Finetuning Vision-Language Model with Reinforcement Learning

Licheng Luo, Kefan Song, Ye Ma
University of Virginia

Abstract

We demonstrated that text-to-image models can be finetuned by RL to improve performance on image captioning task, without requiring human annotated data. We propose two designs of reward functions, one based on diffusion and an image similarity measure and the other based on an image-text retrieval model. We've shown that our method improves the similarity between the image and the generated text on COCO caption validation dataset that is previously unseen to the image-to-text model.

1 Introduction

Current image-to-text models (Wang et al. 2022) relied on human-labeled image caption data. For each pair of training data (X, y), input X is the image and output label y is a brief sentence description of the image, annotated by human. The label is a high-level description of the image, and the length of text label is too short to capture mid-level features or low-level features of the image.

Inspired by autoencoder, we aim to train our image-to-text model to generate text descriptions to contain enough details, with the goal of reproducing the input image from the text descriptions, where the image generation can be performed by a pre-trained diffusion model. The text descriptions of our model play a similar role as the latent space representation in autoencoder, therefore it can be trained in a self-supervised manner that does not require human-labeled caption annotations. We applied reinforcement learning to finetune vision-language models. Our method is able to improve the similarity between the image and the generated text on COCO caption validation dataset. The result demonstrates that vision-language model can be further improved by RL without direct supervision from human annotated labels.

2 Related Works

Autoencoder Autoencoders are a class of neural network models aimed at learning efficient representations (encodings) of input data, typically for the purpose of dimensionality reduction or feature learning. An autoencoder consists of two main components: an encoder and a decoder. The encoder compresses the input data into a lower-dimensional latent representation, and the decoder reconstructs the original input data from this compressed representation. The training process involves minimizing the reconstruction error, which is the difference between the original input and its reconstruction (Hinton and Salakhutdinov 2006). This framework could also be applied to noise removing (Vincent 2010) and generative modeling (Kingma and Welling 2013).

Vision Language Model There exist previous work that aims to alieviate the cost of human labeled data when training vision-language models. BLIP (Li et al. 2022) reduces the reliance on human labeled data by training a captioner to produce synthetic captions given web images, and a filter to

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

remove noisy captions from both the original web texts and the synthetic texts. However, the training of the captioner and the filter is independent, and no reinforcement learning is involved as training the captioner does not utlize the feedback from the filter.

Another line of work on using natural language as the interface for different modalities of data is closely related to our work. De-Diffusion (Wei et al. 2023) applies supervised learning to train an image-to-text model that generates text that can reproduce the input image by a diffusion model. Unlike Large Language Models fine-tuned by RL, De-Diffusion generates text that doesn't follow any grammar rules or sound natural to human. Therefore, the output cannot be considered as better image captions.

Reinforcement Learning with Human Feedback Recent advancements in Reinforcement Learning from Human Feedback (RLHF) have yielded several noteworthy contributions that enhance our understanding and efficacy in the field. A survey delving into RLHF techniques illustrates the broader categorization of reinforcement learning practices that prioritize human-derived insights over traditional reward functions (Kaufmann et al. 2023). Adding to the methodological richness, research on learning from dynamic human choices has introduced a pessimistic outlook on policy optimization, considering the bounded rationality of human agents (Li et al. 2023). Furthermore, the active querying paradigm in RLHF, as explored by Ji et al., propels the field towards greater query efficiency, marking a significant leap in resource-conscious learning (Ji et al. 2024). Comprehensively, Thakur's article on understanding RLHF provides an accessible entry point to the concepts, highlighting its application in training models like InstructGPT and ChatGPT, underscoring the practical implications of this research (Thakur 2021). Collectively, these studies form a robust backbone of literature, driving forward the capabilities and applications of RLHF in contemporary AI systems.

3 Approach

3.1 Reinforcement Learning From AI Feedback (RLAIF)

Following the recent trend of Reinforcement Learning with Human Feedback (RLHF), several studies proposed using other pre-trained AI models to construct the reward function(Black et al. 2024, Lee et al. 2023), which is a technique referred to as Reinforcement Learning From AI Feedback. Our work follows under the broader scheme of RLAIF.

In our proposed structure, we use RL to finetune a vision-language transformer as in (Wang et al. 2022). For the RL algorithm, we will start with Proximal Policy Optimization (PPO) (Schulman et al. 2017) as in other Reinforcement Learning with Human Feedback (RLHF) methods(Ouyang et al. 2022). The reproduced image will not be pixel-wise similar to the original image, therefore a feature similarity (fareid 2023) is calculated as the reward of the PPO agent.

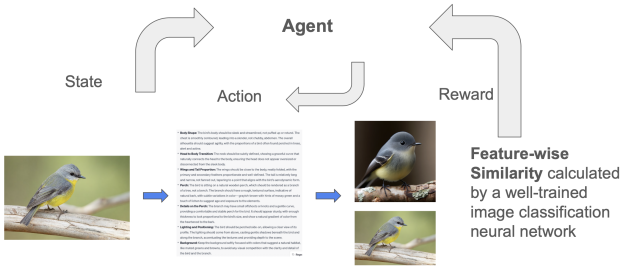


Figure 1: Proposed Structure.

3.2 Pre-trained Vision Language Model as Initial Policy

Under the reinforcement learning framework, the image-to-text model can be regarded as the policy. Since the generated text is regarded as the action, one notable challenge is the exploration over a large action space. To mitigate this issue, we avoid improving the policy starting from random initialization, and only use RL to finetune a pre-trained vision-language model. This technique has been shown to

provide a good enough initial policy (Ouyang et al. 2022), and RLHF only requires limited policy iterations to improve the performance of an LLM to surpass the supervised training LLM on multiple Evaluation Benchmarks.

4 Proximal Policy Optimization in Finetuning LLM

One challenge in finetuning LLM with RL is that training for too many iterations will result in the "reward hacking" problem, where the LLM as a policy will achieve higher reward but fails to generate text that sounds natural to human. Similar issues happens under the supervised learning setting. De-Diffusion (Wei et al. 2023) minimizes a loss function similar to the reward function in our proposal, and is able to generate text that can almost recover the original image. The generated text, however, does not follow any grammar rules or sounds natural to human.

Fine-tuning LLM with Proximal Policy Optimization (PPO) alleviates this issue to some degree. PPO makes conservative steps to update the policy by penalizing the KL divergence of the updated policy and the old policy from last iteration. When applied to Finetuning LLM, we can maintain the naturalness of the updated LLM's output by making sure the updated LLM is close to the pre-trained LLM for all iterations. This is achieved by always setting the old policy as the pre-trained LLM when calculating the KL divergence.

4.1 Diffusion and Image simialrity as Reward Function

Stable-Diffusion (Rombach et al. 2022) and Dall-E (Ramesh et al. 2021) are capable of mapping natural language features to image features and generate images that follow detailed text descriptions. In our proposed structure, we utilize such capability to evaluate the quality of our generated text. To calculate the reward, the generated image is compared with the original image by a similarity measure.

Image Similarity One significant part which directly determines the success of our whole work is image similarity computing. Obviously, we cannot compute the similarity per pixel, since the corresponding pixel may vary even if the two pictures are similar. However, there are several series of works discussed how to address this problem.

(DeepLobe 2021) discusses the fundamental use of the pre-trained CNN classifiers for image similarity and explains how to build a model using these. (Brejon 2021) offers a comparison of cosine similarity performances between VGG16 and ResNet50. These two work give a practical confirmation. Other works like (Appalaraju and Chaoji 2017) talked about deep learning, curriculum learning on similarity computing. There are also some practical implementations like (Core 2024) which gave us inspiration.

Figure 2 represents one of our framework, which obtains the reward by computing the similarity between the inputs and image generated according to the outputs (texts).

Framework Our end-to-end LLM Tuning starts by inputting an image into two identical LLMs. We fix the gradients of one to serve as a reference. The other LLM is used to produce text output. To evaluate the similarity between this text and the original input image, we first use a Diffusion model to generate a new image, and then calculate the similarity between this new image and the original output. We treat this similarity as a reward and input the KL divergence between the output distributions of the two LLMs as regularization, ensuring that the LLM after tuning does not significantly differ from the original.

Figure 3 is a demonstration that indicates how this framework works. For two texts describing the same input image, the images generated according to these texts will be feed into a similarity-retrieving network, where the more similar one will obtain more reward. Then the LLM could update the gradient according to this reward. In this case, the third image obtain higher reward, which means the text over it is more precise than the first one.

4.2 Image-Text Retrieval Score as Reward Function

Image-Text Retrieval In practice, we found that the image-text retrieval score can serve as a more reliable reward function. Recent work on text-to-image retrieval models based on vision-language

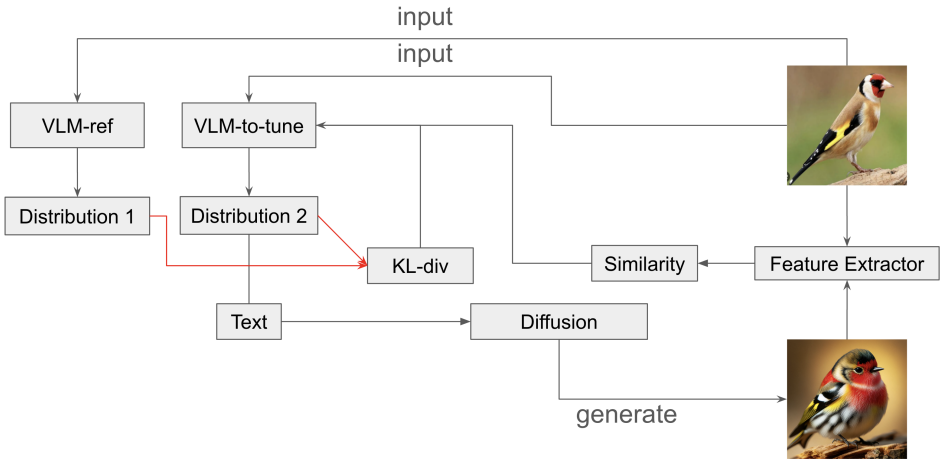


Figure 2: Framework with reward model solved by diffusion and image similarity

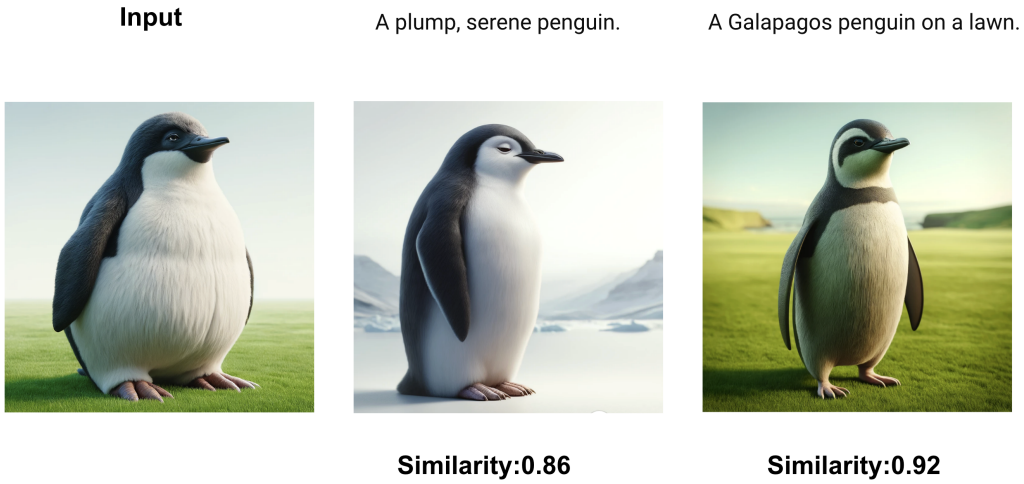


Figure 3: A demonstration indicating how different text-outputs effects reward.

models such as BLIP (Li et al. 2022) achieve the state-of-the-art performance on the image-to-text retrieval task and text-to-image retrieval task. The successful performance on these tasks requires building a common representation space for images and texts, and measuring the similarity of the image-text pairs(Zhang et al. 2023). Instead of using the similarity measure to retrieve the most relevant image from a collection of images, here we repurpose the similarity calculated by these image-text retrieval models as the reward function for finetuning our image captioning model (As shown in Figure 4).

5 Evaluation

5.1 Experimental Setup

For the pre-trained model, we are using a vision-language model composed of a vision transformer encoder and GPT-2 decoder that is pre-trained for the task of image captioning. The maximum length of output is 20. The temperature of model is set to 1. For the reward function, We use pre-trained Stable Diffusion as the text-to-image model and VGG pre-trained on Imagenet data for calculating the feature similarity of the two images.

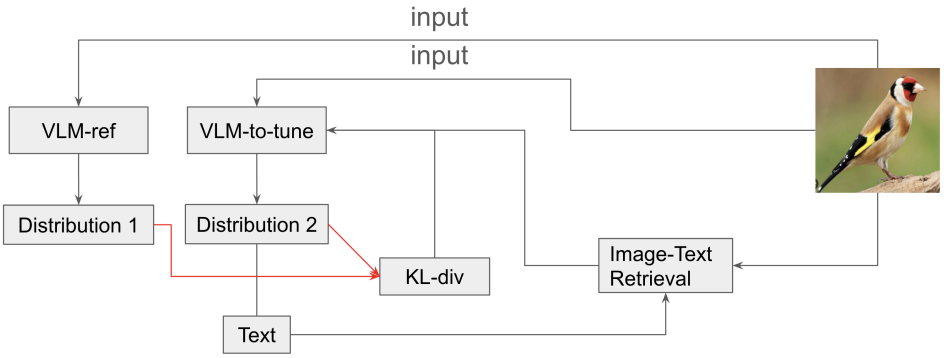


Figure 4: Framework with reward model solved by image-text retrieval model

For finetuning based on the diffusion image similarity as the reward function, because of the denoising process in training the stable-diffusion model, diffusion will output many different images for the same text input. Therefore, the estimated reward will have a large variance, which is detrimental to RL algorithms. To address this issue, for a given text as input, we run the stable-diffusion 5 times to obtain a monte-carlo estimation of the reward, which helps to stablize training.

The PPO for finetuning LLM is configured with batch size 10, steps 1, epochs 1, and learning rate 1e-6, which are all decreased substantially from the default configuration, which learns too fast and results in the generation of nonsensical text.

For finetuning based on the image-text retrieval model as the reward function, we set the coefficient of KL divergence penalty to 0.2 to keep training the model long enough for many iterations without generating non-sensical words, as the default coefficient of KL divergence penalty is too small. The learning rate is set to 3e-6.

6 Results

6.1 Training on Diffusion and Image Similarity as the Reward Function

The experiments on using the diffusion image similarity reward function show that the reward increases as training steps increase as shown in Figure(5). However, such improvement in reward could be a result of reward mis-specification. From the authors' qualitative evaluation, the updated model does not necessarily generate a better description. In many cases, the model mistakes the main object of the image (such as describing a cat image as dog), and still achieves a higher reward.

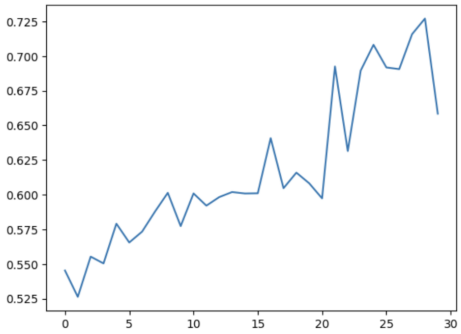


Figure 5: The Reward Plot of training on a single image using the diffusion image similarity reward function

6.2 Training on Image-Text retrieval score as the Reward Function

We randomly picked 1 image, 8 images and 128 images from COCO Caption Validation 2017 dataset. The score is the cosine similarity between the image features and the text features measured by the image-text retrieval model.

Table 1: Comparison of cosine feature similarity between pre-trained and RL fine-tuned ViT-GPT2 models across different numbers of images. Bold denotes the top performance.

Model	1 Image	8 Images	128 Images
Pre-trained ViT-GPT2	0.4310	0.6103	0.6210
RL Fine-Tuned ViT-GPT2	**0.4442**	**0.7862**	**0.6580**

The experiments on using the image-text retrieval score shows that we are able to achieve a higher reward when trained on 1 image, 8 images or 128 images together, where the reward is averaged across all images.

Figure 6: The reward plot of training to improve the caption of one image from the COCO caption validation dataset using the image-text retrieval reward function. The fine-tuned language model can generate a caption that has a higher reward than the initial pre-trained language model

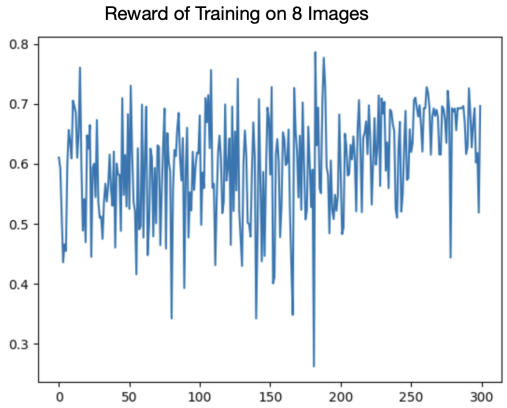


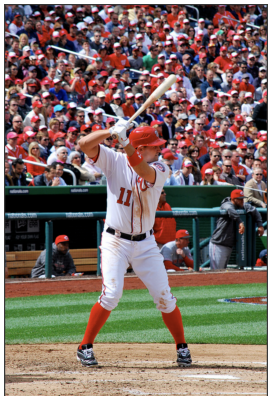
Figure 7: The plot of the average reward of fine-tuning the vision language model on 8 images using the image-text retrieval reward function. After 300 iterations, the policy achieves a higher reward, which means the generated captions and the image share more features than the pre-trained model.

Figure (8) shows the reward plot of training on a dataset of 128 images from COCO caption 2017 validation dataset. Initial observations from the plot suggested an absence of clear uprising trend. However, the model improves performance after 16 iterations. The performance measured by the

cosine similarity score is superior than the pre-trained model when evaluated collectively on the dataset of 128 images.

Figure 8: The plot of the average reward of fine-tuning the vision language model on 128 images using the image-text retrieval reward function. After 16 iterations, the model improved the average reward from 0.621 to 0.658.

Pre-trained ViT-GPT2: a living room with a couch and a table
RL finetuned ViT-GPT2: a living room with a couch a table and chairs and a table



Pre-trained ViT-GPT2: a baseball player holding a bat on a field
RL finetuned ViT-GPT2: a baseball player holding a bat stands near a crowd of spectators at a baseball field

7 Conclusion

In this work, we proposed using Reinforcement Learning to train an image-to-text model based on two designs of the reward function to achieve a better performance on the image captioning task. A pre-trained vision-language model is employed to mitigate the exploration of large action-space problem. The reward is designed using diffusion and image similarity or using the image-text retrieval score. We've shown that our method can improve the image-to-text model's performance on unseen images after a few iterations. Our result demonstrates that by finetuning a vision language model with RL, we can improve the model performance without human-annotated labels.

References

Appalaraju, S. and Chaoji, V. (2017). Image similarity using deep cnn and curriculum learning. *arXiv preprint arXiv:1709.08761*.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. (2024). Training diffusion models with reinforcement learning.

Brejon, L. (2021). Compute similarity between images using cnn. `https://github.com/lbrejon/Compute-similarity-between-images-using-CNN`.

Core, T. (2024). Transfer learning and fine-tuning. `https://www.tensorflow.org/tutorials/images/transfer_learning`.

DeepLobe (2021). Image similarity using deep cnn: Theory to code. `https://deeplobe.ai/blog/image-similarity-using-deep-cnn/`.

fareid (2023). Image similarity using cnn feature embeddings. *Medium*.

Hinton, G. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Ji, K., He, J., and Gu, Q. (2024). Reinforcement learning from human feedback with active queries. *arXiv preprint arXiv:...*

Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. (2023). A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., and Prakash, S. (2023). Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Li, Z., Yang, Z., and Wang, M. (2023). Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms.

Thakur, A. (2021). Understanding reinforcement learning from human feedback (rlhf): Part 1. WB.

Vincent, P. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. (2022). Git: A generative image-to-text transformer for vision and language.

Wei, C., Liu, C., Qiao, S., Zhang, Z., Yuille, A., and Yu, J. (2023). De-diffusion makes text a strong cross-modal interface.

Zhang, X., Niu, X., Fournier-Viger, P., and Dai, X. (2023). Image-text retrieval via preserving main semantics of vision.