FUSING VISION AND LANGUAGE MODELS TO GENERATE SEQUENCE OF RECIPE IMAGES FROM STEPS

Anonymous authors
Paper under double-blind review

ABSTRACT

There has been a lot of work on using generative models for generating text descriptions given an image, demonstrating the power of pretrained large language models. There has also been several work on generating a sequence of text from a sequence of images, highlighting the effectiveness of fusing vision and language models to output text. In this work, we examine the effectiveness of fusing image and language models to generate a sequence of recipe images corresponding to the individual steps. We brainstorm different ways to fuse textual embeddings derived from each step to the encodings from the image, and empirically determine which is best. We also determine the relative importance of image and text encoders.

1 INTRODUCTION

State of the art vision and language models, particularly Stable Diffusion and GPT, have revolutionized text and image generation. There is an ongoing effort to combine these advances to create vision-language models (VLMs), a branch of multi-modal learning that uses the synergy between text and vision models to address tasks that involve an input of both text and image. These include image captioning, text-guided image generation and visual question-answering. However, one relatively unexplored direction is image sequence generation from a well defined sequence of steps. Such a model, if successful, is beneficial for generating recipe images from a sequence of steps, which can be a useful aid for those following such manuals. It can also be used to generating a picture book from a story. Our primary contributions include:

- Empirical study across different fusion methods when image and text encodings are separate.
- Empirical study across different image and text encoder modules to determine relative importance of text and image embeddings.
- Demonstrate that VLMs provide stronger qualitativve and quantitative results over baseline Stable Diffusion that ignores image conditioning
- Novel adaptation to loss term to ensure better style consistency across images (l2 loss btwn gen and prev).

2 METHOD

We brainstorm different ways to fuse the two embeddings. Our general strategy is depicted in the figure below.

Concatenation Here, the output of the text and image encoders are concatenated and passed to a feed-forward neural network to produce an output that is then reshaped to the desired size for stable diffusion conditioning.

Pooling Image and Text Embedding Instead of taking the last hidden state from the image encoder, we take the pooled output embedding, project it to match the size of the text embedding, and add the image encoding to each embedding in the text sequence. We then pass this through a feed-forward network and reshape the output. We call this strategy *Pooling Image Embeddings*. The reverse case, where we pool text embeddings, then project and add it to the image embeddings, is called *Pooling Text Embeddings*.

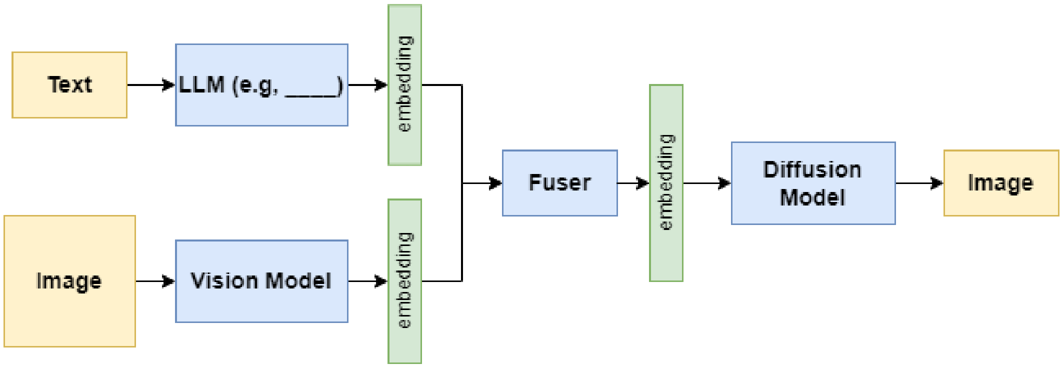

Figure 1: Our method optimizes the standard denoising objective to learn a set of embedding vectors while keeping the model parameters fixed.

Attend to Image and Attend to Text This utilizes the cross attention between image and text as the conditional embedding. In the case of attending to image, the queries are text embeddings, the keys and values are image embeddings. Vice versa for attending to text.

3 Experimental Results

We use the YouCookII dataset, prepared by the University of Michigan. In our experiments, we fine-tune the CompVis/stable-diffusion-v1-4 Stable Diffusion UNet and train the Fuser module. We keep the image and text encoders frozen. We run the stable diffusion fine-tuning for 100000 steps, and use a learning rate of 3e-07, for both UNet and Fuser. For sampling, we use a DDPM scheduler, a guidance scale of 7.5, and 50 iterative refinement steps. The table below compares the performance of the different fusers on Clip Comparison Score, PSNR and SSIM. Table 1 shows some qualitative results when using CLIP image and text encoders. Table 2 shows Clip Comparison score using the best fuser method found from Table 1.

Table 1: Test accuracy of Models with different Fuser Mechanisms trained on the VAE-Processed data.

Fuser	Clip Comparison Score ↓	PSNR	SSIM
Baseline (Stable Diffusion)	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Concatenation	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Text Pooling	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Image Pooling	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Attend to Image	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Attend to Text	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

Table 2: Clip Comparison Score of Image and Text Model Combinations usng xyz Fuser Method

Vision Model	Text Model					
	a	b	c	d	e	f
a	0.0052	0.0031	0.0070	0.0075	0.0092	0.0048
b	0.0061	0.0075	0.0071	0.0065	0.0062	0.0078
c	0.0078	0.0087	0.0057	0.0079	0.0094	0.0060
d	0.0040	0.0061	0.0084	0.0062	0.0060	0.0053
e	0.0075	0.0056	0.0056	0.0056	0.0066	0.0097
f	0.0054	0.0030	0.0088	0.0085	0.0056	0.0091
g	0.0077	0.0073	0.0059	0.0075	0.0049	0.0053
h	0.0062	0.0057	0.0099	0.0068	0.0087	0.0067
i	0.0061	0.0069	0.0101	0.0065	0.0080	0.0071

4 Conclusion

We present an empirical study of fusion methods for VLMs in the task of image sequence generation. We also propose a technical novelty in the training pipeline to ensure style consistency between the images. This empirical study lays the foundation for more advanced neural architectures and loss functions to improve performace on image sequence generation from text. We show the effectiveness of strong pretrained image and language models, and present a simple method for fusing that is superior to stable diffusion and other obvious fusing methods. We leave more advanced fusion methods to further work, and hope to inspire further research in image sequence generation. It is crucial, in fact, some may say it is vital, to understand that **this is a draft. It is missing numeric results (right now, conclusions are derived from qualitative results) and references.**

References

5 APPENDIX

You may include other additional sections here. However, please be mindful that the spirit of the Tiny Papers track is for papers to be short. Avoid overly-long appendices.