Finetuning LLMs for Automatic Concept to TTI Prompt Generation (Student Abstract)

Jeremy Rutter, Maneesh Reddy Chamakura, Justin Delgado, Gene Louis Kim

University of South Florida, Department of Computer Science and Engineering
{jeremyrutter, maneeshreddyc, justind4,genekim}@usf.edu

Abstract

Our work explores bridging the gap between large language models and text-to-image models to create a tool for quickly and easily generating high quality images from a given concept. In our experiments we successfully improved image quality with only a preliminary utilization of the available resources for finetuning.

Introduction

Large language models (LLMs), such as GPT-3 (Brown et al. 2020) and LLaMA (Touvron et al. 2023), provide the ability to flexibly manipulate text and complete text-based tasks. Text-to-image (TTI) models, such as Stable Diffusion (Rombach et al. 2021), can generate high-quality images from text descriptions. We aim to bridge the capabilities of these two technologies to automatically generate TTI high-quality prompts for user-provided image subjects. In this paper we investigate whether LLMs finetuned on caption databases can produce higher quality images without any human intervention.

TTI systems are sensitive to subtle changes in the prompt—word choice, capitalization, and punctuation can all lead to dramatic changes in the image generation quality. This relates to how neural AI models as a whole are sensitive to how examples fit into their training distributions. As input features stray away from those of the training distribution, even in seemingly superficial ways (e.g., modifying a single pixel), neural model performance can drop dramatically (Athalye et al. 2018). This inspires the method used here. We hypothesize that finetuning an LLM on the same data used to build a TTI model will align its generations closer to the TTI model training distribution and lead to higher quality images.

The automatic high-quality image generator that results from this marriage to technologies has wide-ranging uses, e.g., producing images to accompany a short story, generating image assets for a game, etc. This reduces the barrier for new users of TTI systems by circumventing the need to learn effective prompting for generating high-quality images. By eliminating the need to prompt, this system can also directly take in results from NLP systems such as extracted named entities, event structures, and automatic summaries to generate high-quality visualizations of various aspects of source texts.

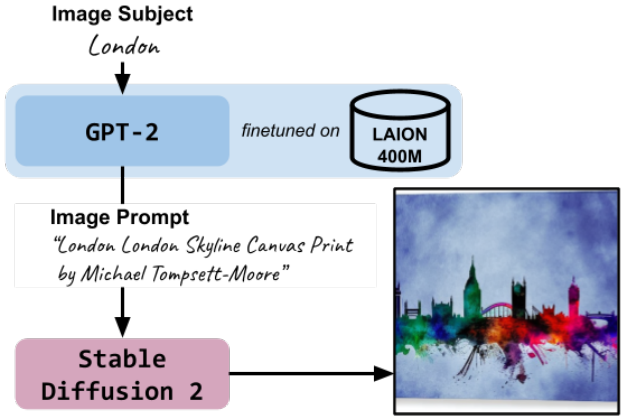

Figure 1: System diagram with an example from our system.

This leads to the overarching task of producing high-quality images from only a subject, where we cannot modify the TTI component of the system. There are two interrelated goals in this production task, which we form into our evaluation metrics: (1) the quality of the produced image and (2) the faithfulness of the produced image to the input subject. To an extent, these metrics are competing since high-quality images for certain subjects may be difficult to generate by TTI models.

Method

The task of processing text is done by a variety of current models which we considered such as BERT (Devlin et al. 2019) and GPT2 (Radford et al. 2019).

Although LLMs are not trained for any specific task, they can be specialized through finetuning.

By comparing existing text generation LLMs on the bases of complexity, accessibility for finetuning, and computational cost, we decided that GPT2 would be most beneficial. We used the 400m LAION dataset (Schuhmann et al. 2021) to finetune GPT2. This dataset is sufficiently large for finetuning, at 400 million caption-image pairs, and was used

to train Stable Diffusion. The LAION dataset was filtered to remove explicit images, however it was not filtered to remove explicit text. We performed explicit text filtering using a blacklist of 835 words in addition to basic preprocessing steps, e.g., removing empty strings.

We chose the largest model of GPT2, `gpt2-xl`, as it produced the highest quality prompts in our initial experiments. We froze the first 42 layers and trained the last 6. We found that freezing fewer layers would lead to incoherent generations and freezing more layers would not.

Analysis of the initial output revealed that the output resembled the middle of a caption rather than one complete caption. Thus, we finetuned the model on a modified version of the dataset that added starting and ending tags to each caption to signify that these captions are complete. Both models were trained on the first 450,000 captions in the dataset with 10% of the input being used for evaluation.

Experiments

To test our methodology, we conducted an experiment comparing output from our two finetuned models, along with the base GPT2 model as a control. Since our goal is to generate prompts using predefined subjects, we extracted candidates from the LAION dataset. We implemented a noun-phrase extractor and scored them according to frequency, word sense ambiguity, and length from 300,000 captions of the dataset separate from our finetuning data. This produced a list of high scoring noun-phrases that would be representative of concepts that frequently appear in captions. We manually filtered these results to retain only noun-phrases that are likely to be an image main subject. The 10 highest scoring concepts were used. For the caption generation length, we estimated the mean token length on a set of 300,000 captions using the GPT2 tokenizer. Using these parameters, we generated prompts from each of the 3 models. We used stable diffusion 2.0 as our TTI model to produce a set of images. For each model and each concept, we generated 5 prompts to account for variability for a total of 150 images.

We evaluated the models along two dimensions: image quality and faithfulness to the provided noun-phrase concept. The evaluation was organized into two surveys each containing 75 images, and responses were measured on a Likert scale (Likert 1985 - 1932) from 1 to 4 for these metrics. Each survey was taken by four different participants and participants were the authors and members of the same lab (2 undergraduate students, 3 graduate students, 1 professor). Two participants took both surveys. Participants were unaware of which model produced which image.

Results

The mean ratings of the survey questions are shown in Table 1. We find that finetuning gives a boost ($+0.15$ points) in image quality at the cost of image faithfulness (-0.30 points). Adding eos and bos tags leads to small boost in both image quality and faithfulness above basic finetuning. We also calculate the interannotator agreement (IAA) of our survey to validate the quality of our questions. We use a weighted Cohen's κ statistic as recommended by Jakobsson and Westergren (2005) for ordinal data. We get $\kappa = 0.386$ for image quality and $\kappa = 0.546$ for faithfulness. The relatively weak correlation for image quality suggests that our survey questions could be improved, or that a different method is required for definite ratings along this subjective dimension.

Model	No FT	FT No Tags	FT w/ Tags
Q	2.81	2.96	3.04
$\sigma(Q)$	0.97	0.95	0.87
F	3.02	2.72	2.77
$\sigma(F)$	0.98	1.21	1.11

Table 1: Mean ratings for image quality (Q) and faithfulness to the concept (F) for each model: the baseline GPT2 model (No FT), finetuning without bos/eos tags (FT No Tags) and finetuning with tags (FT w/ Tags).

The impact of finetuning on image quality and faithfulness is represented in our example in Figure 1, with the generated image sacrificing a degree of likeness to the original concept of 'London' to incorporate an artistic style. Note however that the model manages to generate an image of high quality without the need for a prompt created by an expert in this task.

Conclusion & Future Work

We described a method for finetuning an LLM for improving automatic TTI prompt generation and showed results from a small, initial experiment validating this approach. We found that this method improves image quality even with only 0.1% of the available captions. This comes at the cost of losing some concept faithfulness, pointing to necessary future work in maintaining image faithfulness before running this method on the full dataset. One approach that could be explored is explicitly tagging a main topic for captions in the dataset so the model generates prompts with a main topic aligning with the given concept. For concepts that produced prompts that were unfaithful, finetuning on a larger sample could reduce the degree to which the model overfits to specific stylistic patterns and co-occurring concepts that are not representative of the input concept.

Acknowledgments

We would like to thank the reviewers for their thoughtful feedback on this work. We would also like to thank the other members of our research lab—Mahammed Kamruzzaman and Abdullah Al Monsur—for their participation in the model evaluations.

References

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing Robust Adversarial Examples. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 284–293. PMLR.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell,

A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Jakobsson, U.; and Westergren, A. 2005. Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences*, 19(4): 427–431.

Likert, R. 1985 - 1932. *A technique for the measurement of attitudes / by Rensis Likert.* Archives of psychology ; no. 140. New York: [s.n.].

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *Proceedings of the 1st Data-Centric AI Workshop*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.