequently, results in the subject displaying inconsistent behavior with their statement.

2.2.3 Practical Deceit Detection for Law Enforcement

For practical purposes, law enforcement officials are suggested to follow a set of guidelines supported by psychological findings. It is strongly recommended that, during an interview, instead of focusing a single behavior investigators should first construct a behavioral outline for the subject, gauging their responses and expressions which can be used for comparison with the anomalous behavior and expressions. This is a crucial step in deceit detection as people's speech and behavior can vary drastically [41, 6]. It is advised that the interviewer enables the subject to relax by conversational methods such as rapport building and other dyadic strategies to create psychological comfort, ensuring an optimal baseline can be created [42, 43]. It's also crucial for the interviewer to ensure that this comparison is made under identical circumstances and about the same topics, while the interviews compared should have been occurred within a short time span [17, 44]. Next, the context of the subject's behavior is another factor for consideration, as behavioral anomalies at the beginning of the interview are likely of little value compared to those occurring post-query [18, 6].

Expert also argue that an effective communication is a key factor for a successful interrogation: The interrogator must formulate the questions carefully as to not invoke altered nonverbal behavior masking the true sentiment. For example curiosity as the theme of the query is advantageous compared to suspicion, which engenders tension and stress in the suspect as a cause of the content of the question and the suspect's guilt rather than the tone of the question or the suspect's

emotions. [42, 30]. Some investigators argue there are four viable opportunities during an interrogation to detect irregular behavior that could indicate deceptive intent. These occur when the query is directed to the suspect, when the suspect processes the question, when the suspect answers the question, and finally during the pregnant pause which can be created by the investigator post-reply [44]. Lastly, and perhaps the most importantly, officials argue that there is no singular indicator of deception. Rather, accurate assessment should be based on how nonverbal cues change over time from a person’s baseline and how they combine with the individual’s statement [8, 45, 46, 47].

2.3 Political Misinformation and Deception

2.3.1 Roots of Political Misinformation

One of the earliest studies on political misinformation defines misinformation as faulty information held firmly by individuals. In this study, despite the erroneous nature of their beliefs, the subjects displayed high confidence that their views were correct [48, 49]. When large portions of the population hold these beliefs, it’s argued shared misconceptions can influence the collective stance of the public, undermining the argument that individual ‘errors’ are cancelled out in the aggregate [48, 49, 50].

Political information processing is influenced by truth bias as a result of political affiliation [51] where people tend to process political messages with respect to their political affiliation and its strength [52, 53]. Some believe initial affective response of the human mind biases subsequent information processing in the direction of the this affect [54, 49] which can explain the truth bias. However, there’s evidence the reasoning process is altered by existence of new information [55, 56, 57, 58, 59, 60, 49] which suggests people’s priorities in decision-making vary based on context.

Politicians also contribute to the political misinformation process through their

dishonest behaviors. In the US, politicians are considered among the most deceptive people [61, 62]. The public thinks politicians do not answer questions [63, 64], displaying over refusal or off-topic responses frequently [65]. There’s even evidence that political misinformation can be induced by third parties such as 2016 U.S. Presidential Elections and the Brexit referendum [66].

2.3.2 Fact-Checking Journalism

Fact-checking describes an emerging genre of journalism which reports on the accuracy of a claim or a text from the media [67]. It entails both fighting against misinformation through verification of public statements and potentially confronting the authors about the veracity of their claims. Amongst the earliest practitioners in U.S. are FactCheck.org, *the Washington Post*’s Fact Checker, and Politifact; and since then, this has spread to many countries around the globe [68, 67].

In general, methodology of fact-checks embodies a scientific ethic that any observer should be able to reproduce the analysis and the verdicts [67]. However, these methodologies vary with respect to employing/quoting outside experts or using rating systems indicating a scale of accuracy of the claims [69, 70, 67], etc. It’s also important to note that fact-checking organizations aren’t always consistent in reaching the same conclusions. Studies display a lack of agreement in regards to verdicts and reconciling different rating systems amongst fact-checkers [71, 72] which can impact the trust towards fact-checks.

Researchers also study whether or under which conditions fact-checking is effective. For example, fact-checks can be useful tool in educating the public [73, 74]; helping people understand unfamiliar or fictitious political candidates [75, 76], business people or companies [77]; and emerging health issues [78]; correcting public-officials’ behavior and policies [69, 79, 80, 81]. Alas, there’s also evidence that fact-checks may not be effective in certain circumstances such as in public education [82, 83] and addressing confirmation bias [84, 85].

2.4 Multimodal Deceit Detection

Psychological foundations of deceit has enabled many computer scientists to study deceit detection as a data-driven problem utilizing a variety of modalities such as facial expressions, body posture, and text features extracted from conversation transcripts. The general approach in deceit detection is to extract features in a multimodal fashion, then applying fusion at the end.

From a visual standpoint, initial attempts at automatic deceit detection (ADD) preferred traditional image features such as the local binary patterns [86] shown to be a powerful solution for many vision tasks. As the research in computer vision advanced, off-the-shelf components such as OpenFace [87, 88, 89] replaced visual feature extraction which offers pretrained deep learning models and complex algorithms implemented for many visual tasks such as facial landmark detection, action unit (AU) predictions, facial orientation, etc. OpenFace is capable of providing such information at significant confidence levels, so it's no surprising it's become so very widely-used in the literature [90, 91]. While rare, there have also been attempts to employ transfer learning from various domains such as 3D object classification as well [92], however, research focusing on transfer learning from similar affective / behavioral tasks into deceit detection has not yet been sufficiently studied.

From a language-based perspective, language models that offer insight into the psychology of the language use such as LIWC [93] are quite popular and have been used extensively [94, 90, 91]. Specifically, LIWC offers dictionaries which reflect a certain psychological aspect such as emotions or social concerns. These dictionaries can be used to calculate percentage agreement of a given statement with the words existing within a dictionary of some psychological property. These percentages can then be transformed into feature representations by the researchers. N-gram based language models are another candidate for language models which are simplistic and easy to use [94, 90, 91]. With the advancements of neural networks and introduction of *word2vec* [95, 96] algorithm, a shift towards neural network-based language models can be observed in the literature.

Nowadays, prominent actors in language tasks are the transformers, however, these architectures have not yet been studied extensively in the context of deceit detection.

For speech analysis, various tools are developed for extracting vocal characteristics associated with deceptive intent. Some of these are STRAIGHT [97], voice activity detection (VAD) [98], OpenSmile [99], etc. Specifically, Mel-Frequency Cepstral Coefficients (MFCC) [100, 101] is a powerful tool in describing the characteristics of a speech signal. This property of the MFCC features have naturally attracted many researchers [102, 103].

Lastly, despite the evidence suggesting there’s a correlation between deceptive intent and physiological changes, other involuntary behavioral responses, etc. there’ve been a limited number of studies investigating the feasibility of these modalities. EEG data [102], gaze information [102], and bodily movements [104] are amongst modalities that have been studied to some extent. However, to our knowledge, no study is yet to provide a comprehensive view of biological and behavioral modalities on deceit detection.

2.5 Attention, Self-Attention and Transformers

Attention is a method employed in deep learning that mimics human cognition. In computer vision tasks specifically, attention performs a variety of functions such as determining relevant input positions [105] and channel contributions [106], capturing temporal dependencies [107], etc. In attention, every input sample v_i is assigned a scalar weight w_i which determines the contribution of sample v_i to the learning goal $\hat{y}$. In general, a simple attention framework can be expressed as follows:

$$\hat{y} = \sum_i w_i v_i \text{ where } \sum_i w_i = 1 \text{ and } \forall w_i \in [0, 1]. \quad (2.1)$$

By assigning attention weights to input samples, much like how human attention focuses on bits of information deemed significant, the model assigns a higher priority to samples that are deemed more influential to the learning goal as indicated by the loss function. In other words an attention filter boosts the activation of segments of the input feature map if there's high correlation with the learning goal, and shrinks the activations otherwise.

Generally speaking, there's no singular way to compute attention weights. A common practice is to set weights w_i as probabilities, so attention weights are usually computed by a dot product followed by Softmax function for converting scalars into probabilities. However, it's also possible to model attention as a tensor or a series of linear algebraic operations that output a tensor. For this reason, neural networks are also a popular choice for learning attention filters.

Self-attention, which is an attention variation, is used to construct a representation of a sequence by mapping relations between its parts and positions. First developed for language tasks [108], with the introduction of transformers in [109] self-attention has gained popularity in the literature. Here, the encoded input representation is viewed as $(\mathbf{Key}, \mathbf{Value})$ pairs of length n . Next, the previous output is viewed as a *Query* and the next output is obtained by mapping the query to the pairs above. In this context, self-attention can be expressed as a scaled dot-product [109]:

$$\text{SelfAttn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.2)$$

Here, Q, K, V are matrices representing query, key, and values; $\frac{1}{\sqrt{d_k}}$ is a scaling factor to mitigate the vanishing gradient issue caused by the Softmax function. Alternatively, instead of computing a single attention response, multiple responses can be computed in parallel and concatenated afterwards. Then, multiheaded attention can be expressed as follows[109]:

$$\begin{aligned} \text{Multihead}(Q, K, V) &= [h_1; h_2; \dots; h_x]W^0 \\ \text{where } h_i &= \text{SelfAttn}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.3}$$

Here, W_i^Q, W_i^K, W_i^V, W^0 are learned transformations. Compared to single-headed attention, multi-headed attention allows the model to jointly attend to information from different representation subspaces at different positions [109].

2.6 An Alternate Strategy for Feature Importance Assessment

While attention is a viable way to infer feature importance, there have been other methods developed for explainability of trained models based on game theory. One of these prominent methods is the Shapley value method, which describes a cooperative game consisting of multiple players and attempts to identify how much each player contributes to the collective, which is measured by the Shapley values. While there have been several such proposals based on Shapley values [110, 111, 112], there are also many issues to be addressed regarding their usage.

Mathematically, variables with *no interventional effect* is an issue to consider [113]. Changes in these variables alone do not change the output of the function. So, while such variables can be left-out, there may be statistical interdependencies with other features. Next, computing Shapley values requires modelling exponential-number of distributions which makes the process computationally expensive [113]. Also, selecting a viable set of arbitrary features should result in a meaningful explanation suitable for our purpose, which further complicates things. Alternatively, interventional value functions may be employed, but such methods rely on out-of-distribution samples for evaluating a model [113]. In other words, calculations depend on never-before-seen-in-training samples despite the fact that predictions for these aren't necessarily relevant to the task at hand. Moreover, the Shapley value satisfying all three axioms (namely *symmetry*,

dummy, and *additivity*) may be unique, yet constraining the value this way limits which models and methods can be explained intuitively by Shapley values.

For human-centric issues, based on the framework of [114], there are three issues to be considered here. First, Shapley values may be regarded as a set of answers to contrastive questions comparing the outcomes to expected values, which may not be possible to be observed at all [113]. Alternatively, Shapley values can represent contrastive statements for why a certain feature is important with respect to the improvement the feature provides. However, as argued in [115], some explanations are more meaningful than others. It follows that not every explanation may be satisfactory, implying that Shapley explanations may not be sufficient for some cases. Second, as formulated by *the counter-factual explanation problem* [116], one of the motivations for explainability is to obtain a set of guidelines for achieving a certain outcome. Shapley value-based methods, however, do not explicitly offer such guidelines. Also, the knowledge that a certain feature plays a significant role in a given model does not guarantee that changes in this variable will lead to a favorable outcome [113]. Lastly, Shapley value-based explanations are generally used for normative evaluation of either the model itself or understanding its decisions [117, 113]. However, since there's no obvious way to convert Shapley values to statements regarding model behavior, relying on such observations may feed the confirmation bias of the user. This suggests that common intuition for the benefits of interpretability may be based on faulty assumptions, and that the questions interpretability studies are expected to answer should be concretely specified and verified [113]. Shapley value-based methods, while they can help form these questions and lead to the answers, they may not provide direct answers to these questions, which is exemplified in [118].

Chapter 3

Methodology

We define the problem of apparent deceitful intent prediction as follows: Let $X = \{X_v, X_l, X_s\}$ represent a multimodal input sample where X_v, X_l, X_s indicate data from visual, vocal, and speech modality-specific data extracted from an input video clip. Note that there's no relation between temporal length of the clip and the length of the corresponding transcript. Therefore, while visual and vocal modalities are aligned and thus share the same temporal length N_1 , number of tokens in speech data is represented as N_2 . Modality-specific data can then be expressed as a sequence as $X_m = \{X_{m,1}, X_{m,2}, \dots, X_{m,N}\}$ where $m \in \{v, l, s\}$ and $N \in \{N_1, N_2\}$. The goal is to classify a sample based on its deceptive content with labels $y \in \{y_0, y_1\}$ where y_0 is *deceptive* and y_1 is *truthful*. For this purpose, we synthesize modality embeddings F_x for $x \in \{v, l, s\}$. The predicted label $\hat{y}$ is then obtained as follows:

$$\hat{y} = \mathcal{M}(F_v, F_l, F_s) \quad (3.1)$$

where $\mathcal{M}$ is the fusion function and $\hat{y} \in \{y_0, y_1\}$.

Our approach aims to automatically infer apparent deceptive intent of a subject through multimodal data analysis. Initially, face tracking and normalization

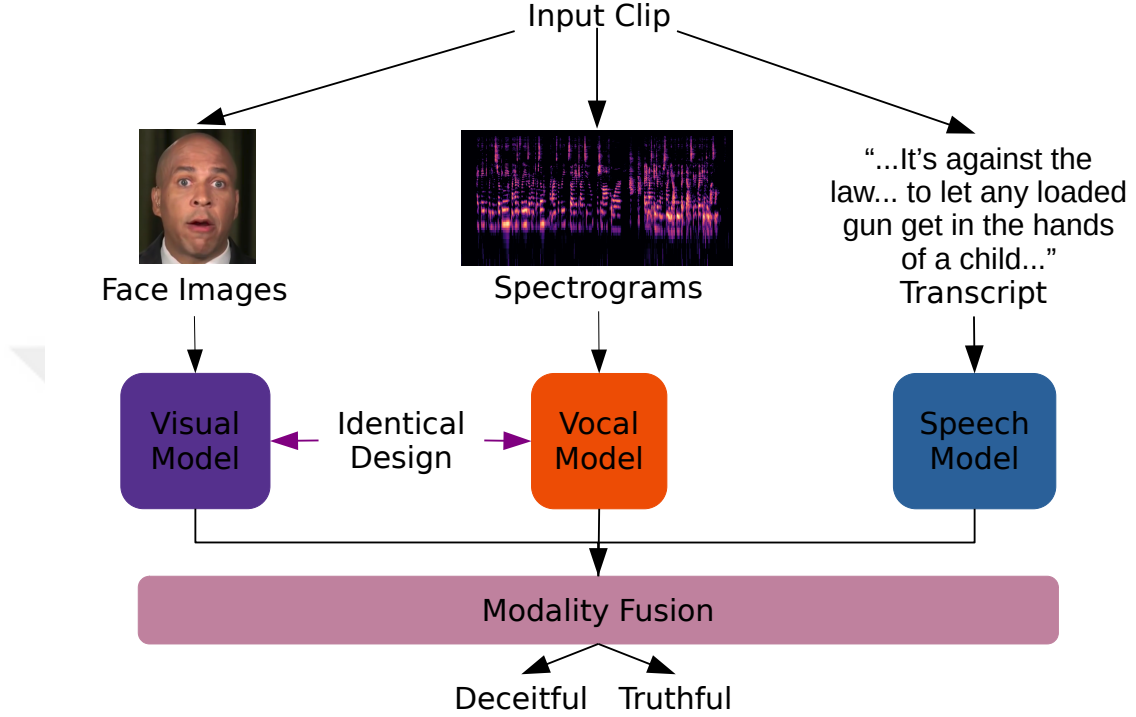
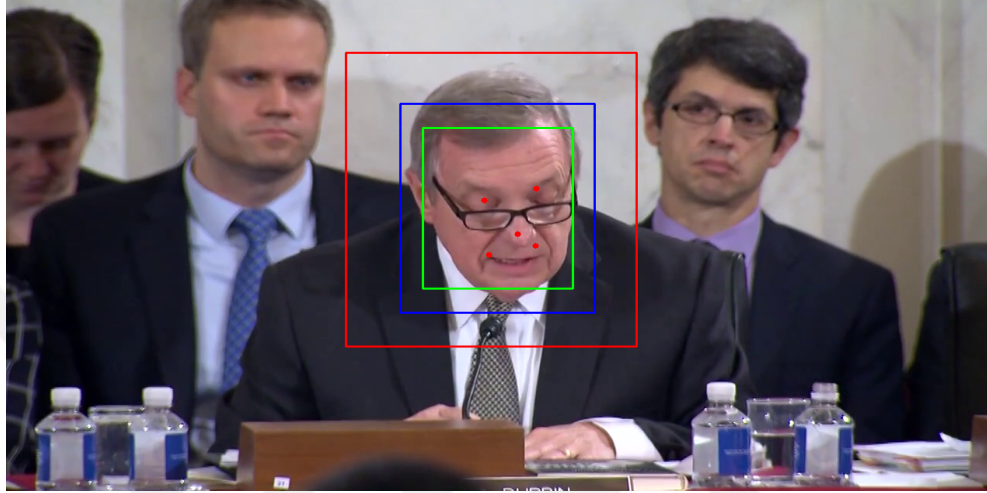


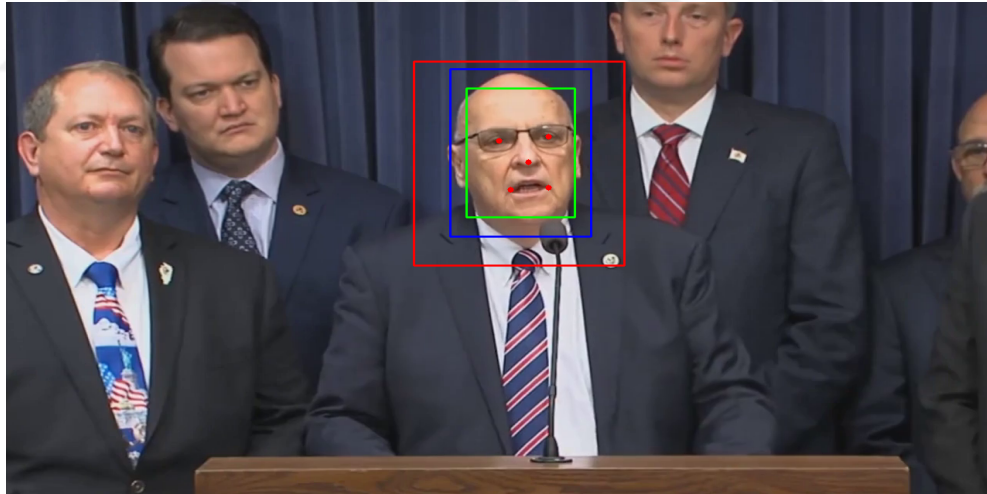
Figure 3.1: Overview of the proposed method.

is applied to extract a frontal facial region where these regions are converted into a series of temporally-ordered color images. Subject’s vocal information as well as their transcribed speech are also extracted similarly. Vocal data is transformed into spectrograms and the transcribed speech is filtered from stop words and similar grammatical elements. These modalities are processed separately by individual modality-specific models that are trained through gradient descent and each returns a singular modality-specific representation. Following the fusion step, our framework outputs a single, discreet binary prediction which is either *deceitful* or *truthful*. This flow is visualized in Figure 3.1. We assume the input X consists of three unique channels of input which are visual X_v , vocal X_s , and speech X_l as denoted by $X = \{X_v, X_l, X_s\}$.

3.1 Face Tracking and Normalization



(a)



(b)

Figure 3.2: Example case of a crowded scene. Red bounding box is manually drawn around the subject of interest to filter out other identities from the image. Green and blue bounding boxes indicate the detected face regions where red dots are the landmarks detected.

For a detailed analysis based on facial dynamics, we extract 5-point facial landmarks (centers of the irises, tip of the nose, corners of the mouth) and facial

bounding boxes of faces in frame rate (FPS)-normalized, visual-channel separated videos. This information is obtained from the MTCNN[119] network which employs hierarchical, cascaded convolutional networks for joint face detection and face alignment.



(a) Before



(b) After

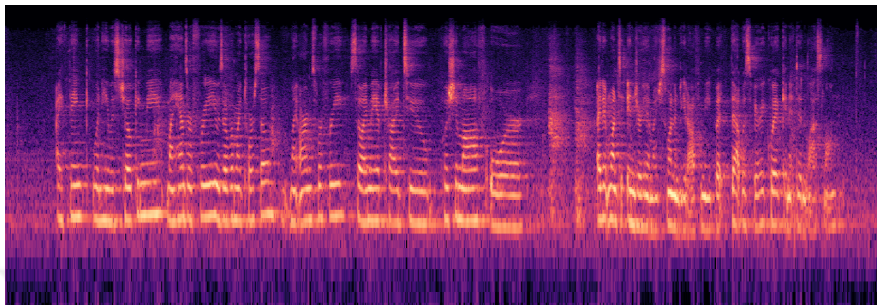
Figure 3.3: Example histogram equalization output. Image brightness is overall increased, eliminating darker spots from the subject's face.

In order to accurately detect the person of interest and their target faces in crowded scenes (frames where multiple individuals are visible alongside the person

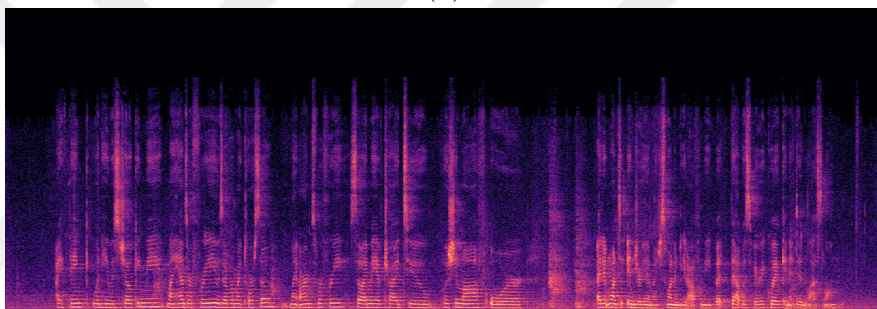
of interest), we manually selected a bounding box within which the person of interest’s face resides (see Figure 3.2). After the facial region is extracted, we enlarge the bounding box by a factor of 0.3 and apply histogram equalization to eliminate darker facial regions negatively influencing the performance of facial representation learning (see Figure 3.3). These images are resized to have 128×128 spatial dimensions and saved as RGB color images for further processing.



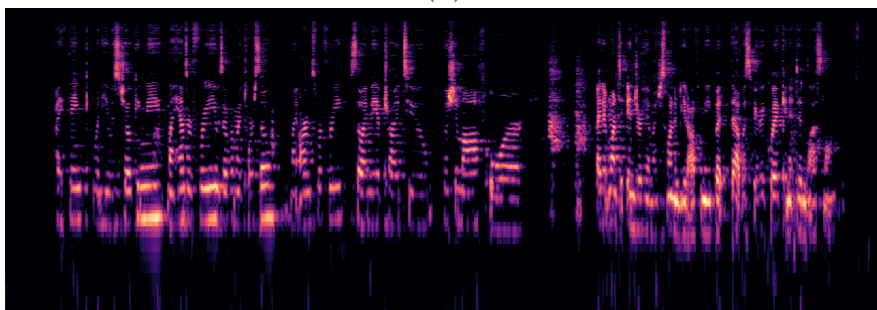
3.2 Voice Representation



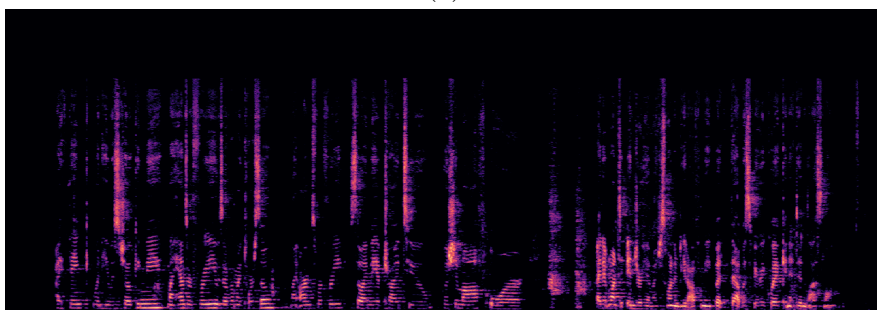
(a)



(b)



(c)



(d)

Figure 3.4: Example spectrograms before (a,b) and after (c,d) vocal separation and gain in log-(a,c) and mel-(b,d) scale.

Similar to face tracking, vocal channels extracted from the input clips are represented as spectrograms. In order to compensate for cases where significant background noise exists, we apply vocal separation algorithm to and decibel (dB) gain in sequence to sample rate-normalized vocal signal (See Figure 3.4). This process clears the background noise from the input, rendering subject’s vocal information ready to be processed. Afterwards, we compute log-scale spectrograms from the filtered signal and save them for further processing.

3.3 Speech Representation

Speech data is realized as transcribed speech of the target individual. While it’s possible to directly employ these transcripts, usually they contain informal language elements, as well as missing phrases. To compensate for these issues we manually cleanup each transcript so that language slang, stop words, etc. are removed before processing as well as punctuation. We also manually correct the transcripts when necessary to complete missing phrases. Having done so, we employ a pretrained RoBERTa [120] transformer as both a vectorizer and a transfer learning backbone. This step allows our language model to be as deep as required without having to deal with the risk of overfitting caused by the sparsely available data. At the end, each language token is converted to a 768-sized vector which is a fixed-sized parameter.

3.4 Visual Model

In visual model our goal is to transform a temporally-ordered sequence of face images into a single latent vector F_v that encodes relevant information for inferring apparent deceptive intent. Our design is inspired from recurrent convolutional networks (RCNN). RCNN is commonly employed in the literature for many visual tasks, especially for video processing [121, 122, 123]. In traditional RCNN architecture convolution units, usually followed by nonlinear activation and batch

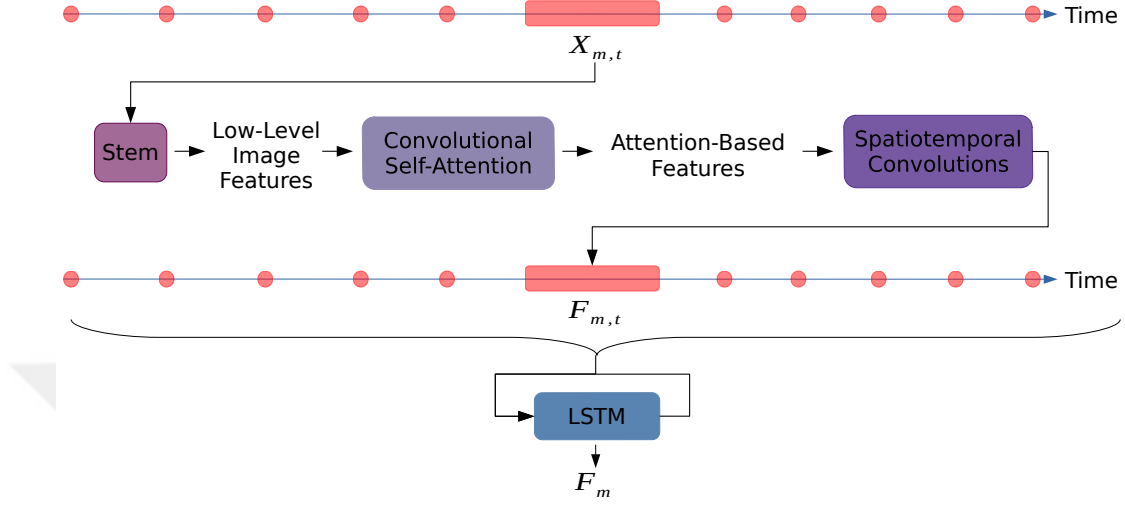


Figure 3.5: Visualization of visual model. Timesteps $X_{m,t}$ extracted from input clips are processed in sequence into intermediate features $F_{m,t}$ as shown, which are then fed to the recurrent unit to output modality representation F_m .

normalization, are used to extract spatial features from an input batch of frames. These frames are generally extracted from a larger video and is referred to as a timestep. Next, the extracted features which is commonly a single feature vector is passed on to a recurrent unit for learning temporal relations between individual timesteps. Nowadays, LSTMs are preferred thanks to their ability to successfully handle long dependency chains.

While the traditional RCNN is shown to be successful on many video tasks, it's heavily reliant on convolution operations to extract relevant image features. For this reason, attention is crucial for improving the performance of the model and the quality of the extracted features. Researchers suggested various techniques for computing visual attention to improve convolutional features [124, 125]. However, cascading convolutions with attention layers overlooks the possibility that these two tasks might be interdependent. Moreover, many of these attention proposals can lead to overfitting in case of sparse data available, due to the number of parameters required to realize the design.

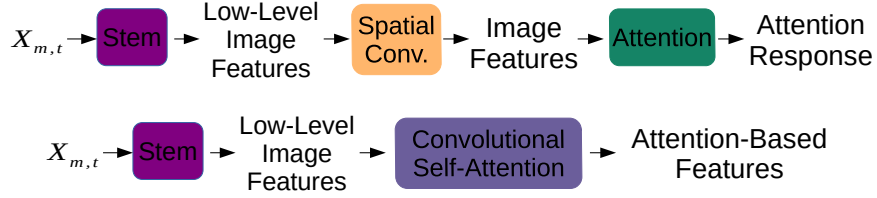


Figure 3.6: Visualization of convolutional self-attention. Unlike the traditional attention (top row), convolutional self-attention is a single-step operation that outputs attention-based representations from its input (bottom row). This not only explores the interdependency of representation learning and attention, but also is an efficient operation since it employs as much as 3 times less number of parameters than its competitors.

3.4.1 Convolutional Self-Attention

Inspired by these observations we propose applying self-attention for joint representation learning and attention computation. As mentioned previously, self-attention is an attention variation that relates different positions and parts of the input with each other. While self-attention is mainly used for language tasks, [126] proposed a variation that allowed joint representation learning and attention performed at filter level applicable for vision tasks. This way, attention becomes an inherent property of feature extraction and learned kernels become attention filters, while filter outputs are attention responses. This not only explores interdependence of representation learning and attention computation, but also reduces the number of parameters required to formulate the operation. What’s more, self-attention based representation learning shares many properties of convolutions such as hierarchically enlarged receptive fields, etc.

To exploit this, we implemented a self-attentional representation learning layer what we refer to as a *convolutional self-attentional* unit. Convolutional self-attention performs a single-headed attention computation at filter level with learnable kernels and biases, mimicking convolution operations (see Figure 3.6). This layer takes as input a feature map $\in \mathbb{R}^{b \times t \times w \times h}$ where b, t, w, h indicate batch size, number of channels, width and height of the feature map respectively. In this *single-headed attention computation with positional embeddings*, attention output for each pixel $x_{i,j}$ with k -extend local neighbourhood $ab \in \mathcal{N}_k(i, j)$ is as follows:

$$\begin{aligned}
y_{ij} &= \text{ConvSelfAttn}(x_{ij}) \\
&= \sum_{a,b \in \mathcal{N}_k(i,j)} \Omega_{ij}^{ab} v_{ab} \\
&= \sum_{a,b \in \mathcal{N}_k(i,j)} \text{Softmax}_{ab}(q_{ij}^T k_{ab} + q_{ij}^T r_{a-i,b-j}) v_{ab}
\end{aligned} \tag{3.2}$$

In the equations *queries* $q_{ij} = Qx_{ij}$, *keys* $k_{ij} = Kx_{ij}$, and *values* $v_{ij} = Vx_{ij}$ are linear transformations of the pixel in position ij and its neighbourhood; $Q, K, V \in \mathbb{R}^{d_{out} \times d_{in}}$ are learned linear transformations; $r_{a-i,b-j}$ is the concatenated positional row and column embeddings. We employ h -headed attention layers and concatenate the responses channel-wise at the end. In order to preserve spatial dimensions, padding is applied to the input before computing attention responses. Each attention head employs a kernel size of k . Note that h, k are hyperparameters.

Comparison to convolution operation in terms of number of parameters. As denoted in [126], a convolution operation may be expressed by the following equation:

$$\begin{aligned}
y_{ij} &= \sum_{a,b \in \mathcal{N}_k(i,j)} W_{i-a,j-b} x_{ab} \\
\mathcal{N}_k(i,j) &= \{a,b \mid |a-i| \leq k/2, |b-j| \leq k/2\}
\end{aligned} \tag{3.3}$$

Note that $W \in \mathbb{R}^{k \times k \times d_{out} \times d_{in}}$ is reused for generating the output for all pixel positions ij which enforces translation equivariance. While for convolutional self-attention parameter count is independent of the size of spatial extent, for convolutions this quantity grows quadratically with spatial extend. For example, for $d_{in} = d_{out} = 128$, the authors claim that a convolution layer with $k = 3$ has the same computational cost as an attention layer with $k = 19$. This is so since, unlike convolutions $Q, K, V \in \mathbb{R}^{d_{out} \times d_{in}}$ are computed independent from the spatial extend k . Additionally, in their experiments, the authors in [126] show that a fully attention-based design outperforms a ResNet-50 baseline while having 12% floating point operations (FLOPS) and 29% fewer parameters.

Table 3.1: Comparison of a single-headed convolutional self-attention unit and a single convolution layer based on variations in parameter counts as kernel spatial extend and in/out channels change.

Num. Channels	Kernel Extend	ConvSelfAttn Param.	Conv. Param.
$d_{in} = d_{out} = 128$	$k = 3$	49536	147584
$d_{in} = d_{out} = 128$	$k = 5$	49792	409728
$d_{in} = d_{out} = 128$	$k = 7$	50048	802944
$d_{in} = d_{out} = 128$	$k = 11$	50560	1982592
$d_{in} = d_{out} = 128$	$k = 13$	50816	2769024
$d_{in} = d_{out} = 128$	$k = 17$	51328	4735104
$d_{in} = d_{out} = 32$	$k = 5$	3232	25632
$d_{in} = d_{out} = 64$	$k = 5$	12608	102464
$d_{in} = d_{out} = 128$	$k = 5$	49792	409728
$d_{in} = d_{out} = 256$	$k = 5$	197888	1638656
$d_{in} = d_{out} = 512$	$k = 5$	788992	6554112

We display the advantage of convolutional self-attention over convolutions in Table 3.1. Here, we compare a single-headed convolutional self-attention unit to a single convolution layer where number of input and output channels, as well as kernel spatial extend change. In this table, we first observe that as the kernel extend increases convolution layer shows a significant increase in parameter count while parameter count for the convolutional self-attention unit is relatively fixed, which is supported by the previous argument that parameter count for a convolutional self-attention unit is independent from kernel spatial extend. Changes in convolutional self-attention parameters can be explained by the fact that we’re employing positional embeddings, which is effected by the kernel extend. Next, we observe that parameter count for the convolutional self-attention unit grows almost quadratically with respect to number of channels. This also supports the argument that convolutional self-attention unit is parameterized by $Q, K, V \in \mathbb{R}^{d_{out} \times d_{in}}$, whose sizes are dependent on number of channels of the input.

Overall, there are two main advantages to using convolutional self-attention units. First, convolutional self-attention units perform joint spatial representation learning and attention computation, which happens on a filter-basis. It follows

that convolutional self-attention units explore the interdependence of both tasks while simultaneously boosting the representation quality and contributing to the explainability of the trained model. Second, convolutional self-attention units, while they show a quadratic change in their parameters as number of channels fluctuates, are not effected by kernel spatial extend; which simplifies the model significantly. This ensures that the resulting model is less prone to overfitting, which is a desirable quality for deception since available data is extremely sparse.

3.4.2 Design Details

For the visual model, we combined spatial and spatiotemporal representation learning in our design. For spatial representation learning we employ a convolutional stem and a series of convolutional self-attention layers. The convolutional stem is commonly employed in the literature [127, 128] which seeks to learn low-level image features such as edges and corners, while convolutional self-attention layers attempt to identify relevant facial regions and generate a representation for the deception task. The convolutional stem, as the name implies, consists of a convolution unit followed by nonlinear activation and pooling. Respectively, convolutional self-attention layers are implemented according to Section 3.4.1, and each unit is followed by nonlinear activation and batch normalization. In our design, we employ convolutional self-attention units separated by pooling operations. As mentioned before, the outputs are image features consisting attention responses.

For spatiotemporal representation learning we employ spatiotemporal convolutions followed by nonlinear activation and batch normalization, where the inputs are attention responses computed previously. Spatiotemporal representation learning, while not common in RCNN design, is a crucial step in our design. It not only studies relationships between consequent attention responses, but also addresses the issue of long temporal dependency chain. Recurrent units are known to fail at capturing long temporal dependencies. By employing spatiotemporal

representation learning, we learn temporal dependencies between timestep representations instead of every single frame within the input video. This stabilizes the training, eliminates the vanishing gradient problem, and renders studying long videos possible. After spatiotemporal representation learning, each input timestep $X_{m,t}$ is converted to a single feature vector denoted $F_{m,t}$.

Lastly, we employ a recurrent unit which is realized as an LSTM to learn temporal dynamics between consecutive timestep encodings $F_{m,t}$. It processes every timestep representation and outputs a single modality encoding F_m as follows:

$$\begin{aligned} hs_{v,t}, out_{v,t} &= \text{LSTM}(F_{m,t}, hs_{v,t-1}), \forall t \in \{0, \dots, N_1\} \\ F_m &= \mathcal{F}_v(hs_{v,N_1}) \end{aligned} \quad (3.4)$$

where $hs_{v,t}$ and $out_{v,t}$ are LSTM hidden state and output vectors at time t , $F_{m,t}$ is timestep representation at time t , $hs_{v,t-1}$ is the hidden state vector at time $t-1$, and $\mathcal{F}_v$ is the modality-specific encoding function realized as a neural net. $hs_{v,0}$ is initialized with random numbers. Since each video is processed separately, we employ the hidden state representation as a context vector for the entire clip to output modality representation.

Visual model is optimized by minimizing the cross-entropy loss as:

$$Q_v^* = \arg \min_{Q_v} -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^2 \mathcal{Y}_{i,c} \log(p_{i,c;v}) \quad (3.5)$$

where Q_v is the parameters of visual model, N indicates the number of training samples, $\mathcal{Y}$ indicates whether truth label c is correct for sample i and $p_{i,c;v}$ is the predicted probability that observation i belongs to truth label c , based on visual information.

During preliminary experiments, we studied many designs that combined convolutional self-attention units and other deep learning operations in various ways. Due to the computational cost of cross-validating these designs for each fold, based on preliminary experiments we've fixed the following setup for our model:

The convolutional stem employs 7×7 kernels. Convolutional self-attention units employ 5×5 kernel and 2x2 padding, uses 4-headed attention, and are stacked as groups of 2, separated by maxpooling. Spatiotemporal convolutions employ 3×3 kernels. Two such units are stacked on top of each other, followed by non-linear activation and batch normalization, separated by a pooling unit. Since kernel sizes, paddings, and other hyperparameters we employ are commonly used in the literature [127, 128], extensive validation for these hyperparameters is not required. The recurrent unit is a 1-layered, one-directional LSTM. Lastly, the timestep length for visual data is to 5 frames without overlap since we observed no clear benefit in introducing timestep overlap.

3.5 Voice Model

Recall that input data to the voice model is the spectrograms constructed from vocal data extracted from an input video clip (see Section 3.2). Voice and visual models, due to similar nature of their inputs, share the same architectural design. Two models run in parallel where voice model is optimized as follows:

$$Q_s^* = \arg \min_{Q_s} -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^2 \mathcal{Y}_{i,c} \log(p_{i,c;s}) \quad (3.6)$$

where Q_a denotes the weights of the voice model, and $p_{i,c;s}$ is the predicted probability that observation i belongs to truth label c , based on voice data. For voice model timestep length is set to 5 frames with no overlap.

3.5.1 Speech Model

Visualization of speech model is provided in Figure 3.7. Recall that speech data is the transcribed speech of the subjects and are tokenized by a pretrained RoBERTa transformer (see Section 3.3). Then, each token is converted to a 768-dimensional

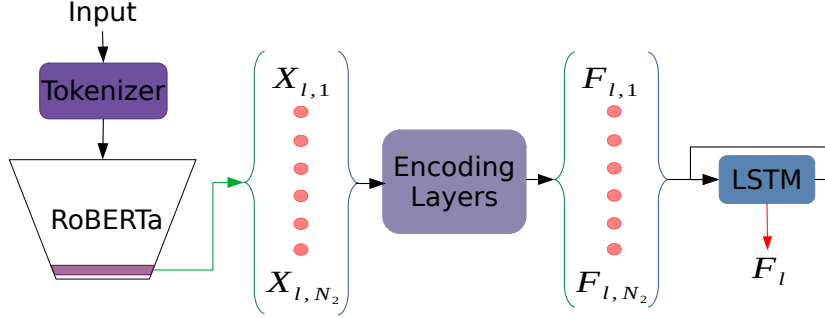


Figure 3.7: Visualization of the speech model.

vector. This results in a $(N_2, 768)$ -dimensional feature matrix where individual rows correspond to token representations denoted by $X_{l,t}$. These features are then processed by fully-connected encoding layers which output intermediate representations denoted by $F_{l,t}$. The recurrent unit outputs a modality embedding denoted by F_l similar to the visual model.

Speech model is optimized by minimizing cross-entropy loss, similar to Equations 3.5 and 3.6, as follows:

$$Q_l^* = \arg \min_{Q_l} -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^2 \mathcal{Y}_{i,c} \log(p_{i,c;l}) \quad (3.7)$$

where Q_l represents the parameters of the speech model, and $p_{i,c;l}$ is the predicted probability that observation i belongs to truth label c , based on speech data.

During preliminary experiments, we tested many language models as well as various pretrained/finetuned transformers within our design. Observations suggest that training a language model from the scratch performs significantly worse than employing a pretrained transformer, while computation cost of finetuning a transformer is too great to adequately validate the language model. Thus, we fixed our backbone as a pretrained RoBERTa whose weights are frozen. Instead, we added embedding layers and a recurrent unit on top to mitigate the issue

of frozen backbone. We do not employ cross validation for the design of this architecture as transformers are shown to be successful in many language tasks [109, 120].

3.6 Modality Fusion

After modality-specific representation learning, we obtain three latent vectors F_v, F_s, F_l from visual, vocal, and speech models respectively. These representations are assumed to reflect apparent deceptive intent based on information from individual modalities, so, for better classification performance fusion can be applied to merge information channels, outputting a single apparent deceptive intent prediction as denoted in Equation 3.1.

Preliminary experiments showed that end-to-end training the entire network by fusing modality encodings F_v, F_s, F_l into a singular representation and performing classification over this feature vector failed to correctly identify deceptive intent. As we investigated further, we noticed that individual models require unique optimization conditions, which renders such a fusion attempt ineffective. Thus, we perform modality-specific training based on optimization goals given in Equations 3.5, 3.6, and 3.7. Fusion, as a consequence, is performed at decision-level as either hard-voting or soft-voting according to Equation 3.1. These fusion techniques are defined as follows:

$$\hat{y} = \begin{cases} \text{mode}\{C_v(F_v), C_s(F_s), C_l(F_l)\} & \text{if hard-voting} \\ \arg \max \frac{\sum_{j \in \{v, s, l\}} \text{Softmax}(F_j)}{3} & \text{if soft-voting} \end{cases} \quad (3.8)$$

where $C_j(F_j) = \arg \max \text{Softmax}(F_j)$ for $j \in \{v, s, l\}$.

Chapter 4

Datasets

4.1 Real-Life Trial Dataset

We employ the Real-Life Trial (RLT for short) [90] dataset for measuring the performance of the proposed design. RLT is a multimodal dataset consisting of 121 videos and 56 unique subjects recorded in high-stake court trials from contemporary U.S. popular criminal justice history. *High-stake* trials describe cases where the suspect faces the risk of long jail times or similar outcomes. In this dataset, data samples consist of a video clip and the transcript of the suspect’s statement manually extracted by the authors. Ground truths are annotated based on the trial verdict. Despite the dataset containing behavioral annotations displayed by the subjects, we did not employ them in our design.

For preprocessing, we first normalize each clip to 20 FPS. This is crucial for ensuring consistency between data samples during face tracking, as some clips contain significantly smaller frame rate than others. We then split the clips into visual and auditory streams that are temporally aligned with each other. Vocal data extracted from video clips are sample-rate normalized 44.1 kHz, which is important for consistency across the dataset. Spectrograms based on sample-rate normalized signals are extracted once for each 0.2 second interval. Lastly,

transcribed speeches of subjects, which is the input to our speech model and provided by authors, are manually corrected in order to complete missing phrases and fix typos in these transcripts. We also eliminate punctuation, stop words (a, the, is, are, etc.), and filler words (um, uh, okay, like, etc.).

Lastly, note that for 7 videos in RLT dataset we could not detect the subjects of interest and their facial regions automatically using the strategy described before (see Section 3.1). This leaves 114 videos in total available for processing. Distribution of these clips are 57 *truthful* and 57 *deceitful*, while the number of unique subjects in total is 50.

4.2 PoliDB

In this section, we describe the recently-collected PoliDB dataset which is collected in order to address the data sparsity and annotation bias problems that are frequently observed in deceit detection literature. Even though deceit is a commonly-occurring phenomena in daily life, accurately capturing these instances is difficult because annotating these events require domain knowledge and most likely subject’s cooperation. While nowadays multimedia platforms allow access to lots of data that can be utilized to study deceptive behaviors, localizing deceptive behaviors require immense manual efforts from the researchers alongside either preexisting knowledge or domain expertise.

Moreover, while in daily life subjects might be willing to cooperate or admit their deceptive intent, which can occur in what’s usually named as low-stake scenarios, in high-stake scenarios such as court trials the subjects have a strong incentive to deceive the listeners and thus would be non-cooperative. This renders annotations to be completely dependent on the researcher’s expertise and efforts in inferring deceptive intent. However, this introduces significant bias in annotation efforts and developed models can be negatively impacted by such a phenomena.

To address these problems, we introduce the novel database *PoliDB*, which is a multimodal, low-stakes, in-the-wild database factually annotated through fact-checks performed by Politifact. In the remainder of this section we describe how the annotations are determined, how data acquisition is performed from the internet, and how the raw data is processed to produce the samples.

4.2.1 Politifact and Its Fact-Check Procedure

Politifact¹ is a fact-checking organization that’s been active for almost two decades. They employ a truth-rating system called Truth-O-Meter [129], which gives a verdict according to the following six categories:

- **True:** The statement is accurate and there’s nothing significant missing.
- **Mostly True:** The statement is accurate but needs clarification or additional information.
- **Half True:** The statement is partially accurate but leaves out important details or takes things out of context.
- **Mostly False or Barely True:** The statement contains an element of truth but ignores critical facts that would give a different impression.
- **False:** The statement is not accurate.
- **Pants-on-Fire:** The statement is not accurate and makes a ridiculous claim.

When selecting statements to analyse, Politifact considers a few criterion as follows: (1) Statement in question needs to be rooted in a verifiable fact. (2) Statement should have some significance to justify the research process. (3) A contagious statement with a far-reaching effect on the public deserves a through fact-checking. (4) Public interest in veracity of a claim is a bonus incentive. (5)

¹politifact.com

The party holding power in U.S. and persons known to make faulty claims are targeted more often [129].

Politifact makes their verdicts according to information available at the time, and publishes corrections further on when necessary. The fact-check process generally includes a review of what other fact-checkers have found previously; a thorough search on popular search engines such as Google and Bing; a search of online databases; consultation with a variety of experts; a review of publications and a final overall review of available evidence [129]. A more detailed explanation can be found in [130].

Politifact prioritizes first-hand sources such as government reports or academic studies, and publications include links to publicly available sources when possible. In case of a necessity to utilize news report made by other media organizations, these instances are clearly indicated to be unverified by Politifact. Before publication, each article is subjected to a peer-review process that takes multiple iterations of clarifications and re-evaluations. Lastly, a majority-voted decision by the editors finalizes the verdict after considering (1) similar past verdicts, (2) whether the statement is literally true and backed by evidence, (3) if the speaker provided necessary evidence to prove their claims, (4) whether the statement is open to interpretation [129].

4.2.2 Statement Acquisition From Politifact

An example Politifact article is shown in Figure 4.1. The title states the subject’s name, the date of the claim, the location of the claim or the media it’s published on, and contains the truth verdict as an image. The rest of the article explains the origin of the claim and their research in detail. The subject’s name and the truth verdict are immediately available from this section alone, however, the statement is sometimes summarized or paraphrased in the title. For this reason, extracting the exact statement occasionally requires manually locating the subject’s claim from the article. In some rare occurrences, even the article body may not contain the exact phrasing. In this case, more manual labor is

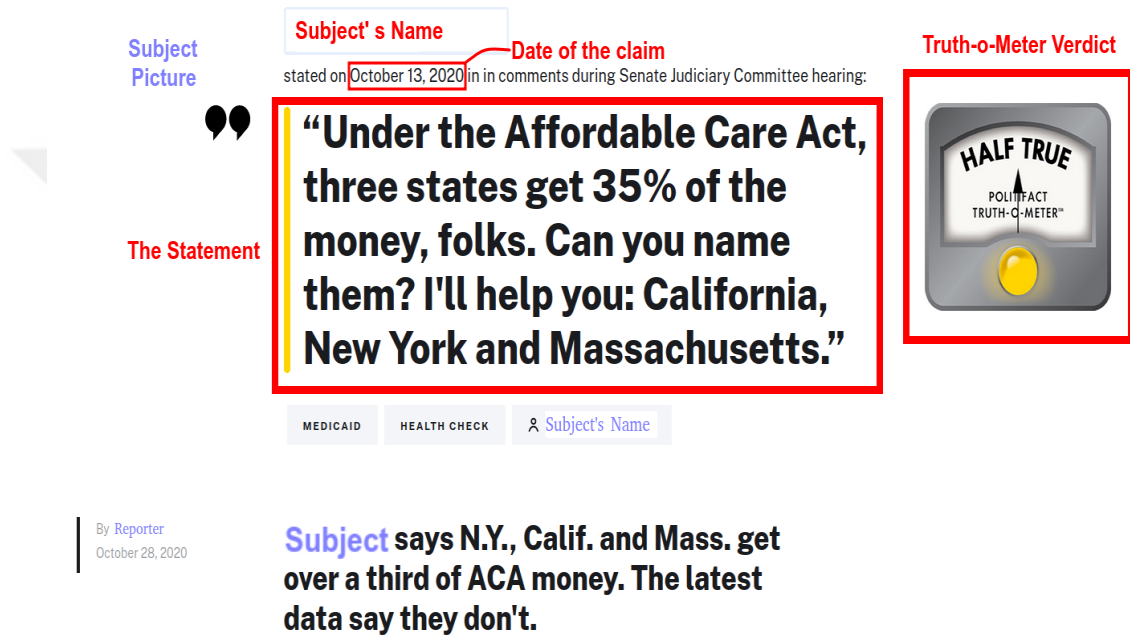


Figure 4.1: Sample screenshot of a Politifact article from our database. The subject’s name, the statement text, the truth verdict, and the date of the claim are emphasized with red markings. The subject’s and the reporter’s identities are censored for privacy concerns.

required to search the internet for finding the statement body and the media in which the phrasing takes place.

The verdict, provided as an image, is a rating from Politifact’s Truth-o-Meter system explained in Section 4.2.1. We converted the 6-category rating system into ordinal labels as follows: *True* is ordinal class 5, *Mostly True* is ordinal class 4, *Half True* is ordinal class 3, *Mostly False* or *Barely True* is ordinal class 2, *False* is ordinal class 1, and *Pants-on-Fire* is ordinal class 0. For our purposes, classes *True* and *Mostly-True* are considered **truthful**; and classes *Barely True*, *False*, and *Pants-on-Fire* are considered **deceitful**. We omit class *Half True* from this version of the dataset since it resides on the boundary of the two class and

is likely to cause severe confusion within the employed models.

4.2.3 Data Acquisition Procedure

For reasons explained in Section 4.2.2, obtaining the exact statement from a Politifact article is challenging and often requires manual labor. There're many difficulties experienced during this process since the claim may appear in any media or platform, and the online content containing the utterance may reside anywhere. Moreover the media of interest may be deleted, missing completely, locked behind a pay-wall, available only to users residing in a specific geographical location; or may be recorded at a lower resolution than what we aim for. Despite these challenges, a general algorithm for data acquisition can be described as follows:

First step is to obtain the subject, date, and exact statement of a claim from Politifact or other news media available. This necessitates obtaining the body of the claim from the Politifact article, and querying various summaries of the body on Google for the exact statement whenever necessary.

Second step is locating the source media. After acquiring the statement, queries constructed by paraphrasing the body are searched on Google to locate the source media. These queries usually contain some keywords we deem relevant in locating the source, or simply consists of the subject's name and the date of occurrence. Occasionally, such broad queries won't provide satisfactory search results. When this happens we switch to more focused, site-specific searches for domains known to host web-scale media. Some of these are: C-Span² hosts media related to U.S. politics such as presidential debates. YouTube³ hosts a great variety of videos about including but not limited to politics, journalism, cable TV broadcasts, live streams, etc. Internet Archive⁴ consists of what's referred to as a wayback machine that archives past media on the internet within its domain for

²<https://www.c-span.org/>

³<https://www.youtube.com/>

⁴<https://archive.org/>

future access. Some cable TV platforms have their corresponding online services that host the channel’s broadcast for online access, so these are also useful sources for our purpose.

Third step is extracting timestamps. When the media is located, we attempt to extract the timestamps within which the exact statement is phrased. For platforms such as YouTube, this requires matching the statement body to the transcript provided alongside the video that’s timestamped automatically. For others, more manual labor is necessary: there are some claims which had to be extracted from hours-long recordings with no additional information.

Fourth step is determining subject’s age at the time of phrasing. Most subjects’ dates of birth are publicly available on the internet, and we determine the apparent age by subtracting year of birth from the year in which the statement is phrased. This information is obtained for future use.

Last step is the cropping of intervals. Having normalized each video to have 30 FPS, intervals marked by the timestamps are cropped using *ffmpeg*⁵ package.

⁵<https://ffmpeg.org/>

4.2.4 Database Details



Figure 4.2: Sample screenshots from PoliDB. Our goal in PoliDB is to ensure every video is recorded in a similar fashion to views here: the subject is centered in the camera’s field-of-view, the camera has a close-up frontal view of the subject, and video quality is sufficiently high to enable capturing facial expressions. Information related to subjects’ identities are censored with red rectangles.

PoliDB statistics are displayed in Table 4.1 where example screenshots from the database can be found in Figures 4.2, 4.3, and 4.4. PoliDB is a multimodal, low-stakes, in-the-wild database factually annotated through fact checks. It’s multimodal since samples consists of video clips and associated transcripts of the statement; moreover, if one so desires, vocal and speech data can be extracted from the videos automatically. We classify PoliDB as a low-stakes database which is clearly justified since one can see a clear absence of repercussions for our subjects in case of untruthful claims and naturally, the subjects lack the motivation to convince the listeners. PoliDB is also an in-the-wild database since samples are recorded outside laboratory environment and in daily life situations. We further claim human-bias in PoliDB is minimal since fact-checks are employed to assess veracity of subjects’ statements. Note that these annotations are not based on opinions of individuals but rather legitimate sources transparently and publicly available within each article. Unlike other studies where annotator agreement is required, each label is backed by what can be named a scientific endeavour and therefore contains minimal human bias. Lastly, note that we refrain from choosing subjects based on their identities or political affiliations, and will attempt to expand this dataset in the future.



Figure 4.3: Sample screenshots from PoliDB, continued.

Table 4.1: Statistics of PoliDB.

Name	Quantity
Subject Count	52
Video Count	101
Gender Distribution	37 Male - 15 Female
Video Distribution	50 <i>Deceitful</i> - 51 <i>Truthful</i>
Age Distribution	42 - 79

For preprocessing, we follow the same approach to what’s described for RLT (see Section 4.1). However, videos are normalized to 30 FPS instead of 20. Vocal data is sample-rate normalized to 44.1 kHz as well. For speech data, we directly follow the same procedure.

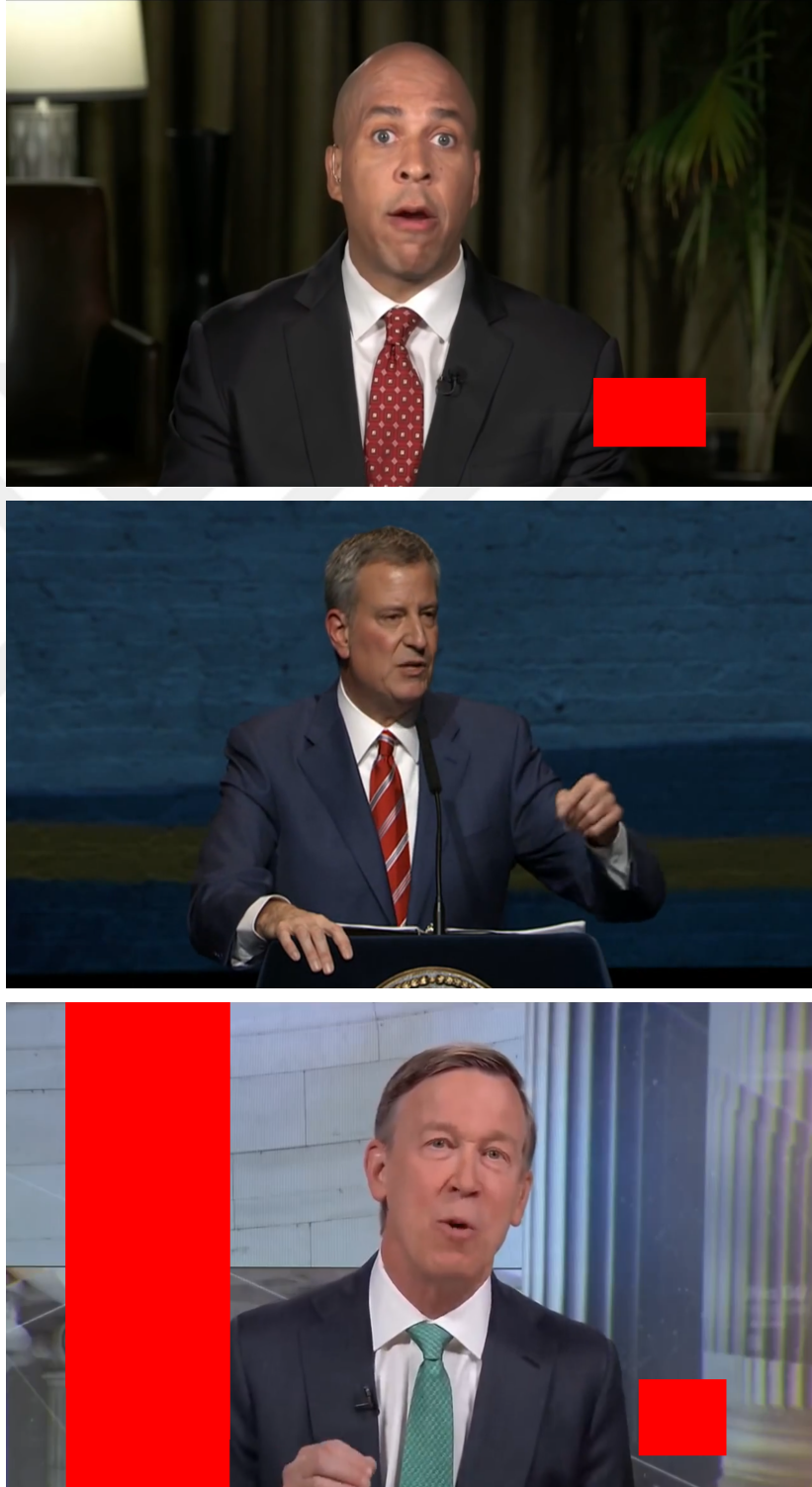


Figure 4.4: Sample screenshots from PoliDB, continued.

Chapter 5

Experiments & Results

5.1 Experimental Setup

5.1.1 Experimental Setup for RLT Dataset

We employ *leave-one-subject-out* (*LOSO* or *LOPO*) cross-validation scheme. In this setting, data collected from the same subject is collectively assigned to the same fold to prevent information leak caused by subject overlaps. The number of folds is equal to the number of subjects. Fold structure is fixed and is reused across experiments. In order to apply fusion strategies mentioned in Section 3.6 we perform modality-specific training according to Equations 3.5, 3.6, and 3.7. As mentioned previously, due to computational cost of validating feature extractor designs, we do not employ cross validation to validate some design choices such as layer arrangements and number of convolutional self-attention layers. Instead, we’re employing architectures that worked best in preliminary experiments.

In Table 5.1, we display the hyperparameters and optimization algorithms considered during cross validation for all modalities. In this table, momentum is only relevant for *RMSprop* algorithm. Learning rate decay determines the algorithm used for reducing learning rate as the training progresses. In this

Table 5.1: Optimization algorithms and hyperparameters considered for cross validation.

Hyperparameter	Considered Values
Optimization Algorithm	{Adam, RMSprop}
Initial Learning Rate	$\{1e^{-4}, 5e^{-5}, 1e^{-5}\}$
Weight Decay	$\{1e^{-3}, 1e^{-4}, 1e^{-5}\}$
Momentum	{0.99, 0.9}
Learning Rate Decay	{Step, Plateau}
Learning Rate Decay Rate (f)	{0.8, 0.5}
Learning Rate Decay Step (e)	{3, 5}
Max Training Epochs	{10, 20, 30}

sense, *Step* indicates learning rate is multiplied by a factor f every e th epoch. *Plateau* indicates the learning rate is multiplied by a factor f whenever validation loss stops improving.

5.1.2 Experimental Setup for PoliDB

For PoliDB, we employ 10-fold cross-validation scheme based on subject identities in order to prevent information leak caused by subject overlaps. Fold structure is created based on subject identities such that data from the same subject is contained within the same fold. This arrangement is fixed and is reused across all experiments. Similar to RLT, we apply modality-specific training and apply a fusion step at the end to merge information channels. Here, we follow the same fixed model design explained before (see Section 5.1.1). Hyperparameters and optimization algorithms considered for cross-validation is the same as RLT, which is given in Table 5.1.

5.2 Results

5.2.1 Results on RLT Dataset

5.2.1.1 Comparison of Visual Attention Proposals

Table 5.2: Comparison of attention proposal performance on RLT dataset based on correct classification rate (CCR).

Method	CCR
Squeeze-and-Excite[124]	0.646
Squeeze-and-Excite[124] + Spatial Attn.	0.711
CBAM [125]	0.745
Squeeze-and-Attention [131]	0.801
Convolutional Self-Attention	0.983

In this section we compare classification performance of the visual model when various attention proposals for representation learning are employed. In Table 5.2, we display these results. We observe that methods incorporating spatial attention as well as channel-wise attention is superior to methods that only employ channel-wise attention. For instance, *Squeeze-and-Excite* attention, when enhanced with spatial attention that applies its channel-wise attention proposal to spatial attention, shows an %7 percent increase in CCR. In comparison, *CBAM* and *Squeeze-and-Attention*, which is designed for both channel-wise and spatial attention, achieves even higher CCR. However, *convolutional self-attention* achieves the highest CCR while using significantly less parameters than others, thanks to the independence of their parameters from kernel spatial extend as argued in Section 3.4.1. Hence, convolutional self-attention units not only improve the learned representations as the experiments show, but also contribute to preventing overfitting within our design.

Table 5.3: Effect of employed language backbone/tokenizer on speech model performance, based on correct classification rate (CCR).

Method	CCR
word2vec [95, 96]	0.383
BERT [109]	0.839
RoBERTa [120], finetuned	0.901
RoBERTa [120], pretrained + LSTM	0.968

5.2.1.2 Influence of Language Information on Speech Model

Here, we display the effect of language information obtained by vectorization and language backbone on classification performance of the speech model in Table 5.3. We tested 4 alternative language models: The first one employed *word2vec*[95, 96] algorithm which performed poorly likely due to the size of available speech data. Next, we separately implemented a BERT[109] transformer as well as finetuned a RoBERTa[120] transformer with no recurrent unit on top. While these methods showed significant improvement over the previous baseline, training was unstable. We believe this is due to the size of the speech data as well, since transformers require lots of data for training. Lastly, we tested a pretrained RoBERTa and custom embedding layers with an LSTM on top, as proposed in Section 3.5.1. This model was more stable to train due to frozen transformer weights, and achieved higher CCR despite the lack of speech data available. We believe this is due to the backbone providing information from a larger language task that we failed to capture from RLT dataset.

5.2.1.3 Reliability of Modality Fusion

Correct classification rates (CCR for short) for RLT fusion experiments are given in Table 5.4. According to this table, the best performing modality is the face model and it has the highest contribution. When mixed with other modalities on a binary basis we observe a clear decrease in accuracy and a minor improvement

Table 5.4: Correct classification rates (CCR) calculated for each fusion experiment performed on RLT dataset.

Modality	Fusion Type	CCR
Visual	-	0.983
Speech	-	0.968
Vocal	-	0.975
Visual + Speech	Hard Voting	0.981
Visual + Speech	Soft Voting	0.981
Visual + Vocal	Hard Voting	0.974
Visual + Vocal	Soft Voting	0.974
Speech + Vocal	Hard Voting	0.974
Speech + Vocal	Soft Voting	0.974
Visual + Speech + Vocal	Hard Voting	0.978
Visual + Speech + Vocal	Soft Voting	0.985

when mixed with both, indicating conflicts could arise whenever multiple modalities are employed. Furthermore, most cases display minimal difference between soft and hard voting (except in 3-way fusion), supporting the claim that voting is a non-optimal strategy for fusion. Investigating additional fusion techniques is left for future work.

5.2.1.4 Comparison to Other Methods

Table 5.5: Comparison of classification performance of various studies performed on RLT dataset based on reported correct classification rates (CCR).

Method	CCR
Pérez-Rosas <i>et al</i> , 2015a [90]	0.752
Pérez-Rosas <i>et al</i> , 2015b [132]	0.821
Şen <i>et al</i> , 2020 [91]	0.830
Wu <i>et al</i> , 2018 [133]	0.877
Krishnamurthy <i>et al</i> , 2018 [92]	0.961
Venkatesh <i>et al</i> , 2019 [103]	0.970
Ours	0.985

Here, We report performance of some methods from the literature that employ RLT dataset. Note that since experimental setup for RLT dataset isn't fixed, the

results shown in Table 5.5 are not one-to-one comparable. Specifically, while we employed LOSO cross-validation (see Section 5.1.1), other methods follow different cross-validation strategies. Nonetheless, our approach the highest correct classification rate on RLT dataset. We also observe almost all modalities and fusions surpass previous studies, yet the best score is achieved through three-way fusion of visual, textual, and auditory modalities. However, considering the limitations of RLT dataset in terms of sparsely available data, these results are subject to change on larger datasets and requires additional testing for confirmation.

5.2.1.5 Visual Analysis of Self-Attention Maps



Figure 5.1: Example self-attention response overlays for RLT dataset.

Figure 5.1 shows a selection of self-attention map overlaps extracted for RLT dataset. The highlights are attention outputs and mark areas with higher activations. In other words, brighter areas contribute significantly more than darker areas to learning goal. We believe there’s a correspondence between learned attention filters and facial action units defining facial muscle movements [134],

which have been shown to capture apparent deceptive intent [7, 135, 18].

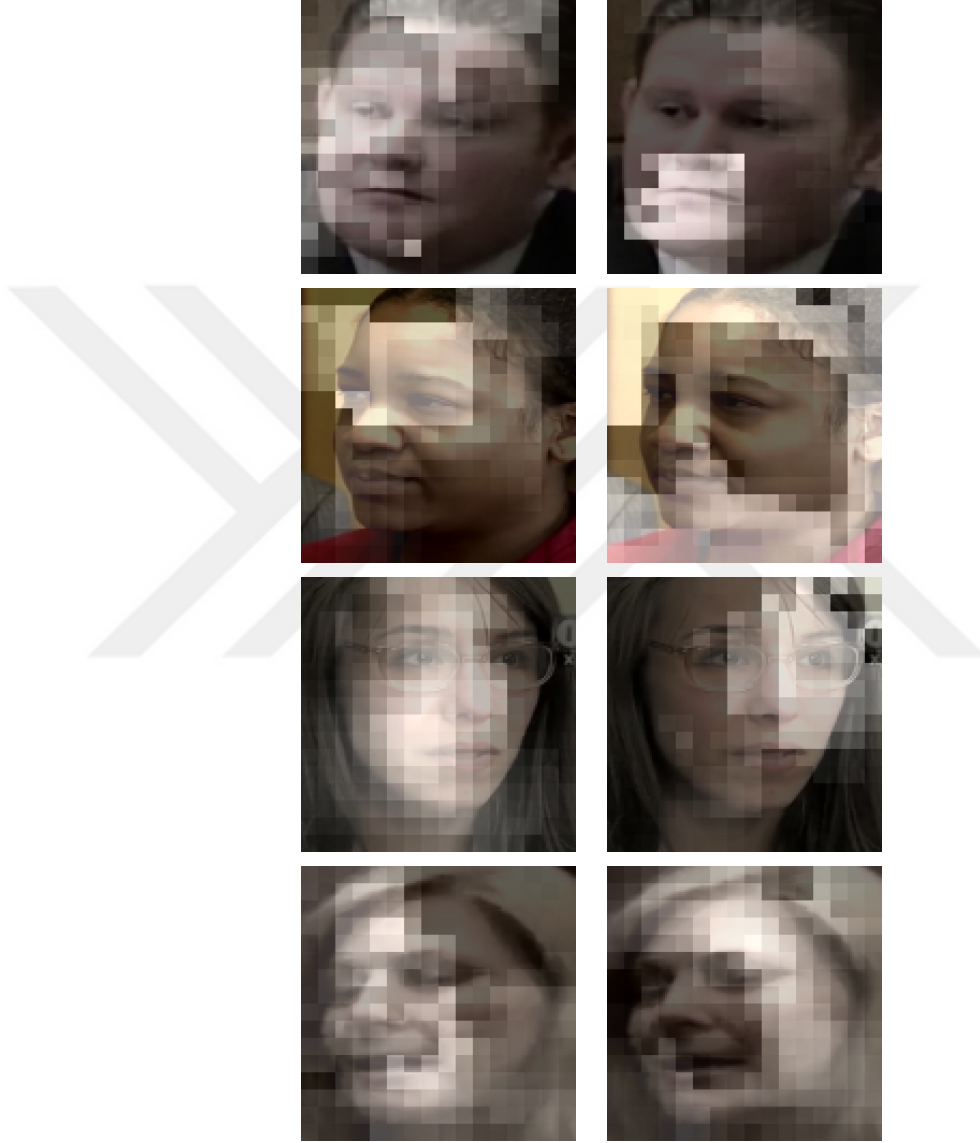


Figure 5.2: Example self-attention response overlays for RLT dataset, continued.

Figure 5.3 supports this by showing how action unit (AU) predictions overlap with regions having high attention activations. However, according to Figures 5.1 and 5.2, there’s also some noise in attention computations. We suspect this is caused by the model’s limitations in dealing with variations in facial orientation, lighting, and the image resolution which are inherent problems of RLT dataset.

5.2.1.6 Deceptive Cues in RLT Dataset

In this section, we attempt to identify nonverbal and language-based queues that we believe our design focuses on. Note that nonverbal cues referred to in this section are based on [17, 18]. Similarly, language-based queues are based on [23, 25, 18].

We noticed many of the stereotypical nonverbal indicators of deceit exist in both deceptive and truthful videos. Rapid blinking, changes in bodily and facial orientation, stops and pauses, vocal changes, etc. are example patterns we observed not only in most clips but also during deceptive and truthful segments according to model predictions. In some cases, deceptive subjects displayed behaviors that do not agree with stereotypical patterns which could be explained based on attempted behavioral control that liars deliberately avoided showing stereotypical deceptive signs.

Interestingly, manual inspection suggests there’s also a correlation between strong emotions such as anger or annoyance and apparent deceptive intent. We noticed subjects were more aggressive in their facial expressions, displayed more bodily movements, and employed unusual mouth movements (such as stretching one side of the mouth, pushing the chin forward) compared to their truthful baseline, whenever they experienced strong emotions. We believe this causes behavioral leaks as a consequence of emotional activation, and is one of the nonverbal indicators the model focuses on. However, note that nonverbal deceptive cues are highly subjective, thus, it’s difficult to identify every nonverbal indicator.

In terms of verbal indicators, truthful subjects were generally more informative in their statements. They described events clearly and in general in a chronological order. They provided more details, both related and superfluous (*‘...My mother was a single mother. I have two older sisters. I have one older brother...’*). They also performed spontaneous corrections (*‘...Um I lived, we lived until I was approximately in fifth grade...’*) and self-deprecation more frequently than liars. Deceptive subjects, on the other hand, repeated some sentences multiple times (for a stronger effect, we supposed); rarely provide many spatial, temporal,

or contextual information regarding the event or answered the question directly. Instead, they employed non-prompted negation ('*I could've killed people ... Look at all the ammo I had ... I weren't trying to kill nobody.*') and equivocation ('*We withheld the evidence...*' doesn't explain which pieces of evidence were undisclosed or why), provided extraneous info (for example, explaining in great detail the activities performed in a religious organization), utilized many stop words (which could indicate a major cognitive overload).



5.2.1.7 Apparent Deceptive Intent as a Continuous Function of Time

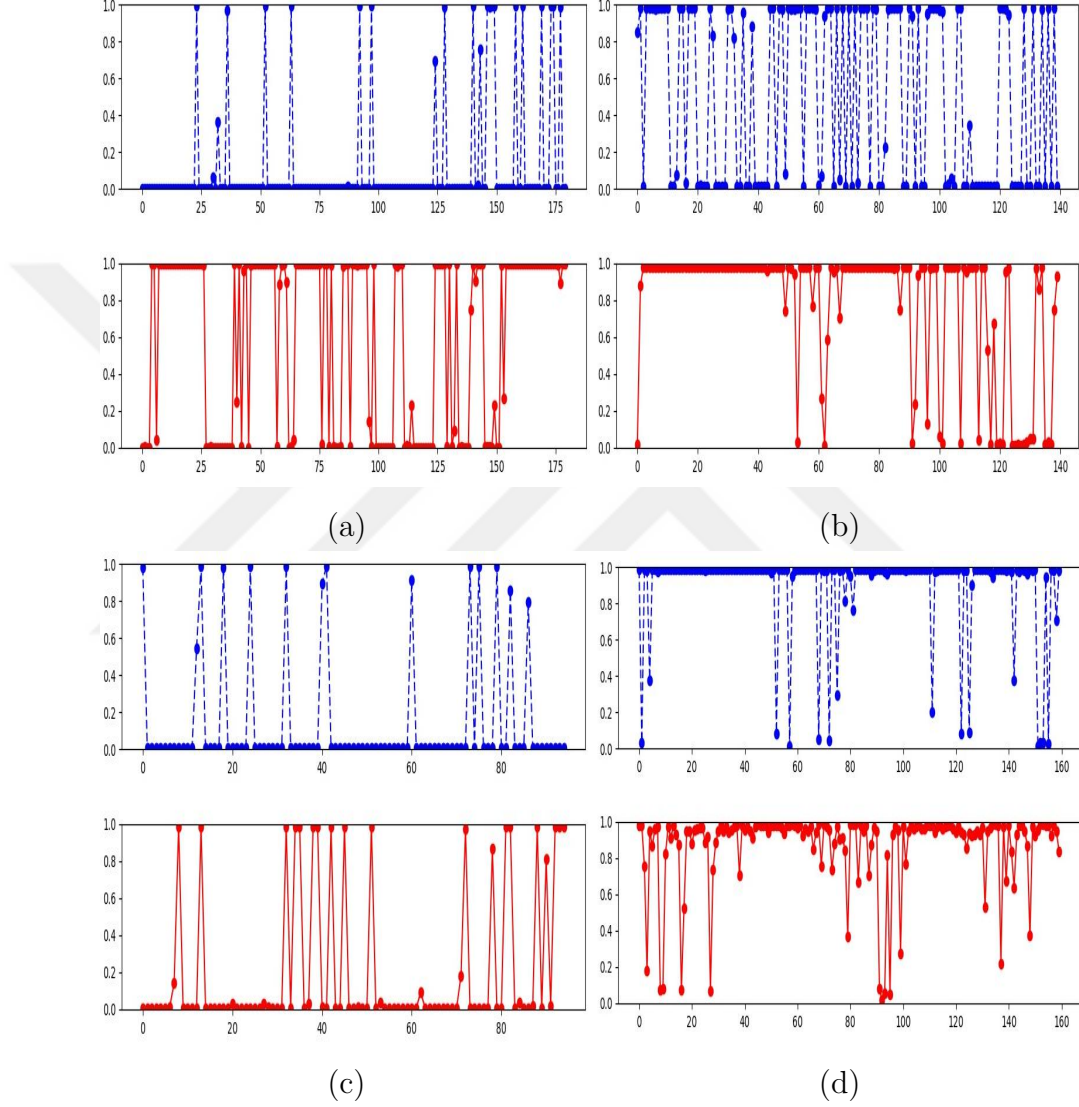


Figure 5.4: Timestep-wise (blue plots) and cumulative (red plots) predicted deceit levels throughout selected videos from RLT dataset. Blue plots indicate deceptive predictions based on the current timestep, while red plots indicate predictions based on the history of observed timesteps. Deceptive behavior levels are represented as real numbers in range $[0, 1]$: The higher a value is the more deceptive a subject appears to be. *Best viewed with zoom.*

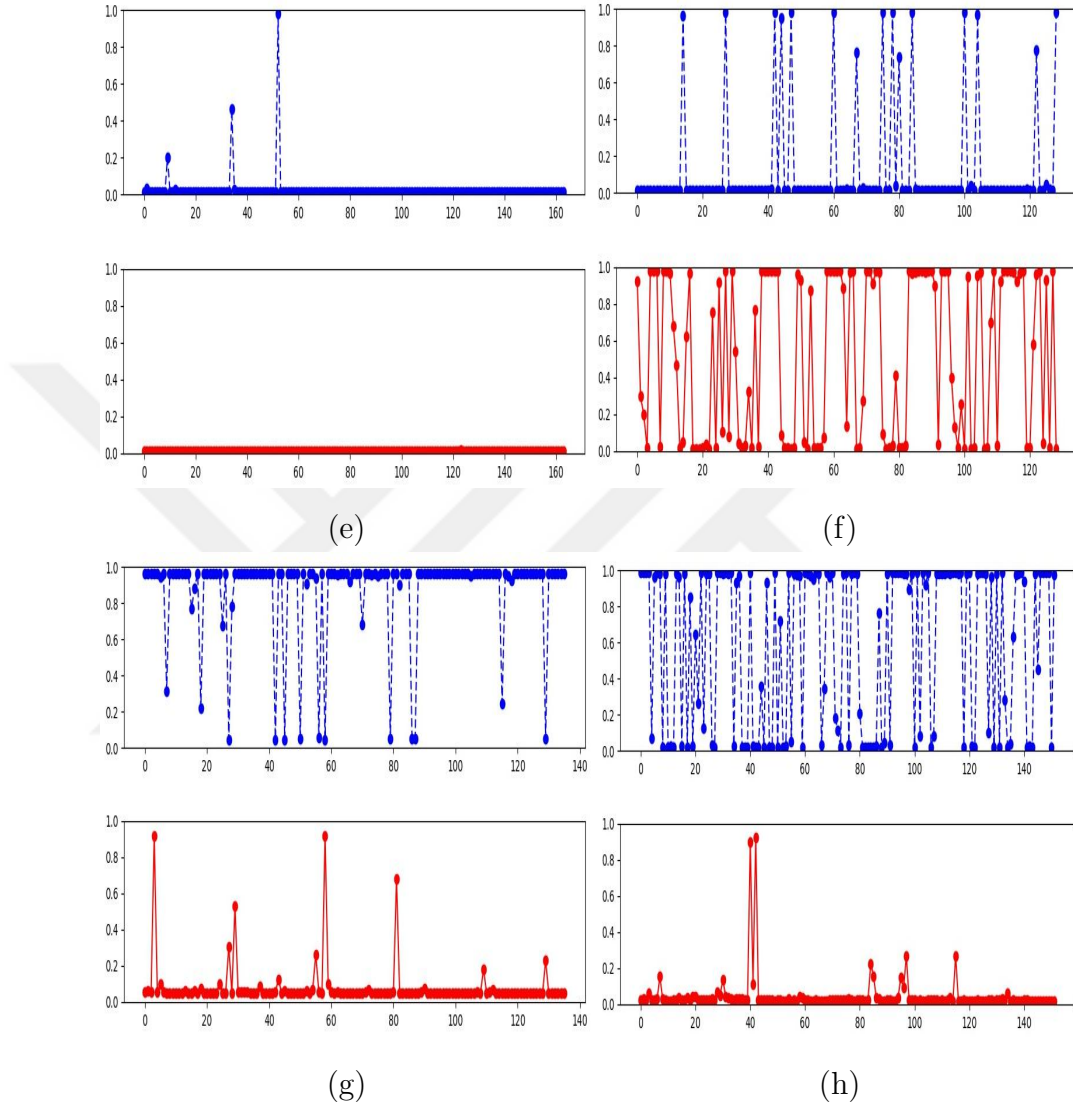


Figure 5.5: Deceit plots, continued.

According to Figures 5.4 and 5.5, we note that apparent deceptive intent is a continuous function of time. There are intervals within which subjects display high levels of deceptive intent while others seem truthful to the trained model. This suggests that generalizing top-level deceptive labels to every interval for a given clip can be misleading.

Next, we observe that focusing on specific behavioral indicators out-of-context can be an unreliable way to induce deceptive intent. Criminal psychology argues

observed behaviors deviating from the subject’s baseline in the context of the conversation is a reliable way to detect deceptive intent [8, 47]. This suggests, contrary to common belief, singular instances of abnormal behavior does not explicitly indicate deceptive intent. We show the validity of these claims in Figures 5.4 and 5.5: Many stereotypical deceptive cues are visible in both deceptive and truthful videos, as well as deceptive and truthful segments within the same clips. These behavioral patterns, according to our observations, doesn’t explain fluctuations in cumulative plots so it’s unlikely that the model exploit existence or lack of these cues. We believe the model detects abnormalities and compares them to subject’s observed behavioral baseline to infer apparent deceptive intent. This is also supported by the relation between cumulative and timestep plots: Timestep-wise and cumulative predictions do not always agree on the predicted apparent deceptive intent. Even when most timesteps are deceptive when studied individually, overall attitude can turn truthful when the timesteps are analysed altogether.

Note that while there are indications suggesting fluctuations in cumulative plots have behavioral explanations, they may have also been caused by the temporal dependency chain. This would explain why apparent deceptive intent predictions are rarely assigned intermediate values, and why deceptive trend can change drastically in favor of recent observations. Despite these shortcomings it’s our belief that measuring and recording changes in apparent deceptive intent is useful for determining subject-specific and inter-subject behavioral queues to verify psychological findings and accurately interpret behavioral patterns in the context of the recording.

5.2.2 Results on PoliDB

Table 5.6 displays fusion experiments performed on PoliDB using the experimental setup detailed in Section 5.1.2. We first notice similarities between RLT and PoliDB in terms of fusion experiments. In both cases, fusion seems to create confusion and decrease classification performance in most cases compared to

modality-specific models. While this can be attributed to voting acting as a poor fusion technique, for PoliDB the effect is much more prominent than RLT. When pairwise experiments are compared to 3-way fusion, we observe that visual model seems to have the highest contribution to final predictions, which is followed by speech and vocal models, respectively. This is supported by the fact that when visual model is omitted, there’s as much as 8% decrease in CCR between 3-way and pairwise fusion of vocal-speech models. In a similar fashion, for vocal model, we denote a 3% decrease whereas for speech model, this quantity is 2.5%. It’s our belief that since PoliDB offers high-quality visual data while vocal and speech data of the subjects lacks in length (typically one to two short sentences per statement), visual assessment of deceptive intent plays a more prominent role in deceit detection which is nicely performed thanks to convolutional self-attention units.

Table 5.6: Correct classification rates (CCR) calculated for each fusion experiment performed on PoliDB dataset.

Modality	Fusion Type	CCR
Visual	-	0.660
Speech	-	0.653
Vocal	-	0.649
Visual + Speech	Hard Voting	0.649
Visual + Speech	Soft Voting	0.648
Visual + Vocal	Hard Voting	0.657
Visual + Vocal	Soft Voting	0.657
Speech + Vocal	Hard Voting	0.592
Speech + Vocal	Soft Voting	0.592
Visual + Speech + Vocal	Hard Voting	0.677
Visual + Speech + Vocal	Soft Voting	0.678

5.3 Error Analysis

5.3.1 Sources of Visual Error

In this section, we determine multiple possible sources of visual error within our framework. Note that these sources exist as a consequence of employing RLT

and PoliDB and are inherent properties of these datasets.

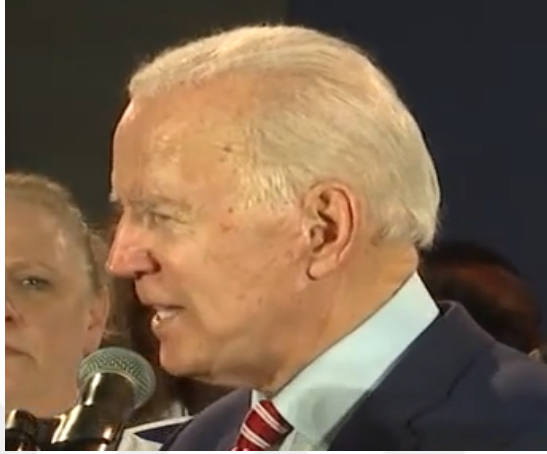


Figure 5.6: Example extreme head orientation.

First, face tracker discussed in Section 3.1 assumes that subject is directly facing the camera. While there's a margin of tolerance to this assumption which allows the model to handle some problematic cases, extreme examples such as the one is provided in Figure 5.6 cannot be handled well. In this figure, almost half of the the subject's face is invisible to the camera which makes landmark detection very difficult.



Figure 5.7: Example blur caused by rapid camera motion.

Second, rapid camera motion can cause a blur effect, limiting the visibility of facial characteristics of the subject. An example is provided in Figure 5.7 where this blur is clearly visible.



Figure 5.8: Example problematic viewpoint.

Third, while PoliDB seeks videos having the subject at the center of the camera view, people may reside at a distance to the camera or they may not be at the center of the view. These cases such as Figure 5.8 limits the model's ability to capture facial details.



Figure 5.9: Example complex background.

Fourth, a complex background such as Figure 5.9 can negatively effect visual representation learning since the subject may not be accurately separated from the background.



Figure 5.10: Example visual artifacts and glare effect.

Fifth, as shown in Figure 5.10, visual artifacts and non-uniform lighting can be a hindrance as well.

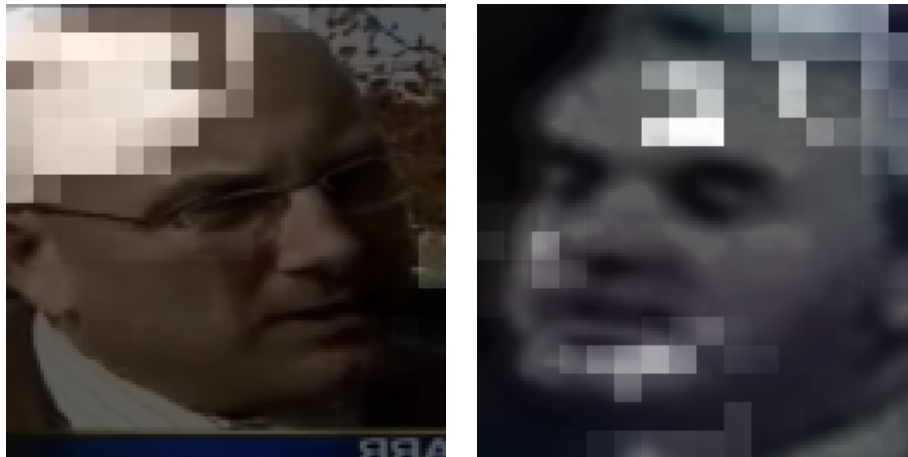


Figure 5.11: Example negative effect of visual quality on visual representation learning.

Lastly, visual quality or resolution can be factor here as well. Since visual

quality directly affects how visible visual characteristics are, low-quality videos tend to result in non-informative visual representations, as shown in Figure 5.11.

5.3.2 The Human Factor

Psychological findings suggest human bias can complicate studies, such as the well-known Hawthorne effect [136]. Similarly, considering humans are hardly better than random chance at detecting deceitful intent, relying on annotator agreement amongst manual annotators whose expertise in psychology is often questionable introduces unnecessary noise caused by human bias and annotation bias/inconsistencies into the problem. For RLT, we believe similar problems occur due to annotation strategy reported in [90]. According to the paper, truth labels for individual samples are heavily reliant on the accuracy of the trial verdict. In this sense, any potential mistake made in the trial can influence the annotations directly. Moreover, since individual samples are annotated by the researchers based on trial verdict, human factors are at play here as well.

Unlike RLT, it’s our opinion that PoliDB contains minimal human error due to reasons explained before (see Section 4.2.4). This can be attributed to the fact-checks the annotations are based on, since we outsource the annotation procedure to Politifact directly. While there might be other issues causing error, we do not believe human bias is an issue for PoliDB.

5.3.3 Curse of Dimensionality and Generalizability

In the context of this study, we suspect overfitting could exist within our model. One reason is because subjects do not contribute equally to each class. Contributions of individuals vary significantly as well. In this sense, the models could favor specific individuals more than those who appear rarely. Another issue relates to the fact that available data for deceit scenarios is extremely sparse. Moreover,

low spatial resolution within available clips as well as limited amount of transcripts available for language-based training could also incentive memorization. While we attempted to limit the number of parameters for our models, as well as employed strategies such as weight decay, learning rate scheduling, and early stopping, it's likely overfitting still occurs within our framework. For this reason, in order to obtain a better comparison of our method with the state of the art, in the future this study should be repeated by employing a larger dataset. Since in the future we plan on expanding PoliDB to contain more unique subjects and videos, PoliDB is a good such candidate for further testing whether overfitting occurs within the proposed design.

5.3.4 Interpretability

Considering most affective problems are studied from an inter-disciplinary standpoint, obtaining quantitative measures is often not enough unlike many machine learning problems. Therefore, it's important to provide explainable results from a non-statistical perspective as well[137]. To illustrate this, complementary to the test results reported previously, we attempted to associate observed behavioral patterns with apparent deceptive intent to provide an insight into what the trained model detects. The discussion can be found in Sections 5.2.1.6 and 5.2.1.7. It is our belief that future studies focusing on deceit detection should strive towards improving the state of the art while achieving explainable results, similar to what we attempted here.

5.3.5 Representation

Related to interpretability, representation refers to the ability of the learned hyperspace to encapsulate psychological criteria that relates to the task at hand or is meaningful in some other way which can be used as intermediary features [137]. For instance, learning facial action units before predicting deceitful intent is an example training strategy which falls inline with this concept. Representation

contributes to interpretability of the learned hyperplane and can stabilize training by introducing intermediary loss and a stronger gradient signal as a consequence. Despite these promises, rarely in affective computing literature representation is considered an important goal/criteria when designing methodologies partially because introducing representation makes it difficult to realize the model design and employ hyperparameter optimization [137]. For these reasons we refrained from introducing representation to our architecture, whose effects at the moment are unknown and subject to future analysis.

5.3.6 Explainability and Real-Life Usage

Neural networks are black-box models. Therefore, blindly trusting model predictions can be dangerous: In a famous example, scientists in [138] attempted to identify COVID-19 cases from chest X-ray images. According to the paper, they found out that the trained model relied on artifacts at image borders and reported correct classification rates even when most of the X-ray image was occluded. A similar occurrence is certainly possible in our case: While we attempted to decipher what learned representations correspond to, we are unable to ascertain with certainty. Moreover, human-centric data is usually extremely sparse, which incentivizes memorization [137]. So, not every design may not work well in practice; such as the shortcomings of iBorderCtrl [139] in regards to its societal implications, as well as bias within facial recognition systems towards some ethnicities [140, 141].

It's our belief that in a real-life prototype, models for any human-centric task needs to be explainable and should be trained on a diverse dataset. For these reasons, we don't believe our approach can be applied to real-life scenarios, as well as to scenarios consisting individuals from a multi-cultural and multi-ethnic background. This is due to subjects in both RLT and PoliDB being from United States and speaking English as their mother tongue. Following [142], designing a UI expressing the decision-making process of the model via visualizations could

be a solution for improving interpretability in the future, however, addressing these problems is out of the scope of this paper.



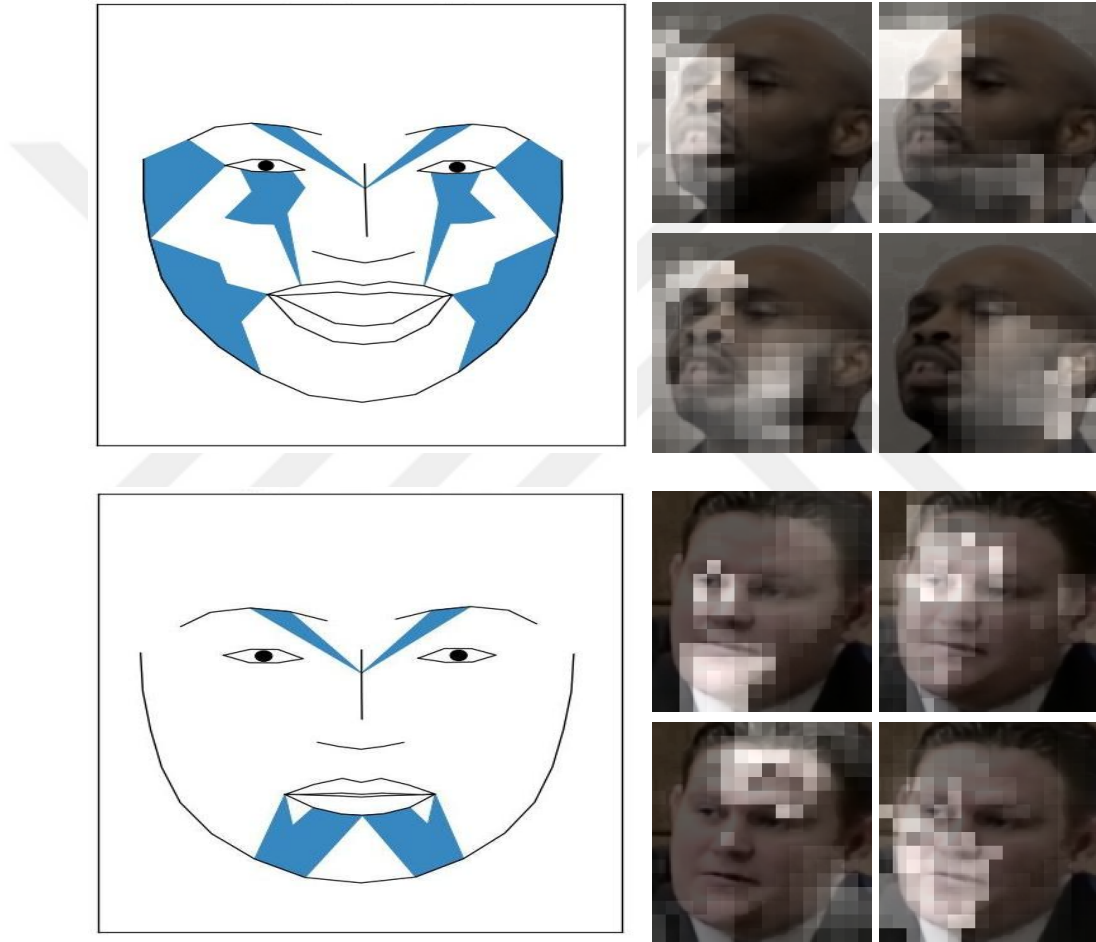


Figure 5.3: Attention overlays and action unit (AU) correspondence. Action unit predictions (left) are visualized over a template face to show the overlap between AUs and attention maps (right). AUs visible in the top row are *brow lowerer*(AU4), *cheek raiser*(AU6), *upper lip raiser*(AU10), *lip corner puller*(AU12), *lips parting*(AU25), and *jaw drop*(AU26). In the bottom row *brow lowerer*(AU4), *dimpler*(AU14), *lip tightner*(AU23), and *lip sucking*(AU28) are visible.

Chapter 6

Conclusion and Future Work

In this work, we showed that convolutional self-attention units are promising alternatives to traditional attention proposals that are preceded by convolutions; as well as transformers are shown to be applicable to deceit detection tasks. We offer both a quantitative and a qualitative analysis into the proposed design and validate its results on both high- and low-stake cases.

Convolutional Self-Attention. For visual representation learning, we proposed use of convolutional self-attention for attention-based representation learning which is performed as a single-step operation instead of a two-step sequence of feature extraction-attention computation which is the norm in the literature. Experiments showed that convolutional self-attention outperformed other attention proposals while using as significantly less parameters compared to others.

Vocal Representation Learning. We explore representing vocal features extracted from a video clip as a sequence of spectrograms which is processed by the same convolutional self-attention based representation learning model. We note that vocal features seemed to contribute less than visual modality, which might be related to this design choice. Further experimentation is required to test the validity of these features.

Speech-Based Representation Learning. We explore use of transformers

as both a tokenizer and a language-model backbone to realize language-based deception task by exploiting a larger language task. Our experiments showed that transformers are successful at this task, yet more experimentation is required on a larger dataset for further confirmation.

Modality Fusion. We performed modality-specific training and performed voting-based fusion on visual, vocal, and speech modalities to merge information channels. We note that decision-based fusion, while useful for many tasks, caused confusion in the final decision and failed to significantly improve the overall performance. This implies that other fusion strategies need to be studied in the future.

Comparison with other methods. We compare our design to other studies performed on the RLT dataset. Since experimental setup of RLT isn't fixed, doing one-to-one comparisons with these predecessors isn't possible; however, we report the highest correct classification performance on RLT with our design. Further study is required to reaffirm these reports.

Self-Attention Maps We visualize the attention responses to the original images in order to visualize the areas the attention focuses on. The results suggest that attention maps coincide with facial action units which are shown to be reliable features for inferring deceptive intent.

Deceptive Cues. We manually determine deceptive cues in RLT that our models focus on, both visual-based and language-based deceit detection. Visual cues seemed to come from behavioral leaks caused by emotional activation of the subjects, and is done in a performative manner with the subject's truthful baseline. Language-based features seemed to focus on phenomena such as non-prompted negation, equivocation, spontaneous corrections, etc.

Deception As a Continuous Function. We visualized the continuous deceptive intent predictions to show that while video annotations may be deceitful, not every timestep in a clip agrees with the label: Subjects seemed to have varying levels of deceptive intents in all videos. Further study is required to exploit

these observations.

Studying Deceptive Cues In Context. We observed that focusing on singular instances of abnormal behavior, similar to the common misconception about deceptive cues in public, is an unreliable way to infer deceptive intent. Our model studies these cues in context, which is in-line with findings in criminal psychology.

Future Work. One of the main limitations for this work is the amount of data available for deceit detection task. Since RLT is quite sparse, reliably training very deep networks as well as extensive testing isn't possible. Moreover, since RLT lacks a unified experimental procedure, one-to-one comparison with previous methods cannot be made. For these reasons, we plan on expanding PoliDB to contain many unique subjects and videos in hopes of incentivising future research on deceit detection task. Specific to this study, we plan on testing the proposed approach on this expanded PoliDB and report the results. This will be performed based on an experimental procedure that will be fixed when new data collected, so future research can make reliable comparisons with other methods.

Bottleneck of the Technology and Its Ethical Implications. As a final note, we would like to emphasize the bottlenecks of the technology and its ethical implications for real-life use. As discussed before, the proposal design is a black box model. This implies that, due to lack of subject diversity, our system may in fact be overfitting to a certain set of characteristics observed in a specific culture. This begs the question, is it ethical to employ the system in real-life? High classification accuracy may not translate directly to real deception. There could also be serious consequences to reckless use such as a decrease in public trust towards the democracies and elected officials, weaponizing the system for political gains and misinformation campaigns, enhancing Deep Fakes for nefarious purposes, etc. are among a few of these issues. Therefore, before employing our method careful considerations of its ethical implications should be made, and human supervision during its usage is strongly recommended.

Bibliography

- [1] L. K. Guerrero, P. A. Andersen, and W. A. Afifi, *Close encounters: Communication in relationships*. Sage Publications, 2017.
- [2] B. M. DePaulo, D. A. Kashy, S. E. Kirkendol, M. M. Wyer, and J. A. Epstein, “Lying in everyday life.,” *Journal of personality and social psychology*, vol. 70, no. 5, p. 979, 1996.
- [3] D. S. McGee and D. J. Cegala, “Patient communication skills raining for improved communication competence in the primary care medical consultation,” 1998.
- [4] E. Donovan-Kicken, T. D. Guinn, L. K. Romo, and L. D. Ciceraro, “Thanks for asking, but let’s talk about something else: Reactions to topic-avoidance messages that feature different interaction goals,” *Communication Research*, vol. 40, no. 3, pp. 308–336, 2013.
- [5] R. Adelson, “Detecting deception,” *APA Monitor on Psychology*, vol. 35, no. 7, pp. 70–73, 2004.
- [6] B. D. Fitch, *The truth about lying: What investigators need to know*. FBI Law Enforcement Bulletin (LEB), FBI, Jun. 10, 2014. [Online]. Available: <https://leb.fbi.gov/articles/featured-articles/the-truth-about-lying-what-investigators-need-to-know>. [Accessed: Feb. 11, 2022].
- [7] P. Ekman and W. V. Friesen, “Nonverbal leakage and clues to deception,” *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

- [8] D. Matsumoto, H. S. Hwang, L. Skinner, and M. Frank, "Evaluating truthfulness and detecting deception," *FBI L. Enforcement Bull.*, vol. 80, p. 1, 2011.
- [9] S. A. McCornack and M. R. Parks, "Deception detection and relationship development: The other side of trust," *Annals of the International Communication Association*, vol. 9, no. 1, pp. 377–389, 1986.
- [10] T. R. Levine, "Truth-default theory (tdt) a theory of human deception and deception detection," *Journal of Language and Social Psychology*, vol. 33, no. 4, pp. 378–392, 2014.
- [11] J. A. Daly and J. M. Wiemann, *Strategic interpersonal communication*. Routledge, 1994.
- [12] A. Vrij and M. Baxter, "Accuracy and confidence in detecting truths and lies in elaborations and denials: Truth bias, lie bias and individual differences," *Expert evidence*, vol. 7, no. 1, pp. 25–36, 1999.
- [13] T. R. Levine, L. N. Anders, J. Banas, K. L. Baum, K. Endo, A. D. Hu, and N. C. Wong, "Norms, expectations, and deception: A norm violation model of veracity judgments," *Communications Monographs*, vol. 67, no. 2, pp. 123–137, 2000.
- [14] J. Fast, *Body Language: The Essential Secrets of Non-Verbal Communication*. MJF Books, 1970.
- [15] F. E. Inbau, J. E. Reid, J. P. Buckley, and B. C. Jayne, *Criminal interrogation and confessions*. Jones and Bartlett, 4 ed., 2004.
- [16] R. Wiseman, C. Watt, L. ten Brinke, S. Porter, S.-L. Couper, and C. Rankin, "The eyes don't have it: Lie detection and neuro-linguistic programming," *PloS one*, vol. 7, no. 7, p. e40259, 2012.
- [17] A. Vrij, *Detecting Lies and Deceit: Pitfalls and Opportunities*. John Wiley & Sons, 2 ed., 2008.
- [18] P. Ekman, *Telling lies: Clues to the Deceit in the Marketplace, Politics, and Marriage*. W.W. Norton & Company, revised ed., 2009.

- [19] P. Ekman and M. Frank, “Lies that fail,” in *Lying and Deception in Everyday Life* (M. Lewis and C. Saarni, eds.), pp. 184–200, New York: Guilford Press, 1993.
- [20] P. Ekman and M. O’Sullivan, “Who can catch a liar?,” *American psychologist*, vol. 46, no. 9, p. 913, 1991.
- [21] B. M. DePaulo and R. L. Pfeifer, “On-the-job experience and skill at detecting deception 1,” *Journal of Applied Social Psychology*, vol. 16, no. 3, pp. 249–267, 1986.
- [22] D. Matsumoto and H. S. Hwang, “Evidence for training the ability to read microexpressions of emotion,” *Motivation and emotion*, vol. 35, no. 2, pp. 181–191, 2011.
- [23] A. Vrij, “Criteria-based content analysis: A qualitative review of the first 37 studies.,” *Psychology, Public Policy, and Law*, vol. 11, no. 1, p. 3, 2005.
- [24] T. Sandoval, D. Matsumoto, H. C. Hwang, and L. Skinner, *Exploiting verbal markers of deception across ethnic lines: An investigative tool for cross-cultural interviewing*. FBI Law Enforcement Bulletin (LEB), FBI, Jul. 8, 2015. [Online]. Available: <https://leb.fbi.gov/articles/featured-articles/exploiting-verbal-markers-of-deception-across-ethnic-lines-an-investigative-tool-for-cross-cultural-interviewing->. [Accessed: Feb. 11, 2022].
- [25] D. Matsumoto, H. C. Hwang, and V. A. Sandoval, “Ethnic similarities and differences in linguistic indicators of veracity and lying in a moderately high stakes scenario,” *Journal of Police and Criminal Psychology*, vol. 30, no. 1, pp. 15–26, 2015.
- [26] P. Ekman, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Henry Holt and Company, 2003.
- [27] S. Porter, L. Ten Brinke, and B. Wallace, “Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity,” *Journal of Nonverbal Behavior*, vol. 36, no. 1, pp. 23–37, 2012.

- [28] A. Vrij, P. A. Granhag, and S. Porter, “Pitfalls and opportunities in non-verbal and verbal lie detection,” *Psychological science in the public interest*, vol. 11, no. 3, pp. 89–121, 2010.
- [29] J. Navarro, “Clues to deceit: a practical list,” *A & A Printing, Tampa*, 2011.
- [30] M. B. Hood and L. J. Hoffman, *Current state of interview and interrogation*. FBI Law Enforcement Bulletin (LEB), FBI, Nov. 6, 2019. [Online]. Available: <https://leb.fbi.gov/articles/featured-articles/current-state-of-interview-and-interrogation>. [Accessed: Feb. 11, 2022].
- [31] F. Goldman-Eisler, “Psycholinguistics: Experiments in spontaneous speech,” 1968.
- [32] J. Navarro, “Detecting lies vs. detecting truth: Serious implications,” *Psychology Today*, 2011.
- [33] C. F. Bond, A. Omar, A. Mahmoud, and R. N. Bonser, “Lie detection across cultures,” *Journal of nonverbal behavior*, vol. 14, no. 3, pp. 189–204, 1990.
- [34] S. Moston and T. Engelberg, “Police questioning techniques in tape recorded interviews with criminal suspects,” *Policing and Society: An International Journal*, vol. 3, no. 3, pp. 223–237, 1993.
- [35] B. C. Jayne and J. P. Buckley, “The reid technique,” in *The Investigator Anthology: A Compilation of Articles and Essays about the Reid Technique of Interviewing and Interrogation*, Chicago: John E. Reid and Associates, 2 ed., 2014.
- [36] S. M. Kassin, “False confessions: Causes, consequences, and implications for reform,” *Current Directions in Psychological Science*, vol. 17, no. 4, pp. 249–253, 2008.
- [37] C. Clarke and R. Milne, *A national evaluation of the PEACE Investigative Interviewing Course*. Home office London, 2001.

- [38] S. M. Kassin, S. C. Appleby, and J. T. Perillo, "Interviewing suspects: Practice, science, and future directions," *Legal and Criminological Psychology*, vol. 15, no. 1, pp. 39–55, 2010.
- [39] R. M. Paulo, P. B. Albuquerque, and R. Bull, "The enhanced cognitive interview: Towards a better use and understanding of this procedure," *International Journal of Police Science & Management*, vol. 15, no. 3, pp. 190–199, 2013.
- [40] B. M. DePaulo, S. E. Kirkendol, J. Tang, and T. P. O'Brien, "The motivational impairment effect in the communication of deception: Replications and extensions," *Journal of nonverbal Behavior*, vol. 12, no. 3, pp. 177–202, 1988.
- [41] B. M. DePaulo and H. S. Friedman, "Nonverbal communication," in *The Handbook of Social Psychology* (D. T. Gilbert, S. T. Fiske, and G. Lindzey, eds.), vol. 1,2, pp. 3–40, New York: McGraw-Hill, 4 ed., 1998.
- [42] J. R. Schafer and J. Navarro, *Advanced interviewing techniques: Proven strategies for law enforcement, military, and security personnel*. Charles C Thomas Publisher, 2016.
- [43] N. E. Dunbar, M. L. Jensen, J. K. Burgoon, B. A. K. J. Robertson, K. M. Kelley, and L. Harvill, "A dyadic approach to the detection of deception," in *Quality in Government and Business Symposium*, p. 17, Citeseer, 2011.
- [44] J. Navarro, *Perspective: Detecting deception*. FBI Law Enforcement Bulletin (LEB), FBI, Aug. 1, 2012. [Online]. Available: <https://leb.fbi.gov/articles/perspective/perspective-detecting-deception>. [Accessed: Feb. 11, 2022].
- [45] M. G. Frank, "Thoughts, feelings, and deception," in *Deception: Methods, Motives, Context, and Consequences* (B. Harrington, ed.), pp. 55–73, California: Stanford University Press, 2009.
- [46] M. G. Frank, M. O'Sullivan, and M. A. Menasco, "Human behavior and deception detection," in *Handbook of Science and Technology for Homeland Security* (J. G. Voeller, ed.), New York: Wiley and Sons, 2007.

- [47] A. Vrij and S. Mann, “Police use of nonverbal behavior as indicators of deception.,” 2005.
- [48] J. H. Kuklinski, P. J. Quirk, J. Jerit, D. Schwieder, and R. F. Rich, “Misinformation and the currency of democratic citizenship,” *The Journal of Politics*, vol. 62, no. 3, pp. 790–816, 2000.
- [49] J. Jerit and Y. Zhao, “Political misinformation,” *Annual Review of Political Science*, vol. 23, pp. 77–94, 2020.
- [50] R. Y. Shapiro and B. Page, “The rational public: fifty years of trends in americans’ policy preferences,” 1992.
- [51] M. A. Hogg and S. A. Reid, “Social identity, self-categorization, and the communication of group norms,” *Communication theory*, vol. 16, no. 1, pp. 7–30, 2006.
- [52] L. Dancey and P. Goren, “Party identification, issue attitudes, and the dynamics of political debate,” *American Journal of Political Science*, vol. 54, no. 3, pp. 686–699, 2010.
- [53] J. Dennis, “Political independence in america, part ii: towards a theory,” *British Journal of Political Science*, vol. 18, no. 2, pp. 197–219, 1988.
- [54] M. Lodge and C. S. Taber, *The rationalizing voter*. Cambridge University Press, 2013.
- [55] D. J. Ahler and G. Sood, “The parties in our heads: Misperceptions about party composition and their consequences,” *The Journal of Politics*, vol. 80, no. 3, pp. 964–981, 2018.
- [56] J. G. Bullock, “Elite influence on public opinion in an informed electorate,” *American Political Science Review*, vol. 105, no. 3, pp. 496–515, 2011.
- [57] J. N. Druckman, E. Peterson, and R. Slothuus, “How elite partisan polarization affects public opinion formation,” *American Political Science Review*, vol. 107, no. 1, pp. 57–79, 2013.

- [58] S. Feldman, L. Huddy, and G. E. Marcus, *Going to war in Iraq: When citizens and the press matter*. University of Chicago Press, 2015.
- [59] S. J. Hill, “Learning together slowly: Bayesian learning about political facts,” *The Journal of Politics*, vol. 79, no. 4, pp. 1403–1418, 2017.
- [60] D. P. Redlawsk, A. J. Civettini, and K. M. Emmerson, “The affective tipping point: Do motivated reasoners ever “get it”?,” *Political Psychology*, vol. 31, no. 4, pp. 563–593, 2010.
- [61] D. E. Clementson, “Effects of dodging questions: How politicians escape deception detection and how they get caught,” *Journal of Language and Social Psychology*, vol. 37, no. 1, pp. 93–113, 2018.
- [62] K. B. Serota, T. R. Levine, and F. J. Boster, “The prevalence of lying in america: Three studies of self-reported lies,” *Human Communication Research*, vol. 36, no. 1, pp. 2–25, 2010.
- [63] P. Bull and K. Mayer, “How not to answer questions in political interviews,” *Political Psychology*, pp. 651–666, 1993.
- [64] S. Harris, “Evasive action: How politicians respond to questions in political interviews,” *Broadcast talk*, vol. 7699, 1991.
- [65] D. Clementson and W. P. Eveland, “When politicians dodge questions: An analysis of presidential press conferences and debates,” *Mass Communication and Society*, vol. 19, no. 4, pp. 411–429, 2016.
- [66] D. Pogue, “How to stamp out fake news,” *Scientific American*, vol. 316, no. 2, pp. 24–24, 2017.
- [67] L. Graves and M. A. Amazeen, “Fact-checking as idea and practice in journalism,” in *Oxford Research Encyclopedia of Communication*, 2019.
- [68] L. Graves and F. Cherubini, “The rise of fact-checking sites in europe,” 2016.
- [69] M. A. Amazeen, “A critical assessment of fact-checking in 2012,” 2013.

- [70] L. Graves, “Boundaries not drawn: Mapping the institutional roots of the global fact-checking movement,” *Journalism Studies*, vol. 19, no. 5, pp. 613–631, 2018.
- [71] M. A. Amazeen, “Revisiting the epistemology of fact-checking,” *Critical Review*, vol. 27, no. 1, pp. 1–22, 2015.
- [72] M. Marietta, D. C. Barker, and T. Bowser, “Fact-checking polarized politics: does the fact-check industry provide consistent guidance on disputed realities?,” in *The Forum*, vol. 13, pp. 577–596, De Gruyter, 2015.
- [73] B. E. Weeks, “Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation,” *Journal of communication*, vol. 65, no. 4, pp. 699–719, 2015.
- [74] D. G. Young, K. H. Jamieson, S. Poulsen, and A. Goldring, “Fact-checking effectiveness as a function of format and tone: Evaluating factcheck.org and flackcheck.org,” *Journalism & Mass Communication Quarterly*, vol. 95, no. 1, pp. 49–75, 2018.
- [75] M. A. Amazeen, E. Thorson, A. Muddiman, and L. Graves, “Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats,” *Journalism & Mass Communication Quarterly*, vol. 95, no. 1, pp. 28–48, 2018.
- [76] K. Fridkin, P. J. Kenney, and A. Wintersieck, “Liar, liar, pants on fire: How fact-checking influences citizens’ reactions to negative advertising,” *Political Communication*, vol. 32, no. 1, pp. 127–151, 2015.
- [77] R. J. Pingree, D. Brossard, and D. M. McLeod, “Effects of journalistic adjudication on factual beliefs, news evaluations, information seeking, and epistemic political efficacy,” *Mass Communication and Society*, vol. 17, no. 5, pp. 615–638, 2014.
- [78] E. K. Vraga and L. Bode, “I do not believe you: How providing a source corrects health misperceptions across social media platforms,” *Information, Communication & Society*, vol. 21, no. 10, pp. 1337–1353, 2018.

- [79] J. Rutter, *High impact factchecking*. Full Fact, Mar. 4, 2016. [Online]. Available: <https://fullfact.org/blog/2016/mar/high-impact-fact-checking/>. [Accessed: Feb. 11, 2022].
- [80] D. MacGuill, *FactCheck: Was Willie O’Dea really not aware of concerns over Lariam, while he was minister?* TheJournal.ie, Oct. 13, 2016. [Online]. Available: <https://www.thejournal.ie/willie-o-dea-lariam-defence-forces-irish-troops-3024987-Oct2016/>. [Accessed: Feb. 11, 2022].
- [81] A. Mantzarlis, *Fact-checking under President Trump*. Poynter, Nov. 10, 2016. [Online]. Available: <https://www.poynter.org/fact-checking/2016/fact-checking-under-president-trump/>. [Accessed: Feb. 11, 2022].
- [82] R. K. Garrett, E. C. Nisbet, and E. K. Lynch, “Undermining the corrective effects of media-based political fact checking? the role of contextual cues and naïve theory,” *Journal of Communication*, vol. 63, no. 4, pp. 617–637, 2013.
- [83] B. Nyhan and J. Reifler, “When corrections fail: The persistence of political misperceptions,” *Political Behavior*, vol. 32, no. 2, pp. 303–330, 2010.
- [84] J. Shin and K. Thorson, “Partisan selective sharing: The biased diffusion of fact-checking messages on social media,” *Journal of Communication*, vol. 67, no. 2, pp. 233–255, 2017.
- [85] D. B. Margolin, A. Hannak, and I. Weber, “Political fact-checking on twitter: When do corrections have an effect?,” *Political Communication*, vol. 35, no. 2, pp. 196–219, 2018.
- [86] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

- [87] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, “Openface: A general-purpose face recognition library with mobile applications,” tech. rep., CMU-CS-16-118, CMU School of Computer Science, 2016.
- [88] T. Baltrušaitis, P. Robinson, and L.-P. Morency, “Openface: an open source facial behavior analysis toolkit,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, IEEE, 2016.
- [89] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 59–66, IEEE, 2018.
- [90] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, “Deception detection using real-life trial data,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 59–66, 2015.
- [91] U. M. Sen, V. Perez-Rosas, B. Yanikoglu, M. Abouelenien, M. Burzo, and R. Mihalcea, “Multimodal deception detection using real-life trial data,” *IEEE Transactions on Affective Computing*, 2020.
- [92] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, “A deep learning approach for multimodal deception detection,” *arXiv preprint arXiv:1803.00344*, 2018.
- [93] J. W. Pennebaker, M. E. Francis, and R. J. Booth, “Linguistic inquiry and word count: Liwc 2001,” *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.
- [94] T. Fornaciari and M. Poesio, “Automatic deception detection in italian court cases,” *Artificial intelligence and law*, vol. 21, no. 3, pp. 303–340, 2013.
- [95] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

- [96] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in neural information processing systems*, vol. 26, 2013.
- [97] H. Kawahara, T. Takahashi, M. Morise, and H. Banno, “Development of exploratory research tools based on tandem-straight,” in *Proceedings: AP-SIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, pp. 111–120, Asia-Pacific Signal and Information Processing Association, 2009 Annual . . . , 2009.
- [98] Z.-H. Tan and B. Lindberg, “Low-complexity variable frame rate analysis for speech recognition and voice activity detection,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 798–807, 2010.
- [99] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.
- [100] P. Mermelstein, “Distance measures for speech recognition, psychological and instrumental,” *Pattern recognition and artificial intelligence*, vol. 116, pp. 374–388, 1976.
- [101] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.
- [102] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, “Bag-of-lies: A multimodal dataset for deception detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.
- [103] S. Venkatesh, R. Ramachandra, and P. Bours, “Robust algorithm for multimodal deception detection,” in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 534–537, IEEE, 2019.

- [104] S. Van der Zee, R. Poppe, P. J. Taylor, and R. Anderson, “To freeze or not to freeze: A culture-sensitive motion capture approach to detecting deceit,” *PloS one*, vol. 14, no. 4, p. e0215000, 2019.
- [105] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, “Spatial transformer networks,” *Advances in neural information processing systems*, vol. 28, 2015.
- [106] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 510–519, 2019.
- [107] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.
- [108] J. Cheng, L. Dong, and M. Lapata, “Long short-term memory-networks for machine reading,” *arXiv preprint arXiv:1601.06733*, 2016.
- [109] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- [110] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [111] A. Datta, S. Sen, and Y. Zick, “Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems,” in *2016 IEEE symposium on security and privacy (SP)*, pp. 598–617, IEEE, 2016.
- [112] C. Frye, C. Rowat, and I. Feige, “Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 1229–1239, 2020.
- [113] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, “Problems with shapley-value-based explanations as feature importance measures,” in *Proceedings of the 37th International Conference on Machine*

- Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 5491–5500, PMLR, 13–18 Jul 2020.
- [114] A. D. Selbst and S. Barocas, “The intuitive appeal of explainable machines,” *Fordham L. Rev.*, vol. 87, p. 1085, 2018.
 - [115] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial intelligence*, vol. 267, pp. 1–38, 2019.
 - [116] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the gdpr,” *Harv. JL & Tech.*, vol. 31, p. 841, 2017.
 - [117] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, “Explainable machine learning in deployment,” in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 648–657, 2020.
 - [118] H. J. Weerts, W. van Ipenburg, and M. Pechenizkiy, “A human-grounded evaluation of shap for alert processing,” *arXiv preprint arXiv:1907.03324*, 2019.
 - [119] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
 - [120] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
 - [121] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634, 2015.
 - [122] N. McLaughlin, J. M. Del Rincon, and P. Miller, “Recurrent convolutional network for video-based person re-identification,” in *Proceedings of*

- the *IEEE conference on computer vision and pattern recognition*, pp. 1325–1334, 2016.
- [123] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natara-jan, “Recurrent convolutional strategies for face manipulation detection in videos,” *Interfaces (GUI)*, vol. 3, no. 1, pp. 80–87, 2019.
 - [124] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.
 - [125] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.
 - [126] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, “Stand-alone self-attention in vision models,” *arXiv preprint arXiv:1906.05909*, 2019.
 - [127] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
 - [128] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
 - [129] A. D. Holan, *The Principles of the Truth-O-Meter: PolitiFact’s methodology for independent fact-checking*. Politifact, Feb. 12, 2018. [Online]. Available: <https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/>. [Accessed: Feb. 11, 2022].
 - [130] A. D. Holan, *7 steps to better fact-checking*. Politifact, Aug. 20, 2014. [Online]. Available: <https://www.politifact.com/article/2014/aug/20/7-steps-better-fact-checking/>. [Accessed: Feb. 11, 2022].

- [131] Z. Zhong, Z. Q. Lin, R. Bidart, X. Hu, I. B. Daya, Z. Li, W.-S. Zheng, J. Li, and A. Wong, “Squeeze-and-attention networks for semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13065–13074, 2020.
- [132] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. Linton, and M. Burzo, “Verbal and nonverbal clues for real-life deception detection,” in *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2336–2346, 2015.
- [133] Z. Wu, B. Singh, L. S. Davis, and V. Subrahmanian, “Deception detection in videos,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [134] P. Ekman and W. V. Friesen, “Facial action coding system,” *Environmental Psychology & Nonverbal Behavior*, 1978.
- [135] P. Ekman, “Darwin, deception, and facial expression,” *Annals of the new York Academy of sciences*, vol. 1000, no. 1, pp. 205–221, 2003.
- [136] R. McCarney, J. Warner, S. Iliffe, R. Van Haselen, M. Griffin, and P. Fisher, “The hawthorne effect: a randomised, controlled trial,” *BMC medical research methodology*, vol. 7, no. 1, pp. 1–8, 2007.
- [137] R. Reiter-Palmon, T. Sinha, J. Gevers, J.-M. Odobez, and G. Volpe, “Theories and models of teams and groups,” *Small Group Research*, vol. 48, no. 5, pp. 544–567, 2017.
- [138] G. Maguolo and L. Nanni, “A critic evaluation of methods for covid-19 automatic detection from x-ray images,” *Information Fusion*, vol. 76, pp. 1–7, 2021.
- [139] J. Sánchez-Monedero and L. Dencik, “The politics of deceptive borders: ‘biomarkers of deceit’ and the case of iborderctrl,” *Information, Communication & Society*, vol. 25, no. 3, pp. 413–430, 2022.
- [140] C. Metz, “Who is making sure the ai machines aren’t racist,” *The New York Times*, vol. 15, 2021.

- [141] J. C. Wong, “Google reportedly targeted people with ‘dark skin’ to improve facial recognition,” *The Guardian*, 2019.
- [142] B. Shneiderman, “Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems,” *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 4, pp. 1–31, 2020.

