Probing the Prompting of CLIP on Human Faces

Anonymous ACL submission

Abstract

Large-scale multimodal models such as CLIP (Radford et al., 2021) have caught great attention due to their generalization capability. CLIP can take free-form text prompts, but the performance varies with different text prompt manipulations, which is considered unpredictable. In this paper, we conduct a controlled study to understand how CLIP perceives images with different forms of text prompts, particularly on human facial attributes. We find that (1) using the prompt starter "a photo of" can guide the model to allocate higher attention weights to human faces, leading to better classification performance; (2) CLIP model is better at aligning information from shorter text prompts, as additional textual details shift away the attention from key words; (3) properly adding punctuation or removing stop words in the text prompt can shift attention to target information. Our practice on facial attributes shed light on the design of reliable text prompts for CLIP in other tasks.

1 Introduction

Recently foundation models such as CLIP (Radford et al., 2021) and GPT-3 (Brown et al., 2020) have caught great attention. These foundation models benefits from pre-training on large scale unlabeled text data from the Internet and can extract semantic meaning from free-form text prompts. As one of the most representative models, CLIP utilizes image data and text prompts to extract useful visual and textual information and align similar images and text by finding their correlation.

The pre-trained CLIP model can serve as zero-shot learners for downstream applications including classification (Choudhury et al., 2021; Bujwid and Sullivan, 2021), image retrieval (Stefanini et al., 2021), image generation (Xia et al., 2021; Patashnik et al., 2021; Karras et al., 2020), etc. Specifically, Shen et al. (2021) shows that incorporating CLIP can improve performance on vision-and-language tasks including Visual Question Answering (Zhou et al., 2020), Visual Entailment (Xie et al., 2019), and Vision-and-Language Navigation (Anderson et al., 2018; Ku et al., 2020).

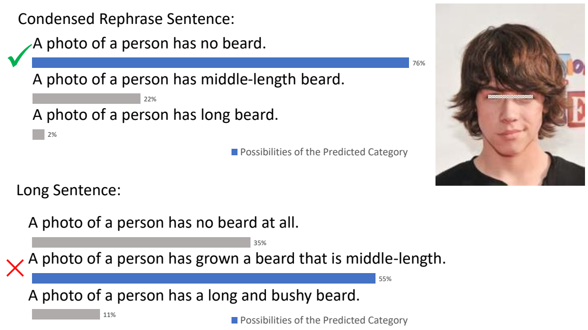

Figure 1: Example of CLIP prompts on a face image. In the beard classification task, for the same portrait on the left, different text prompt designs could have a serious impact on the classification results of CLIP. CLIP correctly predicts the ground truth from shorter prompts but makes a wrong matching on longer prompts.

The flexible prompting ability of CLIP is the key to its success on zero-shot classification tasks. For instance, Radford et al. (2021) used "a photo of {class}" for image classification. Nonetheless, when the carefully designed text prompts are manipulated or rearranged, the CLIP model will perceive the images in very different ways. As shown in Figure 1, two sets of text prompts lead to very different predictions for the same portrait, even though both refer to similar semantic meanings. The sensitivity to prompt manipulation leads to a discrepancy of prediction outcomes or even performance degradation. In contrast, when humans read a sentence that either skips a few words or is randomly rearranged, it is very likely that they can still understand the corrupted sentence and relate it to the correct images (Hahn and Keller, 2016). In consequence, it is crucial to understand and interpret *how CLIP perceives the input image and text prompt* and *how well CLIP performs with* *manipulated text prompts*.

To answer these questions, we conduct controlled experiments on prompt starters, shortened prompts, word orders, and non-semantic tokens to probe the effect of different prompt manipulations of the CLIP model. The CelebA-Dialog dataset (Jiang et al., 2021) provides text annotations of facial attributes at different granularity levels, which is a perfect testbed for our task. Therefore, we experiment with facial images by disentangling different facial attributes and quantitatively assessing the impact of different text prompts on CLIP. Recent works (Agarwal et al., 2021; Wang et al., 2021) have unveiled the bias issues of the CLIP model on human faces but they did not investigate the cause and effect of prompt manipulation on facial attributes.

In this work, we try to understand the explicit effect of different prompt manipulations to facial attributes understanding, and conduct a series of experiments on CelebA-Dialog (Jiang et al., 2021), aiming to answer the following research questions:

1. How does CLIP perceive the sentence starter in the text template (see Section 3)?
2. Do length and order of the text prompt affect the evaluation (see Section 4)?
3. Does non-semantic tokens, like punctuation and stop words, really matter in text prompts (see Section 5)?

2 Settings

Model Our goal is to understand how CLIP perceives the world and how it is different from human. Therefore, we did not apply any modification or task-specific fine-tuning and only used the pre-trained model.[1] The CLIP model can take images and personalized text prompts as input and encode them into the same representation space. The cosine similarity can be used to measure how the image is similar to the text prompt. For classification tasks, we select the text prompt with the highest similarity score as the prediction to the target image.

Dataset We used CelebA-Dialog (Jiang et al., 2021) as our image dataset, which is a large-scale visual-language face dataset annotated with five fine-grained facial attributes and the corresponding textual descriptions. We use the original validation set consisting of 19,864 images for all the experiments. We select four attributes for evaluation, including Eyeglasses, Bangs, Smiling, and Beard. For each attribute, the original CelebA-Dialog dataset contains six degrees. We expect more accurate classification results so that the effect of different text prompts can be observed more clearly. Thus, we grouped six degrees into three classes for all attributes. For instance, we categorize eyeglasses attribute into no eyeglasses, eyeglasses, and sunglasses.

Metric Image-text matching is essentially a classification problem. We use F1 score to evaluate the classification performance.

Visualizing attention heatmap We utilize the attention tool proposed by Chefer et al. (2021). The model aggregates attention heads by integrating the gradients and attention maps to average across attention heads for each attention layer and then aggregates the attention through several layers. The visualization result is generated by relevancy maps for each interaction between text prompts and face images.[2]

3 Prompt Starter Helps CLIP Focus

When designing the text prompts, CLIP (Radford et al., 2021) suggests using "a photo of `{label}`" as the sentence starter. To determine the effect of this design, we applied such a template to the text description drawn from CelebA-Dialog dataset (Jiang et al., 2021). We treat the full description with the prompt starter as a baseline. Table 1 part A shows the performance in each task decreased when sentence starter were removed from the text prompt.

To help reason this discovery, we plot the average attention map of all images and the heat difference between with and without sentence starter in Fig 2. We plot the difference map by subtracting the heatmap without using a sentence starter (induces worse F1 score) from the one with a sentence starter (induces better F1 score). We observe that the difference on the human face is positive and that on the background is negative in general. With sentence starter, CLIP focuses more on nose and mouth than the unrelated background. In the

[1] The pre-trained CLIP model is released at `https://github.com/openai/CLIP`.

[2] The attention visualization tool is available at `https://github.com/hila-chefer/Transformer-MM-Explainability`.

	Prompt	Example	bangs	glasses	smile	beard
	Full (Baseline)	*A photo of a person with thin or thick frame sunglasses.*	42.68	71.48	54.11	40.73
(A)	Removing Sentence Starter	*A person with thin or thick frame sunglasses.*	37.69 (-4.99)	60.45 (-11.03)	53.34 (-0.77)	16.37 (-24.36)
(B)	Condensed Rephrase*	*A photo of a person with sunglasses.*	49.55 (+6.87)	88.53 (+17.05)	60.19 (+6.08)	46.05 (+5.32)
	Random Order	*person photo with sunglasses of thick frame or A thin.*	20.03 (-22.65)	37.85 (-33.63)	25.07 (-29.04)	27.81 (-12.92)
	Randomly Skipping Words	*A photo of with thin frame.*	13.46 (-29.22)	18.33 (-53.15)	14.27 (-39.84)	11.21 (-29.52)
(C)	Adding Punctuation	*A photo of a person with thin or thick frame "sunglasses".*	43.85 (+1.17)	77.55 (+6.07)	59.87 (+5.76)	43.04 (+2.31)
	Adding Random Punctuation	*A photo of a person with "thin or" thick frame sunglasses.*	40.13 (-2.55)	69.15 (-2.33)	43.81 (-10.3)	39.62 (-1.11)
	Removing Stop Words	*A photo of a person thin thick frame sunglasses.*	43.11 (+0.43)	76.29 (+4.81)	57.53 (+3.42)	43.31 (+2.58)

Table 1: F1 scores for different text manipulations over four facial attributes. Full is the baseline text prompt from CelebA-Dialog dataset (Jiang et al., 2021). Part (A) corresponds to section 3, an experiment to show the effect of removing the sentence starter template. Part (B) corresponds to section 4 and shows the effect of using shorter text prompt, condensed rephrase only keeps the key information to the classification and keeps grammatical correctness. Random order shuffles the text to see if word order matters. Randomly skipping words randomly drop words in text prompts. Part (C) corresponds to section 5. Adding punctuation in correct spot can boost the performance, while adding random punctuation distracts the attention. Remove stop words discards all the words that do not contain key information. We observe that condensed rephrase consistently dominates the accuracy over four facial attributes.

facial attributes classification task, It is helpful to use the sentence starter to restrict the scope to the human face and enforce CLIP to focus on the relevant area. Moreover, we conjecture this conclusion can also be applied to other tasks such as "a photo of {class}" in object detection.

4 Impact of Length and Order

Short prompts beat long prompts A complete description of a person's face contains more detailed information about the facial attributes than a shortened version. Given the full description, human readers make better classification decisions. In this experiment, we want to know if such a property holds when CLIP perceives text prompts.

We designed the condensed rephrased template by shortening baseline description. Such a template keeps the key information to the classification and ensures grammatical correctness. Table 1 part B shows that the numerical results on facial attribute classification, given the shortened text prompts. The results of the condensed rephrase template show using such a shortened text prompt can significantly improve F1 scores in all four tasks. When classifying the glasses attribute, the shortened template has an improvement of 17.05%. Although detailed descriptions were missing, the model here will not waste the attention weights on trivial information.

We show the color-coded attention heatmap examples of these text prompts in Fig 3. When CLIP perceives the text prompt, a darker color means higher attention weight and vice versa. The band example heatmap shows that the model did not have any attention weight on the negative word "no" and wasted a portion of attention on the trivial descriptions when using the full prompts as input.

Word order matters Here we want to figure out how word order and missing words in sentences affects the model. Table 1 part B shows performance of CLIP model given a random order text prompt. The performance dropped in all four classification tasks. The average F1 score of bangs classification is 22.65% lower than baseline. Despite the poor performance, the performance over the four tasks still share a similar trend as the baseline setup. Without word order, we found CLIP model behaves similar to human, neither can extract information accurately, but can still make rough guesses.

Table 1 part B also shows randomly removing words in the text prompts. Here key words can be removed during the manipulation and causes the model performs entirely random.

5 Non-semantic Tokens

Punctuation and stop words are non-semantic tokens in a sentence. However, they can help human

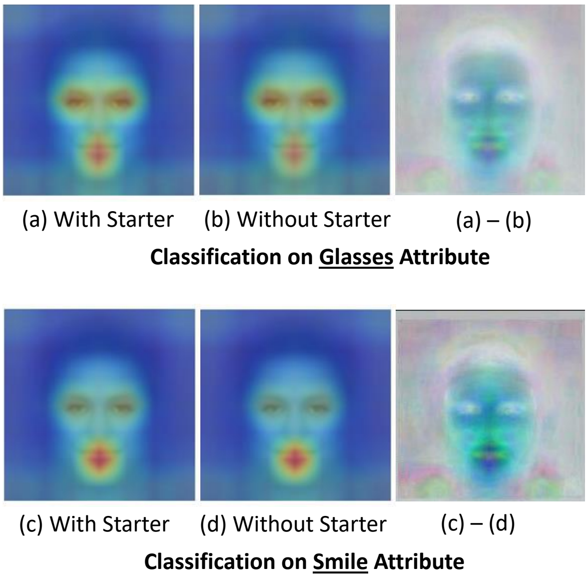

Figure 2: Average image attention visualization. Top row: classification on Glasses attribute; bottom row: classification on Smile attribute. (a) is the average attention heatmap over all the testing images with the prompt starter; (b) is the average attention heatmap without the prompt starter. (a)−(b) is the difference between (a) and (b): blue color represents positive values (more attention from (a) than (b)) and red color represents negative values (less attention from (a) than (b)). The prompt starter makes CLIP focus more on human faces rather than the background.

Bangs:

a photo of a person with long bangs. Can see 0% of the forehead. <Full>
a photo of a person with long bangs. <Condensed Rephrase>
a photo of a person long bangs. Can see 0% forehead. <No-Stopwords>
a photo of a person with 'long' bangs. Can see 0% of the forehead. <Punctuation>

Smile:

a photo of a person who is not smiling at all. The mouth is closed. <Full>
a photo of a person with no smile. <Condensed Rephrase>
a photo of a person not smiling. mouth closed. <No-Stopwords>
a photo of a person who is 'not smiling' at all. The mouth is closed. <Punctuation>

Glasses:

a photo of a person with thin or thick frame sunglasses. <Full>
a photo of a person with sunglasses. <Condensed Rephrase>
a photo of a person thin thick frame sunglasses. <No-Stopwords>
a photo of a person 'with' thin or thick frame sunglasses. <Punctuation>

Beard:

a photo of a person who has a bushy beard that is long. <Full>
a photo of a person with a long beard. <Condensed Rephrase>
a photo of a person bushy beard long. <No-Stopwords>
a photo of a person who has a bushy 'beard that is long.' <Punctuation>

Figure 3: Average text attention heatmap of different text manipulations over four facial attributes. Given the same set of images, a darker color coded text means CLIP pays higher attention to the word, and vice versa. The bracket after text prompts indicate the types of text manipulations, correspond to experiments in Table 1.

readers understand a sentence. In this section, we explore the effect of adding punctuation or removing stop words in text prompts to the CLIP model.

Punctuation helps. To understand the effect of punctuation, we designed two experiments. The first one is manually inserting quote marks into keywords and emphasizing their importance. The second one is randomly inserting quote marks.

In the first experiment, text prompts might not seem grammatically correct, which we previously show not a required constraint, in section 4. Table 1 part C shows that adding punctuation to keywords boosts performance in all four classification tasks; in glasses classification, the F1 score increased 6.07% from the baseline. As an ablation study, the second experiment shows that randomly adding quote marks does not help and even reduces overall performance.

Stop words hurt. To understand the effect of stop words, we evaluated removing all the stop words in the text prompt, and Table 1 part C shows the numerical results. This manipulation causes some prompts to fail to hold grammatical correctness. We were surprised to find that removing stop words shows that such a setup can also increase the performance in all four tasks compared to the baseline. In the glasses and smile classification tasks, the improvement is 4.81% and 3.42%, respectively. As Fig 3 shows with both shortened version, CLIP model pays more attention to the keywords like "band", "smile", "sunglasses", and "beard". However, only removing stop words in text prompts, CLIP still focuses on the trivial descriptions.

From the experiment results in this two setting, we find that a shortened version of text prompt even without grammatical correctness can enforce model to pay higher attention on key words, and leads to performance increase.

6 Conclusion

CLIP allows designing personalized text prompts for a vast range of tasks. While the zero-shot transfer capability is powerful, it is important to rethink how does CLIP understand text prompts and what really matters in prompt engineering. In this work, we compare the performance of a variety of text manipulations and interpret how CLIP perceives them accordingly. We expect the controlled experiment on facial attribute recognition can motivate the practice on other vision and language tasks.

References

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sebastian Bujwid and Josephine Sullivan. 2021. Large-scale zero-shot image classification from rich and diverse textual descriptions. *ArXiv*, abs/2103.09669.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. *arXiv preprint arXiv:2103.15679*.

Subhabrata Choudhury, Iro Laina, C. Rupprecht, and Andrea Vedaldi. 2021. The curious layperson: Fine-grained image recognition without expert labels. *ArXiv*, abs/2111.03651.

Michael Hahn and Frank Keller. 2016. Modeling human reading with neural attention. *arXiv preprint arXiv:1608.05604*.

Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. 2021. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *EMNLP*.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and D. Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. *ArXiv*, abs/2103.17249.

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*.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can CLIP benefit vision-and-language tasks? *CoRR*, abs/2107.06383.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2021. From show to tell: A survey on image captioning. *ArXiv*, abs/2107.06912.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Assessing multilingual fairness in pre-trained multimodal representations.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Towards open-world text-guided face image generation and manipulation. *ArXiv*, abs/2104.08910.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *ArXiv*, abs/1901.06706.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. *ArXiv*, abs/1909.11059.