An NLP Baseline for Image-Informed Mask Filling

Alex Lyman
University of Pennsylvania
alyman@seas.upenn.edu

Abstract

Vision-language models have made huge strides in recent years. While they achieve impressive results on tasks like visual question answering and image captioning, they can have difficulty in fine-grained settings. As part of the News Unmasked competition at the FGVC workshop at CVPR 2023, we train a model to predict masked words in New York Times headlines given the image accompanying the article. We show that an image-agnostic NLP approach performs comparably to methods including information from SOTA image-captioning systems. We release our model at https://huggingface.co/Qilex/roBERTaNYTheadlines

1. Introduction and Related Work

This report details our entry to the News Unmasked competition at the FGVC workshop at CVPR. From the competition description: "The News Unmasked competition aims to explore the limitations of large image-language models in understanding the relationship of an image with a headline... Since headlines and images often work together to communicate an emotion to a reader, the competition aims to understand how the images are related with the semantic characteristics of the text (headline)...

In this competition, participants are expected to predict masked words in headlines associated with a given image, considering the subject, image context, the emotional impact of the image, etc. The challenge explores the effectiveness of large models in generating headlines, with the aim of improving our understanding of the impact of image choice on headline perception and vice versa. Such an understanding of relationships between language and images are important for the application of large models.

The dataset consists of images and their associated news section, paired with headlines that have a few words masked. The goal is to predict the missing words in the headlines." [15]

In recent years, Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP). [2] Using transformer-based [11] architecture, these models leverage the power of massive pretraining to reach state-of-the-art results on many NLP tasks such as extractive question answering and text classification.

More recently, transformer-based joint image-language models have enabled similar results on tasks including visual question answering, few-shot image classification, and image captioning. [1] While implementation details differ, most of these models share some sort of joint embedding space between the image and the natural language components. These image-language models can suffer in fine-grained contexts, as image-language models are often trained on coarser tasks, meaning that out-of-the box image captions often fail to capture relevant and necessary information. [10] We perform a series of experiments to determine the effectiveness of image captions on this mask-filling task.

2. Methods

2.1. Dataset

The competition uses a subset of the the N24 News dataset [13] with roughly 49000 training samples and 12000 test samples. Each sample consists of a news headline, an accompanying image, and the section in which the article was published.

Of the test samples, roughly half of them are pre-masked and used for competition evaluation. We use the remaining samples as the test set for our experiments.

Our synthetic test set consists of 6546 sentences. Of these samples, 20% contain two masks, for a total of 7901 masks. These masks are randomly applied to non-punctuation tokens in the sentences.

2.2. Approach

Since this competition is a mask-filling task, we use RoBERTa, [7] a language model trained using the masked language modeling objective. This presents an advantage over autoregressive language models, as both the previous and following contexts are taken into account.

Using the principle of domain adaptation, [9] we fine-

tune the RoBERTa-large checkpoint on the train set of news headlines. Domain adaptation increases the likelihood of in-domain generations (words likely to appear in news headlines) and decreases the likelihood of out-of-domain generations.

For each masked word, we generate 10 top-k candidates with the domain-adapted model. These candidates often include punctuation and special tokens from the language model. Because RoBERTa uses a subword tokenizer, many of the candidate generations are subwords (i.e., ify). We first compare the candidate generations against a list of punctuation, special tokens, and common subwords, filtering out any inappropriate candidates. Then we perform a second pass, comparing the candidate generations against the vocabulary of spaCy's en-core-web-lg model [3] to ensure real words are generated. We finally choose the top remaining candidate generation as the best potential fit for the mask.

After perfoming a qualitative analysis on image captions generated by three SOTA image-captioning models, (GIT-large, COCA, and BLIP) [5, 12, 16] we choose BLIP-large to generate captions for each image.

Large Language Models have shown sensitivity to prompt structure, with both prompt tuning [4] and prefix tuning [6] causing significant upticks in results simply by appending useful information to the beginning of a prompt. We perform a four-way study to determine the optimal prompt.

1. Caption + section + masked sentence
2. Section + masked sentence
3. Caption + masked sentence
4. Masked sentence alone

We concatenate the information using natural language. For example, a constructed prompt containing section information uses the following template:
"A news article published in the _____ section. The headline is: _____"

2.3. Implementation Details

We use Pytorch [8] and the Transformers library [14] to implement our model. Pre-trained models are fetched from the HuggingFace hub via the transformers library.

Domain adaptation is performed on the RoBERTa-large checkpoint using roughly 41000 sentences from the competition train set, with the remainder being used for validation. We train for two epochs at a learning rate of 2e-5. We seed our training run for reproducibility.

2.4. Evaluation

We evaluate our model using the cosine similarity between semantic embeddings from the candidate generations and ground truth answer. The embeddings are acquired from spaCy's en-core-web-lg model. [3] The cosine similarity ranges from -1 (least similar) to 1 (most similar). We sum the cosine similarities and divide by the number of masks to compute an accuracy percentage.

Prompt Structure	Score
Caption + section + sentence	**58.50%**
Section + masked sentence	**58.50%**
Caption + masked sentence	58.07%
Masked sentence alone	57.66%

Table 1. Results on prompt study.

3. Results

As can be seen in Table 1, prefixing the prompt with extra information results in a small increase in score. Caption information alone seems to be slightly helpful, though not as helpful as section information.

Interestingly, providing information on the section in which the article appeared is as useful to the model as providing both section information and an image caption. This suggests that information from generic image captions may not meaningfully contribute in fine-grained contexts.

Because there was no meaningful difference in performance between the caption+section+sentence group and the section+sentence group, we use the principle of Occam's razor, and use the section+sentence approach for our contest submission.

4. Discussion

While this contest entry was by no means an exhaustive probe of vision-language models, our experiment suggests that generic image captions generated by SOTA vision-language models do not necessarily provide meaningful information on fine-grained image tasks. These findings are not necessarily new, [10] but they do reinforce the notion that a more advanced VQA approach would be needed for meaningful prefix tuning.

Our experiments also operate under the assumption that prefix- and prompt-tuning apply equally well to smaller masked language models, when these techniques have been shown mostly to apply to larger autoregressive models. A more thorough study is needed, both using diverse language model architectures as well as improving the prompts for vision language models.

5. Conclusion

Given the task of image-informed mask filling in a fine-grained news context, we show that an image-agnostic mask filling pipeline performs comparably to one involving generic text descriptions of the image.

Using the principle of domain adaptation, we fine-tune a masked language model to predict masked words in news headlines. We perform an experiment, using prefix-tuning to provide the model with additional information and find that generic image captions do not cause a meaningful improvement in performance over an image-agnostic approach.

References

[1] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. *CoRR*, abs/2102.02779, 2021. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 1

[3] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 2

[4] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 2

[6] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, Aug. 2021. Association for Computational Linguistics. 2

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 1

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[9] Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. 1

[10] Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre. Image captioning for effective use of language models in knowledge-based visual question answering. *Expert Systems with Applications*, 212:118669, 2023. 1, 2

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1

[12] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022. 2

[13] Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. N24news: A new dataset for multimodal news classification, 2022. 1

[14] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 2

[15] Srishti Yadav. News unmasked 2023, 2023. 1

[16] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 2