ClipCap: CLIP Prefix for Image Captioning

Under the supervision of
Prof. Svetlana Lazebnik

Presented By

Swaraj Thakre (sthakre2)
Pallaw Kumar (pallawk2)
Neha Jain (nehaj4)

Table of Contents

Executive Summary **3**

Introduction **4**

Problem Statement and Motivation 4

Related Works 4

Brief Summarization of the Approach 4

Details of the Approach **5**

Dataset Information 5

Training the Baseline Models 6

Experiments 6

Results **7**

Quantitative evaluation 7

Inference on Validation Dataset 8

Inference on Online(internet) Images 9

Discussion and Conclusions **9**

Main Insight 9

Failure Cases for Best Model 10

Potential Solutions 11

Statement of individual contribution **12**

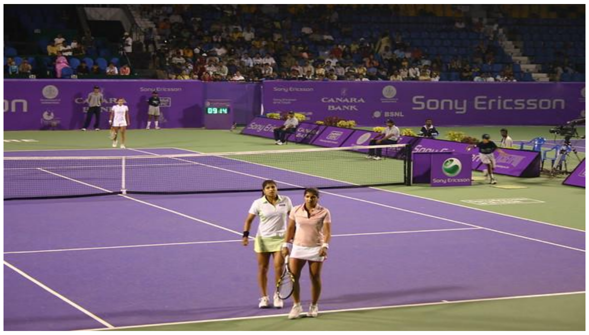
References **12**

Executive Summary

Traditional image captioning methods often use Convolutional Neural Networks (CNNs) to analyze images and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for language generation. However, these models are resource-intensive, requiring considerable training time, a large number of trainable parameters, and massive datasets.

The paper "ClipCap: CLIP Prefix for Image Captioning"[1] proposes a more efficient image captioning method. This approach employs the pre-trained vision-language model CLIP to streamline the captioning process, along with a mapping network for compatibility with the Large Language GPT-2 model. In addition to implementing the models from the ClipCap paper, we executed two supplementary experiments to enhance the overall performance of the image captioning framework. The first experiment involved the integration of the advanced GPT-Neo language model as a substitute for the GPT-2 model. The second experiment involved the use of a more sophisticated transformer-based mapping network. We assessed the performance using BLEU, METEOR, CIDEr, and ROUGE-L scores on the COCO validation set, offering a robust evaluation of the model's efficiency and effectiveness.

The most effective models utilized the MLP-based mapping network combined with fine-tuned GPT2 and the Transformer-based mapping network integrated with frozen GPT2. However, models with the Transformer-based mapping network with frozen GPT-Neo and the Transformer-based mapping network (16 layers) combined with frozen GPT2 did not surpass the performance of the baseline models, mainly due to limited computational resources preventing fine-tuning. We discuss each model inference result, main insights and several failure cases of the models where the generated captions did not accurately represent the images. Further, this findings highlight the challenges in image captioning and the need for further refinements in the approach.



MLP-based mapping network + fine-tuning GPT2	*A couple of women standing on top of a tennis court*
Transformer-based mapping network + frozen GPT2	*Two women standing on a tennis court with rackets*
Transformer-based mapping network + frozen GPT-Neo	*Two women standing next to each other on a tennis court*
Transformer-based mapping network (16 layers) + frozen GPT2	*Two female tennis players on a track in a sports arena*

Introduction

Problem Statement and Motivation

In image captioning, the task is to provide a meaningful and valid caption for a given input image in a natural language. Many approaches have been proposed for image captioning that utilize CNN to understand image contents and find out objects in an image while RNN or LSTM is used for language generation. These models are resource hungry. They require extensive training time, a large number of trainable parameters, and a massive dataset, which limit their practical applicability. While we came across several image captioning approaches, we found the proposed paper titled “ClipCap: CLIP Prefix for Image Captioning”(2021)[1] which uses a different strategy. This approach proposes a lightweight captioning method that leverages a pre-trained vision-language model called CLIP to simplify the captioning process.

Related Works

The CLIP model [2], with its dual encoders for image-text representation, has been a breakthrough in vision-language tasks. The authors leverage this, uniquely using only CLIP's image encoder for image captioning. Traditional image captioning models have been limited by the need for additional annotations or extensive pre-training. The authors' approach mitigates this by utilizing CLIP's visual embeddings and the auto-regressive language model, GPT-2 [3], for caption generation. Notably, this work draws from studies [4,5,6] that use vision-language pre-training for a shared latent space. Unlike these studies that require additional supervision or lengthy pre-training, the authors' method capitalizes on CLIP's inherent joint representation, simplifying the process significantly.

Brief Summarization of the Approach

This paper presents an image captioning approach that utilizes the state-of-the-art CLIP model for joint image and natural language encoding, alongside a mapping network that makes the output compatible with the GPT-2 language model. Two types of mapping networks were introduced in the paper: the MLP and Transformer-based mapping networks. We initially replicated the baseline methods delineated in the paper to validate the proposed results. Furthermore, we conducted two experimental studies aimed at enhancing the baseline framework: one integrating the GPT-Neo as the language model, and another implementing a Transformer mapping network with an expanded 16 layers. The objective of these innovative adaptations was to optimize and refine the performance of the baseline image captioning framework.

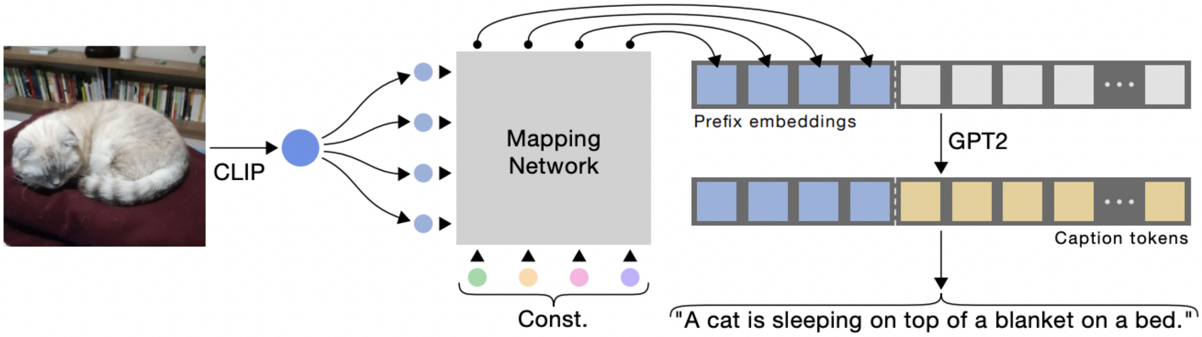


Fig 1: Overview of transformer-based mapping network from the CLIP embedding space and a learned constant to GPT-2. At inference, GPT-2 is used to generate the caption given the prefix embeddings

Details of the Approach

Dataset Information

We chose to use the Microsoft Common Objects in Context (COCO) [7] dataset for training and validating our model. This dataset is popular in the research community because it's comprehensive and provides lots of images and annotations for image captioning. We split the COCO dataset into two subsets - one for training (82,783 images) and one for validation (40,504 images). This way, we could use the validation set to test how well our model generalized from the patterns it learned in the training set. Plus, each image in both subsets had five distinct captions, which was useful for making our language model more robust and reliable.

Approach	Number of epochs	Training Time
MLP-based mapping network + fine-tuning GPT2	10	6h
Transformer-based mapping network + frozen GPT2	10	10h
Transformer-based mapping network + frozen GPT-Neo	10	12h
Transformer-based mapping network (16 layers) + frozen GPT2	10	14h

Table 1: Table describing the number of epochs and the training time required by each model

Training the Baseline Models

The image captioning approach proposed in this paper harnesses the advanced capabilities of the CLIP (Contrastive Language-Image Pre-Training) model. This revolutionary tool can concurrently encode both images and natural language into a shared high-dimensional vector space. Through the use of the pre-trained CLIP encoder, the paper generates an embedding for the input image that encapsulates its visual features, aligning them with their corresponding semantic representation.

A key innovation introduced in this paper is the mapping network, which serves as an intermediary between the CLIP-generated embedding and an embedding compatible with the GPT-2 language model. This translated embedding becomes a learned prefix for the language model, guiding it to generate captions that are semantically and syntactically connected to the input image. The paper introduces two types of mapping networks:

1. **MLP (Multilayer Perceptron)**: This network consists of a single hidden layer (2 fully connected layers). The first layer takes the visual features as input and projects them into a lower-dimensional space. The second layer takes these projected visual features along with the textual input, producing a joint embedding that effectively captures the intricate relationship between the two modalities.

2. **Transformer-based mapping network**: This network has two distinct inputs - the visual encoding of CLIP and a learned constant input. The constant input plays a dual role. Firstly, it extracts meaningful information from the CLIP embedding through a sophisticated process of multi-head attention. Secondly, it learns to adapt the fixed language model to the new data, ensuring that the model stays relevant and effective. For the setup, the paper uses the CLIP embedding with (K = 10) constant tokens and applies 8 multi-head self-attention layers with 8 heads each.

In the paper's baseline method, a pre-trained CLIP image encoder is used to generate embeddings from the input images. These embeddings are then transposed into a format that the GPT-2 language model can interpret via the mapping network. The transformed embeddings, serving as a learned prefix for the language model, enable the generation of captions that hold semantic and syntactic relevance to the input image. The paper employs GPT-2 as the language model in this baseline approach due to its proven efficiency and robustness.

Experiments

We conducted two distinct experimental studies to refine the performance of the baseline image captioning framework.

The first experiment involved integrating GPT-Neo as the language model while retaining the same setup for the rest of the components. GPT-Neo, a recent model, delivers performance equivalent to GPT-3 models of similar sizes, but with more efficient resource utilization. As opposed to the GPT-2 model used in the baseline method, which contains 117 million parameters, the GPT-Neo model boasts 125 million parameters. Training was executed over 10 epochs on Google Colab, utilizing the A100 GPU's computational prowess. The performance of this revised framework was evaluated on the COCO validation set, with BLEU, METEOR, CIDEr and ROUGE-L scores among the metrics used for assessment.

The second experiment featured a transformer mapping network with 16 layers, an increase from the 8 layers in the baseline method. This was done with the aim to generate more precise embeddings for the GPT2 language model, which also serves as a prefix for image caption generation. Similar to the first experiment, training was conducted over 10 epochs on Google Colab, using an A100 GPU, and performance was evaluated on the COCO validation set using a range of metrics, including BLEU, METEOR, CIDEr and ROUGE-L scores. The enhanced transformer mapping network in this setup was expected to significantly bolster the overall image captioning performance of the system.

Results

Quantitative evaluation

As depicted in Table 2, the models featuring the MLP-based mapping network combined with fine-tuned GPT2 and the Transformer-based mapping network integrated with frozen GPT2 yield the most superior performance. Due to computational constraints, the metrics in Table 2 were derived from a subset of 5k images out of the total 40k image validation set. Notably, our experimental models - the Transformer-based mapping network with frozen GPT-Neo and the Transformer-based mapping network (16 layers) combined with frozen GPT2 - did not surpass the performance of the baseline models. This was primarily attributed to the limited computational resources, which prohibited us from fine-tuning these models.

Approach	MLP-based mapping network + fine-tuning GPT2	Transformer-based mapping network + frozen GPT2	Transformer-based mapping network + frozen GPT-Neo	Transformer-based mapping network (16 layers) + frozen GPT2
Bleu@1	0.7852	0.8003	0.4624	0.5919
Bleu@2	0.6353	0.6534	0.3134	0.4047
Bleu@3	0.4978	0.5096	0.1988	0.2668

Bleu@4	0.3866	0.3902	0.1229	0.1745
METEOR	0.2987	0.2933	0.1751	0.1940
CIDEr	0.5927	0.5915	0.4240	0.4419
ROUGE-L	1.3068	1.2924	0.4662	0.5779
Inference Time (ms per image)	192 ms	264 ms	288 ms	276 ms

Table 2: Quantitative Evaluation of all the models on COCO validation subset of 5k images

Inference on Validation Dataset

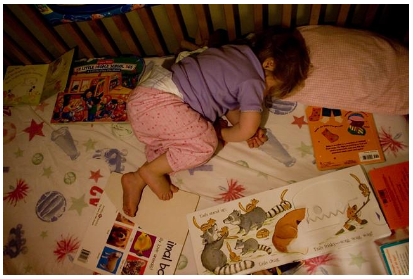
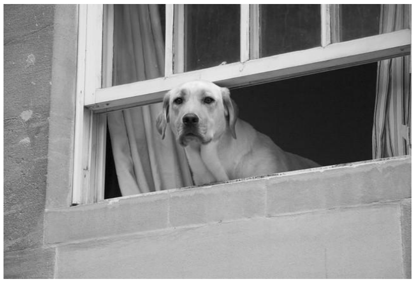
Approach		
MLP-based mapping network + fine-tuning GPT2	A baby sleeping on a bed with lots of books	A dog that is looking out of a window
Transformer-based mapping network + frozen GPT2	A child sleeping on a bed with a book	A black and white dog looking out a window
Transformer-based mapping network + frozen GPT-Neo	A child is laying on the floor with a book	A brown and white dog is sitting on top of a window
Transformer-based mapping network (16 layers) + frozen GPT2	A child is sleeping on the bed of a book	A dog is looking out of the window of a building

Table 3: Generated captions via each model for the images from the COCO validation dataset.

Inference on Online(internet) Images

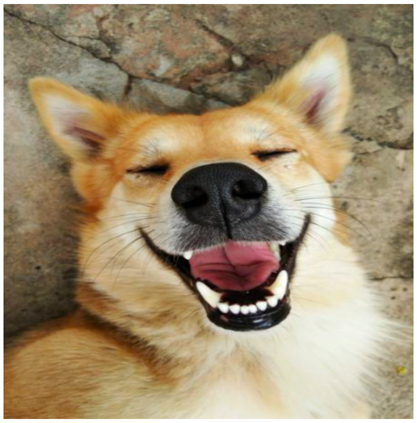
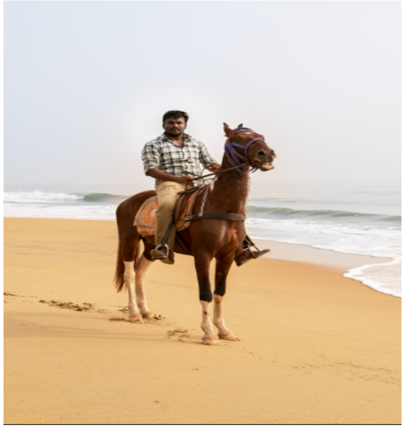
Approach		
MLP-based mapping network + fine-tuning GPT2	a close up of a dog smiling at the camera	A man riding a brown horse on top of a sandy beach
Transformer-based mapping network + frozen GPT2	A dog with a smile on its face	A man riding a horse on the beach
Transformer-based mapping network + frozen GPT-Neo	A picture of a dog with it's teeth out	A man in a blue shirt is riding on top of a horse
Transformer-based mapping network (16 layers) + frozen GPT2	A close up picture of a dog with a smile on his face	A man sits on the back of a horse on the beach

Table 4: Generated captions via each approach for random images from the internet.

Discussion and Conclusions

Main Insight

Mapping network is used to bridge the gap between the image embeddings generated by the CLIP image encoder and the embeddings compatible with the Large Language Model.

Fine-tuning the Large Language Model(GPT-2/ GPT-NEO) enables us to utilize a simpler Mapping network such as MLP, as it allows the model to learn intricacies that are specific to the dataset. Conversely, if we keep CLIP and the Large Language Model frozen and solely train the mapping network, we would require a complex mapping network (transform) to learn the dataset's intrinsic features.

Nevertheless, if we have computation in hand, we can fine-tune CLIP, the Mapping network, and the Large Language Model simultaneously to enhance performance. This end-to-end fine-tuning approach facilitates a more comprehensive understanding of the data by the model.

Failure Cases for Best Model

(MLP-based mapping network + fine-tuning GPT2)

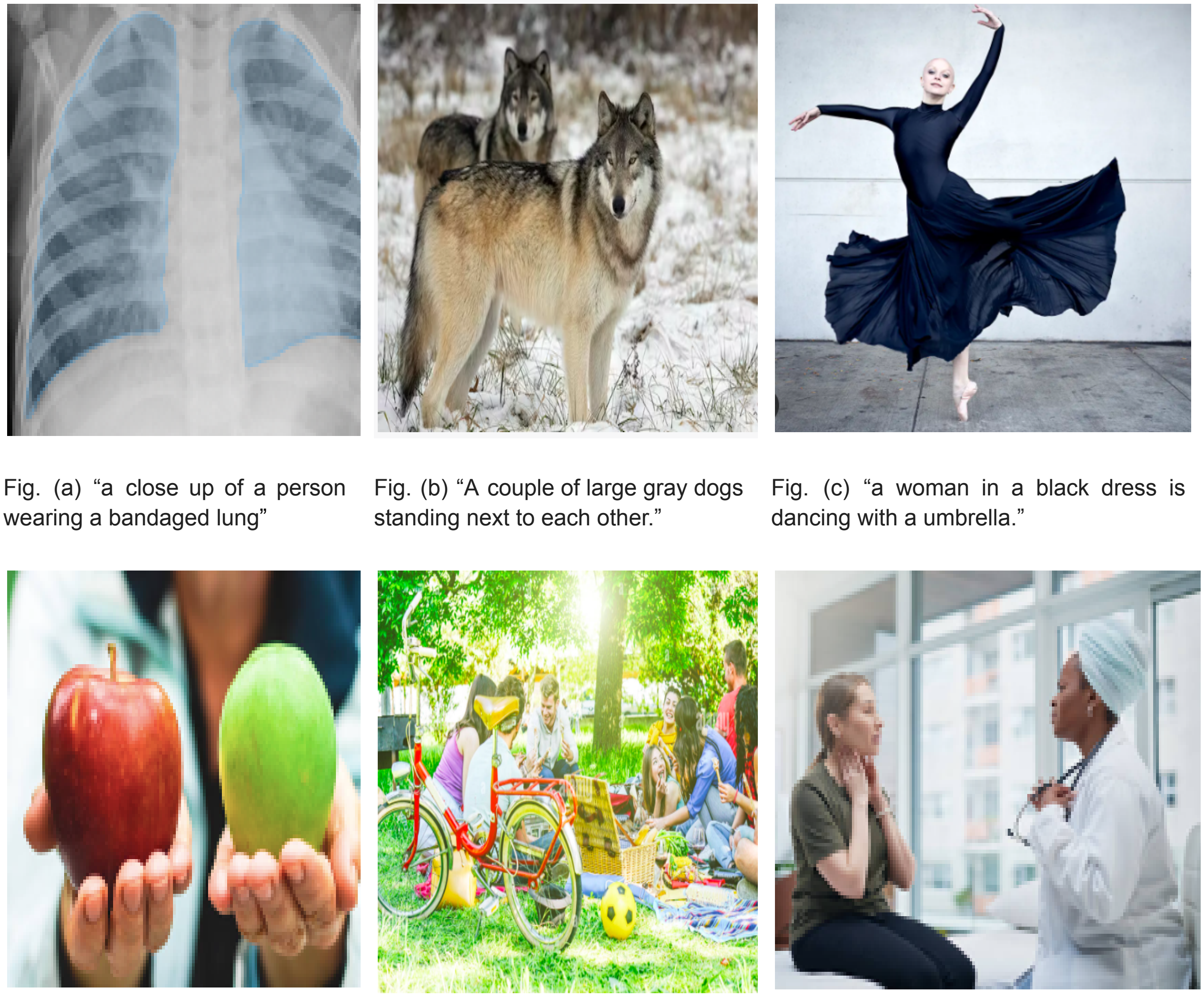


Fig. (a) “a close up of a person wearing a bandaged lung”

Fig. (b) “A couple of large gray dogs standing next to each other.”

Fig. (c) “a woman in a black dress is dancing with a umbrella.”

Fig. (d) “a close up of a person holding an apple and an apple.**”**

Fig.(e) “A group of people sitting in the grass next to a yellow bicycle.”

Fig. (f) “A woman in a hospital gown talking to another woman.”

Fig. 2: Captions generated from MLP-based mapping network + fine-tuning GPT2 model

Although our best-performing model (MLP-based mapping network + fine-tuning GPT2) successfully generates meaningful captions, below are some cases where the best-performing model fails.

1. **Out-of-distribution sample**: The model is trained on the COCO dataset. For an image that falls outside its training distribution, the model fails to generate accurate and meaningful captions. For example, in Fig. 2-(a) an image showing a chest X-ray generated as “a close-up of a person wearing a bandaged lung”.

2. **Image Ambiguity**: The model fails in case of ambiguity when an image contains ambiguous objects or multiple possible interpretations. An example of it is shown in image Fig. 2-(b) where the caption for two wolves is generated as “a couple of large gray dogs standing next to each other”. Consequently, the model failed to distinguish between dogs and wolves.

3. **Requires Visual Perception Interpretability**: The model lacked direct visual perception. This limitation led to inaccurate captions, especially when it had to comprehend complex visual scenes, Fig. 2-(c) shows a woman in a black dress is dancing but the model generated the caption “a woman in a black dress is dancing with an umbrella”. The model failed to interrelate that the curling dress is visually matching like an open umbrella but is not an umbrella.

4. **Image lacking Visual Content**: The model fails to interpret direct access to the visual content, resulting in captions that are not precisely aligned with the visual details and fails to describe specific visual attributes accurately. As an example, Fig. 2-(d) shows a person holding a red apple in one hand and a green apple in the other hand. The model could not describe the specific visual attributes of the objects in the image.

5. **Image lacking Coherence**: The model generates captions for images lacking coherence and fails to provide a consistent narrative structure, as shown in Fig. 2-(e) a group of friends having a picnic in a park. The model's inability to infer a lack of coherence generated a caption "a group of people sitting in the grass next to a yellow bicycle." that clearly lacks a logical narrative progression.

6. **Image sensitive to model bias**: The model is sensitive to biases present in the training data. This bias is reflected in the generated captions, potentially leading to stereotyping or unfair associations in the generated descriptions, as shown in Fig. 2-(f). A doctor wearing a white coat, a stethoscope around their neck, and interacting with a patient. Due to language biases in the training data, the model generated a biased caption: "A woman in a hospital gown talking to another woman.".

Potential Solutions

There are several potential solutions to consider in order to enhance the performance and quality of the generated captions:

1. **Fine-tuning:** The CLIP model can be fine-tuned on a specific image captioning dataset to improve its performance for the task. By training CLIP with captioned image data, it can learn to

better align visual and textual information, leading to more accurate and contextually relevant captions.

2. **Using a larger and more diverse dataset:** The more data that a model is trained on, the better it will be able to generalize to new images.

3. **Using a better mapping network:** The mapping network is responsible for converting the visual representation of an image into a textual representation. A better mapping network will allow the model to generate more accurate and informative captions.

4. **Using a more powerful language model:** GPT-2 is a powerful language model, but there are even more powerful models available. Using a more powerful language model like GPT-3 will allow the model to generate more creative and informative captions.

Statement of individual contribution

- All: Understanding the architecture, Setting up the hardware, acquiring the appropriate dataset for the use-case, identifying the failure cases, and contributed to the report.
- Neha Jain: Training the baseline models and setting up evaluation to test the model performance and inference code for all model
- Pallaw Kumar: Training the clip model and training the transformer-based mapping network (16 layers) + frozen GPT2
- Swaraj Thakre: Training the clip model and training the Transformer-based mapping network + frozen GPT-Neo

References

[1] Mokady, R., Hertz, A., & Bermano, A. H. (2021). ClipCap: CLIP Prefix for Image Captioning. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2111.09734

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[4] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In European Conference on Computer Vision, pages 121–137. Springer, 2020.

[5] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.02265, 2019.

[6] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pretraining for image captioning and vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 13041–13049, 2020.

[7] Chen, X. (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. ArXiv.Org. https://doi.org/10.48550/arXiv.1504.00325

[8] Code: https://github.com/rmokady/CLIP_prefix_caption