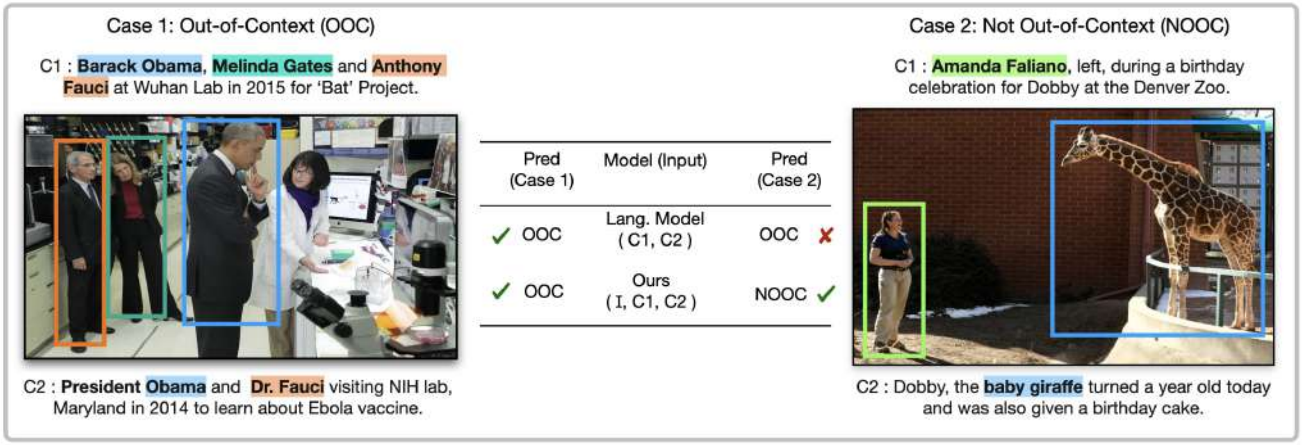
COSMOS: Catching Out-of-Context Misinformation with Self-Supervised Learning

Shivangi Aneja[1] Chris Bregler[2] Matthias Nießner[1]
[1]Technical University of Munich [2]Google AI


Case 1: Out-of-Context (OOC)
Case 2: Not Out-of-Context (NOOC)

C1 : Barack Obama, Melinda Gates and Anthony Fauci at Wuhan Lab in 2015 for 'Bat' Project.

C2 : President Obama and Dr. Fauci visiting NIH lab, Maryland in 2014 to learn about Ebola vaccine.

Pred (Case 1)	Model (Input)	Pred (Case 2)
✓ OOC	Lang. Model (C1, C2)	OOC ✗
✓ OOC	Ours (I, C1, C2)	NOOC ✓

C1 : Amanda Faliano, left, during a birthday celebration for Dobby at the Denver Zoo.

C2 : Dobby, the baby giraffe turned a year old today and was also given a birthday cake.

Figure 1: Our method takes as input an image and two captions from different sources, and we predict whether the image has been used out of context or not. We show that it is critical to the task to ground the captions w.r.t. image, and it is insufficient to consider only the captions; e.g., a language-only model would incorrectly classify the right image to be out of context. To this end, we propose a new self-supervised learning strategy allowing to make fairly accurate out-of-context predictions.

Abstract

Despite the recent attention to DeepFakes, one of the most prevalent ways to mislead audiences on social media is the use of unaltered images in a new but false context. To address these challenges and support fact-checkers, we propose a new method that automatically detects out-of-context image and text pairs. Our key insight is to leverage grounding of image with text to distinguish out-of-context scenarios that cannot be disambiguated with language alone. We propose a self-supervised training strategy where we only need a set of captioned images. At train time, our method learns to selectively align individual objects in an image with textual claims, without explicit supervision. At test time, we check if both captions correspond to same object(s) in the image but are semantically different, which allows us to make fairly accurate out-of-context predictions. Our method achieves 85% out-of-context detection accuracy. To facilitate benchmarking of this task, we create a large-scale dataset of 200K images with 450K textual captions from a variety of news websites, blogs, and social media posts. The dataset and source code is publicly available here[1].

1. Introduction

In recent years, the computer vision community as well as the general public have focused on new misuses of media manipulations such as DeepFakes [12, 13, 29, 31] and how they aid the spread of misinformation in news and social media platforms. At the same time, researchers have developed impressive media forensic methods to automatically detect these manipulations [35, 27, 22, 46, 50, 11, 1, 2, 21, 42, 3, 10]. However, despite the importance of DeepFakes and other visual manipulation methods, one of the most prevalent ways to mislead audiences is the use of unaltered images in a new but false or misleading context [14]. Fact checkers refer to this as out-of-context use of images, where an image appears on with two (or even more) online sources with different and contradictory captions.

The danger of out-of-context images is that little technical expertise is required, as one can simply take an image from a different event and create highly convincing but potentially misleading messages. At the same time, it is extremely challenging to detect misinformation based on out-of-context images given that the visual content by itself is not manipulated; only the image-text combination creates misleading or false information. In order to detect these out-of-context images, several online fact-checking initiatives have been launched by news rooms and independent

[1] https://shivangi-aneja.github.io/projects/cosmos/

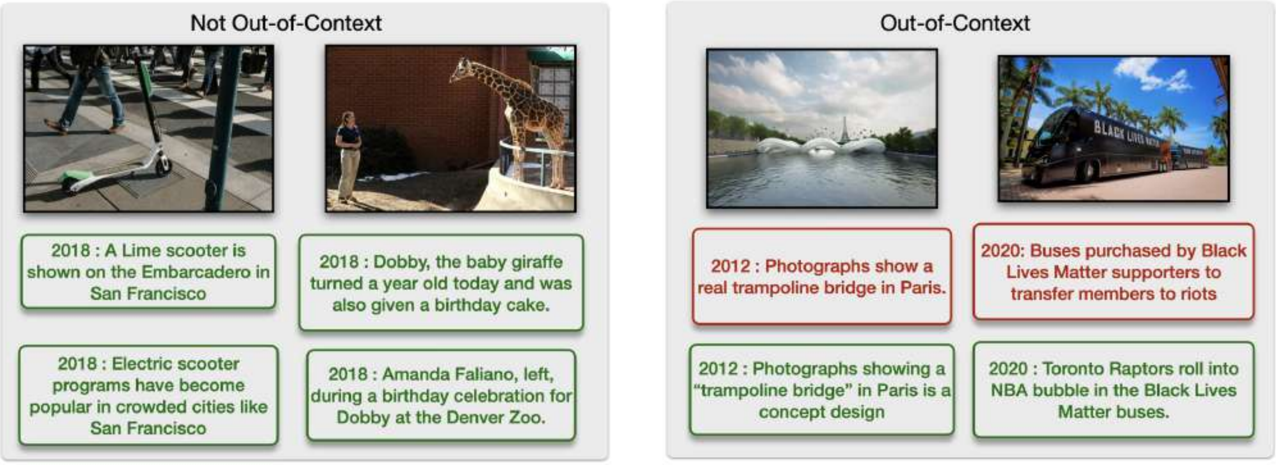

Not Out-of-Context
Out-of-Context
2018 : A Lime scooter is shown on the Embarcadero in San Francisco
2018 : Dobby, the baby giraffe turned a year old today and was also given a birthday cake.
2012 : Photographs show a real trampoline bridge in Paris.
2020: Buses purchased by Black Lives Matter supporters to transfer members to riots
2018 : Electric scooter programs have become popular in crowded cities like San Francisco
2018 : Amanda Faliano, left, during a birthday celebration for Dobby at the Denver Zoo.
2012 : Photographs showing a "trampoline bridge" in Paris is a concept design
2020 : Toronto Raptors roll into NBA bubble in the Black Lives Matter buses.

Figure 2: Examples from our dataset where images from social media and online news were used out of context (right) and those which were not (left). (Red) denotes false captions and (green) shows the true captions along with year published.

organizations, most of them currently being part of the International Factchecking Network [32]. However, they all heavily rely on manual human efforts to verify each post factually, and to determine if a fact-checking claim should be labelled as "out-of-context" or not. Thus, automated techniques can aid and speed up verification of potentially false claims for fact checkers.

Seminal works along these lines focus on predicting the veracity of a claim based on certain evidence like subject, context, social network spread, prior history, etc. [44, 39]. However, these methods are limited only to the linguistics domain, focusing on textual metadata to predict the factuality of the claim. In particular, language-only analysis cannot accurately identify many out-of-context scenarios, as shown in Figure 1 – the grounding of which objects in an image the language refers to is essential towards understanding whether there is an out-of-context situation.

We define out-of-context use of images as presenting the image as an evidence of untrue or unrelated event(s). If the two captions refer to same object in the image, but are semantically different, i.e. correspond to different events, then it indicates out-of-context use of image. However, if the captions correspond to the same event irrespective of the object(s) they captions describe, the it is defined as not-out-of-context.

Note that a not-out-of-context scenario makes no conclusions regarding the veracity of the statements. To automatically detect these cases, we propose a new data-driven method that takes an image and two text captions as input. As output, we predict whether the two captions referred to the image are out-of-context or not.

The core idea of our method is a self-supervised training strategy where we only need captioned images; we do not require any explicit out-of-context annotations which would be potentially difficult to annotate in large numbers. We can then establish the image captions from the data as matches, and random captions from other images as non-matches. Using these matches vs non-matches as loss function, we are able to learn co-occurrence patterns of images with textual descriptions to determine whether the image appears to be out-of-context with respect to textual claims. During training, our method only learns to selectively align individual objects in an image with textual claims, without explicit out-of-context supervision. At test time, we are able to correlate these alignment predictions between the two captions for the input image. If both texts correspond to same object but their meaning is semantically different, we infer that the image is used out-of-context.

In order to train our approach, we create a large-scale dataset of over 200K images with their corresponding 450K textual captions (some images appear with various captions, although not a necessary requirement) from a variety of news websites, blogs, and social media posts. We further manually annotated a subset of 1700 triplet pairs (an image and 2 captions) for benchmarking purposes only. In the end, our method significantly improves over alternatives, reaching over 85% detection accuracy.

In summary, our contributions are as follows:

- This paper proposes the first automated method to detect out-of-context use of images.
- We introduce a self-supervised training strategy for accurate out-of-context prediction while only using captioned images.
- We created a large dataset of 200K images with 450K corresponding text captions from a variety of news websites, blogs, and social media posts.

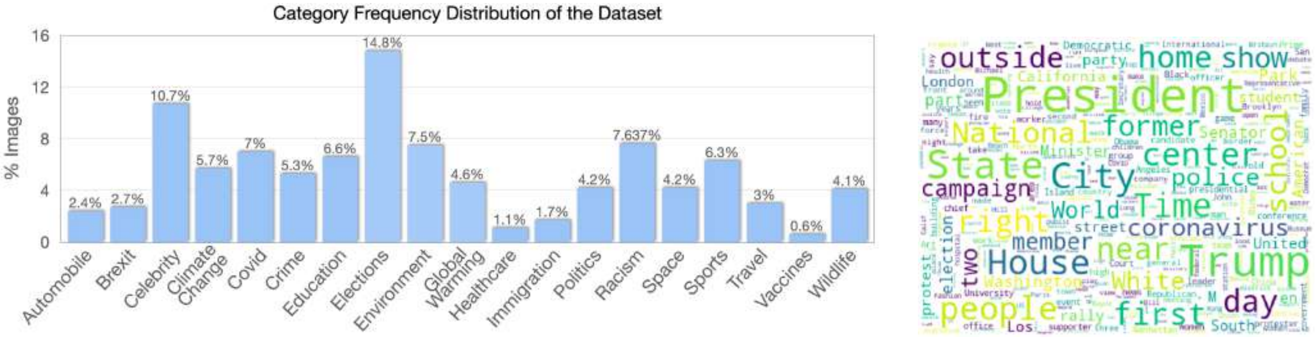


Figure 3: High-level overview of our dataset: (left) category-wise frequency distribution of the images; (right) word cloud representation of captions and claims from the dataset.

2. Related Work

Fake News & Rumor Detection. Fake news and rumor detection methods have a long history [33, 20, 23, 24, 49, 34, 25, 26] and with the advent of deep learning, these techniques have accelerated in progress. Most fake news and rumor detection methods focus on posts shared on microblogging platforms like Twitter. Kwon et al. [20] analyzed structural, temporal, and linguistic aspects of the user tweets and modelled them using SVM to detect the spread of rumors. Liu et al. [23] proposed an algorithm to debunk rumors in real-time. Ma et al. [26] examined propagation patterns in tweets and applied tree-structured recursive neural networks for rumor representation learning and classification. Tan et al. [38] detect neural fake news by exploiting visual and semantic inconsistencies in the news article.

Automated Fact-Checking. In recent years, several automated fact-checking techniques [44, 39, 15, 6, 28, 5, 40] have been developed to reduce the manual fact-checking overhead. For instance, Wang et al. [44] created a dataset of short statements from several political speeches and designed a technique to detect fake claims by analyzing linguistic patterns in the speeches. Vasileva et al. [40] proposed a technique to estimate check-worthiness of claims from political debates. Atanasova et al. [6] propose a multitask learning technique to classify veracity of claim and generate fact-checked explanations at the same time based on ruling comments.

Verifying Claims about Images. Both fake news detection and automated fact-checking techniques are extremely important to combat the spread of misinformation and there are ample methods available to tackle this challenge. However, these methods target only textual claims and therefore cannot be directly applied to claims about images. To detect the increasing number of false claims about images, few methods [18, 48, 37, 45, 51, 19] have been proposed recently. For instance, Khattar et al. [19] learn a variational autoencoder based on shared embedding space (textual and visual) with binary classifier to detect fake news. Jin et al. [18] use attention-based RNNs to fuse multiple modalities to detect rumors/fake claims. Since these techniques are supervised in nature, they require large amounts of labelled data, which is difficult to obtain, especially for false claims. We, however, propose a self-supervised method to achieve the goal.

3. Out-of-Context Detection Dataset

Our dataset is based on images from news articles and social media posts. We gather images from a wide variety of articles (Fig. 3), with special focus on topics where misinformation spread is prominent.

3.1. Dataset Collection

We gathered our dataset from two primary sources, *news websites*[2] and *fact-checking websites*. We collect our dataset in two steps: (1) First, using publicly available news channel APIs, such as from the New York Times [4], we scraped images along with the corresponding captions. (2) We then reverse-searched these images using Google's Cloud Vision API to find other contexts in which the image is shared. Note that the second step is not required for training, but we collect these captions for benchmarking as well as increased dataset size. Thus, we obtain captioned images that we can use to train our models. Note that we do not consider digitally-altered/fake images; our focus here is to detect misuse of real photographs. Finally, we currently aim to detect conflicting-image-captions in the English language only.

[2] New York Times, CNN, Reuters, ABC, PBS, NBCLA, AP News, Sky News, Telegraph, Time, DenverPost, Washington Post, CBC News, Guardian, Herald Sun, Independent, CS Gazette, BBC

3.2. Data Sources & Statistics

We obtained our images primarily from *news channels* and a fact-checking website (*Snopes*). We scraped images on a wide variety of topics ranging from *politics, climate change, environment, etc* (see Fig. 3). For images scraped from *New York Times*, we used publicly available Article Search developer API [4], and for other new sources, we wrote our custom scrapers. For images from news channels, we scraped corresponding image captions from <*figcaption*> tag and *alt text* attribute, and for Snopes, we scraped text written in the <*Claim*> header, under the *Fact Checks* section of the website. In total, we obtain 200K train images and 1700 test images; see Tab. 1.

Split	Primary Source	No. of Images	Context Annotation
Train	News Outlets[2]	160K	✗
Val	News Outlets	40K	✗
Test	News Outlets, Snopes	1700	✓

Table 1: Statistics of our out-of-context dataset.

Train Set: For training, we used images scraped from news websites. We consider several news sources[2] to gather the images. In total, we gathered around 200K images with 450K captions, 20% of which we use as the validation set.

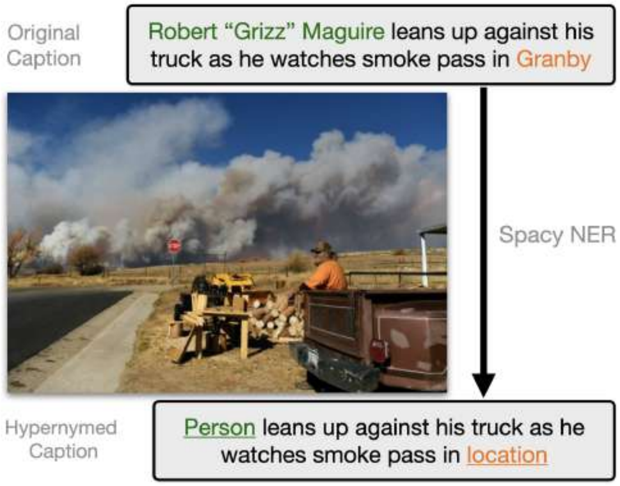
Test Set: At test time, we use the images from the fact-checking website *Snopes* along with news websites. We collected 1700 images with two captions per image. We build an in-house annotation tool to manually annotate these pairs with out-of-context labels. On average, it takes around 45 seconds to annotate every pair, and we spent 100 hours in total to collect and annotate the entire test set. We ensured an equal distribution of both out-of-context and not-out-of-context images in the test split.

4. Method

We consider a dataset of captioned images, where images may have more than one associated caption; however, we do not have any mapping for the objects referenced by the captions nor labels for which captions are out-of-context. We notice that in typical out-of-context use of images, different captions often describe the same object(s) but with a different meaning. For example, Fig. 2 shows several fact-checked examples where the two captions mean something very different, but describe the same parts of the image. Our goal is to take advantage of these patterns to detect scenarios and identify images used out-of-context.

4.1. Text Pre-processing

Since the captions used in our dataset are scraped from news websites, most captions contain proper nouns such as a person's name, city/country, venues, etc., which is hard for a model to interpret and thus makes it difficult to learn correct grounding. Hence, we used Spacy Named Entity Recognizer (NER)[3] to replace named entities in all the captions with their hypernyms. An example is shown in Figure 4. Note that we always input these cleaned and hypernymed captions to our model for all our experiments.



Robert "Grizz" Maguire leans up against his truck as he watches smoke pass in Granby

Person leans up against his truck as he watches smoke pass in location

Figure 4: Text Pre-processing: we pre-process captions to replace named entities in the image with their corresponding hypernyms. For instance, the person's name "Robert Grizz Maguire" is replaced with the hypernym *Person* and the town "Granby" is replaced with the hypernym *location*

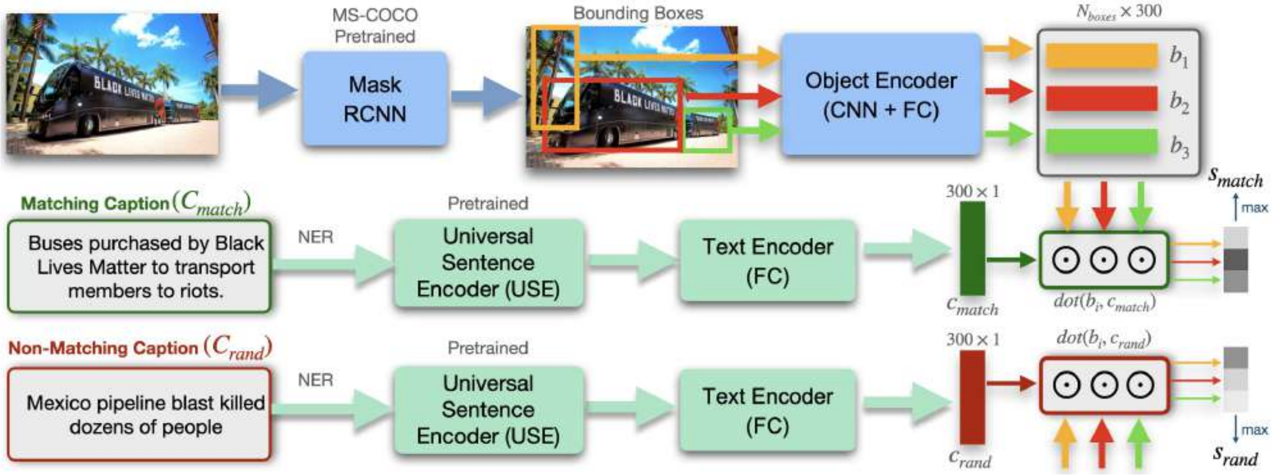
4.2. Image-Text Matching Model (Training)

The core of our method is a self-supervised training strategy leveraging co-occurrences of an image and its objects with several associated captions; i.e., we propose training an image and text based model based only on a set of captioned images. We thus formulate a scoring function to align objects in the image with the caption. Intuitively, an image-caption pair should have a high matching score if visual correspondences for the caption are present in the image, and a low score if the caption is unrelated to the image. To infer this correlation, we first use a pre-trained Mask-RCNN [16] to detect bounding boxes of objects in the image.

For each detected bounding box, we then feed the corresponding object regions to our Object Encoder, which uses a ResNet-50 [17] backbone from a pre-trained Mask-RCNN followed by RoIAlign, average pooling, and two fully-connected layers. As a a result, for each object, we obtain a 300-dimensional embedding vector.

In parallel, we consider the corresponding (pre-processed) image caption C_{match}, and sample a random caption from a different image in the dataset, C_{rand}. The

[3] https://spacy.io/api/entityrecognizer



MS-COCO Pretrained
Mask RCNN
Bounding Boxes
Object Encoder (CNN + FC)
$N_{boxes} \times 300$
b_1
b_2
b_3
s_{match}
max
Matching Caption (C_{match})
Buses purchased by Black Lives Matter to transport members to riots.
NER
Pretrained
Universal Sentence Encoder (USE)
Text Encoder (FC)
300×1
c_{match}
$dot(b_i, c_{match})$
Non-Matching Caption (C_{rand})
Mexico pipeline blast killed dozens of people
NER
Pretrained
Universal Sentence Encoder (USE)
Text Encoder (FC)
300×1
$dot(b_i, c_{rand})$
c_{rand}
max
s_{rand}

Figure 5: Self-supervised training of our method. First, a Mask-RCNN [16] backbone detects up to 10 object boxes in the image whose regions are embedded through our Object Encoder, providing a fixed-size embedding for each object. In parallel, two captions – one that appeared originally with the image C_{match} (matching caption) and another caption sampled randomly C_{rand} (non-matching caption) – and encoded using the Universal Sentence Encoder model (USE) [8]. The sentences embeddings are then passed to a shared Text Encoder that embeds them in the same multi-modal space. Similarities between object-caption pairs are computed with inner products (grayscale indicating score magnitude) and finally reduced to scores following Eq. 1.

captions are fed into a pre-trained sentence embedding model. Specifically, we use the Universal Sentence Encoder (USE) [8], which is based on a state-of-the-art transformer [41] architecture and outputs a 512-dimensional vector. We then process this vector with our Text Encoder (ReLU followed by one FC layer), which outputs a 300-dimensional embedding vector for each caption (to match the dimension of the object embeddings).

We then compare the visual and language embeddings with a dot product between the i-th box embedding b_i and the caption embedding c as a measure of similarity between image region i and caption C. The final image-caption score S_{IC} is obtained through a max function:

$$S_{IC} = \max_{i=1}^{N}(b_i^T c), \quad N = \#bboxes. \tag{1}$$

Our objective is to obtain higher scores for aligned image-text pairs (i.e., if an image appeared with the text irrespective of the context) than misaligned image-text pairs (i.e., some randomly-chosen text which did not appear with the image). We train the model with max-margin loss (Eq. 2) on the image-caption scores obtained above (Eq. 1). Note that we keep the weights of the Mask-RCNN [16] backbone and the USE [8] model frozen, using these models only for feature extraction.

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N} \max(0, (S_{IC}^r - S_{IC}^m) + \text{margin}), \tag{2}$$

where S_{IC}^r denotes the image-caption score of the random caption and S_{IC}^m the image-caption score for the matching caption. We refer to this model as *Image-Text Matching Model*; the training setup is visualized in Fig. 5. Note that during training, we do not aim to detect out-of-context images, but rather learn accurate image-caption alignments.

4.3. Out-of-Context Detection Model (Test Time)

The resulting Image-Text matching model obtained from training now provides an accurate representation of how likely a caption aligns with an image. In addition, as we explicitly model the object-caption relationship, the max operator in Eq. 1 implicitly gives a strong signal as to which object was selected to make that decision, thus providing spatial knowledge from the image. At test time, we consider an image and two captions that it appeared with, which may or may not be semantically similar. The Image-Caption1-Caption2 (I, C_1, C_2) triplet is used to predicts whether the image was used out-of-context with respect to the captions. Based on the evidence that out-of-context pairs correspond to same object in the image (c.f. Fig. 2), we propose a simple rule to detect such images, i.e., if two captions align with same object(s) in the image, but semantically convey different meanings, then the image with its two captions is classified as out-of-context. More specifically, we make use of the pre-trained model as follows:

(1) Using the Image-Text Matching model, we first compute the visual correspondences of the objects in the image for both captions. For each image-caption pair $\{I, C_j\}$, we choose the object box B_{I,C_j} with the highest score S_{I,C_j}

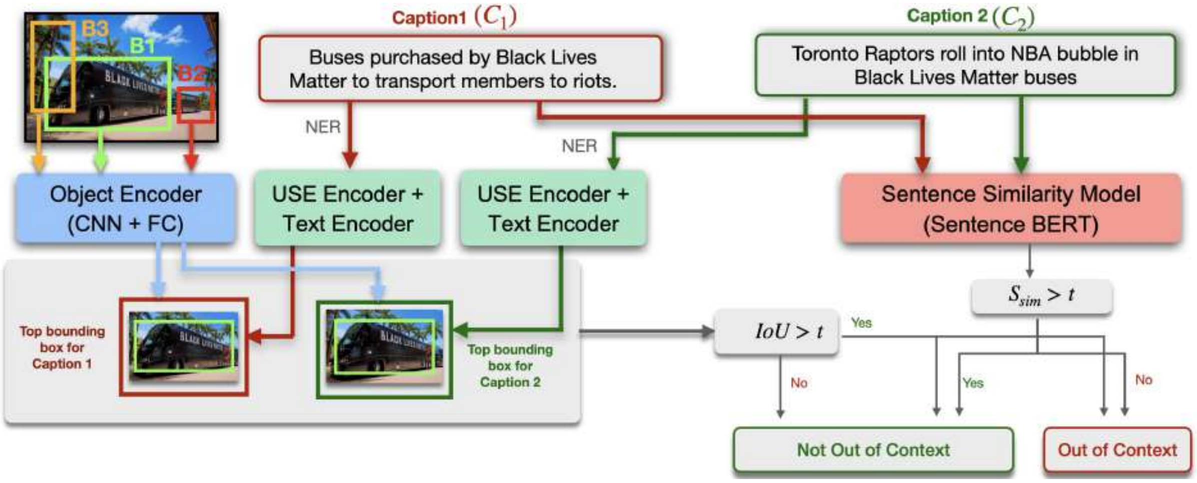

Caption1 (C_1)
Buses purchased by Black Lives Matter to transport members to riots.
Caption 2 (C_2)
Toronto Raptors roll into NBA bubble in Black Lives Matter buses
B3
B1
B2
NER
NER
Object Encoder (CNN + FC)
USE Encoder + Text Encoder
USE Encoder + Text Encoder
Sentence Similarity Model (Sentence BERT)
$S_{sim} > t$
Top bounding box for Caption 1
Top bounding box for Caption 2
$IoU > t$
Yes
No
Yes
No
Not Out of Context
Out of Context

Figure 6: Test time out-of-context detection. We take as input an image and two captions; we then use the trained Image-Text Matching model where we first pick the highest scoring object (based on Eq. 1) for both the captions. If the IoU between them > threshold t_i, we infer that image regions overlap. If the image regions overlap, we compute textual overlap S_{sim} with a pre-trained Sentence Similarity model SBert [43] and if $S_{sim} < t_s$, it implies that the two captions are semantically different, thus implying that the image is used out of context.

by Eq. 1 (strong alignment of caption with the object).

(2) We leverage a state-of-the-art SBERT [43] model that is trained on a Sentence Textual Similarity (STS) task. The SBERT model takes two captions C_1, C_2 as input and outputs a similarity score S_{sim} in the range $[0, 1]$ indicating semantic similarity between the two captions (higher score indicates same context):

$$S_{sim} = \text{STS}(C1, C2) \tag{3}$$

As a result, SBERT provides the semantic similarity between two captions, S_{sim}. In order to compute the visual mapping of the two captions with the image, we use the IoU overlap of the top bounding box for the two captions. We use thresholds $t_i = t_s = 0.5$ for all our experiments, both for IoU overlap and text overlap. If the visual overlap between image regions for the two captions is over a certain threshold $IoU(B_{I,C_1}, B_{I,C_2}) > t_i$ and the captions are semantically different ($S_{sim} < t_s$), we classify them as out-of-context (OOC). A detailed explanation is given in Fig. 6 and is as follows:

$$\text{OOC} = \begin{cases} \text{True}, & \text{if IoU}\big(B_{IC_1}, B_{IC_2}\big) > t \;\; \& \\ & \quad S_{sim}(C1, C2) < t \\ \text{False}, & otherwise \end{cases} \tag{4}$$

5. Results

5.1. Visual Grounding of Objects

Quantitative Results. Our model is trained in a self-supervised fashion only with matching and non-matching captions. To quantitatively evaluate how well our model learns the visual grounding of objects, we use the RefCOCO dataset [47] which has ground truth associations of the captions with the object bounding boxes. Note, however, that this evaluation is not our final task, but gives important insights into our model design. We experiment with three different model settings: (1) *Full-Image*, where an image is fed as input to the model and directly combined with the text embedding. (2) *Self-Attention*, where an image is fed as input to the model but combined with text using self-attention module. (3) *Bbox*, where only the detected objects are fed as input to model instead of full image. For this experiment, we use the ILSVRC 2012-pre-trained ResNet-18 [17] backbone to encode images and a one-layer LSTM model to encode text. Words are embedded using Glove [30] pretrained embeddings. Tab. 2 shows that using object-level features (given by bounding boxes) gives the best performing model. This is unsurprising, as object regions can provide a richer feature representation for the entities in the caption compared to the full image; but we also significantly outperform a self-attention alternative.

Qualitative Results. We visualize grounding scores in Fig. 7 from applying our image-text matching model to several image-caption pairs from the test set. The results indicate that our self-supervised matching strategy learns sufficient alignment between objects and captions to perform out-of-context image detection.

5.2. Out-of-Context Evaluation

Which is the best Text Embedding? To evaluate the effect of different text embeddings, we experiment with: (1) Pre-

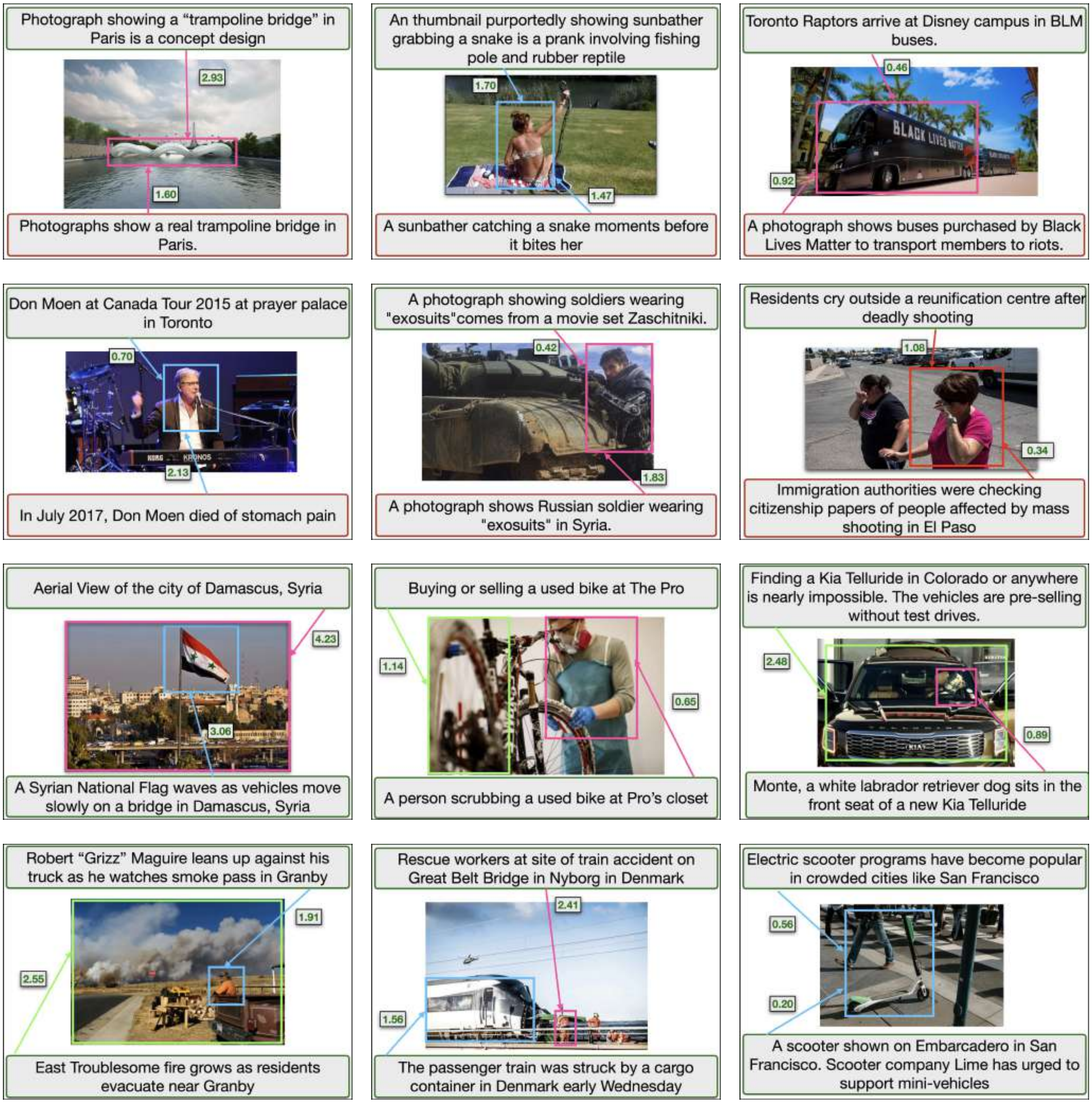

Photograph showing a “trampoline bridge” in Paris is a concept design
2.93
1.60
Photographs show a real trampoline bridge in Paris.
An thumbnail purportedly showing sunbather grabbing a snake is a prank involving fishing pole and rubber reptile
1.70
1.47
A sunbather catching a snake moments before it bites her
Toronto Raptors arrive at Disney campus in BLM buses.
0.46
0.92
A photograph shows buses purchased by Black Lives Matter to transport members to riots.
Don Moen at Canada Tour 2015 at prayer palace in Toronto
0.70
2.13
In July 2017, Don Moen died of stomach pain
A photograph showing soldiers wearing "exosuits"comes from a movie set Zaschitniki.
0.42
1.83
A photograph shows Russian soldier wearing "exosuits" in Syria.
Residents cry outside a reunification centre after deadly shooting
1.08
0.34
Immigration authorities were checking citizenship papers of people affected by mass shooting in El Paso
Aerial View of the city of Damascus, Syria
4.23
3.06
A Syrian National Flag waves as vehicles move slowly on a bridge in Damascus, Syria
Buying or selling a used bike at The Pro
1.14
0.65
A person scrubbing a used bike at Pro’s closet
Finding a Kia Telluride in Colorado or anywhere is nearly impossible. The vehicles are pre-selling without test drives.
2.48
0.89
Monte, a white labrador retriever dog sits in the front seat of a new Kia Telluride
Robert “Grizz” Maguire leans up against his truck as he watches smoke pass in Granby
1.91
2.55
East Troublesome fire grows as residents evacuate near Granby
Rescue workers at site of train accident on Great Belt Bridge in Nyborg in Denmark
2.41
1.56
The passenger train was struck by a cargo container in Denmark early Wednesday
Electric scooter programs have become popular in crowded cities like San Francisco
0.56
0.20
A scooter shown on Embarcadero in San Francisco. Scooter company Lime has urged to support mini-vehicles

Figure 7: Qualitative results of visual grounding of captions with the objects in the image. The top two rows show the grounding for out-of-context pairs and the bottom two rows show the grounding for pairs which are not out of context. We show object-caption scores for two captions per image. The captions with green border show the true captions and the captions with red border show the false caption. Scores indicate association of the most relevant object in the image with the caption.

trained word embeddings including Glove [30] and FastText [7] embedded via a one-layer LSTM model and (2) the Transformer based Sentence embeddings proposed by USE [8]. The results in Tab. 3 show that even though the match accuracy for all the methods is roughly the same (72%), using sentence embeddings significantly boosts our final out-of-context image detection accuracy of the model by 9% (from 76% to 85%). In addition, we also compare our results with state-of-the-art pretrained language baseline S-BERT [43] and outperform it by a margin of 8%.

Img Features.	Object IoU	Match Acc.
Bbox (GT)	0.36	0.89
Full-Image	0.11	0.63
Self-Attention	0.16	0.78
Bbox (Pred)	**0.27**	**0.88**

Table 2: Ablation of different settings for visual grounding in our self-supervised training setting (no loss on IoU).

Text Embed	Match Acc.	Context Acc.
S-Bert [43]	-	0.77
Glove [30]	0.72	0.76
FastText [7]	0.71	0.78
USE [8]	0.72	**0.85**

Table 3: Ablation with different text embeddings. Top row shows pre-trained S-Bert [43] language baseline evaluated on our test set.

How much training data is needed? Next, we analyze the effect of the available training data corpus for self-supervised training. We experiment with different percentages of training data size (w.r.t. to our full data) in Tab. 4. For these experiments, we use a pretrained USE [8] embedding with a one-layer FC model to encode text. We observed that a larger training set significantly improves the performance of the model. For instance, training with full-dataset (160k images) improves out-of-context detection accuracy by an absolute 13% (from 72% to 85%) compared to a model trained only with 10% of the dataset (16K images), thus benefiting from the large diversity in the dataset. We also notice that a relatively higher match accuracy leads to better out-of-context detection accuracy. This suggests that our proposed self-supervised learning strategy (trained with match vs no-match loss) effectively helps to improve out-of-context detection accuracy.

Dataset Size	Train Images	Match Acc.	Context Acc.
Ours (10%)	16K	0.64	0.72
Ours (20%)	32K	0.65	0.74
Ours (50%)	80K	0.68	0.77
Ours (100%)	160K	0.72	**0.85**

Table 4: Ablation with variations in number of train images w.r.t. our full corpus of 200K (160K train, 40k val) images. Using all available data achieves the best results.

Comparison with alternative approaches. Finally, we compare our best-performing model with other baselines, in particular, methods that work on rumor detection. Most other fake news detection methods are supervised, where the model takes an image and a caption as input and predicts the class label. EANN [45] and Jin et al. [18] were proposed specifically for Rumor/Fake News Classification; however, EmbraceNet [9] is a generic multi-modal classification method. Since neither of these methods perform self-supervised out-of-context image detection using object features (using bounding boxes), an out-of-the-box comparison is not feasible, and we must adapt these methods for our task. Following our training setup (Sec. 4.2), we first train these models for the binary task of image-text matching with the network architecture and losses proposed in their original papers. During test time, we then use Grad-CAM [36] to construct bounding boxes around activated image regions and perform out-of-context detection as described in Sec. 4.3. The results in Tab. 5 show that our model outperforms previous fake news detection methods for out-of-context detection by a large margin of 14% (from 71% to 85%). Overall, we achieve up to 85% out-of-context image detection accuracy.

Method	Match Acc.	Context Acc.
EANN [45]	0.57	0.63
EmbraceNet [9]	0.59	0.68
Jin *et al.* [18]	0.60	0.71
Ours	0.72	**0.85**

Table 5: We compare our method against three state-of-the-art methods. Our method outperforms all other methods, resulting in 85% out-of-context detection accuracy.

6. Conclusions

We have introduced an automated method to detect out-of-context images with respect to textual descriptions. Our key insight is to ground text to the image, as language-only analysis cannot effectively interpret semantically different captions that do not conflict due to referring to different objects in the image. Our approach thus ties two potential captions for an image to corresponding object regions for out-of-context determination, reaching up to 85% detection accuracy. We adopt a self-supervised training strategy to learn strong localization features based only on a set of captioned images, without the need for explicit out-of-context annotations. We further introduce a new dataset to benchmark this out-of-context task. Overall, we believe that our method takes an important step towards addressing misinformation in online news and social media platforms, thus supporting and scaling up fact-checking work. In particular, we hope that our new dataset, which we will publish along with this work, will lay a foundation to continue research along these lines to help online journalism and improve social media.

Acknowledgments

This work is supported by a TUM-IAS Rudolf Mößbauer Fellowship and the ERC Starting Grant Scan2CAD (804724). We would also like to thank New York Times Team for providing API support for dataset collection, Google Cloud for GCP credits, Farhan Abid for dataset annotation, and Angela Dai for video voice over.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network, Dec 2018.

[2] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting World Leaders Against Deep Fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, page 8, Long Beach, CA, June 2019. IEEE.

[3] Shivangi Aneja and Matthias Nießner. Generalized zero and few-shot transfer for facial forgery detection, 2020.

[4] NYT Developer API. Nyt developer api, 2020.

[5] Pepa Atanasova, Preslav Nakov, Lluís Màrquez i Villodre, A. Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James R. Glass. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, 11:1 – 27, 2019.

[6] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online, July 2020. Association for Computational Linguistics.

[7] P. Bojanowski, E. Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[8] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics.

[9] Jun-Ho Choi and Jong-Seok Lee. Embracenet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270, 2019.

[10] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection, 2020.

[11] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018.

[12] Deepfakes, 2017.

[13] Faceswap gan, 2018.

[14] Lisa Fazio. Out-of-context photos are a powerful low-tech form of misinformation, 2020.

[15] Maram Hasanain, Reem Suwaileh, T. Elsayed, A. Barrón-Cedeño, and Preslav Nakov. Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. task 2: Evidence and factuality. In *CLEF*, 2019.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[18] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 795–816, New York, NY, USA, 2017. Association for Computing Machinery.

[19] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. *The World Wide Web Conference*, 2019.

[20] Sejeong Kwon, Meeyoung Cha, K. Jung, Wei Chen, and Y. Wang. Prominent features of rumor propagation in online social media. *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108, 2013.

[21] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020.

[22] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts, 2018.

[23] X. Liu, A. Nourbakhsh, Quanzhi Li, R. Fang, and S. Shah. Real-time rumor debunking on twitter. In *CIKM '15*, 2015.

[24] J. Ma, Wei Gao, P. Mitra, Sejeong Kwon, Bernard J. Jansen, K. Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*, 2016.

[25] J. Ma, Wei Gao, and K. Wong. Detect rumor and stance jointly by neural multi-task learning. *Companion Proceedings of the The Web Conference 2018*, 2018.

[26] Jing Ma, Wei Gao, and K. Wong. Rumor detection on twitter with tree-structured recursive neural networks. In *ACL*, 2018.

[27] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[28] Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. Multi-hop fact checking of political claims, 2020.

[29] Britt Paris and Joan Donovan. Deepfakes and cheap fakes. In *United States of America: Data and Society*, 2019.

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[31] Ivan Petrov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. Deepfacelab: A simple, flexible and extensible face swapping framework, 2020.

[32] Poynter. Poynter : The international fact-checking network, 2020.

[33] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *EMNLP*, 2011.

[34] N. Ruchansky, Sungyong Seo, and Y. Liu. Csi: A hybrid deep model for fake news detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017.

[35] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.

[36] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626. IEEE Computer Society, 2017.

[37] Lanyu Shang, Yang Zhang, Daniel Zhang, and D. Wang. Fauxward: a graph neural network approach to fauxtography detection using social media comments. *Social Network Analysis and Mining*, 10:1–16, 2020.

[38] Reuben Tan, Bryan A. Plummer, and Kate Saenko. Detecting cross-modal inconsistency to defend against neural fake news, 2020.

[39] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[40] Slavena Vasileva, Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, and Preslav Nakov. It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1229–1239, Varna, Bulgaria, Sept. 2019. INCOMA Ltd.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

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

[43] Bin Wang and C.C. Jay Kuo. SBERT-WK: A sentence embedding method by dissecting bert-based word models. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:2146–2157, 2020.

[44] William Yang Wang. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[45] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery*, KDD '18, page 849–857, New York, NY, USA, 2018. Association for Computing Machinery.

[46] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[47] Licheng Yu, Patrick Poirson, Shan Yang, C. Alexander Berg, and L. Tamara Berg. Modeling context in referring expressions. *ECCV*, 2016.

[48] Daniel Zhang, Lanyu Shang, Biao Geng, Shuyue Lai, Ke Li, Hongmin Zhu, Tanvir Amin, and Dong Wang. Fauxbuster: A content-free fauxtography detector using social media comments. In *Proceedings of IEEE BigData 2018*, 2018.

[49] Zhe Zhao. Spotting icebergs by the tips: Rumor and persuasion campaign detection in social media. 2017.

[50] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection, Jul 2017.

[51] Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

Supplemental Material

In this supplemental document, we include additional details regarding the terminology used in our main paper; see Section A. In addition, in Section B, we document details regarding dataset cleanup/annotation and dataset statistics. We also visualize image-caption groundings for additional images from the test set. We then provide experimental details and metrics used for evaluation in our main paper in Section C. Finally, in Section D, we perform ablations with different augmentations on our dataset and experiments on RefCOCO [47] to evaluate which configuration is most effective to learn better object-caption groundings.

A. Definitions

- *Caption:* The fact-checking community uses multiple terms for the accompanying text that describes what is in the image, including the term "caption" or "claim". In this paper, we only use the word *caption* for the textual image descriptions. We do not use the word *claim*.
- *Conflicting-Image-Captions*: Describing falsely what is supported by an image with the aim to mislead audiences is often labelled by the fact-checkers as "misleading context", "out-of-context", "transformed context", "context re-targeting", "re-contextualization", "misrepresentation", and possibly other terms. A news article or social media post that is in question only shows the image with one caption and some fact-checking articles also only show a single instance of a false caption, but most fact-checks show at least two captions, usually one correct caption from the original dissemination of the image, and also the false caption. We describe an image to be *out-of-context* relative to two contradicting captions and we describe these two captions as *conflicting-image-captions*. However, we do not aim to determine which on of the two captions is false or true, or possibly if both captions are false. This is up to future research.

B. Dataset Details

We build a web-based annotation tool based on Python Flask and MySQL database to cleanup and label the samples of our dataset. We first use Google Cloud Translate API[4] library based on Google's language detection to translate non-English captions to English. We then cleanup the captions by removing image/source credits by using our web tool to avoid over-fitting to specific news sources. Finally, we manually annotate 1700 pairs from our Test Set. Note that the images used in our test split are gathered from Fact-Checks section of the Fact-Checking website Snopes[5] and from several other news websites. We ensure equal split of both Out-of-Context and Not-Out-of-Context images in the Test split for a fair evaluation. For every image, up to 4 caption pairs with diverse vocabulary are annotated. Duplicate caption pairs are removed during annotation from Test split. However, we did not clean up up duplicates from the training set. Several samples from our dataset along with the corresponding captions from different sources are shown in Fig 8. Then in Figure 9 and Figure 10 we visualize the image-caption groundings learnt by our the trained model. Word Clouds corresponding to different named entities in the captions from the dataset are shown in Figure 11. Table 6 lists number of captions from the dataset that contain these entities.

Named Entity	% of Captions
Person	0.74
Named Groups	0.20
Geopolitical entity	0.90
Named Locations	0.06
Named Events	0.05
Named Organizations	0.48
Work-Of-Art	0.05
Time	0.02

Table 6: The table documents total percentage of captions from the dataset that contain each of the named entities.

C. Setup & Evaluation Metrics

Experimental Setup: Our dataset consists of 160K training, 40K validation and 1700 test images. Every image is associated with one or more captions in the training set and only 2 captions in the test set. During training, we do not make use of out-of-context labels. We train our model only with *match* vs *no match loss* and evaluate for out-of-context image detection using the algorithm proposed in main paper. For all the experiments, we use a threshold value $t = 0.5$, to compute IoU overlap for image regions as well as for textual similarity.

Hyperparameters: We run all our experiments on a Nvidia GeForce GTX 2080 Ti card. We use a learning rate of 1e-3 with Adam optimizer for our model with decay when validation loss plateaus for 5 consecutive epochs, early stopping with a patience of 10 successive epochs on validation loss. For the baselines, we use their default hyperparameters used in their original papers.

Quantitative Results: In the main paper, we evaluate our method using three metrics explained below:

(1) *Match Accuracy* quantifies how well the caption grounds with the objects in the image. To compute this

[4] https://cloud.google.com/translate

[5] https://www.snopes.com/fact-check/

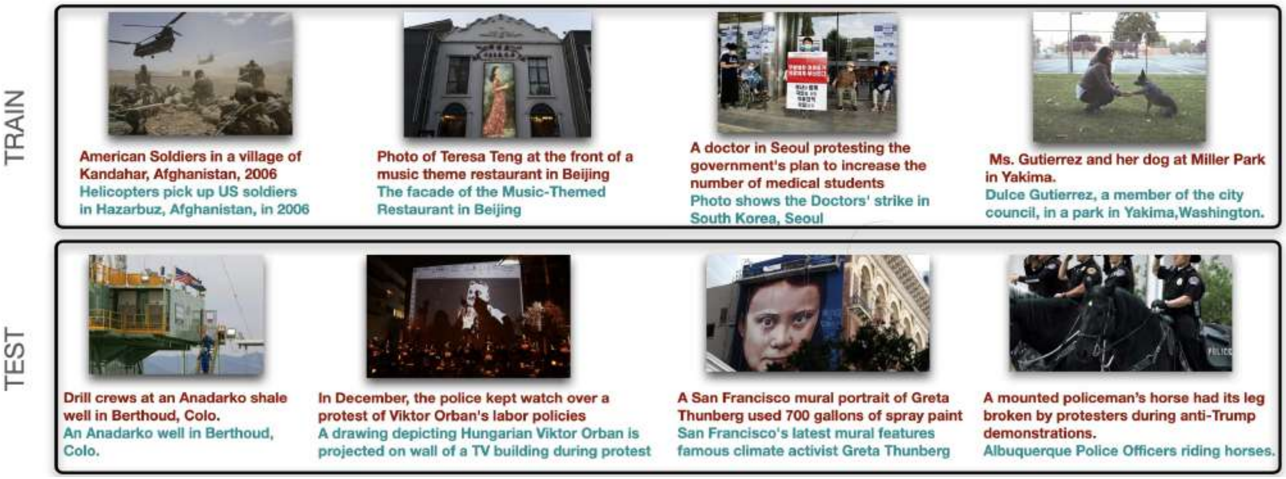

TRAIN
American Soldiers in a village of Kandahar, Afghanistan, 2006
Helicopters pick up US soldiers in Hazarbuz, Afghanistan, in 2006
Photo of Teresa Teng at the front of a music theme restaurant in Beijing
The facade of the Music-Themed Restaurant in Beijing
A doctor in Seoul protesting the government's plan to increase the number of medical students
Photo shows the Doctors' strike in South Korea, Seoul
Ms. Gutierrez and her dog at Miller Park in Yakima.
Dulce Gutierrez, a member of the city council, in a park in Yakima,Washington.
TEST
Drill crews at an Anadarko shale well in Berthoud, Colo.
An Anadarko well in Berthoud, Colo.
In December, the police kept watch over a protest of Viktor Orban's labor policies
A drawing depicting Hungarian Viktor Orban is projected on wall of a TV building during protest
A San Francisco mural portrait of Greta Thunberg used 700 gallons of spray paint
San Francisco's latest mural features famous climate activist Greta Thunberg
A mounted policeman's horse had its leg broken by protesters during anti-Trump demonstrations.
Albuquerque Police Officers riding horses.

Figure 8: Dataset: images and captions from the train set (top row); samples from the test set (bottom row).

numerically, we pair every image I in the dataset with a matching caption C_m and a random caption C_r and compute scores for both the captions with the image. A higher score for matching caption (s_m) compared to random caption (s_r), i.e., $s_m > s_r$ indicates correct prediction.

(2) *Object IoU* evaluates how well the caption grounds with the correct object in the image. For example, if the caption is "a woman buying groceries", the caption should ground with particular "woman" in the image. For this, we select the object with maximum score and compare it with GT object (provided in the dataset) using IoU overlap. Note that we do not have these annotations for our dataset, hence we evaluate this metric on the RefCOCO [47] dataset.

(3) *Out-of-Context (OOC) Accuracy* evaluates the model on the out-of-context classification task described in main paper We collected an equal number of samples for both the classes (Out-of-Context and Not-Out-of-Context) for fair evaluation.

D. Additional Experiments

D.1. Does Data Augmentation Help?

To mitigate over-fitting and improve the overall performance of our model, we apply several augmentations during training. For each of the detected objects in the image, we apply color jitter by varying the hue and saturation by a factor of 0.2 and applied random horizontal flip and random rotate (by angle of 10 degrees). Note that we do not apply these augmentations during inference. And for each of the captions, we pre-process them by replacing named entities with the corresponding hypernyms as shown in the main paper. Text pre-processing is always applied to all the captions in the dataset for all the splits. We analyze the effect of each of these augmentations in Table 7.

Augmentation	Context Acc.
J	0.73
J + R	0.74
NER	0.78
J + R + NER	**0.85**

Table 7: The table illustrates the effect applying different augmentations during training. J (Jitter) and R (Random Rotate) are applied on detected objects and NER (Named Entity Recognition Replacement) is applied on textual caption. We notice that applying all these augmentations together gives us the best performing model

D.2. How to Encode Text?

We evaluate which of the two text encodings (word-level or text-level) learn better object-caption groundings. For word-level encoding, we compute fixed-size embeddings for every word in the caption and combine with every detected object during experiments. For text-level encoding, a single embedding is generated for the entire caption and combined with detected objects in the image. Results are shown in Table 8. We notice that encoding the entire caption as single embedding consistently outperforms word-level encodings given that text-level embeddings contain richer information about certain object attributes like relative positions thereby learning better visual grounding. Hence, we used text-embeddings for all our experiments in the main paper.

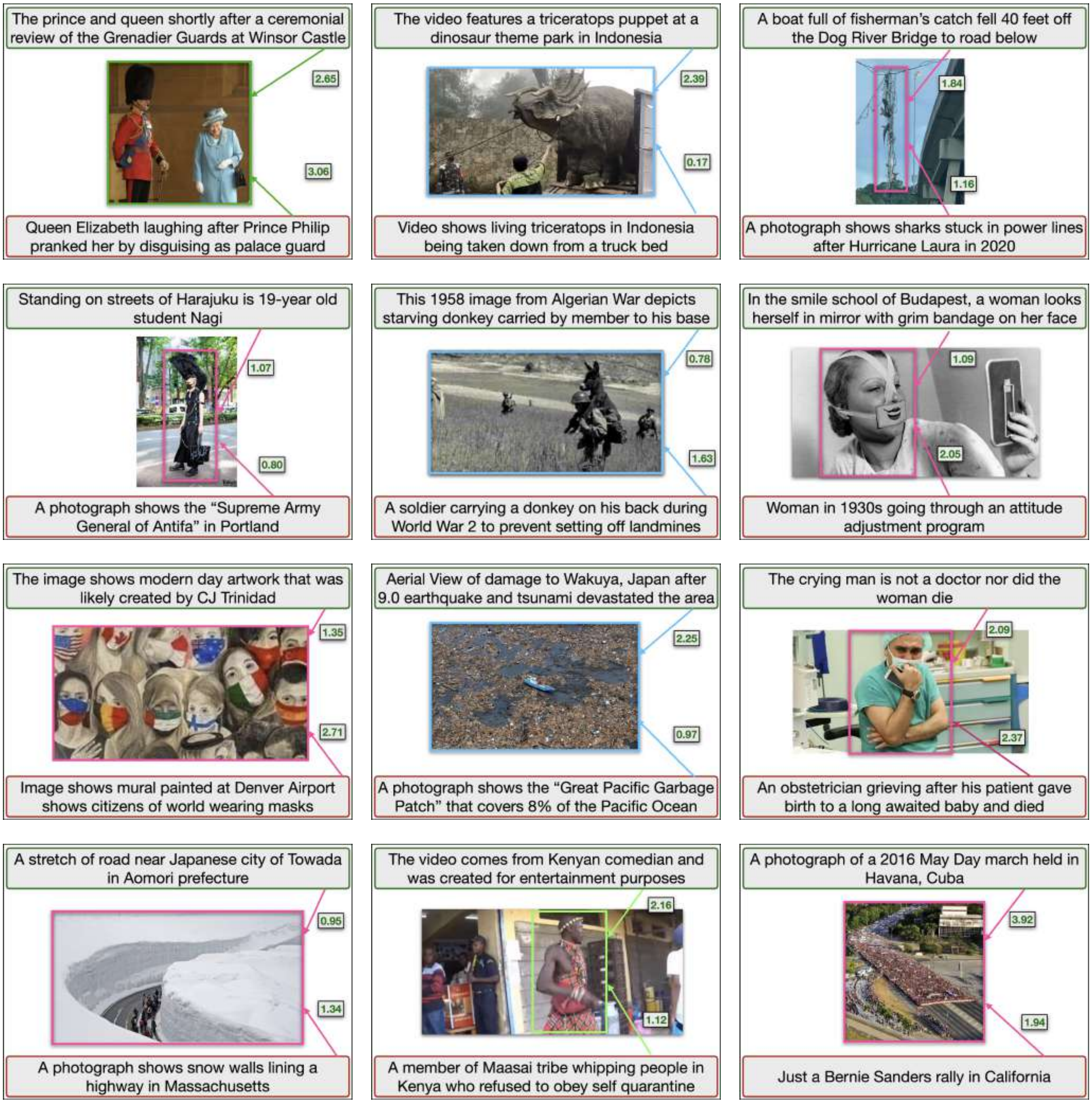

The prince and queen shortly after a ceremonial review of the Grenadier Guards at Winsor Castle
2.65
3.06
Queen Elizabeth laughing after Prince Philip pranked her by disguising as palace guard
The video features a triceratops puppet at a dinosaur theme park in Indonesia
2.39
0.17
Video shows living triceratops in Indonesia being taken down from a truck bed
A boat full of fisherman's catch fell 40 feet off the Dog River Bridge to road below
1.84
1.16
A photograph shows sharks stuck in power lines after Hurricane Laura in 2020
Standing on streets of Harajuku is 19-year old student Nagi
1.07
0.80
A photograph shows the "Supreme Army General of Antifa" in Portland
This 1958 image from Algerian War depicts starving donkey carried by member to his base
0.78
1.63
A soldier carrying a donkey on his back during World War 2 to prevent setting off landmines
In the smile school of Budapest, a woman looks herself in mirror with grim bandage on her face
1.09
2.05
Woman in 1930s going through an attitude adjustment program
The image shows modern day artwork that was likely created by CJ Trinidad
1.35
2.71
Image shows mural painted at Denver Airport shows citizens of world wearing masks
Aerial View of damage to Wakuya, Japan after 9.0 earthquake and tsunami devastated the area
2.25
0.97
A photograph shows the "Great Pacific Garbage Patch" that covers 8% of the Pacific Ocean
The crying man is not a doctor nor did the woman die
2.09
2.37
An obstetrician grieving after his patient gave birth to a long awaited baby and died
A stretch of road near Japanese city of Towada in Aomori prefecture
0.95
1.34
A photograph shows snow walls lining a highway in Massachusetts
The video comes from Kenyan comedian and was created for entertainment purposes
2.16
1.12
A member of Maasai tribe whipping people in Kenya who refused to obey self quarantine
A photograph of a 2016 May Day march held in Havana, Cuba
3.92
1.94
Just a Bernie Sanders rally in California

Figure 9: Qualitative results of visual grounding of captions with the objects in the image for Out-of-Context image-caption pairs. We show object-caption scores for two captions per image. The captions with green border show the true captions and captions with red border show the false/misleading captions. Scores indicate association of the most relevant object in the image with the caption.

D.3. Do Object-Detector Features Help?

We experiment if same backbone network (ResNet-50 [17]) trained on different tasks (Image Classification and Object Detection) can help learn better object-level features ultimately learning better object-caption groundings. Results are shown in Table 9. We notice that Mask-RCNN pre-trained backbone performs better in comparison to ImageNet pre-trained backbone verifying that object detector features help learn better object-caption groundings. Hence, we used Mask-RNN pre-trained backbone as image encoder for our experiments in the main paper.

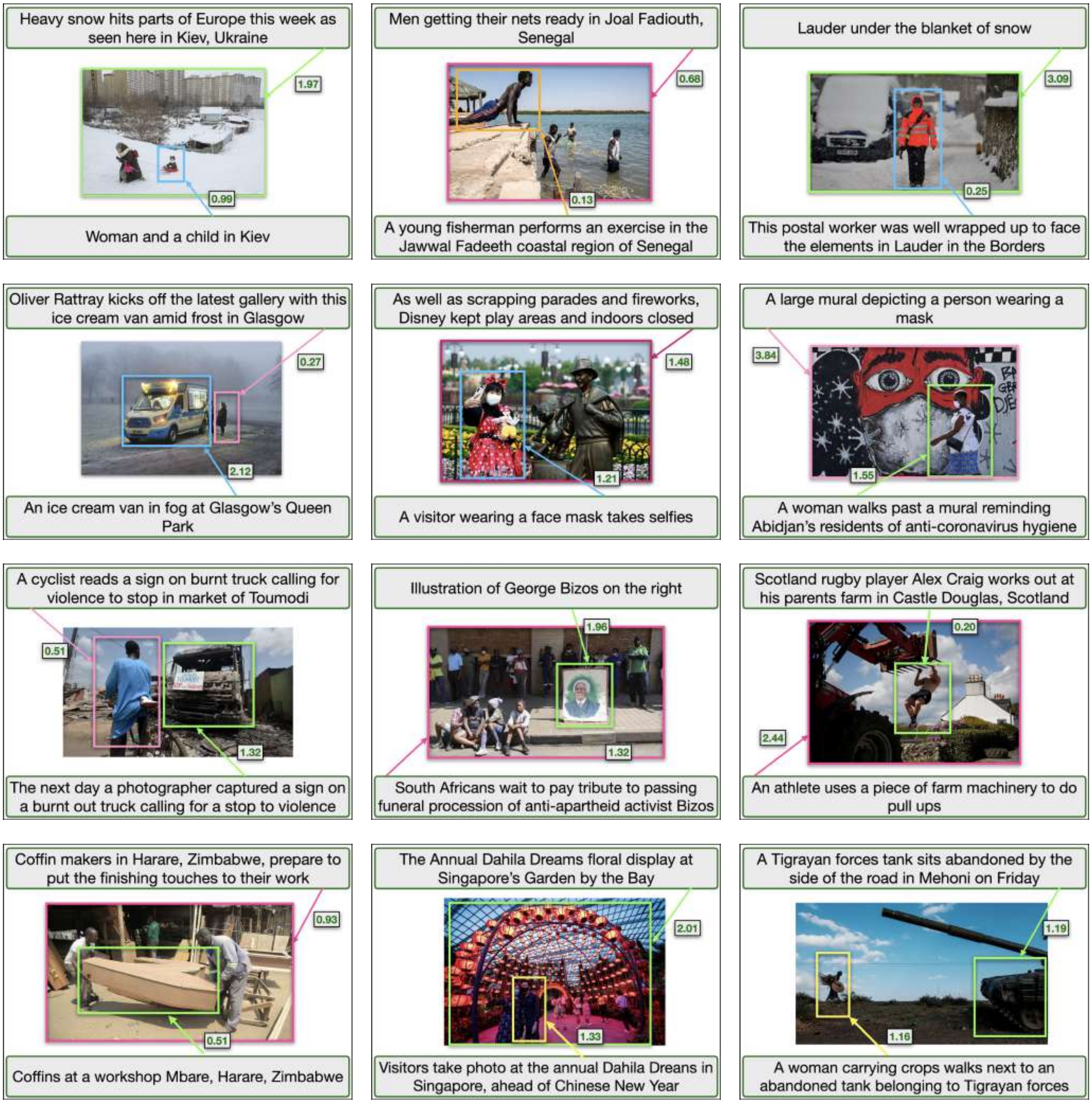
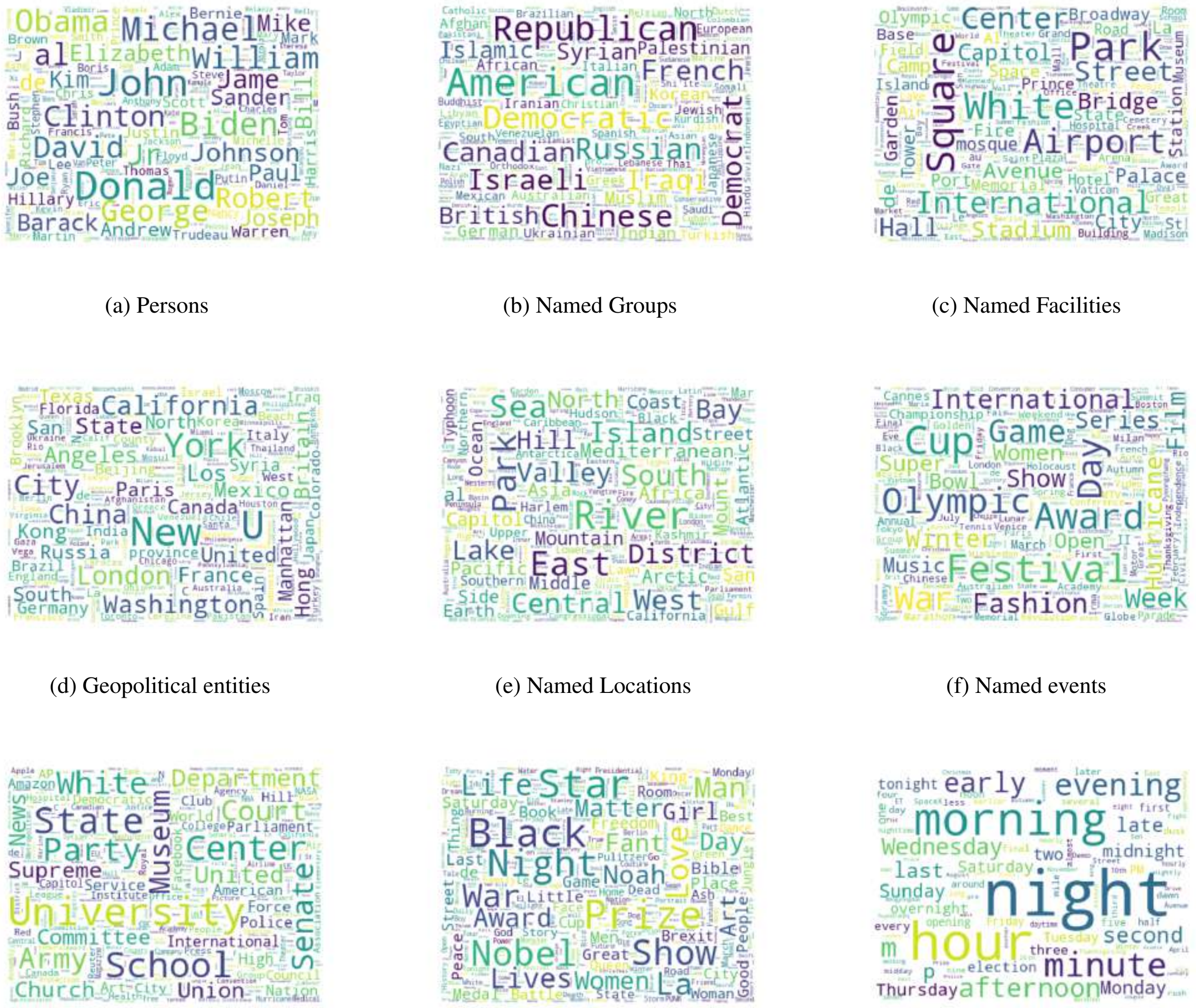

Heavy snow hits parts of Europe this week as seen here in Kiev, Ukraine
1.97
0.99
Woman and a child in Kiev
Men getting their nets ready in Joal Fadiouth, Senegal
0.68
0.13
A young fisherman performs an exercise in the Jawwal Fadeeth coastal region of Senegal
Lauder under the blanket of snow
3.09
0.25
This postal worker was well wrapped up to face the elements in Lauder in the Borders
Oliver Rattray kicks off the latest gallery with this ice cream van amid frost in Glasgow
0.27
2.12
An ice cream van in fog at Glasgow's Queen Park
As well as scrapping parades and fireworks, Disney kept play areas and indoors closed
1.48
1.21
A visitor wearing a face mask takes selfies
A large mural depicting a person wearing a mask
3.84
1.55
A woman walks past a mural reminding Abidjan's residents of anti-coronavirus hygiene
A cyclist reads a sign on burnt truck calling for violence to stop in market of Toumodi
0.51
1.32
The next day a photographer captured a sign on a burnt out truck calling for a stop to violence
Illustration of George Bizos on the right
1.96
1.32
South Africans wait to pay tribute to passing funeral procession of anti-apartheid activist Bizos
Scotland rugby player Alex Craig works out at his parents farm in Castle Douglas, Scotland
0.20
2.44
An athlete uses a piece of farm machinery to do pull ups
Coffin makers in Harare, Zimbabwe, prepare to put the finishing touches to their work
0.93
0.51
Coffins at a workshop Mbare, Harare, Zimbabwe
The Annual Dahila Dreams floral display at Singapore's Garden by the Bay
2.01
1.33
Visitors take photo at the annual Dahila Dreans in Singapore, ahead of Chinese New Year
A Tigrayan forces tank sits abandoned by the side of the road in Mehoni on Friday
1.19
1.16
A woman carrying crops walks next to an abandoned tank belonging to Tigrayan forces

Figure 10: Qualitative results of visual grounding of captions with the objects in the image for Not-Out-of-Context image-caption pairs. We show object-caption scores for two captions per image. The captions with green border show the true captions. Scores indicate association of the most relevant object in the image with the caption.



(a) Persons

(b) Named Groups

(c) Named Facilities

(d) Geopolitical entities

(e) Named Locations

(f) Named events

(g) Named Organizations

(h) Work Of Art

(i) Time

Figure 11: Figure shows Word Clouds corresponding to different named entities used for text pre-processing for the experiments conducted in the main paper. Persons (a) shows names of the person; Named Groups (b) include Nationalities / religious / political groups; Named Facilities (c) include Buildings, airports, highways, bridges, etc; Geopolitical entities (d) include Countries, Cities, States; Named Locations (e) include locations that are not geo-political entities like mountain ranges, bodies of water; Named events (f) include certain famous events like like battles, wars, sports events; Named Organizations (g) include names of Companies, agencies, institutions; Work Of Art (h) includes Titles of books, songs, awards, etc; Time (i) includes times smaller than a day

Method	Embed. Type	Object IoU	Match Acc.
Bbox (*GT*)	Word	0.33	0.85
	Text	**0.36**	**0.89**
Bbox (*Pred*)	Word	0.20	0.80
	Text	**0.27**	**0.88**

Table 8: Ablations study of word vs sentence embeddings on RefCOCO dataset [47]. Text is embedded using Glove [30] pre-trained embeddings. *GT* indicates Ground Truth bounding boxes used for experiments. *Pred* indicates bounding boxes predicted by Mask-RCNN used for experiments.

Object Features	Bbox	Object IoU	Match Acc.
ImageNet	GT	0.39	0.90
Mask-RCNN [16]		**0.45**	**0.92**
ImageNet	Pred	0.32	0.88
Mask-RCNN [16]		**0.38**	**0.91**

Table 9: We evaluate different backbones for our self-supervised training setting.(GT) indicates we used ground truth boxes for experiments and (Pred) indicates we used bounding boxes predicted by pre-trained Mask-RCNN. On average, we have 12.6 bboxes per image (Random Chance = 0.07)