Towards Multimodal Semantic Consistency Analysis of Long Form Articles

Yuwei Chen
ychen69@albany.edu

Ming-Ching Chang
mchang2@albany.edu

University at Albany – State University at New York, Albany NY 12065, USA

Abstract

With the rise of misinformation in news sites and social media, multi-modal machine learning methods that can identify fake news by analyzing inconsistencies within the articles have become increasingly important. Current state-of-the-art methods based on traditional image-caption models can only process captions within 1 to 2 sentences. Existing models struggle on analyzing long articles as they were not trained for such purposes. The main limitation is the lack of fine-grained localized evidence needed for consistency detection. We propose an ensemble method combining a bank of visual detectors and BERT-based NLP models that can effectively compute the consistency among the image(s) and paragraph(s) of texts. Our method is effective in both detecting the standard image-caption pairs and longer form news articles. Our method is able to process longer form of multi-modal media via the localization of fine-grained evidence with modularity and explainability. Evaluation is performed on a MS COCO data subset and a news article benchmark of the DARPA SemaFor program. We achieved 86% AUC on the COCO subset as well as a competitive result within the SemaFor evaluation.

1. Introduction

In the era of deep-fakes, disinformation and misinformation, the ability to evaluate the legitimacy and claims made by articles from published news and social media is crucial. Analyzing the consistency of multimodal articles containing short or long form texts, images with captions, *etc.* under a semantic lens can become a powerful tool for checking fake news. Challenges of such analysis stems from a series of difficulties. (1) News articles typically contain multiple paragraphs within the text body as well as one or more images with captions. In comparison, social media posts such as Tweets are typically short. Semantic consistency analysis among the multiple modalities with variable length poses a great challenge. (2) The style of writing can lead to claims that require multiple jumps among statements to fully understand the context. News articles often contain a writing style that relies heavily on the reader's knowledge of the subject. Semantic analysis in this regard is very different using models trained on image-caption pairs, where the keywords in captions are mostly assumed to be directly visible in the image and require no deduction of leaps in logic to understand the full picture. Due to the above two differences, directly applying image-caption models will result in low performance and poor explainability.

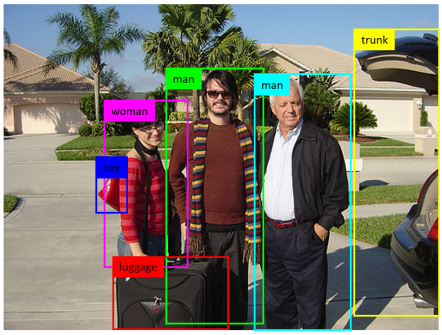


Figure 1: **Direct *vs.* indirect semantic consistency.** Two *men* a a *woman* can be directly observed in the image. Whether these people are a family or they will leave for the airport is not directly verifiable and requires further contexts and cues to determine.

We aim to develop machine learning methods that can effectively explore the the limits of analyzing semantic consistency using no external knowledge base. We assume all knowledge elements gathered for analysis should come solely from the given sample article, and not a knowledge graph that is maintained separately. To this end, a distinction that must be made clear is the difference between **direct** *vs.* **indirect relationships** among modalities in an article. *Direct relationship* refers to keywords in the text domain that can be directly mapped to objects in the corresponding image. In contrast, *indirect relationship* refers to knowledge that cannot be directly verified in the corresponding image. Figure 1 shows an example from the MS COCO Captions dataset [1] with caption "Family group photo before leaving

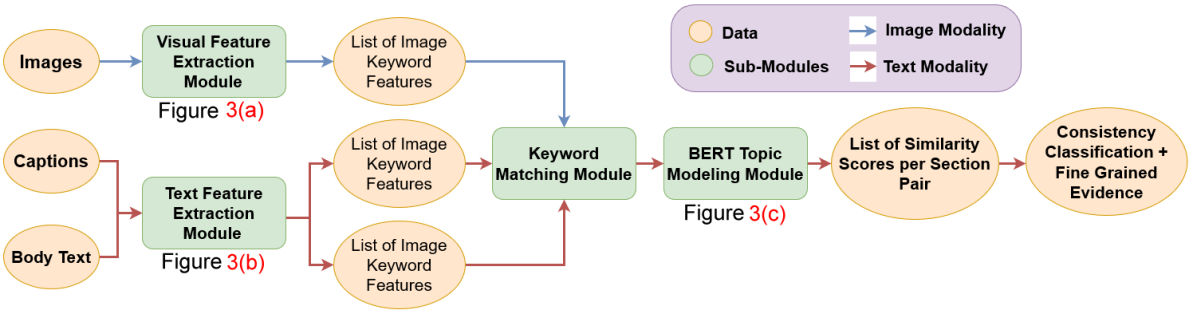


Figure 2: **Overview of the proposed multimodal article semantic consistency analysis pipeline.** The **visual and text feature extraction modules** extract all relevant keywords from the raw text/images. The **keyword matching** module takes the extracted keywords and makes direct keyword matches. The **BERT topic modeling** module takes in any remaining unmatched keywords and and clusters them into topics. Details are provided in Figure 3.

for a flight". The men and woman provide direct relationships, as the two men and a woman can be directly observed by running object detectors. In contrast, *indirect relationship* can not be directly verified whether these people are a family nor are they leaving to the airport. Indirect relationship is not always possible to determine; further contexts and cues are often required. External knowledge base or pre-trained context models can provide possible mappings.

The state-of-the-art text-image analysis methods fall under large data driven end-to-end models, and mostly the task is formulated as a consistency classification problem. Currently the top performing image captioning models include OSCAR [5], VinVL [10], and CLIP [7]. These methods are suitable to process short image captions with length ranging from 1-3 sentences, thus not directly applicable for article semantic consistency analysis. When applied to longer form media with complex semantic scenes, they result in large performance drop. Another limitation is due to the pre-trained model in handling variable writing styles, variable text lengths, and rich semantic contents. The lack of large-scale article datasets with quality manipulated samples limits these end-to-end models.

Our goal is to develop an article semantic consistency analysis method that can effectively handle direct relationships, while it is also competitive against existing methods in handling indirect relationships. We focus on prioritizing evidence generation and topic modeling flexibility. Figure 2 overviews our multimodal article semantic consistency analysis pipeline. Our method consists of three modules in which fine-grained localized evidence is generated, namely (1) initial keyword feature extraction, (2) direct keyword matching, and (3) BERT topic similarity scoring. This design allows us to attend to specific portions of the news article that is either consistent or inconsistent. The architecture allows sub-modules to be replaced or combined with future higher performing modules. Our proposed model addresses the limitation of existing end-to-end methods in processing longer form text. Instead of processing all the text as a whole, we localize the text by separating them into their respective paragraphs/sentences. We then feed the short form text to our specialized feature extractors. Our design distinguishes from other competitive methods in two main features. (1) We do not rely on external knowledge base in article consistency while upholding performance. All inference during consistency analysis comes straight from the sample article in run-time. (2) We retain strong modularity and explainability in evidence generation. The three checkpoints created within our architecture allow our model to generate human interpret-able evidence towards consistency analysis. Our keyword-centered consistency analysis provides localized fine-grained evidence in addition to the binary consistency classification.

We perform evaluation on two relevant datasets: (1) a subset of the MS COCO image captioning dataset [1], and (2) the image and text inconsistency evaluation on news articles provided by the DARPA Semantic Forensics (SemaFor) program. [1] The COCO evaluation demonstrates the effectiveness of our method on traditional short-form image captioning. The SemaFor dataset evaluates longer-form media handling. For each test multimodal article consisting of one or more image(s), caption(s), and text paragraph(s), a binary consistency groundtruth is provided, and ROC-AUC is reported as evaluation metric. Our method achieves a 86% AUC on COCO image captioning and is competitive with other SemaFor performers on the text inconsistency evaluation on the news articles task.

2 Related Works

Most article consistency analysis methods prior to the debut of deep learning processes short articles in the text domain. The work of [3] explores textual similarities between short textual passages by analyzing a combination of linguistic features such as noun-phrases, wordnet-synonyms, and action verbs. The method of [6] performs image-caption consistency analysis.

[1] https://www.darpa.mil/program/semantic-forensics

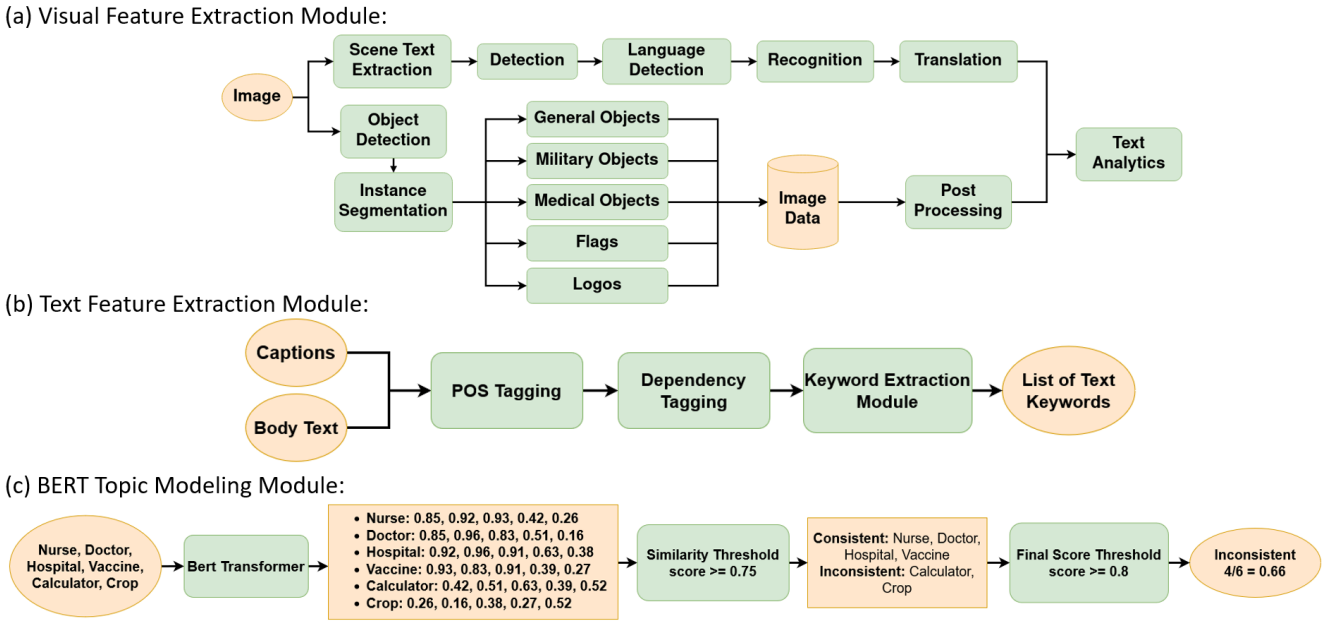


Figure 3: (a) Visual Feature Extraction Module. (b) Natural Language Feature Extraction Module. (c) Workflow for unmatched keyword: Bert Transformer Workflow.

State-of-the-art text-image representation methods such as OSCAR [5] and VinVL [10] use visual detectors to gather semantic visual features. The detected object tokens are then mapped to object image region to the corresponding caption keywords to form triplets (containing word tokens, object tags, and region features of the caption) to train a *Transformer* model [8]. These approaches are limited in their capability in processing longer texts. OSCAR and VinVL can only process texts that are at most around 20 words. OSCAR is trained on primarily MS COCO [1] and Flick30 image captions. The OSCAR model relies heavily on capturing the objects within the caption and image as anchor points. Due to the model's reliance on keyword mappings and carriage limit in the text domain, extending the carriage limit would introduce too much noise in training. OSCAR and VinVL models do not project well to longer form media such as news articles.

The CLIP [7] model became popular for it's great performance on multi-modal tasks. CLIP takes a different approach then OSCAR as it relies solely on two well trained encoders, one to process the image domain and one for the text domain. At training time, CLIP relies heavily on it's preset prompt to place the interest entities in a natural language environment. Many times the prompt would be "A photo of" followed by the main entities in the image. The model would then encode both the image and the mutated prompt, then evaluate the cosine similarity of the image against the mutated prompt. This training strategy limits the models ability to process long text more then OSCAR because the contextual prompts often so not add any viable context during training. Since CLIP mostly focuses on the entity within the short caption, it does not project well to longer form media, where inconsistency is hard to localize. CLIP also struggles in producing any human interpretable evidence for complex longer forms of media. In general, embedding-based end-to-end model would have a hard time providing localized evidence towards it's classifications.

3 Method

We propose a semantically focused dynamic method that extracts a variety of keyword features from articles containing one or more images, captions, and paragraphs. It provides human interpret-able evidence, and has no limitation on topics it can process. Our model architecture can explore the limits of semantic consistency analysis with no external knowledge source.

Given an input article, the containing images, captions, and body text paragraphs are first localized as separate sections for processing. Our vision and natural language submodules extract semantically relevant keywords from each localized section. These extracted keywords are then used to create a dynamic knowledge graphs. Each localized section of the article will be represented in the form of a knowledge graph. These knowledge graphs will then be compared to all other knowledge graphs to create a topic similarity score for each section pair. An article with i images, c captions, and p body paragraphs would be treated as $n = i + c + p$ separate sections and $\frac{n \times (n-1)}{2}$ section pairs for analysis. Once the similarity scores of all section pairs are calculated, we then compute a consistency score

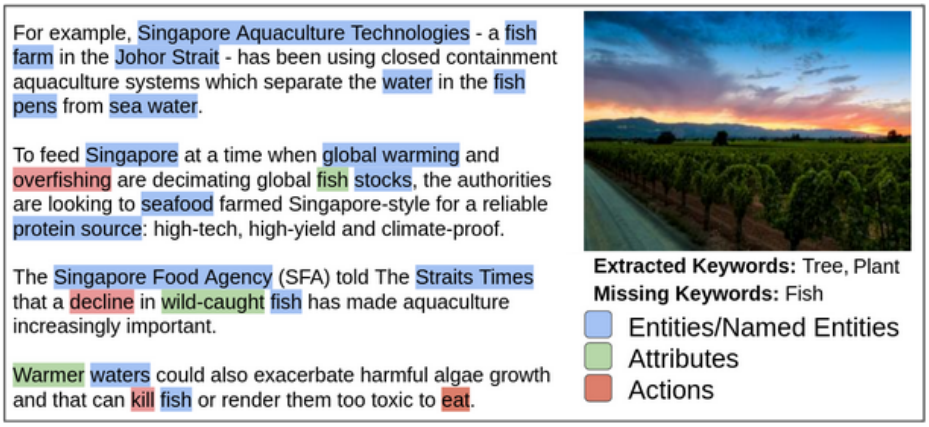


Figure 4: **Article consistency analysis example.** The detected entities are highlighted in blue, attributes in green, and actions in red. The following keywords from text are extracted: *Singapore* (location), *fish* (entity), *sea* (entity), *water* (entity). Image-to-text inconsistency is detected, as the image does not contain *fish*, and the image shows a *plant farm*.

for the whole article. This is done through a simple summation process of what percentage of the article is semantically consistent. Figure 2 overviews our pipeline, and Figure 3 shows detailed steps.

Much like OSCAR, we employ a series of visual detectors to gather visual keyword features. We then employ a series of natural language models to extract relevant keywords from the texts through a combination of speech tagging and dependency parsing. We differentiate ourselves from existing methods by solely relying on information present in the given sample article. Given this architecture, we do not need to rely on any external topic dependent corpus. This allows our model to process any topic and give relevant evidence towards consistency analysis.

In our method, each paragraph is treated as a separate localized article to aid in fine-grained evidence generation. Each localized paragraph is made into a knowledge graph containing the main entities and attributes from said paragraph. These knowledge graphs are compared with each other to compute an initial similarity score. Any unmatched keywords are sent to the BERT topic modeling module to classify the consistency of the remaining words. Through this process we compute a similarity score for each paragraph pairing. This allows us to localize any detected inconsistencies within the article.

Our architecture is distinct from existing fully end-to-end works like OSCAR and CLIP in that, we are able to take all the fine-grained keyword features that was extracted and compare similarity dynamically. Our model is more similar to how a human would comprehend an article. It is not reliant on large amount of training data, as it compares keyword features dynamically. With less extracted features, our method is still able to provide some level of consistency analysis with fine-grained evidence towards the prediction.

3.1 Similarity Calculation Within Section Pairs

In our method, the similarity score between each pair of sections (image-caption, image-paragraph, *etc.*) is calculated via the matching of keywords extracted in the sections. This keyword matching process consists of two steps. (1) Direct keyword match is first performed between the knowledge graph of each section. Once all keywords have gone through the initial direct matching stage, an initial similarity score is generated for each section pair. (2) Any unmatched keywords between the two sections are then passed to the BERT-based topic modeling module. BERT will then generate similarity scores for all unmatched keywords. The direct matches are weighted higher then non-direct matches towards consistency. Based on the combined similarity, instances of consistency and inconsistency are generated. The consistency analysis score of each section pairs are localized with respective keywords as consistency evidence for explainability.

Figure 4 shows an example with details on how an article of 4 paragraph and an image is processed. Similarity scores are calculated for the $(4 \times 3)/2 = 6$ section pairs for consistency analysis. Our model produces the consistency classification as well fine-grained evidence pointing to the exact word/sentence of the detected semantic inconsistency.

Visual Feature Extraction: We assembled and trained a wide range of visual detectors that can detect up to two thousand relevant classes for consistency analysis. These visual classes were picked based on their relevance for our experimental scope. We focused on topics including medical supplies, flags, military vehicles and weaponry. We integrated a set of multilingual scene text detection modules (up to 16 different languages) and logo detection modules, which can effectively extract text- and logo-relevant keywords within the image. Finally, we have also integrated a multi-nation flag detection module [9] that can extract and recognize flags with the country of origin, which is useful

in establishing a location context for cross-referencing. Figure 3(a) depicts our visual feature extraction pipeline.

Text Feature Extraction: We adopt `spaCy` [4] tagging modules to apply POS and dependency parsing on the given texts to extract relevant entities. These keyword features are then extracted and combined into a knowledge graph using a bidirectional graph structure. The bidirectional graph contains two types of nodes, namely *entity* and *action* nodes. Each node contains at least four fields for easy comparisons. These fields are token, type (Named Entity, Location, Flag, *etc.*), part of speech, and the section identification. Entity nodes contain an extra field called list of attributes that stores the entity attributes. Entity nodes are then connected to one another through action nodes, which often contain verbs. Figure 3(b) depicts this natural language feature extraction pipeline.

In an earlier work of our analytics published in [2], we had implemented a **weighted dictionary** to perform direct topic classification. The weighted dictionary contains a series of topic relevant keywords for each chosen topic. Each keyword is given a weight on how likely said keyword would appear in a article of a given topic. These weights were gathered by observing the frequency of which they appear in news articles of a given topic. A list of stop words were used to filter out commonly used words in the English dictionary. The weighted dictionary thus creates a lexicon for each chosen topic. The initial thought was that the dictionary provided a quantitative score for all extracted keywords. This is beneficial to providing a quantitative value to extracted keywords during consistency analysis. However, we decided to remove this module here due to its limitations in the number of topics it can effectively process.

BERT Topic Modeling: We use a BERT-based Transformer language model as shown in Figure 3(c) to perform topic clustering on all text based portions of the given article. A list of unmatched keywords would be feed into the Transformer, where each keyword isolated as its own entity. A similarity score is then calculated based on how similar in the topic each of the unmatched keywords are to each other. All generated similarity scores are then pass through a score threshold of 0.75 to determine if can be considered semantically consistent. A final similarity score will be generated for said pair of paragraphs by summing all keywords that passed the threshold and dividing it by the total number of unmatched keywords. This allows the model to recover from cases where there exists insufficient direct matches in the keywords extracted. It is also intended for cases where the entities mentioned are similar in topic, but not directly similar. Figure 3(c) shows an example, where the unmatched keywords would fall under the topic of health professionals.

The output of our model contains the multiple consistency scores for each occurrence of consistency or inconsistency. These scores allow for better analysis then a single score or binary decision for the whole article. Each consistency score contain a pointer to all knowledge elements used during consistency analysis as well as the localized article regions for those elements within the news article. This output structure allows much more flexibility in analysis as decisions are not finite. Due to the complexity of news articles, the likelihood of an article containing consistent and inconsistent knowledge elements are common. By using this output structure, we are able to isolate these occurrences to better calculate the consistency of the whole news article. This allows our method to maximize explainability for the given knowledge elements present.

Our model architecture is designed with scalability and modularity. It allows easy plugin, replacement, or combination of higher performing sub-modules to expand our model in the future. This will increase the amount of evidence our model can generate and lead to higher explainability. The scalability of our model is high since in low feature extraction environments, our model can still provide consistency analysis towards portions of the news article. The better our feature extractors perform in extracting knowledge elements, the more evidence we can generate for consistency analysis. Therefore, our model scales well as size of the news articles and our model increase.

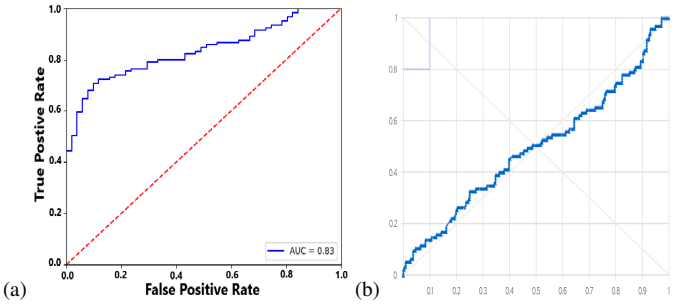
4 Experimental Evaluation

MS COCO captions dataset. We tested our proposed model on a subset of the MS COCO captioning dataset, which contains 10,000 image caption samples that are evenly balanced between manipulated and pristine samples. For the pristine samples, we randomly selected 5,000 image caption samples from [1]. The 5000 manipulated samples were gathered by caption swapping the selected pristine samples.

DARPA SemaFor text-image inconsistency dataset. This dataset was provided by the SemaFor program. To the best of our knowledge, the pristine samples were gathered through scraping real international news sites. The manipulated samples were created manually by taking real news articles and doing a series of manipulations to said articles. These manipulations contain image swaps, GAN generated images, entity swaps, caption or body text generation, etc.

We use ROC-AUC as our **evaluation metric**. Results show that our model accurately captures the direct relationships between image and textual entities in the given samples. Figure 5(a) show the resulting ROC curves on the MS COCO subset, where our method achieved AUC of 86%.

We next report results from the DARPA SemaFor image & text article inconsistency evaluation. This task contains a large dataset of news article gathered from a variety of international news sites. The types of article manipulations range from entity swaps, contextual changes, GAN



(a) (b)

Figure 5: **ROC evaluation of the proposed method** on (a) MS COCO and (b) DARPA SemaFor news article inconsistency dataset.

generated images, *etc*. The magnitude of manipulation ranges from small contextual changes like date or location alternation to larger entity swaps. Figure 6(b) shows that consistency analysis on long form media is very difficult. While the resulting SemaFor ROC curve shows lower performances then the MS COCO, our results are competitive to other performers in the program. Note that our proposed model is comparable in performance while having the upside of explainability and flexibility. We note that many performers in the SemaFor program use versions of OSCAR, VinVL, and CLIP. However, the results show that there is a significant drop in performance for these state-of-the-art models on longer form media. This lower performance can be explained by that news articles are written with a large amount of claims that require the reader to have a baseline working knowledge of the subject. Many entities/subjects presented in news articles can not be physically verified in the provided images, which causes significant confusions for models trained on traditional image captioning tasks.

5 Conclusion

We presented a semantically focused method that can provide fine-grained direct relational evidence towards the decisions it makes. This model is not only effective in capturing direct relationships between the image and text modalities. It can provide fine grained entity based evidence towards it's classifications.

Future work. There are limitations to performing dynamic analysis in the proposed way at run-time. Our model struggles to capture topics that require inference of indirect relationships. In future works we would like to look into improving the scalability of the model, and expand into capturing inference-based relationships. Another aspect of news articles that we haven't taken advantage of is textual ranking. Not all paragraphs in a news article is of equal importance or contains the main idea of the article. We plan to create a hierarchical structure of importance for the body paragraphs to take advantage of this quirk.

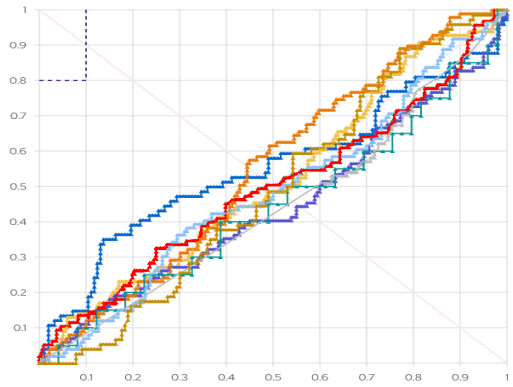


Figure 6: SemaFor Evaluation 2 Comparison. Red curve depicts our ROC curve shown in Figure 5(b). All methods perform poorly in this challenging SemaFor evaluation.

Acknowledgement. This work is supported by the U.S. DARPA Contract HR001120C0123. We thank Matt Turek, Arslan Basharat, Kirill Trapeznikov, and Sam Blazek for providing guidance.

References

[1] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. In *arXiv:1504.00325*, 2015.

[2] Y. Chen and M.-C. Chang. On multimodal semantic consistency detection of news articles with image caption pairs. In *ICCE-TW*, 2022.

[3] V. Hatzivassiloglou, J. L. Klavans, and E. Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *EMNLP*, 1999.

[4] M. Honnibal and I. Montani. `spaCy 2`: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/, 2017.

[5] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. OSCAR: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[6] V. Ordonez, G. Kulkarni, and T. Berg. Im2Text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *NIPS*, 2011.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskeverk. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017.

[9] S.-F. Wu, M.-C. Chang, S. Lyu, C.-S. Wong, A. K. Pandey, and P.-C. Su. FlagDetSeg: Multi-nation flag detection and segmentation in the wild. In *AVSS*, 2021.

[10] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gaok. VinVL: Revisiting visual representations in vision-language models. In *CVPR*, 2021.