

Resilient Data Augmentation Approaches to Multimodal Verification in the News Domain

John Cadigan

john.cadigan@sri.com

Karan Sikka

karan.sikka@sri.com

Meng Ye

meng.ye@sri.com

Martin Graciarena

martin.graciarena@sri.com

Abstract

With the advent of generative adversarial networks and misinformation in social media, there has been increased interest in multimodal verification. Image-text verification typically involves determining whether a caption and an image correspond with each other. Building on multimodal embedding techniques, we show that data augmentation via two distinct approaches improves results: entity linking and cross-domain local similarity scaling. We refer to the approaches as resilient because we show state-of-the-art results against manipulations specifically designed to thwart the exact multimodal embeddings we are using as the basis for all of our features.

1 Introduction

Image-text verification involves determining whether a caption and an image correspond with each other. One of its main applications is detecting misinformation (Müller-Budack et al., 2021).

In the present term, re-purposed media is predominant form of misinformation, so-called cheap-fakes. Investigation of false claims made about Covid-19 found that 59% of misinformation involved repurposing existing media (Brennen et al., 2020). This misinformation is effective because psychologists have found that even the inclusion of related yet non-probative photos can persuade readers to believe and share false facts (Newman and Zhang, 2020).

Text-to-image synthesis poses a more distant, looming threat, especially given the capability for improvements in image-text consistency to lead to more convincing generation. For example, researchers have leveraged the state-of-the-art text-image CLIP embeddings (Radford et al., 2021) to select the more convincing generated images (Ramesh et al., 2021). Likewise, image-text consistency measures can also be included during training of models as in AttnGAN (Xu et al., 2018). With

advancements in consistency models, there are expected advancements in text-to-image synthesis.

By using the NewsCLIPpings dataset (Luo et al., 2021), we are addressing cheap-fake manipulations which are adversarially selected by the very same model we use for all of our features. We show that data augmentation is an effective way to overcome them.

2 Related Work

There has been increasing interest in image-text verification models with several datasets and systems published. The MEIR dataset (Sabir et al., 2018) was created by randomly substituting named entities, person, location or organization, appearing in FLICKR with each other to create manipulated packages including manipulated GPS metadata when applicable. TamperedNews (Müller-Budack et al., 2021) made similar manipulations on entities but used Wikidata to constrain location swaps by geographic distance constraints, people by gender and country of origin, and events by parent class. The COSMOS dataset concerns recontextualization of images with two captions per image (Aneja et al., 2021).

We choose the recently created NewsCLIPpings dataset for several reasons. Foremost, it belongs to the newswire domain with the incumbent challenges of the many entities which appear in them. Furthermore, for every pristine caption, the dataset creators gather mismatched images by various means, including state-of-the-art unimodal and multimodal embeddings, to make the dataset more challenging under threat scenarios of recontextualization of people and places. It is important to have a sufficiently difficult dataset; Luo et al. (2021) report that TamperedNews can be solved by text model alone. Finally, it is designed to be unsolvable by the exact CLIP model we are using (ViT-B/32); for every partition, falsified examples have a higher CLIP image-text similarity score than pristine ex-

	train	val	test
CLIP Text-Image	453128	47248	47288
CLIP Text-Text	516072	53876	54164
Person SBERT Text-Text	17768	1756	1816
ResNet Place	124860	13588	13636
Merged Balanced	71072	7024	7264

Table 1: The number of pairs from NewsCLIPpings by falsification method and training partition

amples 50% of the time.

There have been several relevant approaches to image-text verification. Most similar to our approach, Müller-Budack et al. (2021) attempt to verify news media based on entities: location, scene and people. Unlike our approach, they use embeddings from pretrained visual classifiers, each chosen as relevant to its entity type; for all entities, they compare embeddings from clusters of picture gathered from Wikimedia, Bing and Google to segments of the document image, faces for people and the entire image for location and event entity types. They apply aggregations such as max and quantile to refine these to signals for recovering clean and tampered documents. Tan et al. (2020) detect generated news articles with a model which considers image, captions and article text. Sabir et al. (2018) develop a system which uses features derived from VGG19 trained on image net, word2vec and GPS coordinates as input to multitask models trained to analyze single packages in addition to pairs of packages in order to determine manipulation in social media.

3 Data

The NewsCLIPpings dataset includes an equal number of manipulated and pristine examples for each manipulation type in each training partition. See Table 1.

4 Method

As the base approach to multimodal verification, we use the image-text embedding model CLIP. The first data augmentation approach uses entity linking to gather associated data of entities to verify pictures. The second uses cross-domain similarity local scaling (Conneau et al., 2017).

4.1 Entity Linking

For entity linking, we use the BLINK entity linker (Wu et al., 2019). For each mention, we gather

the top 32 candidate entities from the biencoder and take the max prediction from the cross-encoder if the logit is greater than zero. We use Wikidata property 18 links to gather representative images of the entity in addition to labels and alternative labels for entities as their textual representation. This is depicted in figure 1. In the case of multiple representations for an entity in either modality, we take the most similar with cosine similarity of CLIP embeddings out of all possible pairs as the canonical example. We used the 2020-10-01 dump of English Wikipedia for entity descriptions along with the 2020-11-25 dump of Wikidata for the links and associated images.

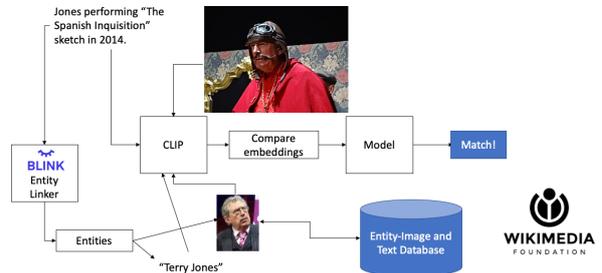


Figure 1: System diagram showing how entity linking augments data

4.2 Cross-domain Similarity Local Scaling

Cross-domain similarity local (Conneau et al., 2017) scaling is an approach meant to combat the hubness issue of embeddings (Dinu et al., 2014). This has been used in text-image retrieval in the MS-COCO dataset (Liu and Ye, 2019). We use embeddings of the 813903 image-caption pairs remaining from the VisualNews dataset (Liu et al., 2020) to improve similarity judgments; this is not additional training data. When it comes time to score the similarity of an image embedding x_i and text embedding y_c , we replace the cosine-similarity score with the following. For each image x_i , we gather its k nearest neighboring captions $\mathcal{N}_c(x_i)$ and take the mean cosine-similarity score:

$$r_c(x_i) = \frac{1}{K} \sum_{y_t \in \mathcal{N}_c(x_i)} \cos(x_i, y_t) \quad (1)$$

Likewise, this is also calculated for the caption in comparison with other images before subtracting both from the cosine score. We calculate nearest neighbors and distances with the efficient GPU implementation of exact nearest neighbor (Johnson et al., 2017).

$$CSLS(x_i, y_c) = 2\cos(x_i, y_c) - r_c(x_i) - r_i(y_c) \quad (2)$$

Artetxe and Schwenk (2018) extends work with an alternative score based on the ratio of cosine similarity and the neighbors which has higher matching performance in finding parallel sentences, multilingual sentence matching:

$$CSLS_{ratio}(x_i, y_c) = \frac{2\cos(x_i, y_c)}{r_c(x_i) + r_i(y_c)} \quad (3)$$

4.3 Features and Model

With representative images and labels for entities appearing in the caption, we use CLIP to compare them. There are 4 possible comparisons between the augmentation data and the reference respectively: image-image, image-caption, label-image, label-caption. Each of these scores is bucketed by entity type, MISC, PER, LOC, and ORG, in order to produce sub features, reasoning that some may be more reliable than others. This produces 16 buckets of comparison and entity types. To produce features, aggregations are applied to these: min, max, sum, median and mean. This creates 80 features. Finally, a count feature is added for each type, a base image-caption CLIP score and the number of caption characters.

For our experiments, we consider logistic regression from scikit-learn (Pedregosa et al., 2011) and LightGBM (Ke et al., 2017). We conducted a series of experiments with the linear classifier to show how features build upon each other. First, we start with the performance of a single-feature logistic regression model based on CLIP image-text similarity alone, LR:IC. Second, we add the entity-based features. Each of these features are normalized to 0 mean and unit variance. Third, we show experiments with LightGBM tuned with FLAML (?) on the Merged validation set; the final number of estimators was adjusted by an order of magnitude (10x) with the proportional decrease in learning rate. The number of estimators was 1670; max number of leaves, 66; minimum samples for a leaf, 12; learning rate, $7.6e-3$; subsample rate, 65%; alpha regularization, 13.0; lambda, 20.3. Finally, to show the improvement from replacing cosine-similarity with CSLS when comparing across modalities, we have parallel experiments for each of those 3 conditions. In our experiments, we gathered the 3 nearest neighbors across modalities; 3 was based on tuning

ROC-AUC on the training and validation set. See Table 2.

5 Results

First, we present the results on the validation merged-set. These are present in Table 3. We can see that entity-based features and LightGBM model improve performance. Finally, in each of these conditions, CSLS improves average performance. See Figure 2 for feature importance of the LGBM model; surprisingly, label similarity was more influential than similarity to images; images of people were useful while those of locations and organizations were less so.

Second, we present comparisons to the results of (Luo et al., 2021) in which the CLIP model was fine-tuned against particular manipulation types in Table 4. We show equal or greater performance in 4 out of 5 subsets. Table 5 shows performance within the Merged partition. In every case except for the Text-Image CLIP manipulations, the system performs much better. Overall, the system performs better on average.

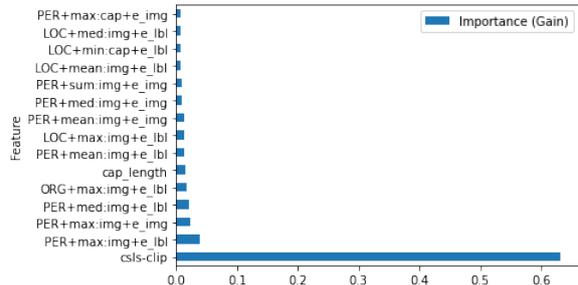


Figure 2: LGBM feature importance from Merged train partition

	cos	CSLS	CSLS-ratio
k			
0	0.713	–	–
1	–	0.742	0.739
3	–	0.742	0.739
10	–	0.739	0.736

Table 2: ROC-AUC for image-text similarity methods as k varies on combined validation and train Merged partition

6 Discussion

The creators of the NewsCLIPPings dataset claim that their dataset is unsolvable by CLIP embed-

	LR:IC	LGBM	caption	image	entity	entity image
False	0.583	0.239	Shibuya at night A typical night in Shibuya Tokyo on a rainy Wednesday night		Shibuya	
True	0.487	0.641	Dave Brubeck at the Newport jazz festival in 1981		Dave Brubeck	
True	0.533	0.341	Carlyle Capital is based in Guernsey		Guernsey, island	
False	0.497	0.779	Dylan Hartley did not train with his England teammates on Tuesday afternoon		Dylan Hartley	

Figure 3: Example predictions from single-feature LR model and full CSLS LGBM model. Bold is correct.

	cos	CSLS
LR:image-caption	0.641	0.666
LR:+entities	0.653	0.674
LightGBM	0.664	0.683

Table 3: Performance on the validation Merged set

Partition	Model: (Luo et al., 2021)	image-caption	+entities	LightGBM
CLIP Text-Image	cos	0.670	0.527	0.578
	CSLS	—	0.514	0.579
CLIP Text-Text	cos	0.694	0.655	0.664
	CSLS	—	0.678	0.686
SBERT-WK Text-Text	cos	0.610	0.628	0.651
	CSLS	—	0.644	0.668
ResNet Place	cos	0.682	0.672	0.674
	CSLS	—	0.688	0.693
Merged Balanced	cos	0.605	0.654	0.663
	CSLS	—	0.672	0.681
Average	cos	0.652	0.627	0.646
	CSLS	—	0.639	0.661

Table 4: Accuracy on test subsets when training and testing only within the specified subset; compare with Table 4 of (Luo et al., 2021)

dings alone. This is largely reflected in our results, but a simple classifier built on just CLIP and a bias term can achieve 52.7-67.2% accuracy depending on manipulation type. This suggest the cutoff was not high enough to be completely unsolvable.

With the addition of entity features and the LightGBM model, accuracy improves. CSLS improves results in nearly every condition. Compared to a fine-tuning approach, the combination of all three achieved 3% higher accuracy on average; 8.5% in the Merged partition.

Analyzing the important features, Figure 2, and

Partition	Model: (Luo et al., 2021)	LR:IC	LR:IC+entities	LightGBM
CLIP Text-Image	cos	0.605	0.475	0.503
	CSLS	—	0.482	0.511
CLIP Text-Text	cos	0.584	0.839	0.834
	CSLS	—	0.861	0.862
SBERT-WK Text-Text	cos	0.605	0.626	0.635
	CSLS	—	0.646	0.646
ResNet Place	cos	0.616	0.676	0.681
	CSLS	—	0.700	0.706

Table 5: Breakdown of accuracy on merged test by manipulation type when using the merged training subset; compare with Table 5 of (Luo et al., 2021)

example errors, Figure 3, point to future improvements. We can see that images of people generally help while other types are less useful. There may be further improvements by improving the quality of textual labels.

7 Conclusion

We have described two distinct methods of data augmentation which improve results in the multimodal verification task despite the adversarial construction of the dataset, without fine-tuning. Future work should consider attention models for integrating relevant data through augmentation.

Acknowledgements. This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

- Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2021. Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*.
- Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.
- J Scott Brennan, Felix Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. Types, sources, and claims of covid-19 misinformation. *Reuters Institute*, 7:3–1.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.
- Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154.
- Fangyu Liu and Rongtian Ye. 2019. A strong and robust baseline for text-image matching. *arXiv preprint arXiv:1906.01205*.
- Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visualnews: A large multi-source news image dataset. *arXiv preprint arXiv:2010.03743*.
- Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv preprint arXiv:2104.05893*.
- Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, Sherzod Hakimov, and Ralph Ewerth. 2021. Multimodal news analytics using measures of cross-modal entity and context consistency. *International Journal of Multimedia Information Retrieval*, pages 1–15.
- Eryn J Newman and Lynn Zhang. 2020. Truthiness: How non-probative photos shape belief. In *The Psychology of Fake News*, pages 90–114. Routledge.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.
- Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. 2018. Deep multimodal image-repurposing detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1337–1345.
- Reuben Tan, Bryan A Plummer, and Kate Saenko. 2020. Detecting cross-modal inconsistency to defend against neural fake news. *arXiv preprint arXiv:2009.07698*.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.
- Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.