Out-of-Context Caption Generation

Anurag Singh
Technical University of Munich

Abstract

With 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 cheap fakes where an image is described using an out-of-context caption to spread misinformation. The main challenges in the task of detecting such out of context multimedia is the availability of large scale datasets due to challenges in collection and annotation of out-of-context multimedia. Several methods employ randomly selecting captions from other news examples to generate training inputs of image, true caption and out-of-context caption for their task. Similarly out of context multimedia generation works also have focused on semantically matching images to captions within the dataset for generating out-of-context multimedia. We aim to address these limitations by introducing a novel task of out-of-context caption generation. Our main contribution are that we build a new model that generates realistic out-of-context captions given an input image and some conditional word tokens as contextual input. We demonstrate that it is possible to control semantics and context of the generated captions using conditional word tokens unlike previous methods which generate out-of-context media by matching semantics. We benchmark our model with several experiments and ablations to analyse the quality of our captioning results.

1. Introduction

With the recent improvements in several computer vision tasks due to success of deep learning, a new threat has developed as manipulation of media using computer vision models has become more realistic [12, 19]. Therefore several works have recently focused on forensic methods for detection of such DeepFakes based media manipulations [2, 1, 26]. One of the most prevalent ways to still spread misinformation is to use out-of-context news as it requires very little to no computing resources compared to DeepFakes [7]. In out-of-context misinformation generation often an unaltered image is used out-of-context with a fake caption to spread misinformation. There has been some recent progress in detection of out-context misinformation [4]. Most of the detection works also propose a method to create these out-of-context captions for training of detection models, often by selecting other captions at random. Others use hypernamed text or similarity methods to select out-of-context captions from a reference set. Often these out-of-context captions are easy to identify for detection models [14].

To address these challenges, we propose the task of out-of-context caption generation i.e. generate an out-of-context caption for an image given some conditional word input. Our task differentiates itself from the task of image captioning since there is no necessity for the generated out-of-context caption to be exactly descriptive of the events or actions in the image. This shall allow out-of-context detection models to be trained in the self-supervised fashion using only an image, true caption and conditional word tokens. As the image and conditional word tokens are used to generate out-of-context caption which can then be used as a triplet input in the training of the downstream detection models [4]. This approach significantly differentiates our method from several state of the art methods for generating out-of-context multimedia [14] which focus on matching existing unmanipulated text and images in the dataset to generate new out-of-context multimedia. One of the major limitations of these semantic matching based approaches to generate out-of-context multimedia are that they are unable to generate captions for out-of-vocabulary words as they match images to the captions within their dataset. We address this issue by using byte-pair encoding based captioning module that is able to process out-of-vocabulary tokens. Also, matching based methods are computationally expensive to be used in practice by detection methods to generate out-of-context captions for their training examples since these methods involve some form of linear search within the dataset to match image to a semantically similar out-of-context caption.

In summary, our main contributions are as follows:

- We propose a new method to that generate realistic out-of-context captions given an input image and some conditional word tokens as contextual input.
- We demonstrate that it is possible to control semantics and context of the generated captions using conditional word tokens.
- Our method addresses the limitations of previous matching based out-of-context multimedia generating methods by generating captions for out-of-vocabulary conditional word tokens.

2. Related Work

2.1. Misinformation Detection

There are some recent works to detect out-of-context captions in cheap fakes [20] [4]. Of which [4] propose, a self supervised way of learning to detect out-of-context news by using object detection and visual grounding of image with two caption pairs. Their method uses a triplet of image, true caption and out-of-context caption to train in a self supervised fashion. The out-of-context caption is randomly selected from the dataset. The also introduce a new large scale dataset for out-of-context news detection containing news examples collected from different news outlets. Each news example contains a image and corresponding news caption. Most of the collected data does not have context annotations. However, a small subset has been manually annotated as in- or out-of-context. The problem proposed by [4] is different from ours as it focuses on analysing an image when it is paired with two different captions. FakenewsNet [21] and Fakeddit [16] focuses on fake news detection that is human-made. While these works have important real world use cases for task of out-of-context multimedia, our focus is on exploring an targeted out-of-context caption where given an image and conditional word tokens we are able to automatically generate a new out-of-context caption. Our method can address the limitations of randomly selecting out-of-context captions from the dataset for training of the detection models by generating a targeted out of context caption for the given example.

2.2. Generating Out-of-Context Multimedia

For generating out-of-context multimedia several research has been done to allow for creation of datasets for matching of out-of-context captions in the reference set to corresponding query images and vice a versa [14] [10] [20]. We try to review several of these datasets in detail. There are two earlier proposed datasets to train for identifying image repurposing or to identify discrepencies in joint semantics of image and text for multimedia i.e. MAIM dataset [10] and MEIR dataset [20]. MAIM simply 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. The underlying assumption to their approach is that for each multimedian image caption pair, there exists another semantically similar pair in the reference set whose caption can be used to replace other caption to create a falsified set. Recent works have produced [15] [14], of which [15] release a dataset that has swapped named entities of people, location and events with other random entities. Authors in [14] argue that this hand crafting from hypernamed text to create out-of-context media generates linguistic biases that are easy to identify. Therefore [14] propose a clip based matching approach for text manipulation. These problem statements are significantly different from ours as we do not assume availability of any reference set, rather for each image some named entities are to be provided that form the conditional word tokens to generate the caption. Also, these matching methods computationally very expensive to generate out-of-context captions for training of detection methods. As obtaining a semantically related out-of=context caption to an image of example they involves some form of linear search within the dataset.

2.3. News Captioning:

In [22] authors aim to create a Neural News generation method which is able to replace the entire real news articles with text generated from Grover [29] and similarly replace real captions with fake captions generated via model. However, they do not aim to mismatch the images which are relevant to article's content. Analysis of images for this method have limited impact as analyzing article and caption 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 [5] [23]. Both these methods take news article and a corresponding image as an input to generate a news caption for the article. Again our task is significantly different from the previous work in news captioning literature since we do not assume any access to complete news article 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 on input image and object features with the conditional tokens to generate an out-of-context captions. The main components of our model architecture are 1) Conditional word tokens 2) Feature Extraction & Detection Backbone 3) Relational Graph 4) Captioning Module.

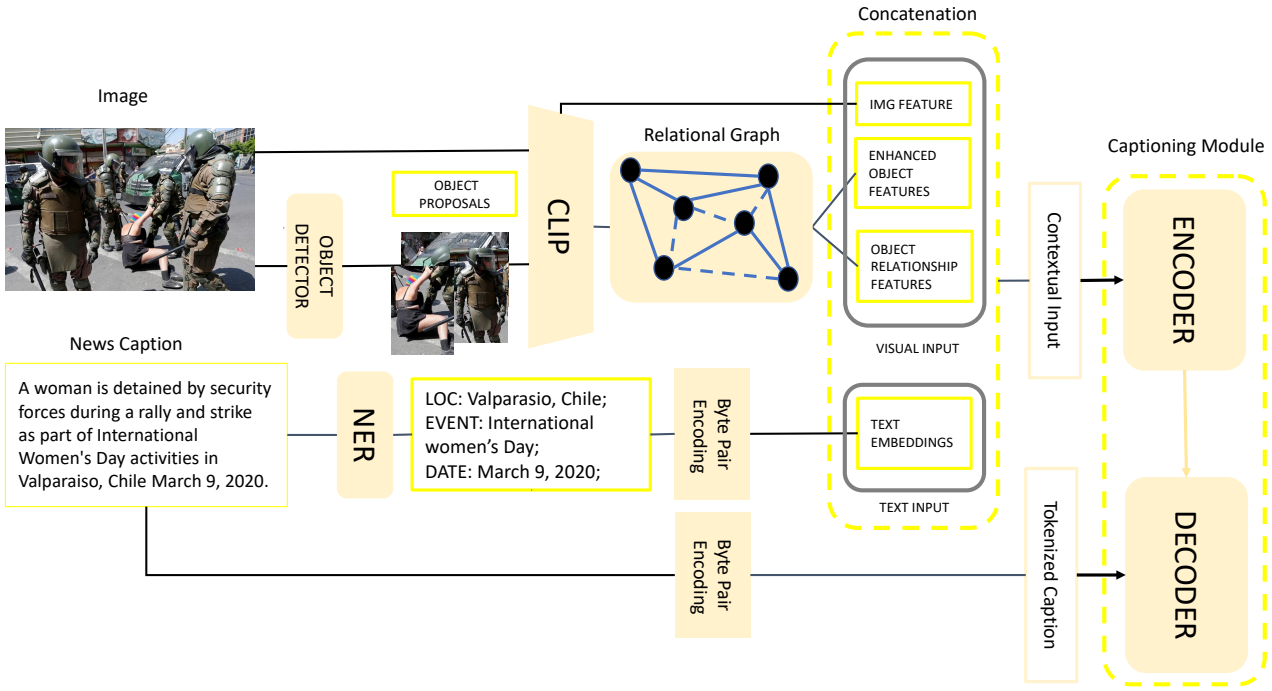

Figure 1: The semantic figure describes image and news caption input that are pre-processed using Object-detection, CLIP and Named Entity Recognition respectively to obtain image and object level features along with named entity dictionary. The object features are enhanced with a relational graph. The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

3.1. Conditional word tokens

As the first step for generating our conditional token texts on which our out-of-context captions is conditioned, we parse the input captions to classify named entities in the text. We use SpaCY NER [9] for processing the captions to identify named entities. This parsing of captions into named entities is only pre-processing of our dataset. We would like to highlight that our model does not need entire caption as input and can rather work using named entities only. The input token is classified into 18 different categories of named entities by SpaCY. This results in a dictionary $NamedEntity$ with keys as $NamedEntityType$ and values as tokens.

3.2. Feature Extraction & Detection Backbone

From the input image in our network we extract image features using a CLIP [3] visual encoder which we denote as $\mathcal{I}$. We also detect all probable objects in the given image using a attention based object detector DETR [6]. The object proposals are extracted from the image using crop based on the predicted bounding boxes. These object proposals are also encoded into object features using CLIP visual encoder.

3.3. Relational Graph

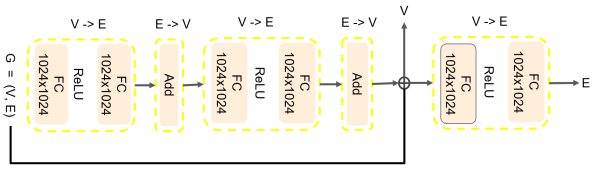

Figure 2: The relational graph module which takes CLIP image encoding from object proposals as input as is represented as the nodes in the graph.

Describing the actions and events in an image often involve understanding of relationship of an object's appearance in context with other objects in the scene. In order 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 ob-

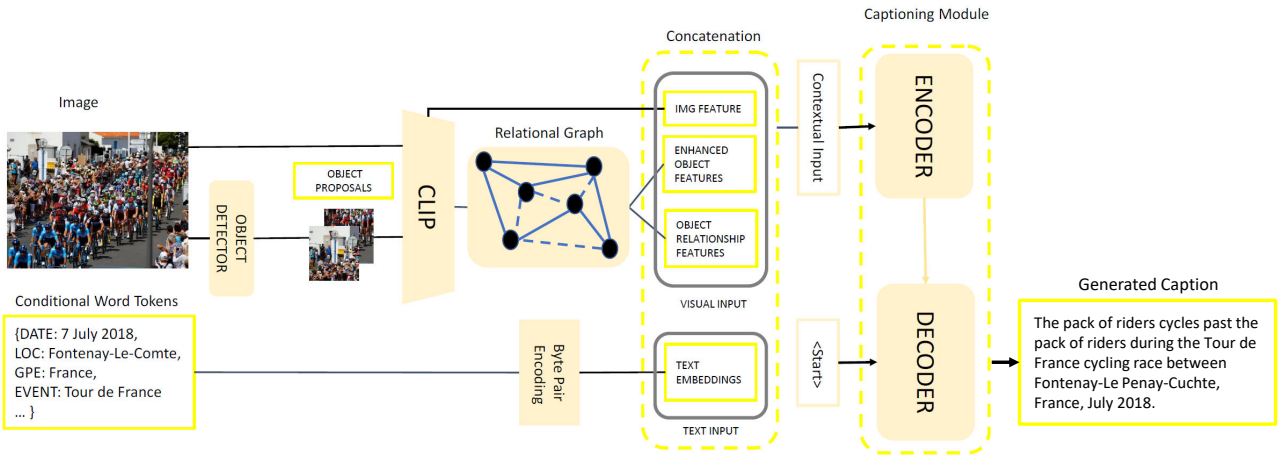

Figure 3: The semantic figure describes test time image and conditional word token input that are pre-processed using Object-detection, CLIP for image and byte pair encoding to convert tokens into text embeddings. Similar to train, this forms input to the encoder of the captioning module. The decoder is conditioned using start token that denotes start of sentence and is used in an auto-regressive fashion to generate caption.

ject 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 represent the object features as nodes and the relationship between the objects are represented using 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 [8] where the message passing at graph step τ is defined as follows:

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

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 a 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 $\mathcal{V}^{\tau}$. We append an additional message passing layer after the last graph step and use the learned message $\mathcal{E}^{\tau+1}$ as the output object relation features.

3.4. Captioning Module

Our captioning module consists of a 3 layer Transformer with 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

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

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

We also consider Byte pair encoding to encode the tokens into word embedding. We use the pre-trained GPT-2 byte-pair encoding for our task. The effectiveness of byte pair encoding for our task is that is allows us to generalize to out of vocabulary words in much more easier manner than other standard word embeddings like word2vec or GLOVE. For our named entity dictionary $\mathcal{D}^{NER}$ which contains named entity class as key and corresponding tokens as values, we compute the byte pair encoding of the key followed by the tokens. We use the byte pair encoding of the $keys$ inorder to allow our model to understand the context of each conditional token. The byte pair encoding of each key and corresponding tokens is concatenated and is followed by the word embedding of end token to allow the model to differentiate between different categories of named entities.

3.5. Decoder

A stack of 3 identical layers constitutes the decoder of the captioning module. Similar to architecture in [24] decoder receives features from the encoder to compute cross attention. In addition to byte-pair encoding decoder also uses GPT-2 pre-trained position encoding for understanding the relative position of the token in the caption. The final layer generates a prediction over the vocabulary to predict the output token. The decoder masks the embeddings

of tokens ahead in the caption while predicting the output of the current token. At the time of inference, we generate the new captions in an auto-regressive way.

4. Implementation Details

We implement our architecture using PyTorch [18] and train end-to-end using ADAM [11] with a learning rate of 1e-3. We train the model for 100 epochs until convergence. We extract maximum of 10 objects for each image and for images with less objects we pad the object features with 0. For the named entity texts we condition on 20 tokens from all named entity classes, 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 byte-pair encoding of GPT-2 to indicate the start and end of the caption.

5. Experiments

Dataset: We use COSMOS dataset [4] of out-of-context news captions which contains images and corresponding news captions from different news outlets. COSMOS consists of 200k number of images with 450k different captions, corresponding to the true captions from news outlets and also captions corresponding to fake news caption. 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 on COSMOS [4] i.e. we use 160K images and corresponding captions to train our model. We use validation set from our model to test our captions since the validation set is much larger with 40k images and captions to test our model than the test set in COSMOS which only contains 1700 images and corresponding captions. We construct a smaller dataset from cosmos for our task where we select single caption corresponding to each image in our training and validation test set.

Metrics: To evaluate the quality of generated captions we used to evaluation metrics commonly used for the tasks of image captioning such as BLEU-4 [17], CiDEr [25], METEOR [25] and ROUGE [13]. 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 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 effectiveness of our method we implement strong baselines using LSTM Seq-2Seq architecture [27] and with attention [28] 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 conditional text act 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 only image features and byte pair encoding with no named entity types 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 also analyse the qualitatively the captions generated by different baseline models. We observe that 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 object relationship between man and the police. This allow the model to include semantics of man being escorted in the caption.

Playing around with the input: We try to qualitatively analyse the quality and how real captions look as we change the conditional word tokens for an unseen image in Fig 5. In the figure, the NER context denotes the named entities from the Spacy that form the conditional word tokens and the corresponding generated out-of-context caption. We observe for different cases how caption changes as we change the named entities in the context. In left figure we observe that adding the word quiet to

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.

the context allows change in semantics of the generated out-of-context caption. We also note that removing word street from the context the generated caption still describes the street from the visual input. Similar observation is also made for the word token Friday. In right Figure we observe that input from the image allows the model to

understand tear gas, protest and police and use these in the generation of out-of-context caption. Further, changing context to India and Delhi allows to control the location of the generated caption.

5.3. Ablations

Effect of Embedding? We analyze the effect of choice of embedding for our task for which we take 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 pre-trained GPT-2 Byte pair encoding based embedding. We observe that for our task byte pair encoding outperforms every other embedding since there are lot of words in the news caption that are out of vocabulary for other embeddings. However, for byte pair encoding the tokenizer breaks out of vocabulary words into tokens within the learned vocabulary thus allowing model to learn meaningful representations for words it has not seen before. In Figure 6 also analyse the qualitative effect

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 [27].

of the choice of embedding on the generated caption. We observe that use of the glove and fastText as embedding results in lot of unknown tokens since these embeddings are trained on different world 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 to handle out of vocabulary tokens easily. Thus generating more meaningful captions in comparison to glove and fasttext for challenging examples. Some limitations of byte pair encoding are also underlined in the figure for transformer backbone caption that it can misconstrue the out of vocabulary words.

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.

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.

Ablation on contextual input The contextual input to the encoder in the captioning module consists of two modalities i.e. *visual* input which includes the image feature and the enhanced object features and the object relationship features, and *textual* input which includes 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.

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

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 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.

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

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.

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

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

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 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 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.

6. Human Evaluation

We conduct a human evaluation to estimate the difficulty of the proposed task for the humans. We aim to access how well are humans able to recognized model generated out of context multimedia and how convincing out model generated multimedia is to humans. In our human evaluation we collect 47 responses from people of different demograph-

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> <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.

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.

ics and age on 30 multimedia examples i.e. news caption pairs. Out of which, 15 are real news examples and rest of the 15 news examples have model generated captions for which a totally different context is provided manually. We ask the subjects of our study to answer for each of the 30 news example 2 different questions (1): Do you believe this news is real (Yes/No) and (2) How confident you are in your evaluation (0-5). Not that we specify to each of the subject that they **must refrain** from using search engines and use their best judge answer these questions. From our 47 respondents, we obtain an average accuracy of 14.83 for first

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.

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.

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 we 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. It is can be noted that lot of fake captions are misclassfied as real by the respondents which demonstrates the hardness of our task for humans.

7. Conclusion

Overall, we show that it is possible to automatically generate and control semantics of caption for an image given some conditional word tokens using limited compute. We present a challenging benchmark to foster the development of defenses against large-scale image re-purposing. From our experimental results, we find that the both visual and textual conditional input significantly help in the quality of the generated captions. We also observe that for the task of generating out-of-context captions byte pair encoding significantly helps in improving the performance as it is able to handle out of vocabulary tokens effectively. We also observe that use of a relational graph helps in identifying the underlying object object relationship and thus enrich the semantics of the caption from the events that could otherwise

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

be ignored by the model.

8. Acknowledgements

I would like to express my 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 Dainel 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] S. Aneja, C. Bregler, and M. Nießner. Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*, 2021.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

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

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

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.