Out-of-Context Caption Generation

Anurag Singh
Technical University of Munich

Abstract

With the increasing influence of social media, online misinformation has grown to become a societal issue. To spread misinformation, adversaries rely on several methods ranging from realistic-looking deepfakes to less compute-intensive methods like cheapfakes. The motivation of our work comes from the threat caused by cheapfakes, where an unaltered image is described using a news caption in a new but false-context. This false-context can be interpreted as an out-of-context caption to spread misinformation. The main challenge in detecting such out-of-context multimedia is the unavailability of large-scale datasets due to annotation of out-of-context multimedia being a cumbersome process. Several detection methods employ randomly selected captions to generate out-of-context training inputs. However, these randomly matched captions are not truly representative of out-of-context scenarios due to inconsistencies between image description and the matched caption. We aim to address these limitations by introducing a novel task of out-of-context caption generation. In this work, we propose a new method that generates realistic out-of-context caption given a visual and textual context. We also demonstrate that semantics of the generated captions can be controlled using the textual context. We also evaluate our method against several baselines and our method improves over image captioning baseline by 6.2% BLUE-4, 2.96% CiDEr, 11.5% ROUGE, and 7.3% METEOR[1].

1. Introduction

With recent improvements in several computer vision tasks, manipulation of media has become more realistic threat [15, 29]. Therefore, many works have recently focused on forensic methods for detection of such deepfakes-based media manipulations [2, 1, 37, 30, 25, 18, 40, 42, 6, 9] However, one of the most prevalent ways to spread misinformation to date are cheapfakes, requiring very little to no computing resources compared to deepfakes [10]. In this work, we focus on a specific type of cheapfakes, where an unaltered image is used out-of-context with a fake caption to spread misinformation. There has been some recent progress in the detection of out-of-context misinformation [5, 17, 4, 13, 31] Most of the detection works also propose a method to create these out-of-context captions for the training of detection models, often by selecting other captions at random [23, 4, 16]. Others use hypernamed text or similarity methods to select out-of-context captions from a reference set [13, 31]. A study in [20] illustrates that these text manipulations lead to linguistic biases which can be detected without image inputs.

To address these challenges, we propose the task of out-of-context caption generation i.e. generating an out-of-context caption for an image given some contextual input. Our task is different from the task of image captioning since unlike image captions news captions are not described by the images alone. To address this, we condition our out-of-context caption on conditional tokens also. The captions generated by our method can by used for downstream tasks such as helping improve out-of-context detection models. The proposed model can also be fine-tuned for the downstream task as a plugged in module. This approach significantly differentiates our method from state-of-the-art methods for generating out-of-context multimedia [20] which focus on matching existing unmanipulated text and images in the dataset to obtain out-of-context multimedia.

Semantics matching-based methods match a query to a caption in their reference set [3, 21, 20]. Therefore, they cannot create captions for unseen named entities outside their reference set. We address this issue using a byte-pair encoding-based captioning module that can process out-of-vocabulary tokens. Additionally, matching-based methods involve some form of linear search within the dataset to match an image to a semantically similar out-of-context caption. Search makes these methods computationally expensive in practice which are also addressed by generating captions instead. In summary, our main contributions are as follows:

- We propose a new method to generate realistic out-of-context captions given an input image and some conditional word tokens as contextual input.

[1] https://github.com/Anurag14/OutofContextCaptioning

- We demonstrate that it is possible to control the semantics and context of the generated captions using conditional word tokens.
- Our method addresses the limitations of the previous matching-based methods by generating captions for out-of-vocabulary conditional word tokens.

2. Related Work

2.1. Misinformation Detection

In recent years, models to detect cheapfakes have started to gain attention [31, 5, 17, 16, 4]. In [5], authors propose self-supervised training method for out-of-context caption detection. Their model utilizes visual grounding of an image objects in the associated captions and the similarity in contexts of different captions for self supervision. The triplets where captions are grounded in same objects but have different contexts are marked as out-of-context pairs. They also introduce a new large-scale dataset for out-of-context news detection containing news examples collected from different news outlets. Each news example has an image and corresponding news caption. Most of the collected data does not have context annotations. FakenewsNet [32] and Fakeddit [24] focuses on detection of human-made fake news.

While these works have important real-world use cases for the task of out-of-context multimedia, our focus is on exploring a targeted out-of-context caption where given an image and conditional word tokens, we can automatically generate a new out-of-context caption. Our method can address the limitations of randomly selecting out-of-context captions from the dataset by generating a targeted out-of-context caption for the example.

2.2. Generating Out-of-Context Multimedia

For generating out-of-context multimedia, matching out-of-context captions in the reference set to corresponding query images and vice versa is done [20] [13] [31]. Matching is performed by building a large-scale dataset that acts as a reference set. We try to review several of these datasets in detail. The authors propose MAIM [13] and MEIR [31] datasets for identifying image re-purposing or discrepancies in the joint semantics of image and text. MAIM takes random captions of other images to create false image caption pairs. MEIR tries to select captions by swapping named entities of people, organizations, and locations in the caption to create false image caption pairs. Recent works [23] [20] have released large scale datasets for generating out-of-context multimedia. Authors in [23] release a dataset that has swapped named entities of people, location, and events with other random entities. Authors in [20] argue that this handcrafting from hyper-named text to create out-of-context media generates linguistic biases that are easy to identify. To overcome this limitation [20] proposes a clip-based matching approach for text manipulation. However, matching methods are computationally expensive as obtaining an out-of-context media for a query involves some form of search within the dataset.

These methods assume that for each multimedia image caption pair, another semantically similar pair exists in the reference set whose caption can replace the original caption. Our problem statement is significantly different, as we do not assume the availability of any reference set. Instead, for each image some input named entities act as conditional word tokens to generate the caption.

2.3. News Captioning:

The task of news captioning involves generating news captions using the news article and corresponding image as input. In [33] authors try to solve the task of fake news captioning by building a model which can replace the entire real news articles with text generated from a large language models like Grover [41]. However, they do not aim to mismatch the images which are relevant to the news article's content. Analysis of images for this method has limited impact as analyzing articles and captions of the news is the best way to ensure good out-of-context detection performance. Apart from generating fake news articles, several works have focused on automating news captioning of real news [7] [34]. Both these methods take a news article and a corresponding image as input to generate a new caption for the article.

Our task is significantly different from the previous work in news captioning literature since we do not assume any access to complete news articles for our task. At the test time, our model does not need captions, and tokenized named entities can act as input to the model.

3. Method

We propose an end-to-end architecture to generate out-of-context captions using an input image and conditional tokens. The main components of our model architecture are 1) Named Entity Recognition 2) Feature Extraction & Detection Backbone 3) Relational Graph 4) Captioning Module.

3.1. Named Entity Recognition (NER)

The first step for generating our out-of-context captions is conditioning them on conditional word tokens. During training, we parse the input captions to classify named entities in the text. We use SpaCY NER [12] for processing the captions to identify named entities. This parsing of captions into named entities is only pre-processing of our dataset. We want to highlight that our model does not need the entire caption as input and can work using named entities only. SpaCY classifies input tokens into 18 different categories of

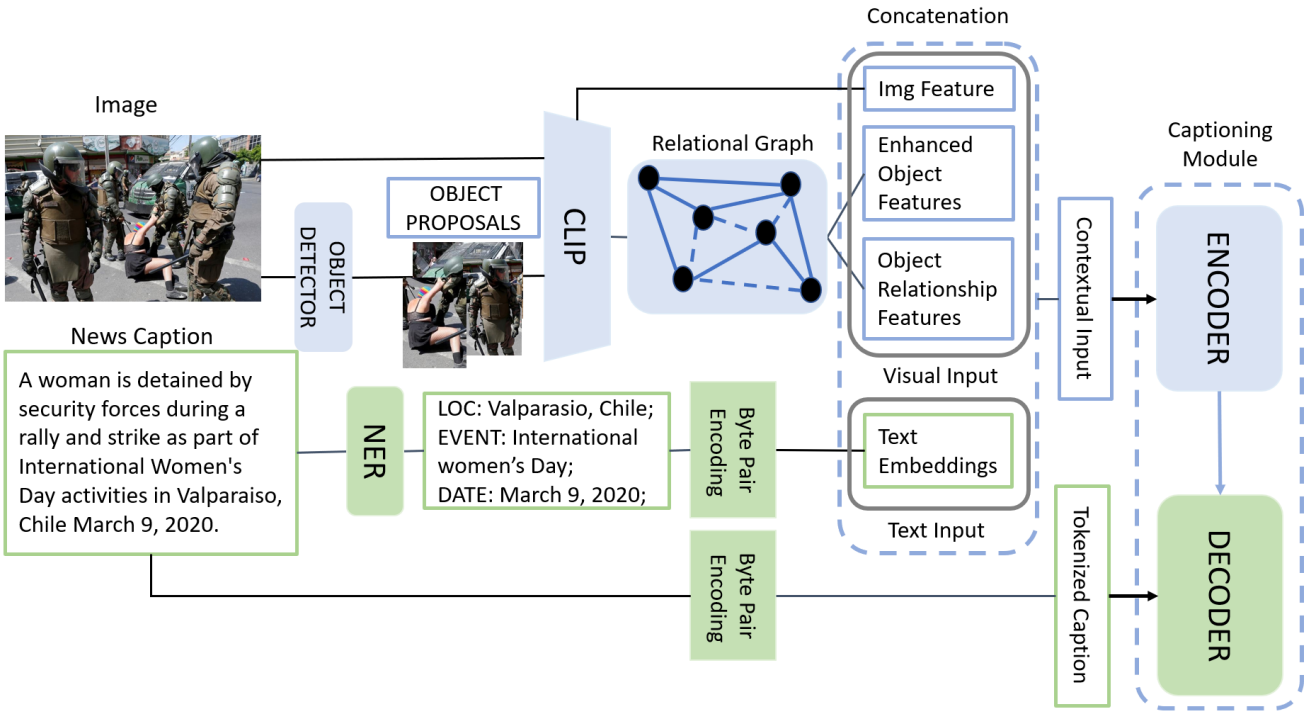

Figure 1: The semantic figure describes image and news caption input that are pre-processed using Object-detection, CLIP, and Named Entity Recognition respectively to obtain image and object level features along with a named entity dictionary. The object features are enhanced with a relational graph. The named entities are encoded into text embedding using byte-pair encoding(BPE). The embeddings act as the input to the encoder of the captioning module. Similarly, the original news caption is tokenized using BPE to form input to the decoder during training, and CE loss is used to optimize the model.

named entities. This results in a dictionary D^{NER} with keys as named entity types and tokens of named entities belonging to corresponding types as values.

3.2. Feature Extraction & Detection Backbone

From the image input, we extract image features using a CLIP [3] visual encode denoted as $\mathcal{I}$. We also detect all probable objects in the given image using an attention-based object detector DETR [8]. The object proposals are extracted from the image using the predicted bounding boxes. These object proposals are encoded into object features using a CLIP visual encoder.

3.3. Relational Graph

Describing the actions and events in an image often involves understanding the relationship of an object's appearance in context with other objects in the scene. To capture the semantics of such actions and events in the generated out-of-context caption, we employ a graph neural network with message passing to obtain enhanced object features and to extract object relation features from the edges of the relationship graph. We create a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where we

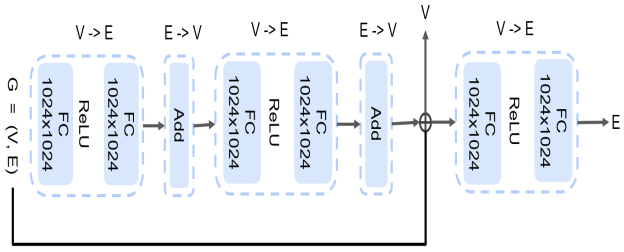

Figure 2: The relational graph module takes CLIP image encoding from object proposals as input that represent nodes in the graph.

represent the object features as nodes and the relationship between the objects is represented by edges between the nodes of the graph. We construct the graph where we consider at most K relationships for an object, in other words, each node as K number of edges. We use standard neural message passing [11] where the message passing at graph step τ is defined as follows:

$$\mathcal{V} \rightarrow \mathcal{E} : e_{i,j}^{\tau+1} = f^{\tau}([v_i^{\tau}, v_j^{\tau} - v_i^{\tau}]) \quad (1)$$

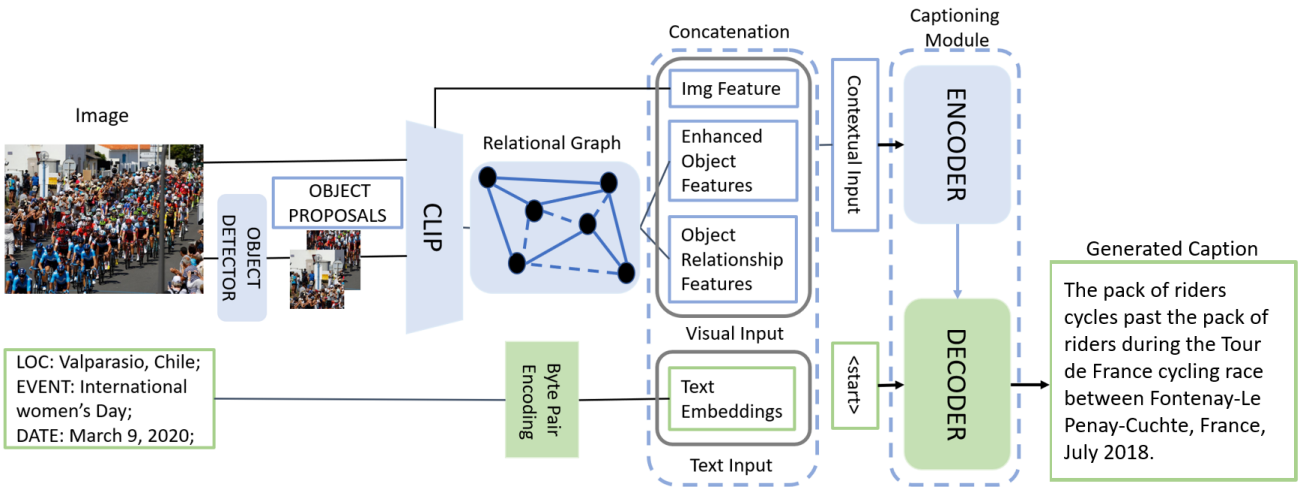

Figure 3: The semantic figure describes the test time image and conditional word token as input to our model. Image is processed using Object-detection and CLIP. Byte pair encoding converts word tokens into text embeddings. These representations form an input to the encoder of the captioning module. We condition the decoder using a start token that denotes the start of a sentence. It then generates a caption in an auto-regressive fashion.

Where for nodes i, j we denote features at step τ as $v_i^\tau \in \mathcal{R}^{1024}$ and $v_i^\tau \in \mathcal{R}^{1024}$ respectively. $e_{i,j}^{\tau+1} \in \mathcal{R}^{1024}$ represents the edge relationship message between nodes i and j in the next graph step $\tau + 1$. The hard brackets $[\cdot, \cdot]$ represent the concatenation of two vectors, and the function f() is an MLP. The node features are aggregated from the messages post every message passing step as $\mathcal{E} \rightarrow \mathcal{V} : v_i^{\tau+1} = \sum_{k=1}^{K} e_{i,k}^{\tau}$. The node features in the last step are summed to the original features via a skip connection and output as enhanced object features $\mathcal{V}^\tau$. An additional message passing layer after the final layer outputs message $\mathcal{E}^{\tau+1}$ as the object relation features.

3.4. Captioning Module

Our captioning module consists of a 3-layer Transformer with an Encoder and Decoder where we feed the visual feature input concatenated to the embeddings of conditional tokens to the Encoder. The byte pair encoded embeddings of the caption tokens are fed as input to the Decoder.

3.4.1 Encoder

We concatenate the image feature vector with the object feature vectors to obtain total visual feature vectors X^{vis}.

$$\mathcal{X}^{vis} = [\mathcal{I}, \mathcal{V}^\tau, \mathcal{E}^{\tau+1}] \tag{2}$$

These total visual feature vectors X^{vis} acts as input to the encoder. We also consider byte-pair encoding (BPE) to encode the tokens into word embedding. We use pre-trained GPT-2 BPE for our task. The effectiveness of BPE for our task is that it allows us to handle out-of-vocabulary words as compared to word2vec[22] or GLOVE[28]. We compute the textual features of the full dictionary $\mathcal{D}^{NER}$. We use the BPE of the named entity types to allow our model to understand the context of each conditional token. For each named entity type, we compute the BPE of named entity type and its corresponding tokens which are then concatenated. This concatenation is followed by the embedding of the end token. This allows the model to differentiate between different categories of named entities. The BPE of full dictionary is obtained by concatenating all the vectors.

3.4.2 Decoder

A stack of 3 identical layers constitute the decoder of the captioning module. Similar to architecture in [35], decoder receives features from the encoder to compute cross attention. In addition to the byte-pair encoding, we also use a pre-trained position encoding from GPT-2. Position encoding helps in understanding the context of the token in the caption. The final layer generates a prediction over the vocabulary to generate the output token. The decoder masks the embeddings of tokens ahead in the caption while predicting the output of the current token.

4. Implementation Details

We implement our architecture using PyTorch [27] and train end-to-end using ADAM [14] with a learning rate of 1e-3. We train the model for 100 epochs until convergence. We extract a maximum of 10 objects for each image. We

pad the object features with 0 for images with fewer objects. We consider token length of 20 for our named entity texts. For the examples with less than 20 tokens, we pad the token embeddings with 0. We use the token length in our model to mask attention on the padding. We truncate input training captions longer than 100 tokens and add $start$ and end tokens from the byte-pair encoding of GPT-2 to indicate the start and end of the caption.

5. Experiments

Dataset: We use the COSMOS dataset [5] of out-of-context news captions which contain images and corresponding news captions from different news outlets. COSMOS consists of 200k images with 450k different captions, corresponding to the true captions from news outlets and also captions corresponding to fake news captions. COSMOS is a dataset designed for self-supervised out-of-context caption detection, therefore, the annotations of true and fake captions are absent.

Train & Val Split: We follow the standard train and val split from COSMOS [5] i.e. we use 160K images and corresponding captions to train our model. We use the val set to test our model since the val set is much larger with 40k examples in comparison to test set which only contains 1700 examples. We construct a smaller dataset from COSMOS for our task where we select a single caption corresponding to each image in our training and validation test set.

Metrics: To evaluate the quality of generated captions, we used evaluation metrics commonly used for the tasks of image captioning such as BLEU-4 [26], CiDEr [36], METEOR [36] and ROUGE [19]. Here BLEU-4 and METEOR allow us to evaluate the effectiveness of the sentence, CiDEr allows us to evaluate the quality of the sentence with respect to the image description, and ROUGE allows us to evaluate the automatic summarization/recall of the generated caption with respect to the caption from which named entities are extracted.

Baselines: To evaluate the effectiveness of our method we implement strong baselines using LSTM Seq-2Seq architecture [38] and with attention [39] by modifying them to our task where we use them as our captioning module. For a fair comparison, we use GPT-2 pretrained byte-pair encoding in all the methods. We feed conditional input to Seq-2-Seq-based architecture with our visual features and textual features as input to the encoder which outputs the context embedding that conditions the LSTM decoder to generate out-of-context caption. We also consider 3 layer Transformer baseline with image features and text encoded using BPE but no named entity types and relationship graph as another strong baseline for comparison.

5.1. Quantitative Results

Method	Bleu-4	Cider	Rouge	Meteor
LSTM	27.73	36.89	29.6	12.8
LSTM+Attention	28.9	40.8	30.2	13.4
Transformer	31.2	47.14	33.0	15.1
Ours	**37.4**	**50.1**	**41.5**	**22.4**

Table 1: Comparison of out-of-context caption descriptions using image captioning metrics with modified LSTM Seq-2-Seq architecture with byte-pair encoding, LSTM +attention with byte-pair encoding

We compare our method to baselines on the COSMOS while training on full train set and evaluating on official val set. The results of our method are presented in the Table 1.

5.2. Qualitative Results

Comparison of baselines: In Figure 4 we qualitatively analyze the captions generated by different baseline models. We observe that the use of the Transformer allows for better inclusion of the textual context in the generated caption. In other words, captions are better conditioned on the input text as dates are correct for the transformer model. We observe that using object features and relationship graph allows for better grounding of the caption in the image as transformer model baseline without object features and relation graph mistakes man for women in the caption. The relation graph allows for the model to discover the object relationship between man and the police. This allows the model to include the semantics of man being escorted in the caption.

Playing around with the input: In Figure 5, we try to qualitatively analyze captions as we change the conditional word tokens for an unseen image. In this Figure, the NER context denotes the named entities extracted from the corresponding caption in the dataset. We observe for different cases how caption changes as we change the named entities in the context. In the left example of Fig 5, we observe that adding the word quiet to the context allows a change in the semantics of the generated caption. Also removing the word street from the context, the generated caption still describes the street from the visual input. A similar observation is also made for the word token Friday. In the right example of the fig 5, input from the image allows the model to understand tear gas, protest, and police and use these in the generated caption. Further, changing the context to India and Delhi allows controlling the location of the event.

Context: 'Israeli', 'Israel', 'Gaza', 'May 14, 2018'
Full context: NORP Israeli; GPE Israel, Gaza; DATE May 14, 2018
Original: Israeli soldiers patrol near a burning field on the Israeli side of the border between Israel and Gaza , may 14 , 2018 .
LSTM: israeli soldiers stand next to a burning barricade *during clashes with israeli troops* at a protest marking the 69th anniversary of nakba, near the jewish settlement of beit el, near the west bank city of ramallah. **LSTM+Attention**: Israeli soldiers take cover from teargas *fired by Israeli troops* during clashes at a military base in the southern Gaza Strip May 14, 2018. **Transformer**: Israeli soldiers take cover as they take part in a military drill in Gaza May 14, 2018.
Ours: Israeli soldiers take cover during clashes on Israel-Gaza border, May 14, 2018.

Context:'Russian', 'Douma', 'Damascus', 'Syria', 'Ghouta', 'January 10, 2016'
Full context: NORP Russian GPE Douma Damascus Syria ORG Ghouta DATE January 10, 2016
Original: A man carries a child that survived from under debris in a site hit by what activists said were airstrikes carried out by the Russian air force in the town of Douma , eastern Ghouta in Damascus , Syria January 10 , 2016.
LSTM: A man carries a child as he walks through rubble during a protest in the rebel held Douma neighborhood of Damascus, Syria.
LSTM+Attention: A man carries his injured son after what activists said was *shelling by civilians* in the besieged town of Douma, eastern Ghouta in Damascus, Syria January 10, 2016.
Transformer: A Russian soldier carries a wounded man in the besieged town of Douma, eastern Ghouta in Damascus, Syria January 10, 2016.
Ours: A man is helped by a Russian helicopter in the besieged town of Douma, eastern Ghouta in Damascus, Syria, January 10, 2016.

Context: 'Syrian', 'Aleppo', 'March 19, 2013'
Full context: NORP Syrian GPE Aleppo DATE March 19, 2013
Original: A woman, affected in what the government said was a chemical weapons attack, breathes through an oxygen mask as she is treated at a hospital in the syrian city of aleppo march 19, 2013
LSTM: A wounded Syrian woman is treated at a hospital after she was injured in a suicide attack in Sanaa March 19, 2013.
LSTM+Attention: A wounded Syrian woman receives treatment at a hospital in the town of al-Faenha, March 19, 2013.
Transformers: A wounded Syrian woman *lies on the floor* after a suicide bomb attack in the Syrian town of Hanuk March 19, 2013.
Ours: A Syrian woman who was injured during a bomb attack, lies in the hospital of the Syrian city of Aleppo March 19, 2013.

Context: 'HSBC', 'Hong Kong', 'last month'
Full context: ORG HSBC; GPE Hong Kong; DATE last month
Original: A man in a face mask inside HSBC's Hong Kong headquarters last month.
LSTM: A *shopping mall* in Hong Kong last month.
LSTM+Attention: A HSBC advertisement in Hong Kong last week.
Transformer: A message room at the HSBC office in Hong Kong last month.
Ours: A worker wearing a mask at the HSBC campus in Hong Kong last month.

Context: 'the Bosphorus', 'Bridge', 'Istanbul', 'Turkey', 'July 16, 2016'
Full context: 'FAC the Bosphorus Bridge; GPE Istanbul, Turkey; DATE July 16, 2016
Original: policemen protect a soldier from the mob after troops involved in the coup surrendered on the Bosphorus bridge in Istanbul , Turkey July 16 , 2016.
LSTM: a migrant from *Bangladesh* leans on a dinghy on a Turkish migrant ship at the port of Piraeus, Greece , October 2, 2015.
LSTM+Attention: Riot police officers detain a man during a protest against the Bosphorus Bridge in Istanbul, Turkey July 16, 2020.
Transformer: A *woman* reacts as she is detained by police after being thrown by protesters at the Bosphorus Bridge in Istanbul, Turkey, July 16, 2016.
Ours: A man is escorted by police officers as he arrives at a protest on the Bosphorus Bridge in Istanbul, Turkey July 16, 2016.

Context: 'Castle Dale', 'Utah', 'Trump'
Full context: 'GPE Castle Dale, Utah; ORG Trump '
Original:A coal-burning power plant in Castle Dale, Utah. A Trump administration plan to regulate coal-fired plants more lightly faces a major legal challenge.
LSTM: The Castle in Utah. The Trump administration has accused the company of importing fossil fuel from *plants and plants*.
LSTM+Attention: The coal-fired Castle Dale, a former Trump coal-fired power plant in Utah, is a serious *chemical-rich puppet*.
Transformers: A coal-fired power plant in Casle Dale, the trump administration has made a plan to significantly reduce power plants in the country.
Ours: A coal-fired power plant in Castle Dale, Utah. The Trump administration has been criticized for the shutdown by federal government.

Figure 4: Qualitative comparison of caption generated by different model baselines. The incorrect attributes being included in the caption are highlighted by underlining in the captions. The green highlighting of the text in the caption denotes the semantics which the model understands from the image input.

5.3. Ablations

Effect of Embedding? We analyze the effect of the choice of embedding for our task for which we take the captioning module to be the LSTM backbone in our experiments. We experiment with different embedding methods in Table 2, where we use Glove, fastText, and

Original: Streets were quiet Friday outside Zuniga's Restaurant & Cakery in San Jose, Calif.

NER Context: 'Friday', 'Zuniga's Restaurant & Cakery', 'San Jose', 'Calif.'
Generated: Streets were set up on Friday at Zuniga's Restaurant & Cakery in San Jose, Calif.

Context: 'quiet', street', 'Zuniga's Restaurant & Cakery','San Jose', 'Calif.'
Generated: A packed street is quiet during a minute's blackout at Zuniga's Restaurant & Cakery in San Jose, Calif.

Context: 'quiet', 'Zuniga's Restaurant & Cakery', 'San Jose', 'Calif.'
Generated: A packed street went quiet in minutes after a blackout caused by the coronavirus pandemic, at Zuniga's Restaurant & Cakery in San Jose, Calif.

Context: 'Harlem', 'New York', 'coronavirus'
Harlem, a neighborhood of New York, has been closed because of the coronavirus outbreak.

Context: 'Friday', 'Harlem', 'New York', 'lockdown'
Generated: A street in Harlem on Friday. New York shops are less likely to be closed because of the coronavirus pandemic.

Original: Tear gas floats in the air during clashes between police and protesters at a demonstration by French health workers in Nantes as part of a nationwide day of actions to urge the French government to improve wages and invest in public hospitals, in the wake of the the coronavirus disease (COVID-19) crisis in France June 16, 2020.

NER context: 'French','Nantes','France','June 16, 2020'
Generated: French riot police apprehend a protester during a demonstration against the government's plans to impose a lockdown in Nantes, France, June 16, 2020.

Context: 'French','climate', 'Nantes', 'France', 'June 16, 2020'
Generated: French riot police use tear gas to disperse protesters during a protest against rising prices of gasoline enforced by state restrictions on the climate change, in Nantes, France, June 16, 2020.

Context: 'worker', 'wage', 'protest', 'violent'
Context: 'worker', 'wage', 'protest', 'violence'
Generated: Workers clash with police turn violent during a protest against government's labor reforms in Venezuela.

Context: 'India', 'worker', 'wages', 'Delhi', 'June 18, 2022'
Generated: A mob of workers fire a tear gas canister during a protest against the India's government handling of the coronavirus disease (COVID-19) in Delhi, June 18, 2022.

Figure 5: Qualitative Comparison of the effect of the conditional word tokens on the semantics of caption generated. The green highlighted words in the generated caption denote the semantics model implicitly learns from the image input.

pre-trained GPT-2 byte-pair encoding-based embedding. We observe that for our task, BPE outperforms every other embedding since other embeddings simply treat out-of-vocabulary as unknowns. However, BPE tokenizer breaks out-of-vocabulary words into tokens within the learned vocabulary, thus, allowing the model to learn meaningful representations for words it has not seen before. In

Method	Bleu-4	Cider	Rouge	Meteor
Glove	20.77	13.92	22.14	8.20
fastText	21.80	15.01	22.30	8.50
BPE	**27.73**	**36.89**	**29.6**	**12.8**

Table 2: Ablation for effect of choice of embedding on performance of LSTM baseline [38].

Figure 6, we also analyze the qualitative effect of different embeddings on the generated caption. We observe that the use of the Glove and fastText as embedding results in a lot of unknown tokens because these embeddings are trained on different word corpus. Using a self-training vocabulary is also not better than Glove since named entities in the unseen test captions are out-of-vocabulary for the train time corpus as well. We observe that byte-pair encoding works best for our task as it allows us to handle out-of-vocabulary tokens. Using byte-pair encoding generates more meaningful captions compared to Glove and fastText for challenging examples. We underline a limitation of byte-pair encoding in the figure for Transformer backbone caption. Given less context and challenging out-of-vocabulary word, BPE can misconstrue it.

Effect of Loss: We analyze the effect of loss used in our model on the output probabilities over the vocabulary on training. Since COSMOS dataset has different category of frequencies for topics, we compare cross entropy effect over with weighted cross entropy and focal loss. For weighted cross entropy, we compute the weights by computing the frequency of each of the 50k tokens in our byte-pair encoding. There is marginal improvement in the performance on considering the imbalance in the tokens within the loss.

Ablation on contextual input The contextual input to the encoder in the captioning module consists of two modalities. Firstly, the *visual* input which includes the

Context: 'Sheila Bridges'
Full context: PERSON Sheila Bridges;
Original: Star sconces: Sheila Bridges, a decorator, earned oohs and aahs for her ceiling fixtures during a recent group video call.
LSTM (Glove): The man of the <unk> <unk> <unk> <unk> <unk> <unk>.
LSTM (FastText): Dr. <unk> , a psychiatrist , said he was not aware of the <unk>.
LSTM (BPE): Sheila *Hancock*, a former prosector, is a novelist and writer who is challenging her influence on the job.
Transformer (BPE): Sheila Bridges, a freelance journalist, said she was "very clear of the same place."
Ours (BPE): Sheila Bridges, a senior editor in the show's office, said that the artist had to wield a stroke.

Context: 'Christian Borle', 'Groff', 'Seymour'
Full context: PERSON Christian Borle, Groff; PRODUCT Seymour
Original: Christian Borle, as a sadistic dentist, goes to work on Groff's Seymour.
LSTM (Glove): <unk> <unk> and <unk> <unk> in game of thrones.
LSTM(FastText): <unk> <unk> and john <unk> in " the <unk>."
LSTM (BPE): Christian Borleff, left, and Groff, in "Help My Friend," a series of upsets, a series of upsets and downs.
Transformer (BPE): Christian Borle, left, and Groff in *"Seym Rak,"* a new musical about the musical.
Ours (BPE): Christian Borle, left, and Groff in a scene from the Seymour theatre.

Figure 6: Qualitative comparison of generated caption with different choice of embeddings for LSTM and Transformer captioning backbones. The red highlighted word token in the caption denotes the out of vocabulary word that is best captured by our final model which is tokenized incorrectly in only transformer backbone and is underlined as incorrect.

Method	Bleu-4	Cider	Rouge	Meteor
CE	37.4	50.1	41.5	22.4
Weighted CE	37.4	51.0	41.2	22.6
Focal	37.8	51.2	41.8	22.9

Table 3: Comparison of Cross Entropy (CE), Weighted Cross Entropy and Focal loss over output probabilities of decoder.

image feature, enhanced object features and the object relationship features. Secondly, the *textual* input which includes BPE of the conditional word tokens and the named entity types. In the table 4 we analyse the effect of different modalities on the performance of our model. We observe that both modalities significantly help the performance of our model and the model performance significantly drops with the removal of any one of these modalities. In fig 7 we demonstrate this qualitatively with the help of an example. In the figure, the removal of textual input from the context limits the model capacity to caption the news with correct named entities. On the other hand, removal of the visual input limits the models capacity to understand the relationships between the input named entities and the generated caption is no longer grounded in the reference images.

Are entity types in textual input helpful? We analyze the effect of the entity types in the textual input. In Table 5 we consider the ablation of the model without

Context: 'Eddie Johnson', 'Chicago', 'Monday'
Full context: PERSON Eddie Johnson; GPE Chicago; DATE Monday
Original: Eddie Johnson, the Chicago police superintendent, speaking to reporters on Monday.
w/o textual: Police officer, left , and police in Atlanta on Saturday.
w/o visual: Eddie Jones , the former mayor of Chicago , was fired on Monday night.
Ours: Eddie Johnson, a police officer, spoke at a news conference in Chicago on Monday.

Context: 'March 7', 'Grant', 'Park', 'Chicago'
Full context: DATE March 7; LOC Grant, Park; GPE Chicago
Original: The stage after a March 7 rally in Grant Park in Chicago.
w/o textual: The Seattle police center, is being built on a sharply avenue in San Francisco, California.
w/o visual: a Chicago police officer stands guard in front of the Grant Park in Chicago
Ours: The The stage on March 7 at Grant Park in Chicago.

Figure 7: Qualitative comparison of generated caption without different modalities of the contextual input. The green highlighted words in the final caption denotes semantics captured by our final model using both of the modalities in the context.

Method	Bleu-4	Cider	Rouge	Meteor
w/o visual	28.0	43.6	30.9	13.2
w/o textual	22.1	32.4	27.1	12.0
Ours	**37.4**	**50.1**	**41.5**	**22.4**

Table 4: Comparison of the effect of visual and textual input on the overall performance of the model

named entity types to with with respect to overall model which is denoted by Named Entity Type + Relational Graph. We observe that removing the named entity types from the text reduced the performance of the model as it removes the input from the model to know how to use the named entities in what context. We also observe that none of the generated captions when including named entity types contain entity types in the text, which underscores that model learns them only as semantics.

Method	Bleu-4	Cider	Rouge	Meteor
w/o NET	33.1	47.0	34.4	17.6
NET	35.2	47.8	37.5	20.1
NET+Relational Graph	37.4	50.1	41.5	22.4

Table 5: Comparison of effect of removal of named entity types (NET) vs removing the named entity types

Is using a relational graph helpful? We analyze the effect of relational graph on the performance of our model. In Table 6 we consider the ablations of our model without relational graph and then within the relation graph we selectively remove the enhanced object features and object relationship/edge features. We observe that the enhanced object features alone do not allow the model to understand and incorporate objects in the sentences as incorporating

Original: children hold olive branches as they look out from the sunroof of a car to be blessed by priests roaming around neighbourhoods to celebrate palm sunday, in Marjayoun, southern lebanon april 5, 2020.
Context: 'Sunday', 'April 5, 2020', 'Marjayoun', 'Lebanon'
Full context: DATE Sunday April 5, 2020; GPE Marjayoun, Lebanon
Ours w/o Relation Graph: People celebrate after a Sunday procession in Marjayoun, Lebanon April 5, 2020.
Ours: A boy waves at car as he departs a street after a shooting at a school on Sunday, April 5, 2020 in Marjayoun, in southern Lebanon.

Figure 8: Qualitative comparison of generated caption with and without relational graph. The red highlighted word token in the caption denotes object relationship that is absent from the w/o relationship graph caption.

objects in the caption requires context and the needs to have relationship between objects learnt as form of object relationship features that are extracted from the edges of the relation graph. In Figure 8 also analyse the qualitative

Method	Bleu-4	Cider	Rouge	Meteor
w/o relational graph	35.2	47.8	37.5	20.1
w/o edge features	35.3	48.0	38.2	19.0
w/o object features	36.0	48.1	39.6	20.5
Ours	**37.4**	**50.1**	**41.5**	**22.4**

Table 6: Comparison of effect of relational graph and also of different components of the relational graph on the overall performance of the model

effect of the relationship graph on the generated caption. We observe that use of the relationship graph allows to better capture the object relationship and thus understand and include semantics of events like waving in the caption.

How much training data is needed? We analyze the effect of increase of training data for the task of out-of-context generation. We experiment with different percentages of training data size (w.r.t. to our full data) in Table 4.

Method	Bleu-4	Cider	Rouge	Meteor
Ours (10%)	24.6	34.04	26.49	11.94
Ours (20%)	28.0	43.6	30.9	13.2
Ours (50%)	31.9	47.2	36.4	17.4
Ours (100%)	37.4	50.1	41.5	22.4

Table 7: Ablation for comparison of different percentages of training data. With respect to training data of 160K images, all context metrics are reported on validation data of 40k images.

6. Human Evaluation

We conduct a human evaluation to evaluate how convincing model-generated multimedia is to humans. In our human evaluation, we collect 47 responses from people of different demographics and ages on 30 multimedia examples i.e. news caption pairs. Out of which, 15 are real news examples and the rest of the 15 news examples have model-generated captions for which a completely different context is provided manually. We ask the subjects of our study to answer for each of the 30 news examples 2 different questions (1): Do you believe this news is real (Yes/No) and (2) How confident you are in your evaluation (0-5). Note that we specified to each of the subjects that they **must refrain** from using search engines and only rely on their best judgment to answer these questions. From our 47 respondents, we obtain an average accuracy of 14.83 for the first question for all 30 examples. We also obtain a median score of 15 for the same question for all 30 examples. This allows us to infer that most of the respondents were at best random in guessing which multimedia examples were out-of-context (fake) and which were not (real). In Figure 9, 12 out of 30 news examples are shown. The figure contains 6 real captions and 6 model-generated caption examples denoted as real and fake respectively. We also provide a statistic of how many respondents misclassify each news example. We find that many fake captions are misclassified as real by the respondents. This demonstrates the hardness of out-of-context multimedia detection task for humans when out-of-context captions are generated by our model.

7. Conclusion

Overall, we show that it is possible to automatically generate and control the semantics of caption for an image, given some conditional word tokens. We present a challenging benchmark to foster the development of defenses against large-scale image re-purposing. From our experimental results, we find that both visual and textual conditional input significantly improve the quality of the generated captions. We also observe that byte-pair encoding helps in improving the captions significantly as it can handle out-of-vocabulary tokens effectively. We also observe that use of a relational graph helps in identifying the underlying object-object relationships. These object relationships enrich the semantics of the caption from the events that could have been ignored by the model.

8. Acknowledgements

I would like to express my sincere gratitude to my primary advisor, Ms. Shivangi Aneja, who guided me throughout this project. Without her guidance and support, it would not have been possible to accomplish this work. I would also like to thank Prof Matthias Nießner and Prof Daniel

FAKE: Ukraine: Ukraine's armed conflict has intensified.
72.3% believe it's **Not-Out-of-Context.**

FAKE: People queue to receive their first aid during Tamil Nadu floods in India.
57.3% believe it's **Not-Out-of-Context.**

FAKE: A mob of students clash with police turn violent during a protest against the India's government handling of the unemployment crisis in Hyderabad, June 22, 2022.
49% believe it's **Not-Out-of-Context.**

FAKE: A woman raises her fist during a rally against police brutality and racism in Raleigh, North Carolina, April 21, 2020.
68.1% believe it's **Not-Out-of-Context.**

FAKE: A protester holds a sign during a rally against the government, demand changes in the abortion rights movement. The group is opposed to abortion rights in the United States.
63.8% believe it's **Not-Out-of-Context.**

FAKE: Supporters of former Prime Minister Nawaz Sharif shout slogans during a protest against China in Hyderabad, Pakistan, on Monday, May 13, 2022.
51% believe it's **Not-Out-of-Context.**

REAL: German chancellor tasting Roasted Wild boar meat.
66% believe it's **Out-of-Context.**

REAL: Prime Minister Narendra Modi of India and President Trump at a September event in Houston called "Howdy, Modi: Shared Dreams, Bright Futures."
40.4% believe it's **Out-of-Context.**

REAL: Books burned during the battle with the Islamic State militants, lie in the library of the University of Mosul.
55.3% believe it's **Out-of-Context.**

REAL: Facebook said on Thursday that it banned the conspiracy theorist Alex Jones and others from its social media services.
57.4% believe it's **Out-of-Context.**

REAL: Waymo, the driverless-car company that was spun out of Google in 2016, registered a Shanghai subsidiary in May.
43% believe it's **Out-of-Context.**

REAL: The Trump administration had sought to change the Migratory Bird Treaty Act to eliminate penalties for energy companies that kill birds "incidentally."
49% believe it's **Out-of-Context.**

Figure 9: Qualitative comparison of 12 news examples from the user study. We denote the model generated captions as fake and original captions for the corresponding images as real. We also provide a statistic of how many respondents misclassify each news example

Cremers for allowing me an opportunity to pursue this topic and also providing the necessary computing resources.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[2] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In *CVPR workshops*, volume 1, page 38, 2019.

[3] O. AI. Radford, alec and kim, jong wook and hallacy, chris and ramesh, aditya and goh, gabriel and agarwal, sandhini and sastry, girish and askell, amanda and mishkin, pamela and clark, jack and others. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[4] T. Akgul, T. E. Civelek, D. Ugur, and A. C. Begen. Cosmos on steroids: a cheap detector for cheapfakes. In *Proceedings of the 12th ACM Multimedia Systems Conference*, pages 327–331, 2021.

[5] S. Aneja, C. Bregler, and M. Nießner. Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*, 2021.

[6] S. Aneja and M. Nießner. Generalized zero and few-shot transfer for facial forgery detection. *arXiv preprint arXiv:2006.11863*, 2020.

[7] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475, 2019.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[9] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15108–15117, 2021.

[10] L. Fazio. Out-of-context photos are a powerful low-tech form of misinformation. *The Conversation*, 14, 2020.

[11] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[12] M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[13] A. Jaiswal, E. Sabir, W. AbdAlmageed, and P. Natarajan. Multimedia semantic integrity assessment using joint embedding of images and text. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1465–1471, 2017.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] P. Korshunov and S. Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

[16] T.-V. La, M.-S. Dao, Q.-T. Tran, T.-P. Tran, A.-D. Tran, and D. T. D. Nguyen. A combination of visual-semantic reasoning and text entailment-based boosting algorithm for cheapfake detection. 2022.

[17] T.-V. La, Q.-T. Tran, T.-P. Tran, A.-D. Tran, D.-T. Dang-Nguyen, and M.-S. Dao. Multimodal cheapfakes detection by utilizing image captioning for global context. In *Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, ICDAR '22, page 9–16, New York, NY, USA, 2022. Association for Computing Machinery.

[18] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

[19] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[20] G. Luo, T. Darrell, and A. Rohrbach. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv preprint arXiv:2104.05893*, 2021.

[21] H. Ma, H. Zhao, Z. Lin, A. Kale, Z. Wang, T. Yu, J. Gu, S. Choudhary, and X. Xie. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18051–18061, 2022.

[22] 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*, 26, 2013.

[23] E. Müller-Budack, J. Theiner, S. Diering, M. Idahl, and R. Ewerth. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 16–25, 2020.

[24] K. Nakamura, S. Levy, and W. Y. Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*, 2019.

[25] H. H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311. IEEE, 2019.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[28] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[29] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, R. Luis, J. Jiang, et al. Deepfacelab: A simple, flexible and extensible face swapping framework. 2020.

[30] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[31] E. Sabir, W. AbdAlmageed, Y. Wu, and P. Natarajan. Deep multimodal image-repurposing detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1337–1345, 2018.

[32] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

[33] R. Tan, B. A. Plummer, and K. Saenko. Detecting cross-modal inconsistency to defend against neural fake news. *arXiv preprint arXiv:2009.07698*, 2020.

[34] A. Tran, A. Mathews, and L. Xie. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13035–13045, 2020.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[37] L. Verdoliva. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

[38] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[39] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[40] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[41] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

[42] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 1831–1839. IEEE, 2017.