On Multimodal Semantic Consistency Detection of News Articles with Image Caption Pairs

Yuwei Chen
ychen69@albany.edu

Ming-Ching Chang
mchang2@albany.edu

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

***Abstract*—Recently multi-modal consistency detection has been proposed as method to combat disinformation. However, state-of-the-art methods lack the fine grained localized evidence needed for consistency detection. Current methods also struggle on longer texts that are present in news articles. We propose an ensemble method that combines a series of visual detectors and BERT based NLP models to compute consistency between modalities. The proposed method is effective in both detecting the standard image-caption pairs and news articles containing multiple paragraphs. Our method can localize and provide fined grained evidence towards its given responses. We evaluate our method on a MSCOCO image-caption subset and image & text inconsistency evaluation of news articles from the U.S. DARPA SemaFor program. We achieved an 83% AUC on the gathered MSCOCO dataset, which shows the effectiveness of our method.**

I. Introduction

With the increase popularity of digital news media, news articles containing disinformation has become the norm. The ability to detect inconsistencies and false claims in news media is emergent. Difficulties in applying multi-modal methods to news media include the handling of lengthy texts and diverse writing styles. News articles usually contains several paragraphs, while typical multi-modal tasks such as image captioning only contain 1-2 sentences. News articles often contain indirect relationships between the images and texts.

A distinction we need to make clear is the difference between **direct** *vs.* **indirect relationships** among modalities. *Direct relationship* refers to keyword(s) in the text domain that can be directly mapped to an object in the image. Refer to the example in Fig. I, where the caption clearly suggests the mapping of keywords"monkey" and "man". In contrast, for *indirect relationship* of caption "Luke is standing next to his monkey" cannot be physically verified from the image, as we can not verify if the monkey in the image is Luke or if the man is Luke. The inference of indirect relationship requires external knowledge to complete the mapping. In this work we mainly focus on capturing direct relationships.

Fig. 1. Man next to monkey

State-of-the-art models such as OSCAR [1] and VinVL [2] is similar to our method in using visual detectors to gather semantic visual features. The model then maps the detected object token and object image region to the corresponding caption keywords. This forms a triplet that is then used to train a *Transformer* model. The limitation of this approach is due to a carriage limit in the text domain; OSCAR and VinVL can only process texts that are at most around 20 words. OSCAR is trained on primarily MS COCO and Flick30 image captions; also due to the model's reliance on keyword mappings, extending the carriage limit would introduce too much noise in training. Due to these limitations, models like OSCAR and VinVL do not project well to longer form media such as news articles.

The CLIP [3] model became popular for good performance on many 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 would first mutate the caption to follow a preset prompt "A picture of" followed by the main entities in the image. The model would then encode both the image and the mutated prompt and perform cosine similarity to see how similar the image is with the prompt. Much like OSCAR, CLIP does not project well to longer form media, mainly due to the length of the training captions and the inability to localize the inconsistency.

In this work, we propose a semantically focused method that can process both short-form captions and long-form text for consistency analysis. Much like OSCAR, we employ a series of visual detectors to gather visual keyword features. We then employ NLP models to extract relevant keywords from the texts through a combination of speech tagging and dependency parsing. Our model is unique in that each paragraph is treated as a separate localized article. In our model, each localized paragraph is made into a knowledge graph containing the main entities and attributes from said paragraph. These knowledge graphs are compared with our weighted dictionary and BERT topic modeling module to classify the topic of each paragraph. Each paragraph topic is then compared to each other to obtain a consistency score for the whole 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 does 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.

We perform evaluation on two relevant datasets: (a) a subset of the MS COCO image captioning dataset, and (b) the image

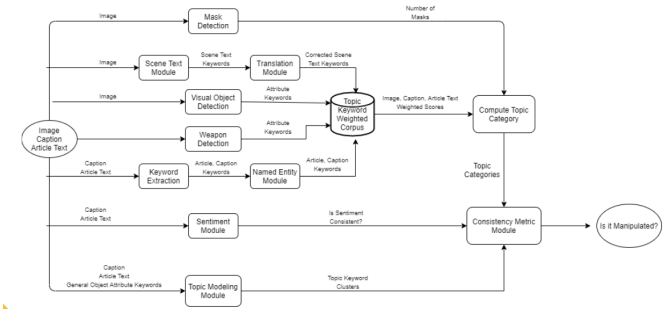

Fig. 2. Semantic keyword matching based pipeline.

and text inconsistency evaluation on news articles provided by the DARPA SemaFor program. COCO evaluation can demonstrate the effectiveness of our method on traditional short-form image captioning. The SemaFor dataset will showcase longer-form media handling. ROC-AUC is reported as evaluation metric. We achieve 83% AUC on COCO image captioning.

II. Method

We propose a semantically focused method that extracts a variety of keyword features from articles containing one or more images, captions, or paragraphs. As new articles contain broad topics, in this paper, we narrow our scope and focus on only military and COVID relation articles.

First, all images, captions, and body texts from the article are localized as separate sections. Our vision and natural language sub-modules then extract semantically relevant keywords. A weighted dictionary is next used for keyword matching. Once all keywords are matched, each component (image, caption, body paragraph) is given a topic classification. These topics are then compared against each other based on the weighted score they received from the dictionary and BERT topic modeling module. A consistency score for the whole article is then generated based on the percentage of the article that is consistent in topic.

Visual Modules: We assembled a series visual detectors that detects up to two thousand relevant classes for consistency analysis. These visual classes were picked based on their relevance for our experimental scope.

Text Modules: We adopt `spaCy` tagging module to apply POS and dependency parsing on the given texts, to extract relevant entities. These extracted features are then passed onto a BERT module for topic modeling.

We use 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. We extracted all relevant entity keywords from 1,000 military and COVID related news articles. A list of stop words were used to filter out commonly used words in the English dictionary. The weighted dictionary is thus created for each chosen topic.

BERT Topic Modeling: We use a BERT based language model to perform topic clustering on all text based portions of the given sample article. Based on the respective topic clusters, we calculate how similar each of the corresponding body paragraphs and captions are to each other. This allows to recover cases where our weighted dictionary fails to assign a topic to certain sections of the article.

III. Results

Dataset Preparation: 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 [4]. The 5000 manipulated samples were gathered by caption swapping the selected pristine samples.

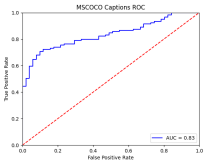

Fig. 3. ROC curve.

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. Fig. III show the resulting ROC curves. Our method achieved AUC of 83% for the gathered dataset.

We also evaluated our method in the DARPA Semantic Forensics (SemaFor) image & text news article inconsistency task. The test set was gathered from a variety of international news sites. The type of news article manipulations ranges from entity swaps, contextual changes, GAN generated images, *etc.* The magnitude of manipulation ranges from small contextual changes like date or location alternation to larger entity swaps. Lower scores of AUC 60% were received on this evaluation compared to the MS COCO experiment. This is mainly due to the indirect relationships present in most news articles.

IV. Conclusion

In this work, we proposed a semantically focused method that can provide fine-grained direct relational evidence towards the decisions it makes. This model is effective in capturing direct relationships between the image and text modalities. In future works we will improve the scalability of the model and expand into inference-based relationships.

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

References

[1] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. OSCAR: Object-semantics aligned pre-training for vision-language tasks. 2020.

[2] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gaok. VinVL: Revisiting visual representations in vision-language models. 2021.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskeverk. Learning transferable visual models from natural language supervision. 2021.

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrenc Zitnick. Microsoft COCO captions: Data collection and evaluation server. 2015.