COSMOS: Catching Out-of-Context Image Misuse with Self-Supervised Learning

Shivangi Aneja[1], **Chris Bregler**[2], **Matthias Nießner**[1]

[1] Technical University of Munich
[2] Google AI

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. We propose a new method that automatically highlights out-of-context image and text pairs, for assisting fact-checkers. Our key insight is to leverage the 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 the 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.

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 (Lu 2018; Paris and Donovan 2019; Petrov et al. 2020) 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 (Rössler et al. 2019; Nguyen, Yamagishi, and Echizen 2019; Li and Lyu 2019; Yang, Li, and Lyu 2019; Zhou et al. 2017; Cozzolino et al. 2018; Afchar et al. 2018; Agarwal et al. 2019; Li et al. 2020; Verdoliva 2020; Aneja and Nießner 2020; Davide Cozzolino and Andreas Rössler and Justus Thies and Matthias Nießner and Luisa Verdoliva 2021). 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 (Fazio 2020). Fact checkers refer to this as out-of-context use of images, where an image is recontextualized with one or 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 message. 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 organizations, most of them being part of the International Factchecking Network. 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 the 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. (Wang 2017; Thorne et al. 2018). 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.

An image serves as evidence of the event described by a news caption. If two captions associated with an image are valid, then they should describe the same event. If they align with the same object(s) in the image, then they are broadly conveying same information (see Fig. 2). Based on these patterns, we define out-of-context use of images as presenting an image as an evidence of *untrue* and/or *unrelated* event(s). If the two captions refer to same object in the image, but are semantically different, i.e. associate the same subject to different events, then it indicates out-of-context use of image (Case 1 from Fig. 1). However, if the captions correspond to the same event irrespective of the object(s) the captions describe, then it is defined as not-out-of-context (Case 2 from Fig. 1) use of images. Note that a not-out-of-context scenario makes no conclusions regarding the veracity of the statements.

Our work differs from standard fake news detection meth-

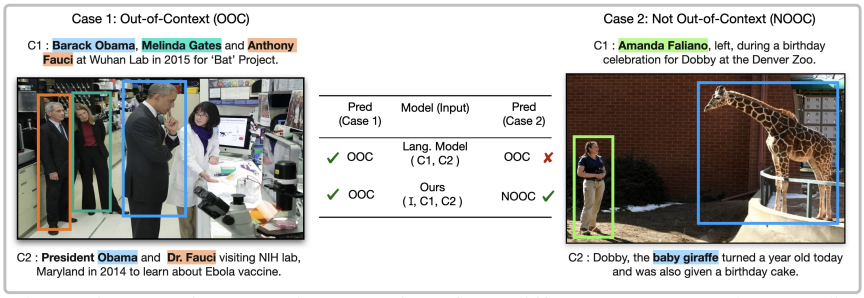

Figure 1: Our method takes as input an image and two captions from different sources, and we predict whether the image has been used out-of-context or not. We show that it is critical to the task to ground the captions w.r.t. image, and it is insufficient to consider only the captions; e.g., a language-only 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.

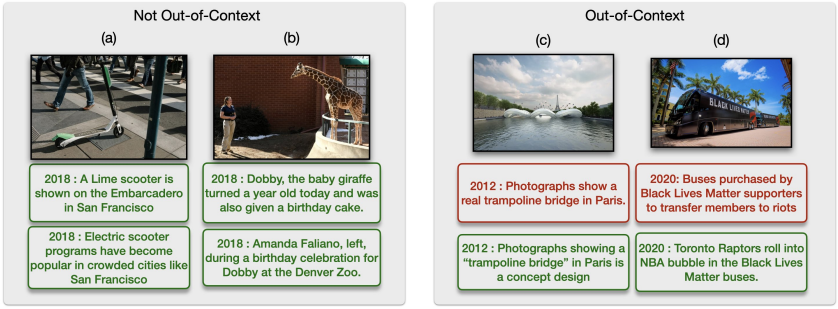

Figure 2: Examples images with associated captions from our dataset. Red denotes fake captions and green shows the real captions along with year published. Left: Multiple captions associated with the image indicating not-out-of-context situation. In these cases, the captions might describe same object (scooter in Fig(a)) or different objects (giraffe Dobby and person Amanda in Fig(b)), but they refer to the same event. Right: For the out-of-context scenario, it is observed that the captions describe the same object (bridge in Fig(c) and buses in Fig(d)) but are semantically different or refer to different set of events.

ods that aim to identify fake news posts with or without images. For the task of fake news detection, the images shared with fake news posts could be *photoshopped/manipulated*, signalling that its a false information. However, for out-of-context image misuse, the images are always *genuine*, which makes the task challenging.

To accelerate the detection of 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 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 semantics are different (for e.g. same person described by the two captions differently in context of event, time, place, etc), we infer that the image is used out-of-context.

Our method detects conflicting image caption triplets which indicates miscontextualization of the image. We do not identify which of the two captions is false or true. We argue that for a given real image, detecting whether its associated caption is false is a challenging task even for human moderators without prior information about image origin. Luo et al (Luo, Darrell, and Rohrbach 2021) verified this with a study on human evaluators (who were instructed not use search engines) where the average human accuracy came out to be roughly 65%. Additionally, false positives/negatives for certain sensitive topics (e.g. terrorism, attacks) can be dangerous and human intervention is required to make a decision, which is why we do not make hard decision on truth value of the caption. In particular, we consider the scenario of assisting fact checkers by highlighting conflicting images-caption triplets to narrow down their search space, which remains one of the main challenges to this day.

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.

Related Work

Fake News & Rumor Detection. Fake news and rumor detection methods have a long history (Qazvinian et al. 2011; Kwon et al. 2013; Liu et al. 2015; Ma et al. 2016; Zhao 2017; Ruchansky, Seo, and Liu 2017; Ma, Gao, and Wong 2018a,b) 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. (Kwon et al. 2013) analyzed structural, temporal, and linguistic aspects of the user tweets and modelled them using SVM to detect the spread of rumors. Ma et al. (Ma, Gao, and Wong 2018b) examined propagation patterns in tweets and applied tree-structured recursive neural networks for rumor representation learning and classification. Tan et al. (Tan, Plummer, and Saenko 2020) 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 (Wang 2017; Thorne et al. 2018; Hasanain et al. 2019; Atanasova et al. 2020; Ostrowski et al. 2021; Atanasova et al. 2019; Vasileva et al. 2019) have been developed to reduce the manual fact-checking overhead. For instance, Wang et al. (Wang 2017) 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. (Vasileva et al. 2019) proposed a technique to estimate check-worthiness of claims from political debates. Atanasova et al. (Atanasova et al. 2020) propose a multi-task learning technique to classify veracity of claim and generate fact-checked explanations at the same time.

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 (Jin et al. 2017; Zhang et al. 2018; Shang et al. 2020; Wang et al. 2018; Zlatkova, Nakov, and Koychev 2019; Khattar et al. 2019) have been proposed recently. For instance, Jin et al. (Jin et al. 2017) use attention-based RNNs to fuse multiple modalities to detect rumors/fake claims.

Split	Primary Source	No. of Images	Context Annotation
Train	News Outlets[1]	160K	✗
Val	News Outlets	40K	✗
Test	News Outlets, Snopes	1700	✓

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

Only a handful of images are used to spread misinformation compared to the amount of images shared on internet every day. This makes it difficult to construct large-scale supervised dataset for the task. Even most real-world supervised fake news detection datasets are sparse (roughly 1K images) in terms of images (Zlatkova, Nakov, and Koychev 2019). An alternate avenue is to synthetically generate fake captions (Luo, Darrell, and Rohrbach 2021). We, however, propose a self-supervised method and train and test on real data to replicate real-world scenario.

Out-of-Context Detection Dataset

Dataset Collection: We gathered our dataset from two primary sources, *news websites*[1] and *fact-checking websites*. We collect our dataset in two steps: (1) First, using publicly available news channel APIs (Times 2020), we scraped images along with 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. The second step is not necessary, but we collect these captions for increased dataset diversity. 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. We currently aim to detect conflicting-image-captions in English language only.

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 (Times 2020), 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.

Train, Val & Test Set: For training, we used images scraped from news websites. We consider several news sources[1] to gather the images. In total, we gathered around 200K images with 450K captions, 20% of which we use in the validation split. 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 verify and annotate

[1]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

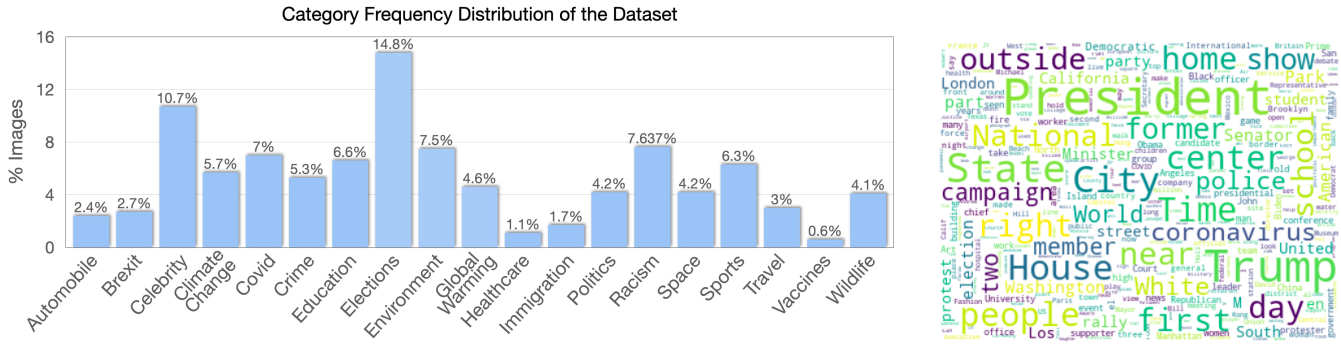

Figure 3: High-level overview of our dataset: (left) category-wise frequency distribution of the images; (right) word cloud representation of captions and claims from the dataset.

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.

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.

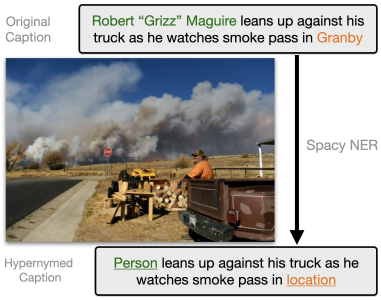

Figure 4: Text Pre-processing: we pre-process captions to replace named entities in the image with their corresponding hypernyms. For instance, the person's name "Robert Grizz Maguire" is replaced with the hypernym *Person* and the town "Granby" is replaced with the hypernym *location*

Text Pre-processing: Since the captions used in our dataset are scraped from news websites, most captions consist of 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 (details in supplemental). Hence, we used Spacy Named Entity Recognizer (NER)[2] 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 matching model for all our experiments. For more analysis, refer to supplemental.

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 (He et al. 2017) 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 (He et al. 2016) 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 captions are fed into a pre-trained sentence embedding model. Specifically, we use the Universal Sentence Encoder (USE) (Cer et al. 2018), which is based on a state-of-the-art transformer (Vaswani et al. 2017) 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 irrespec-

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

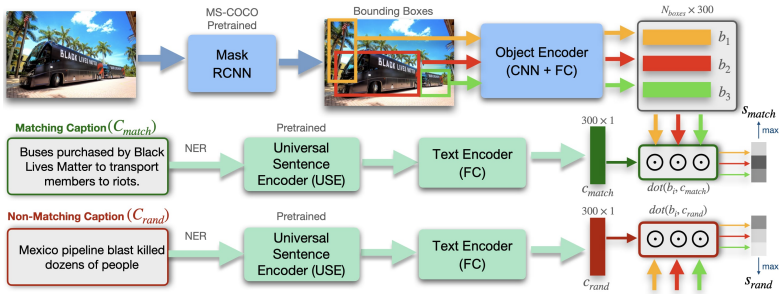

Figure 5: Self-supervised training of our method. First, a Mask-RCNN (He et al. 2017) backbone detects up to 10 object boxes in the image whose regions are embedded through our Object Encoder, providing a fixed-size embedding for each object. In parallel, two captions – one that appeared originally with 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) (Cer et al. 2018). 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.

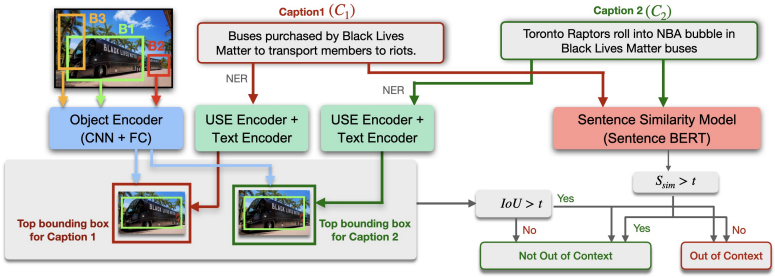

Figure 6: Test time out-of-context detection. We take as input an image and two captions; we then use the trained Image-Text Matching model where we first pick the highest scoring object (based on Eq. 1) for both the captions. If the IoU between them $>$ threshold t_i, we infer that image regions overlap. If the image regions overlap, we compute textual overlap S_{sim} with a pre-trained Sentence Similarity model SBert (Wang and Kuo 2020) and if $S_{sim} < t_s$, it implies that the two captions are semantically different, thus implying out-of-context use of image.

tive 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 (He et al. 2017) backbone and the USE (Cer et al. 2018) model frozen, using these models only for feature extraction.

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N} \max(0, (S^r_{IC} - S^m_{IC}) + \text{margin}), \qquad (2)$$

where S^r_{IC} denotes the image-caption score of the random caption and S^m_{IC} 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.

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

(2) We leverage a state-of-the-art SBERT (Wang and Kuo 2020) 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):

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

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:

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

Results

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 (Yu et al. 2016) 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 (He et al. 2016) backbone to encode images and a one-layer LSTM model to encode text. Words are embedded using Glove (Pennington, Socher, and Manning 2014) pre-trained 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.

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).

Out-of-Context Evaluation

Which is the best Text Embedding? To evaluate the effect of different text embeddings, we experiment with: (1) Pre-trained word embeddings including Glove (Pennington, Socher, and Manning 2014) and FastText (Bojanowski et al. 2017) embedded via a one-layer LSTM model and (2) the Transformer based Sentence embeddings proposed by USE (Cer et al. 2018). The results in Tab. 3 show that even though the match accuracy for all the methods is roughly the same (72%), using USE embeddings (Cer et al. 2018) 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 (Wang and Kuo 2020) and outperform it by a margin of 8%.

Text Embed	Match Acc.	Context Acc.
S-Bert (Wang and Kuo 2020)	-	0.77
Glove (Pennington, Socher, and Manning 2014)	0.72	0.76
FastText (Bojanowski et al. 2017)	0.71	0.78
USE (Cer et al. 2018)	0.72	**0.85**

Table 3: Ablation with different text embeddings. Top row shows pre-trained S-Bert (Wang and Kuo 2020) language baseline evaluated on our test set.

Method	Match Acc.	Context Acc.
EANN (Wang et al. 2018)	0.57	0.63
EmbraceNet (Choi and Lee 2019)	0.59	0.68
Jin *etal* (Jin et al. 2017)	0.60	0.71
Ours	0.72	**0.85**

Table 4: 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.

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 (Wang et al. 2018) and (Jin et al. 2017) were proposed specifically for Rumor/Fake News Classification; however, EmbraceNet (Choi and Lee 2019) 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, 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 GradCAM (Selvaraju et al. 2017) to construct bounding boxes around activated image regions and perform out-of-context detection as described in method section. The results in Tab. 4 show that our model outperforms previous fake news detection methods for out-of-context detection

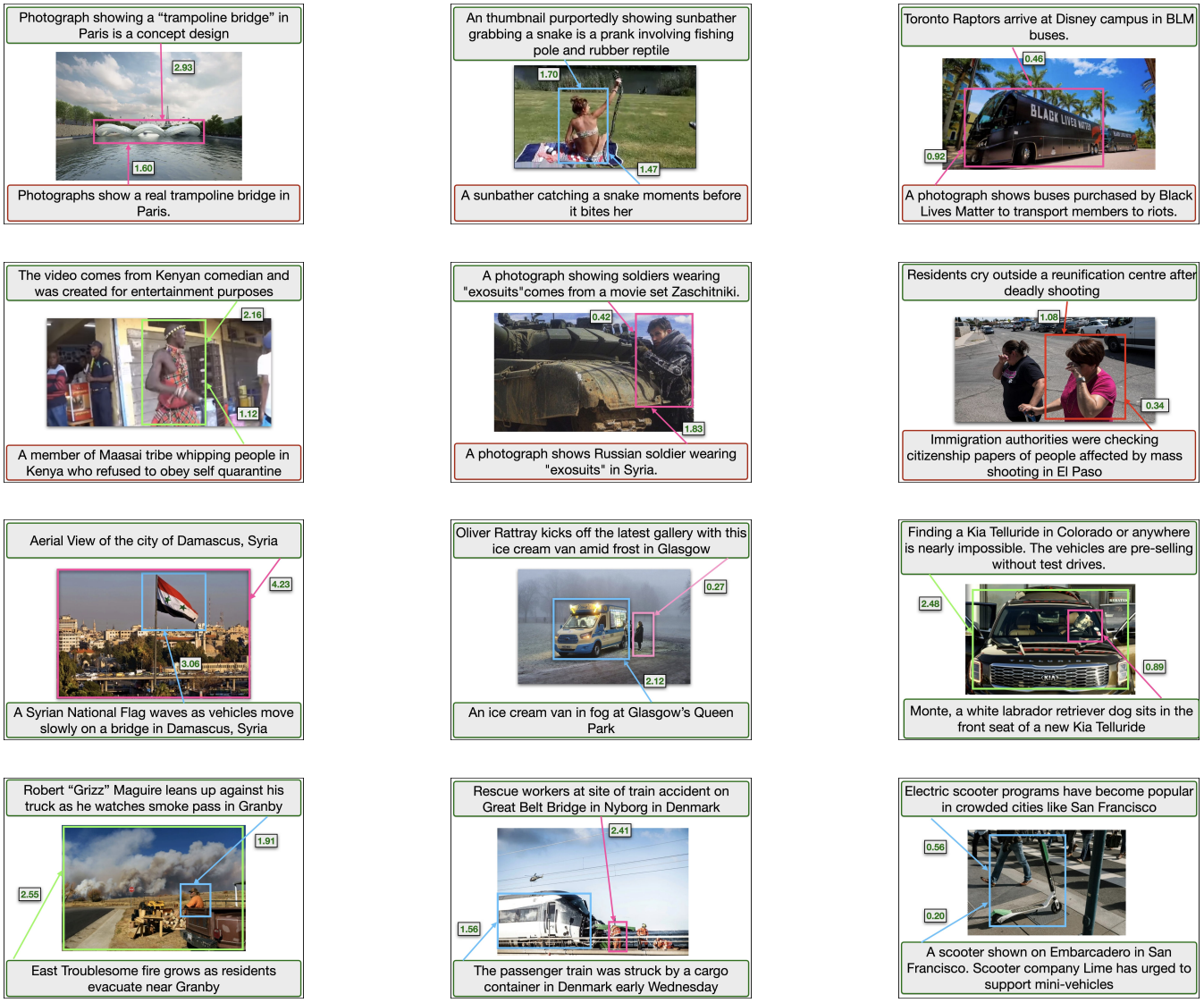

Figure 7: Qualitative results of visual grounding of captions with the objects in the image. The top two rows show the grounding for out-of-context pairs and the bottom two rows show the grounding for pairs which are not out of context. We show object-caption scores for two captions per image. The captions 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.

by a large margin of 14% (from 71% to 85%). Overall, we achieve up to 85% out-of-context detection accuracy.

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 with 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 85% detection accuracy. We adopt 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.

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. MesoNet: a Compact Facial Video Forgery Detec-

tion Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE. ISBN 9781538665367.

Agarwal, S.; Farid, H.; Gu, Y.; He, M.; Nagano, K.; and Li, H. 2019. Protecting World Leaders Against Deep Fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 8. Long Beach, CA: IEEE.

Aneja, S.; and Nießner, M. 2020. Generalized Zero and Few-Shot Transfer for Facial Forgery Detection. In *ArXiv preprint arXiv:2006.11863*.

Atanasova, P.; Nakov, P.; i Villodre, L. M.; Barrón-Cedeño, A.; Karadzhov, G.; Mihaylova, T.; Mohtarami, M.; and Glass, J. R. 2019. Automatic Fact-Checking Using Context and Discourse Information. *Journal of Data and Information Quality (JDIQ)*, 11: 1 – 27.

Atanasova, P.; Simonsen, J. G.; Lioma, C.; and Augenstein, I. 2020. Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7352–7364. Online: Association for Computational Linguistics.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5: 135–146.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; St. John, R.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; Strope, B.; and Kurzweil, R. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–174. Brussels, Belgium: Association for Computational Linguistics.

Choi, J.-H.; and Lee, J.-S. 2019. EmbraceNet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51: 259–270.

Cozzolino, D.; Thies, J.; Rössler, A.; Riess, C.; Nießner, M.; and Verdoliva, L. 2018. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. In *arXiv preprint arXiv:1812.02510*.

Davide Cozzolino and Andreas Rössler and Justus Thies and Matthias Nießner and Luisa Verdoliva. 2021. ID-Reveal: Identity-aware DeepFake Video Detection. In *ICCV*.

Fazio, L. 2020. Out-of-context photos are a powerful low-tech form of misinformation. https://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959. Accessed: 2020-11-31.

Hasanain, M.; Suwaileh, R.; Elsayed, T.; Barrón-Cedeño, A.; and Nakov, P. 2019. Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 2: Evidence and Factuality. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, 795–816. New York, NY, USA: Association for Computing Machinery. ISBN 9781450349062.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. *The World Wide Web Conference*.

Kwon, S.; Cha, M.; Jung, K.; Chen, W.; and Wang, Y. 2013. Prominent Features of Rumor Propagation in Online Social Media. In *2013 IEEE 13th International Conference on Data Mining*, 1103–1108.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5001–5010.

Li, Y.; and Lyu, S. 2019. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Liu, X.; Nourbakhsh, A.; Li, Q.; Fang, R.; and Shah, S. 2015. Real-time Rumor Debunking on Twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*.

Lu, S.-A. 2018. Faceswap Gan. https://github.com/shaoanlu/faceswap-GAN. Accessed: 2020-11-31.

Luo, G.; Darrell, T.; and Rohrbach, A. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6801–6817. Association for Computational Linguistics.

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B. J.; Wong, K.; and Cha, M. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3818–3824.

Ma, J.; Gao, W.; and Wong, K. 2018a. Detect Rumor and Stance Jointly by Neural Multi-task Learning. *Companion Proceedings of the The Web Conference 2018*.

Ma, J.; Gao, W.; and Wong, K. 2018b. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1980–1989. Melbourne, Australia: Association for Computational Linguistics.

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. In *ICASSP 2019 - 2019 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. ISBN 9781479981311.

Ostrowski, W.; Arora, A.; Atanasova, P.; and Augenstein, I. 2021. Multi-Hop Fact Checking of Political Claims. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 3892–3898. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Paris, B.; and Donovan, J. 2019. Deepfakes and Cheap Fakes. In *United States of America: Data and Society*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 14, 1532–1543.

Petrov, I.; Gao, D.; Chervoniy, N.; Liu, K.; Marangonda, S.; Umé, C.; Dpfks, M.; Facenheim, C. S.; RP, L.; Jiang, J.; Zhang, S.; Wu, P.; Zhou, B.; and Zhang, W. 2020. DeepFaceLab: A simple, flexible and extensible face swapping framework.

Qazvinian, V.; Rosengren, E.; Radev, D. R.; and Mei, Q. 2011. Rumor has it: Identifying Misinformation in Microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1589–1599. Association for Computational Linguistics.

Ruchansky, N.; Seo, S.; and Liu, Y. 2017. CSI: A Hybrid Deep Model for Fake News Detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Niessner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1–11.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV*, 618–626. IEEE Computer Society. ISBN 978-1-5386-1032-9.

Shang, L.; Zhang, Y.; Zhang, D.; and Wang, D. 2020. FauxWard: a graph neural network approach to fauxtography detection using social media comments. *Social Network Analysis and Mining*, 10: 1–16.

Tan, R.; Plummer, B. A.; and Saenko, K. 2020. Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2081–2106. Association for Computational Linguistics.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. 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)*, 809–819. New Orleans, Louisiana: Association for Computational Linguistics.

Times, N. Y. 2020. NYT Developer API. https://developer.nytimes.com/docs/articlesearch-product/1/overview. Accessed: 2020-11-31.

Vasileva, S.; Atanasova, P.; Màrquez, L.; Barrón-Cedeño, A.; and Nakov, P. 2019. 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)*, 1229–1239. Varna, Bulgaria: INCOMA Ltd.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, u.; and Polosukhin, I. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6000–6010. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Verdoliva, L. 2020.

Wang, B.; and Kuo, C. J. 2020. SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28: 2146–2157.

Wang, W. Y. 2017. "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)*, 422–426. Vancouver, Canada: Association for Computational Linguistics.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. 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, 849–857. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355520.

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. ISBN 9781479981311.

Yu, L.; Poirson, P.; Yang, S.; Berg, C. A.; and Berg, L. T. 2016. Modeling Context in Referring Expressions. In *European Conference on Computer Vision (ECCV)*.

Zhang, D.; Shang, L.; Geng, B.; Lai, S.; Li, K.; Zhu, H.; Amin, T.; and Wang, D. 2018. FauxBuster: A Content-free Fauxtography Detector Using Social Media Comments. In *Proceedings of IEEE BigData 2018*.

Zhao, Z. 2017. Spotting Icebergs by the Tips: Rumor and Persuasion Campaign Detection in Social Media. https://deepblue.lib.umich.edu/handle/2027.42/138726. Accessed: 2020-11-31.

Zhou, P.; Han, X.; Morariu, V. I.; and Davis, L. S. 2017. Two-Stream Neural Networks for Tampered Face Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. ISBN 9781538607336.

Zlatkova, D.; Nakov, P.; and Koychev, I. 2019. 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)*, 2099–2108. Hong Kong, China: Association for Computational Linguistics.