Multi-Modal Semantic Inconsistency Detection in Social Media News Posts

Scott McCrae, Kehan Wang, Avideh Zakhor

University of California, Berkeley
{mccrae, wang.kehan, avz}@berkeley.edu

Abstract. As computer-generated content and deepfakes make steady improvements, semantic approaches to multimedia forensics will become more important. In this paper, we introduce a novel classification architecture for identifying semantic inconsistencies between video appearance and text caption in social media news posts. While similar systems exist for text and images, we aim to detect inconsistencies in a more ambiguous setting, as videos can be long and contain several distinct scenes, in addition to adding audio as an extra modality. We develop a multi-modal fusion framework to identify mismatches between videos and captions in social media posts by leveraging an ensemble method based on textual analysis of the caption, automatic audio transcription, semantic video analysis, object detection, named entity consistency, and facial verification. To train and test our approach, we curate a new video-based dataset of 4,000 real-world Facebook news posts for analysis. Our multi-modal approach achieves 60.5% classification accuracy on random mismatches between caption and appearance, compared to accuracy below 50% for uni-modal models. Further ablation studies confirm the necessity of fusion across modalities for correctly identifying semantic inconsistencies.

Keywords: Multi-modal · social media · forensics · fusion

1 Introduction

There has been a great deal of attention on misinformation and deepfakes recently, especially with regards to the COVID-19 pandemic and 2020 US Presidential election. There are a variety of methods for detecting both manipulated media, such as Photoshopped images, and machine-generated data, such as images from generative adversarial networks (GANs) [11, 10, 26, 41, 34, 35, 2, 28]. However, these tools tend to focus on a single modality, such as imagery, and look for clues that the image has been manipulated. While these tools are indisputably useful, we are interested in investigating multi-modal analysis, where we attempt to detect manipulations or misinformation using semantic clues from a variety of modalities.

The use of multiple modalities allows us to reason about the semantic content of each source. For instance, a caption describing an out-of-control protest would be inconsistent with a video of a candle-light vigil. On their own, neither modality is manipulated, but together they represent an inconsistency. This can happen

when an attacker attempts to misrepresent some original source. Furthermore, detecting video semantic inconsistencies is important so that attackers cannot evade deepfake detection by only producing video content.

Detecting caption and video inconsistency is challenging because of the abstract relationships among different modalities. The caption in social media posts is not always a literal description of its corresponding video. Our videos cover a wide range of styles and subjects, are not necessarily well-produced, and have imperfect automatically-generated transcripts with no audio descriptions. We hope to strike a balance between perceived human difficulty and the challenge of learning abstract associations between modalities from a small set of noisy data. We adopt a self-supervised random-swapping approach for generating inconsistencies, in line with the random non-matches generated in [3].

In this paper, we introduce a novel classification architecture for identifying semantic inconsistencies between video appearance and text caption in social media news posts. To analyze the semantic alignment of videos and captions, we need three main ingredients. First, we need pristine data as ground truth. Second, we need to extract semantic feature representations from each modality and its constituents, such as transcripts and named entities. Third, we need to jointly reason about semantic content. Each of these components are discussed in turn in the following sections.

2 Related Works

The capabilities of multi-modal systems have advanced rapidly in recent years. Research in multi-modal learning with text and imagery has demonstrated the efficacy of learning modality-specific embeddings [12]. New methods have been developed with the goal of leveraging transformers to jointly process text and imagery [21, 31, 18, 32]. [24] adapts [37] to include text embeddings which are jointly learned with video embeddings, and is trained on a very large corpus of instructional videos [25]. [23] extends joint text and image transformer-based methods to process text and video clips. [19] employs cross-modal transformers with video frame and text embeddings for multi-modal learning. Recent research has shown promising results adapting transformer methods to process videos [6], opening the door to processing video clips which are longer than a few seconds.

A variety of methods have been introduced recently for detecting computer-generated content and semantic inconsistencies. [40] detects neural fake news by modeling a joint distribution over a news article's domain, date, authors, headline, and body. [34] demonstrates the relative ease of detecting GAN-generated images from a variety of state-of-the-art generators at the time of publication. [33] checks for consistency between a news article and its images and captions. [30] attempts to identify and attribute inconsistencies between images and their captions. [22] introduces and evaluates detection methods on a new dataset for the task of identifying semantic inconsistencies between images and captions.

We introduce a new system in the area of multi-modal semantics, reasoning with video appearance, rather than images, in addition to other modalities like

caption and audio transcript. Specifically, we learn a shared semantic embedding for features extracted from video clips, captions, and transcripts. We then use a recurrent architecture to condense information from several video clips, and concatenate the condensed representation with facial recognition and named entity recognition features before making a final classification. In this manner, we can verify semantic consistency between a caption and a video using information on visual and textual semantics.

3 Motivation and Intuition

To study misinformation and multimedia forensics, we want to learn semantic relationships between video and text in real-world social media content. We opted to create our own dataset to study semantic consistency between many modalities. While there are several popular multi-modal datasets [25, 5, 4, 1, 17], datasets designed for tasks such as human activity recognition or video retrieval are not well-suited to our goal of analyzing inconsistent news in social media.

Instead, we aim to develop a method with capabilities similar to [33, 3, 22], extending semantic inconsistency detection to include videos rather than just text and imagery. Motivated by [12], we aim to learn a semantic embedding for each of the video appearance, caption, and transcript modalities in a social media post. Additionally, we include named entity verification methods inspired by [33]. Since automatic transcriptions may contain typos, we aim to verify names between captions and transcripts by learning a character-based embedding of names in each domain. We also perform facial recognition in an offline manner, by building a database of faces for every name identified in our dataset of captions, then comparing facial recognition features for frames in a video with the facial recognition features for names appearing in the accompanying caption. Our database of faces is collected via Google Images, and features are computed with [29], chosen for its high performance and relatively small feature dimension. We leverage models pre-trained on larger datasets in an effort to alleviate issues with the scale of our relatively small dataset. For instance, while [3] reports results on their full dataset and successively smaller versions, the smallest version reported on is an order of magnitude larger than our full dataset.

4 Method

4.1 System Architecture

We propose a multi-modal model with two stages of fusion, shown in Figure 1. Our pipeline begins with data collection. Then, each modality undergoes a feature extraction step. Captions are passed directly to BERT [8] for feature extraction. Audio is transcribed using DeepSpeech (DS) [13], and then the transcription is passed to BERT. Both caption and transcript are run through a Named Entity Recognition (NER) step to extract the names of people. A separate embedding is learned for each of these text features.

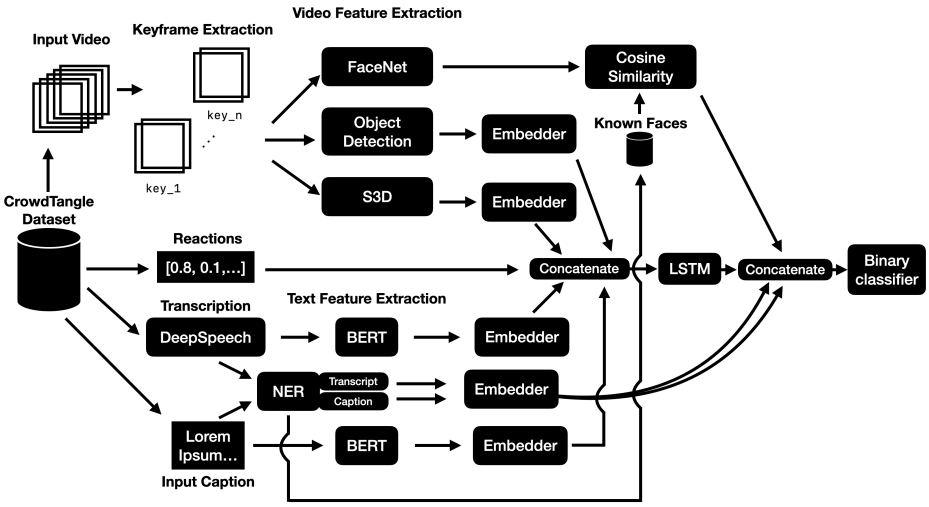

Fig. 1. Our semantic inconsistency detection architecture. Modality-specific feature extraction is run in parallel, and features representing the content of each modality are concatenated with facial verification features in order to perform classification.

Videos are split into clips and undergo several pre-processing steps, described in Section 4.3. We extract both activity recognition and object detection features for each clip, using [14] and [24] respectively, each of which have an additional learned semantic embedding. These embeddings are concatenated with the caption and transcript embeddings.

We also include normalized Facebook reactions to a post as a feature, which we hypothesize provide a measure of sentiment. Normalized reactions are concatenated with the clip, caption, and transcript embeddings. These features are passed to a Long Short-Term Memory (LSTM) [15] module to condense features at the clip level into a summary feature vector for the entire video. We opt to fuse modality features early, before the LSTM, due to the findings of [38].

We add facial verification and name verification features to the fused video, caption, and transcript feature before classification. With all features computed and fused, we make a binary classification using a learned multi-layer perceptron.

4.2 Dataset Design

We construct our dataset using raw data accessed via CrowdTangle (CT) [7], a public insights tool owned and operated by Facebook. The platform can surface public Facebook posts, including sources such as posts by celebrities and news outlets.

Using CT's historical data function, we downloaded all public Facebook posts which had videos in the last decade from the US General Media group, for a total of 647,009 posts. This list of organizations was curated by CT. It ranges from

large, relatively non-partisan sources such as The Associated Press to smaller, more partisan sources such as Breitbart News.

While CT provides access to large amounts of Facebook posts, it has two limitations that impact this project. First, it does not provide labels for whether or not a post contains misinformation. Second, since it does not provide video files, they must be scraped from Facebook using other tools. Therefore, we used CT to source posts to scrape and used the open-source `youtube-dl` tool [39] to scrape video files. Due to this limitation, we were only able to scrape a sample of 4,651 videos.

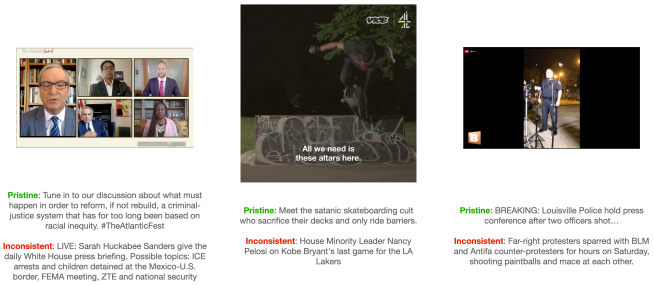

Fig. 2. Example videos and captions from our dataset.

To construct a labelled dataset for multi-modal semantic alignment, we treat the original caption-video post pairs as pristine examples and randomly swap in new captions from other posts to generate inconsistent examples. Examples are shown in Figure 2. In this manner, a pristine example features a real-world video, and associated modalities such as a transcript, and a real-world caption which were intended to relate to each other by the organization which posted them. We assume that pristine examples are semantically consistent across modalities, and that a random swap of caption would result in some amount of semantic mismatch between the new caption and the original video. In practice, half of the examples in our dataset are pristine and half are inconsistent.

We opt to perform swaps on real-world captions rather than creating inconsistencies by generating captions using large language models. This avoids reducing the problem of identifying semantic inconsistencies across modalities to detecting whether or not a caption is synthetically generated. Although some real news posts may include synthetically generated text, such as short reports on financial news [27], we do not attempt to filter out posts which might contain synthetic text. If such synthetic posts are present, they would not be correlated with semantic inconsistency labels due to our random swapping approach.

4.3 Video Pre-Processing

After collecting video data, we standardize video formats for input to our system. Figure 1 illustrates how data flows through our model. Each video is transcoded to a constant resolution of 256×256 pixels and a constant frame rate of 10 frames per second, as in [25], using the FFmpeg utility [9].

Videos from Facebook have a wide range of video lengths, styles, and subjects. In our dataset, the minimum video length is 1 second, the maximum length is 14 hours, and the mean is 8.5 minutes. To handle different video lengths, we adopt a keyframe-based approach. Each video is broken up into a sequence of 32-frame-long clips, with each clip beginning at a keyframe. The clip length was selected based on the recommended parameters of [25].

In practice, we identify keyframes as timestamps in a video where the FFmpeg [9] scene detection filter is triggered, with the scene detection threshold left at the default of 0.4. If no keyframes are detected, which might be the case with videos which are all one shot, we create placeholder keyframes every 32 frames. In this manner, we process as much of a video as possible, even if no keyframes are detected. We choose to use 16 keyframes per video, taking into account that 73% of videos in our dataset have at most 16 keyframes. We did not observe a significant difference in performance between using 8 or 16 keyframes.

Every video is transcribed with DS [13]. Before passing a video's audio stream into DS, we transcode it using FFmpeg to the PCM signed 16-bit little-endian format with a sample rate of 16kHz, apply a highpass filter with cutoff 200Hz, and apply a lowpass filter with cutoff 3kHz. This approach allows us to transcribe the wide range of audio recordings scraped online with an encoding closely matching the training audio for [13]. Below is an excerpt from an example audio transcript with typos generated using DS:

> in ohio on tuesday minnesota senator amicable no time getting back on the campaign trail she picked off with a tour of new hampshire traveling to all ten counties and just thirty hours overcasting a wave of support after snagging the spotlight on tuesday night going head to head against fortune elizabeth warehouses not even the billionaire ...

While our transcripts are mostly correct, they tend to include misspelled names. In this case, misspelled names include "amicable" and "warehouses." The correct names are "Amy Klobuchar" and "Warren." These errors make it difficult to directly compare named entities in captions and transcripts.

4.4 Named Entity Verification

In this section we describe our approaches to verifying named entities using facial verification and text-based comparison of names in captions and transcripts. Our inclusion of named entity verification is motivated by the findings in [33] that named entities can provide strong signals around multi-modal inconsistency.

Facial Verification We define facial verification in this context as checking whether or not people named in the caption of a video actually appear in the video. To accomplish this, we identify people in captions and build a database of representations for them. People are identified with the named entity recognition (NER) feature in the spaCy [16] natural language processing library. Using spaCy's `en_core_web_trf` language model, which implements RoBERTa [20], we run NER on all captions, and take all strings with the `PERSON` label as names of

people. These strings are compiled into a set of people whose names appear in our dataset.

Once all names are identified, we compute a representation for each person. First, we query Google Images for the top 10 results for each name, and consider them ground-truth references for the visual appearance of each name. Having multiple images per name allows us to capture potentially diverse lighting conditions, poses, ages, and camera angles.

Once reference images are collected, we use FaceNet [29] to compute facial recognition features for each image, selected for its relatively small feature size. Figure 1 shows how FaceNet features are used in our model. At inference time, FaceNet features are computed for a video's keyframes. We then take the cosine similarity between the features for names appearing in the caption and the features for each keyframe in the video. In practice, these keyframe features are pre-computed for efficiency. The similarity scores are passed on to our model's classification head to be used alongside features from other modalities.

This approach to person identification has a few drawbacks. The reference images of named entities from Google Images are not manually curated, and multiple people can appear in one single reference image. Additionally, in some cases, an individual might be referenced first by their full name, i.e. "Alice Appleseed," and then only by their first name, "Alice." Our NER approach does not account for this, but it is less of a problem for well-known individuals who can often be uniquely identified by their first or last name, such as "Kanye West" and "Kanye," or "Nancy Pelosi" and "Pelosi."

Name Verification We also compare names in captions to audio transcripts, which provides an extra signal and can alleviate the problem where an individual might be a topic of discussion, rather than a visual subject.

We find that many names in audio transcripts have spelling errors but high phonetic similarity with their corresponding names in the captions. Therefore, to achieve fuzzy name matching, we compute learnable, character-based embeddings for the names which appear in captions and/or transcripts.

Given a string representing a named entity, we convert each character to its lower-case ASCII numerical value and pad to a maximum length of 64 characters. In our dataset, 100% of strings identified as names have at most 64 characters. We then feed this vector into a 2-layer fully connected network, with hidden size 64 and output size 32. These name embeddings are then passed on to our classification head for use along with other modalities, as shown in Figure 1.

By taking in the numerical values of each character of a name, our embedding can learn to match phonetic patterns in names, and the patterns in which DS generates vowels and consonants for sounds in names. Thus, the embedding is able to approximate a textual name to sound conversion.

4.5 Facebook Reactions

Since our data is collected from Facebook, we have access to the Facebook reactions for each post. In Facebook, users can react to a post with: Like, Love, Wow,

Haha, Sad, Angry, and Care. We hypothesize that reactions can provide a coarse measure of the semantics of an entire post, considering all of its modalities.

We take the normalized reactions as an input feature, shown in Figure 1. To normalize reactions, we divide the raw count of each reaction by the total number of reactions a post received, so the model can ignore a post's popularity.

4.6 Ensemble Feature Extraction

We adopt a uni-modal ensemble approach to multi-modal fusion, as shown in Figure 1. To classify whether or not a post is inconsistent, we take as input a video, a transcript, the normalized reactions to the video's pristine post, and a caption. In addition to the named entity verification features described in Section 4.4, we compute features for the caption, transcript, and video clip inputs.

Both the audio transcript and caption are processed using a pre-trained BERT [8] language model, implemented by HuggingFace [36]. When using the language model, inputs are truncated to their first 1024 characters, and split into two sets of characters with length 512 to accommodate the language model's maximum input length. In our dataset, 60% of audio transcripts and 99.97% of captions have at most 1024 characters.

Videos are processed using both a video-understanding network and an object detection network. For video understanding, we use S3D-MIL-NCE (S3D) [24], and for object detection, we use a ResNet50 model [14]. S3D is run on the full 32-frame sequence in each of the video clips, split by keyframe, while ResNet is run on each keyframe. We use the `mixed_5c` output of S3D, as recommended.

4.7 Multi-Modal Fusion

For each modality, we learn an embedding to a semantic latent space, as shown in Figure 1. Each embedding function is implemented as a 2-layer fully connected network, mapping from the output feature space of a feature extraction network to a common 256-dimensional latent space. The learned semantic embeddings for video clips, transcripts, and captions are concatenated and passed through a Long Short-Term Memory (LSTM) [15] module to condense information from the clips into one summary feature vector. This fuses multi-modal content at the clip level, before the output of the LSTM is concatenated with named entity verification features. This fusion approach is motivated by the early fusion methods proposed in [38]. The final combined feature vector is passed to our classification network. Our classifier is a 3-layer fully connected network, with input size 1096, hidden layer sizes 512 and 128, and output size 2.

5 Experiments

5.1 Experimental Design

We train our model with the dataset described in Section 4. We optimize the binary cross-entropy loss function, where our model classifies caption, audio transcript, and video appearance tuples as either pristine or inconsistent.

We report classification accuracy for our experiments, computed as the percentage of examples correctly identified as either pristine or inconsistent in our balanced test set. Our data is split such that 15% of the examples are reserved for the test set, and the other 85% for training and validation.

5.2 Results and Ablation Studies

Table 1. Binary classification accuracy (%) of heavily multi-modal models

	Modality or Feature Removed							
Model	Names & Faces	Caption	Names	Video	Transcript	Faces	Reacts	None
Full	49.8	54.2	52.4	54.7	57.0	56.9	57.4	58.3
No OD	49.9	51.5	54.8	56.5	59.5	59.6	**60.5**	**60.5**

Table 2. Best model confusion matrix (%)

	Predict Pristine	Predict Inconsistent
Pristine Examples	51.0	49.0
Inconsistent Examples	28.6	71.4

Table 3. Binary classification accuracy (%) of uni- and bi-modal models

Modalities Used				
Caption & Video	Video	Caption	Faces	Names
49.6	49.8	49.9	51.7	**53.5**

We perform a variety of ablation experiments to characterize the impact of each modality on the accuracy of our model. The authors are not aware of directly comparable work which detects semantic inconsistencies in the modalities included here, nor directly applicable benchmarks. Results are shown in Table 1, with each modality removed one-by-one. Due to the fact that removing object detection features improved model performance, we perform one-by-one removal ablation studies again, with object detection features always removed. These experiments are referred to as "No OD" models in Table 1. "Removing" a modality refers to removing its features or embeddings from our classifier. For instance, removing video appearance makes the semantic video embeddings inaccessible to our classifier, although facial verification is still performed.

As seen in Table 1, best performance is achieved by using all modalities, except object detection features, and reaches classification accuracy of 60.5%. Table 2 shows the confusion matrix for this model. We observe that the model is more accurate when classifying inconsistent examples. Specifically, it can correctly detect inconsistency 71% of the time, and detects consistency 51% of the time. Table 3 shows results for models using one or two modalities.

We observe that named entities are key to model accuracy, as seen in Table 1, further confirming the importance of named entities demonstrated in [33]. Without facial verification, classification accuracy decreases slightly to 59.6%. Without comparing names between captions and transcripts, classification accuracy falls to 54.8%. Without either consistency check, accuracy falls to 49.9%. We find that named entities are not the only useful information provided by captions. As seen in Table 1, when caption embeddings are removed, accuracy falls to 54.2% and 51.5%, with and without object detection (OD) features, respectively. Combining of semantic embeddings and named entity verification is the best use of the information in the caption modality.

We note that video embeddings from S3D are more important than OD embeddings from ResNet. In fact, removing OD embeddings improves accuracy, while removing S3D embeddings lowers accuracy. When OD embeddings are present, removing S3D embeddings leads to 3.8% lower accuracy, and without OD embeddings, removing S3D embeddings leads to 4% lower accuracy. It could be that S3D features contain much of the relevant OD feature information for our task. Additionally, OD features are not temporally aware. Furthermore, the ResNet50 model we take features from is trained for image classification, which may be too general to be useful for modelling abstract video semantics.

We observe that Facebook reactions do not seem to provide a useful signal.

Finally, we observe that multi-modal fusion is necessary for achieving the best possible accuracy. Removing any one of our modalities decreases performance, with the exception of reactions. No uni-modal model can perform better than random; accuracy for uni- and bi-modal models is shown in Table 3. Caption-only and video-only models achieve 49.9% and 49.8% classification accuracy, respectively, confirming that our dataset does not have linguistic or visual bias. A model combining caption and video clip embeddings achieves 49.6% accuracy, highlighting the importance of incorporating additional modalities and features. A model which solely compares named entities in captions and transcripts achieves 53.5% accuracy, and a model which compares named entities in captions with facial verification features achieves 51.7% accuracy. While named entities are important, they are not sufficient to achieve the best results.

6 Conclusion

We have introduced a novel multi-modal semantic inconsistency detection system for use in real-world social media posts. We demonstrate the importance of making use of modalities beyond video appearance and captions, including transcripts, facial verification, and fuzzy named entity comparison.

We observe that fusion across modalities is key to detecting semantic inconsistencies. We find that named entities provide strong signals for detecting inconsistency, and that verifying named entities using both language-based and visual methods is better than only using one. Semantic consistency checks cannot be fully explained by named entity verification, however, highlighting the need to consider semantic embeddings for language and video.

Future work could explore attributing and characterizing inconsistencies. Modules for explainable facial verification and author attribution could take steps towards addressing this. Our approach would likely benefit from more data, and we are interested in expanding data collection to other social networks. Increasing the size of our dataset might also allow for more challenging inconsistencies during training time.

References

[1] Sami Abu-El-Haija et al. "YouTube-8M: A Large-Scale Video Classification Benchmark". In: *ArXiv* abs/1609.08675 (2016).

[2] Shruti Agarwal et al. "Detecting Deep-Fake Videos from Appearance and Behavior". In: *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. 2020, pp. 1–6.

[3] Shivangi Aneja, Chris Bregler, and Matthias Nießner. "COSMOS: Catching Out-of-Context Misinformation with Self-Supervised Learning". In: *ArXiv* abs/2101.06278 (2021). arXiv: 2101.06278 [cs.CV].

[4] Stanislaw Antol et al. "VQA: Visual Question Answering". In: *International Conference on Computer Vision (ICCV)*. 2015.

[5] A. Araujo et al. "Stanford I2V: A News Video Dataset for Query-by-Image Experiments". In: *ACM Multimedia Systems Conference* (2015).

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. *Is Space-Time Attention All You Need for Video Understanding?* 2021. arXiv: 2102.05095.

[7] CrowdTangle Team. *CrowdTangle*. Facebook, CA, United States, 2021.

[8] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[9] FFmpeg Developers. Version 4.3.1. 2020. URL: http://ffmpeg.org/.

[10] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. "DeepFake Detection by Analyzing Convolutional Traces". In: *CVPR*. June 2020.

[11] David Güera and Edward J. Delp. "Deepfake Video Detection Using Recurrent Neural Networks". In: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2018, pp. 1–6.

[12] Amirhossein et al. Habibian. "Video2vec Embeddings Recognize Events When Examples Are Scarce". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017), pp. 2089–2103.

[13] Awni Hannun et al. *Deep Speech: Scaling up end-to-end speech recognition*. 2014. arXiv: 1412.5567 [cs.CL].

[14] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[15] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667.

[16] Matthew Honnibal et al. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. DOI: 10.5281/zenodo.1212303.

[17] Will Kay et al. "The Kinetics Human Action Video Dataset". In: *ArXiv* abs/1705.06950 (2017).

[18] Gen Li et al. *Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training*. 2019. arXiv: 1908.06066 [cs.CV].

[19] Linjie Li et al. "HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training". In: *EMNLP*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2046–2065.

[20] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[21] Jiasen Lu et al. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. 2019. arXiv: 1908.02265.
[22] G. Luo, T. Darrell, and A. Rohrbach. *NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media*. 2021. arXiv: 2104.05893.
[23] H. Luo et al. "UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation". In: *arXiv:2002.06353* (2020).
[24] Antoine Miech et al. "End-to-End Learning of Visual Representations from Uncurated Instructional Videos". In: *CVPR*. 2020.
[25] Antoine Miech et al. "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips". In: *ICCV*. 2019.
[26] Trisha Mittal et al. "Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20. Seattle, WA, USA: Association for Computing Machinery, 2020, pp. 2823–2832. ISBN: 9781450379885.
[27] Jaclyn Peiser. "The Rise of the Robot Reporter". In: *The New York Times* (Feb. 5, 2019).
[28] A.C. Popescu and H. Farid. "Exposing digital forgeries by detecting traces of resampling". In: *IEEE Transactions on Signal Processing* (2005), pp. 758–767.
[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A unified embedding for face recognition and clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).
[30] Ravi Shekhar et al. "FOIL it! Find One mismatch between Image and Language caption". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, July 2017, pp. 255–265.
[31] Weijie Su et al. *VL-BERT: Pre-training of Generic Visual-Linguistic Representations*. 2020. arXiv: 1908.08530 [cs.CV].
[32] Hao Tan and Mohit Bansal. "LXMERT: Learning Cross-Modality Encoder Representations from Transformers". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019.
[33] Reuben Tan, Bryan A. Plummer, and Kate Saenko. "Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2020.
[34] Sheng-Yu Wang et al. "CNN-generated images are surprisingly easy to spot...for now". In: *CVPR*. 2020.
[35] Sheng-Yu Wang et al. "Detecting Photoshopped Faces by Scripting Photoshop". In: *ICCV*. Oct. 2019.
[36] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *EMNLP: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.
[37] Saining Xie et al. *Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification*. 2018. arXiv: 1712.04851 [cs.CV].
[38] Huijuan Xu et al. "Multilevel Language and Vision Integration for Text-to-Clip Retrieval". In: *AAAI* 33.01 (July 2019), pp. 9062–9069.
[39] *youtube-dl*. Version 2021.01.24.1. 2021. URL: https://youtube-dl.org.
[40] Rowan Zellers et al. "Defending Against Neural Fake News". In: *Advances in Neural Information Processing Systems 32*. 2019.
[41] Hanqing Zhao et al. "Multi-Attentional Deepfake Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 2185–2194.