Bumblebee: Text-to-Image Generation with Transformers

Nathan Fotedar, Julia Wang

CS 224N Winter 2019

Motivation

The problem of generating images from natural language descriptions in an automated fashion is fundamental to a variety of important applications, ranging from computer assisted design to computer generated art. Moreover, it sheds light on the sort of work than can be done at the intersection of natural language processing and computer vision, and multimodal learning in general.

Due to previous success in applying transformers to this problem with the AttnGAN architecture, we aim to explore if the addition of transformers is beneficial for this task. We propose the use of variational autoencoders (VAEs) with transformers.

Previous Work

Current successful approaches to natural image generation have mainly revolved around deep GANs, with the most recent success within the field being AttnGAN, a network that uses a stacked generative and discriminative networks with attention at each level. However, AttnGAN is an extremely complex network, and also falls victim to the many pitfalls of training GANs. Because of the success that has been achieved with GANs, there has been much less research in using VAEs on their own, meaning newer ideas such as transformers have not been explored much in conjunction with VAE models.

Can be transformers be leveraged in the text to image generation task?

Data & Evaluation

We used the MSCOCO Dataset of image and caption pairs.

a pizza with tomatoes, onions and basil on a wooden table.
a pizza on a silver platter and some glasses and plates

Figure 1 | A image/caption pair from MSCOCO.

Evaluation Methods:

We used VAE loss as well as qualitative inspection of images

Methods

Model

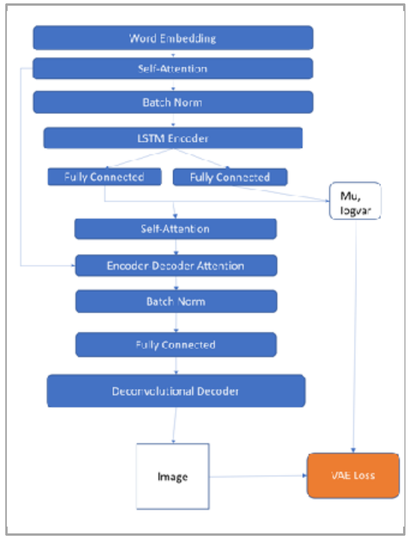
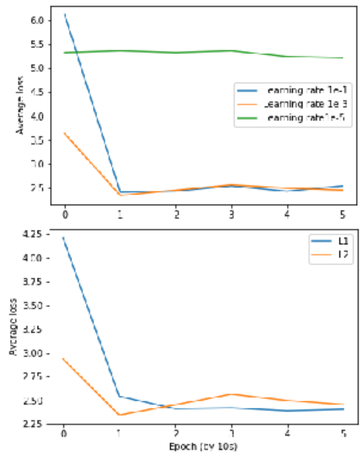


Figure 2 | Our transformers model architecture

Experiments

We tried both L1 and L2 loss and found that L1 performed better. We tried a number of learning rates, hidden sizes, and kernel sizes.



Results

Figure 3 | Images generated by baseline and model after overfit training

To make sure we had models capable of training, we ran overfit tests and were able to produce the images below. Below we have images produced by our baseline (left) and transformer (right) models.

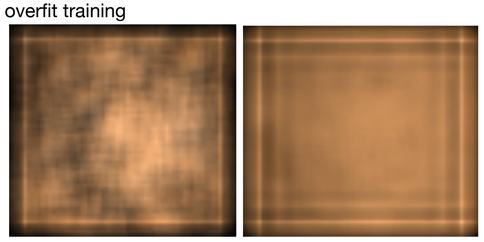


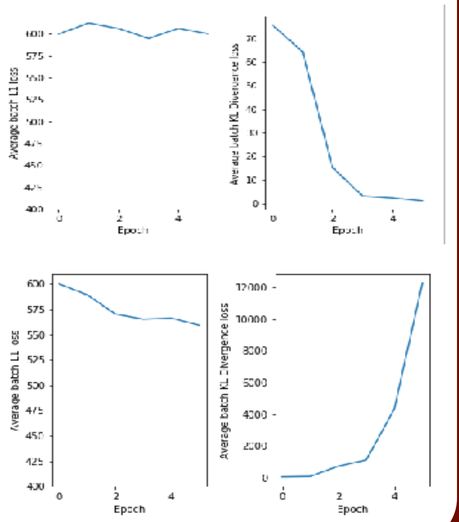
Figure 3 | Images generated by baseline and model

Unfortunately, our model was unable to produce recognizable images. The lack of GANs proved quite detrimental, and our model struggled to generalize. Below are images produced by our baseline (left) and transformer (right) models on the caption "There is a pizza on the table".

Analysis

We also were curious about the effect of the KL Divergence on training. We performed experiments of training with and without KL Divergence, while analyzing both the L1 loss and KL Divergence.

When training with KL Divergence, the model tries to optimize the KL Divergence without changing the loss at all.
When training without KLD the loss has ability to go down slightly, but the KLD explodes; suggesting the model learns some average blurred image.



Conclusions

Ultimately, the marginal improvements made over the baseline were not very significant; issues with the model persisted, and it struggled to generalize. It seems that the VAE loss prompted the model to continue to learn the average image in the dataset it was presented, resulting in largely blurry and nondescript images. While the use of transformers may still be valuable to this task, we have found that the use of GANs is much more critical, and seems to perform a key function in task.

References

[1] Xu, T., et al. (2017). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks.
[2] Vaswani, A., et al. (2017). Attention is all you need.
[3] Kingma, D.P. Welling, M. (2014). Auto-Encoding Variational Bayes.
[4] Karpathy, A. Fei Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions.