Bumblebee: Text-to-Image Generation with Transformers

Nathan A. Fotedar, Julia H. Wang
Department of Computer Science
Stanford University
Stanford, CA 94305
`nfotedar@ jwang22@`

Abstract

While image captioning and segmentation have been widely explored, the reverse problem of image generation from text captions remains a difficult task. The most successful attempts currently employ GANs; however, in this paper we explore a variational autoencoder model with transformers. The motivation for applying transformers to the task of text-to-image generation comes from recent success in applying attention to it (with the AttnGAN model cite. Many researchers have achieved improvements over attention based models by applying transformers, so it seemed that transformers could aid this task as well. The VAE transformer model was ultimately unable to recreate the performance of the more traditional GAN models, and our results suggest that the use of GANs might be key to the text-to-image generation task.

1 Introduction

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. Apart from applications more directly related to human tasks, the ability for a computer to generate art further allows us to explore how much creativity can be expressed a computer, and how this may speak to the general intelligence a computer may be able to posses. Moreover, it sheds light on the sort of work than can be done at the intersection of natural language processing and computer vision. When exploring a task in multi-modal learning such as this one, additional challenges of proper data encoding and proper evaluation metrics arise, but tackling such problems is fundamental if we hope to one day build computer systems which are able to properly mimic the functions of the human mind.

Current state-of-the-art models in text-to-image generation utilize Generative Adversarial Networks (GANs). In particular, recent advances have been made in applying to attention to GANs [5], and it has been shown that such models can map specific parts of natural language to certain regions in the images being generated. Yet, GANs certainly still have their limitations, such as being difficult to train and often creating rather noisy images. As a result, we aim to explore alternative means of generating images from natural language.

Another possible means of data generation that has been explored a bit less in the current literature is Variational Autoencoders (VAE)[3], and in this paper we aim to explore how much success we can achieve in the text-to-image generation task using VAEs. Moreover, there has recently been a great deal of success in adding transformers on top of attention models[4], and as a result, we aim to implement a VAE-based architecture that utilizes attention and transformers to generate images from natural language descriptions.

2 Related Work

Current successful approaches to natural image generation have mainly revolved around deep GANs, with the most recent success within the field being AttnGAN[5], a network that uses a stacked generative and discriminative networks with attention at each level [citation]. However, AttnGAN is an extremely complex network, and also falls victim to the many pitfalls of training GANs, such as modal collapse and tricky hyperparameter tuning.

In terms of variational autoencoders, some current methods have tried to leverage both GANs and VAE networks, such as the incorporation of VAE in the generator model in AttnGAN[5] and others[1]. 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 in conjunction with VAE models. Generally speaking, VAE models alone rely heavily on pixelwise reconstruction loss and therefore are prone to producing heavily blurred images; among other goals, we aim to explore if the addition of transformers can provide some relief on this front.

In recent research, many tasks using attention have received boosts from the addition of transformer models, and have also benefited from the fact that they provide more parallelizable computation. Transformers came to rise in the paper Attention is all you need[4], and they involve applying self-attention to each encoder and decoder, as well as encoder-decoder attention.

Our work is an effort to explore the use of transformers in this particular problem without dealing with the complexities and difficulties in the training of GANs. By using VAEs instead of GANs in our approach, we propose to answer this question in a simpler manner, and also continue to build on the general goal in the field of examining the differences and merits of both GANs and VAEs in solving this problem. We further aim to examine how the use of transformers does or does not benefit the performance of this model.

3 Approach

In our baseline approach, we used a simple variational autoencoder [citation needed] which we adapted for text-to-image generation. We utilized GloVe (Global Vectors for Word Representation) as pretrained word embeddings for our model, but allowed them to continue training as we ran our model. We use a bidirectional LSTM for our encoder, and fully connected layers to generate the mean and standard deviation that we fed into our VAE loss function (to be discussed in more detail further along in the paper). Finally, for our decoder layer, we utilized a fully connected layer followed by a deconvolutional layer. (Figure 1)

Our loss function was VAE loss [3], which was a combination of L1 reconstruction loss as well as KL Divergence. We also made use of the reparametrization technique, which is very often paired with VAE models. [3]

We pulled inspiration from a variety of sources, including the NMT model from our Assignment 4, as well as the general idea of projection text and images to the same semantic space, such as in [2]. For our proposed model, we built our ideas on top of our baseline architecture. We focused the bulk of our efforts on the addition of transformers, which composed of self attention for both the autoencoder and decoder, as well as the encoder-decoder attention. We also added batch normalization to both the encoder and the decoder to help with training, as was done in Attention is All You Need [4]. (Figure 2)

4 Experiments

4.1 Data

We used the 2014 MSCOCO dataset of images and their captions. We mainly trained on a subset of images containing pizza, but in our final testing included a more diverse set of images.

4.2 Evaluation

It is especially difficult to quantify evaluation for this task. Simple mean squared error between two images would not be able to account for differences in semantics. For example, a caption such as "there is pizza on a table" can have an infinite number of correct images – the table could be brown or white, the pizza could have mushrooms or pepperoni, it could be a whole pizza or a mere slice,

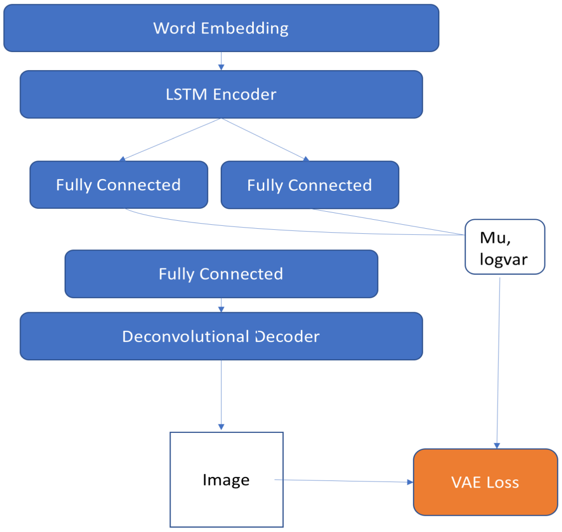


Figure 1: Baseline Model Architecture

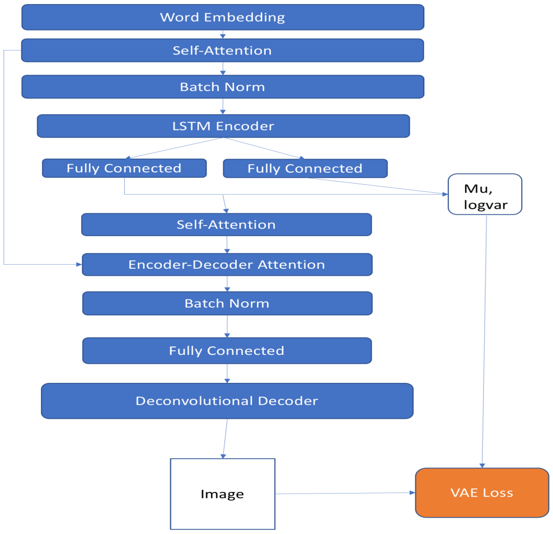
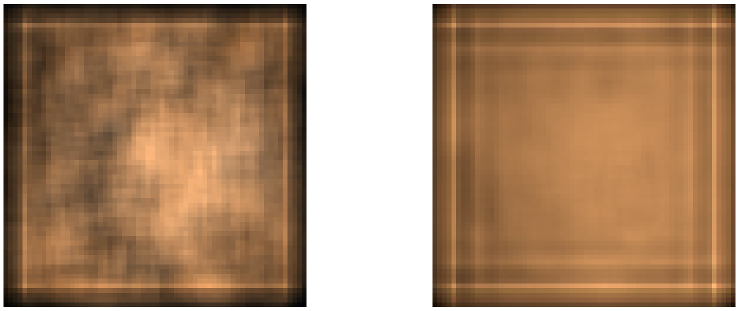


Figure 2: Proposed Model Architecture

and so on. Furthermore, most natural language descriptors focus on the centerpiece of the image and do not provide much information on the background, providing even more possible variance for what a correct image could look like. This is one of the reasons that GANs are useful in this task, as they provide an evaluation metric (inception score [citation needed]) that are not applicable for VAEs. In addition, other types of evaluation metrics (such as DAMSM) are highly involved and include training a separate network. Because of this, there is no viable quantitative metric for our task other than our VAE loss, and we instead relied mainly on qualitative analysis of the generated images. We performed the majority of our experiments by analyzing how much we were able to decrease loss and by noting anywhere loss was getting stuck, and then generated images to get a better idea of what our numbers meant.

4.3 Experiments and Results

Unfortunately, our model was unable to produce recognizable images. Additionally our proposed model had no improvement upon our baseline. In fact, the addition of additional complexity made it harder to train the model. The following images were generated provided the caption "There is pizza on the table". (Figure 3)
We performed a variety of experiments. For loss functions, tried both L1 and L2 loss and found that L1 performed better (Figure 4). For learning rates, we tried a number of options ranging from 0.1 to $1e-7$ (Figure 5). Furthermore, we tried with a variety of hidden sizes [25 to 512] and kernel sizes[3 to 15] (Figure 6). We trained each of our experiments for 60 epochs.



(a) Baseline Model

(b) Transformer Model

Figure 3: Model outputs for baseline and transformer models.

We can see that L1 loss performed better. If the learning rate was too small, the model loss would not go down, however all other learning rates would have the loss drop dramatically in the beginning and then plateau. Increasing hidden size helped marginally, but not enough to remedy our issues.

5 Analysis

We performed overfit tests to ensure that our model was trainable. On the overfit tests for both baseline and proposed models, we were able to reproduce an image (note that the blurriness is largely due to our compression of images). (Figure 7) We see that our models are trainable on one image to five captions pairs, but unable to generalize to a more diverse dataset.
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.

We see that when training with KL Divergence, the model simply tries to optimize the KL Divergence without changing the loss at all. However, when training without KL Divergence the loss has ability to go down slightly, but the KL Divergence explodes; suggesting the model learns some average blurred image.

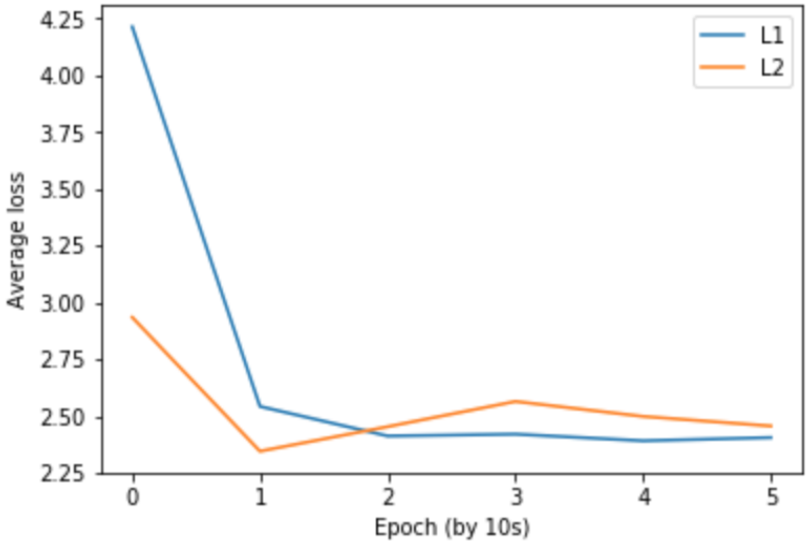


Figure 4: Average epoch loss for different types of loss functions.

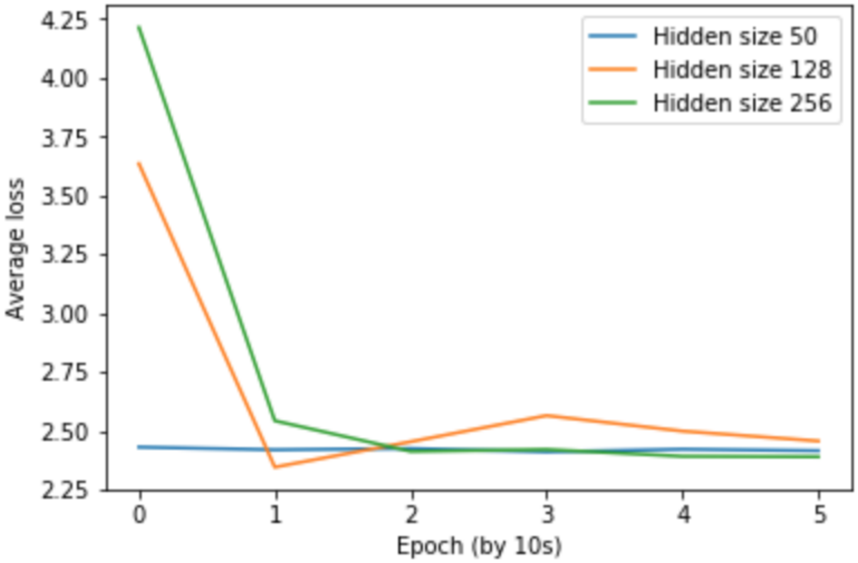


Figure 5: Average epoch loss for different hidden sizes.

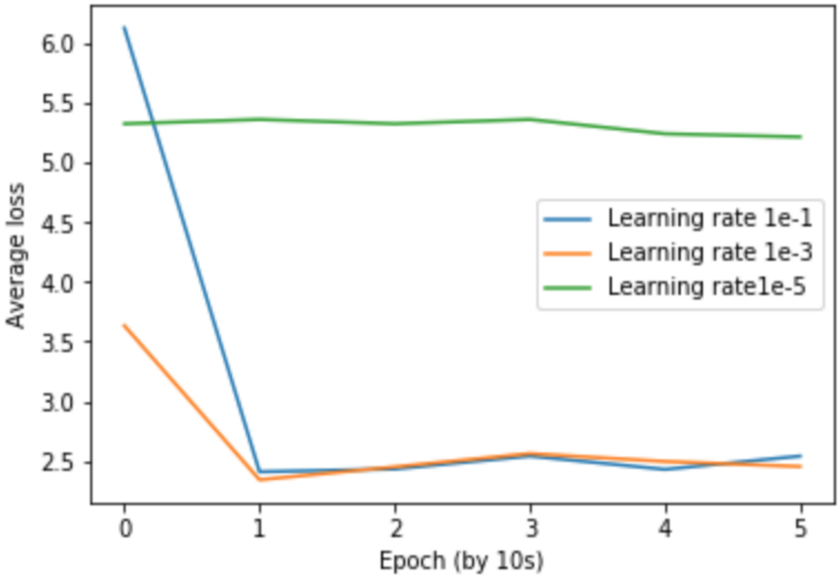
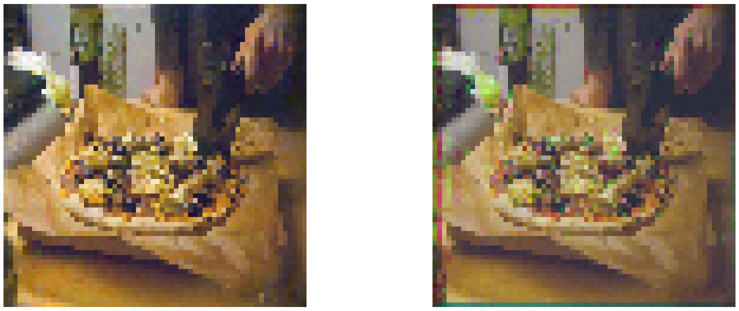


Figure 6: Average epoch loss for different learning rates.



(a) Baseline Model

(b) Transformer Model

Figure 7: Overfit tests for baseline and transformer models.

6 Conclusion

Ultimately, the marginal improvements made over the baseline were not very significant. Overall, 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 non-descript 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 the text-to-image generation problem. The accessibility of a meaningful evaluation metric as well as the generation of further training data gave GAN-centered models a significant boost that we could not achieve using purely VAEs. The lack of a helpful metric to train on did not provide as much information on the use of transformers as we had hoped; however, we suspect that if transformers were added on top of GANs, such as on top of AttnGAN, the model would benefit. This is certainly an area where future would could and should be done.

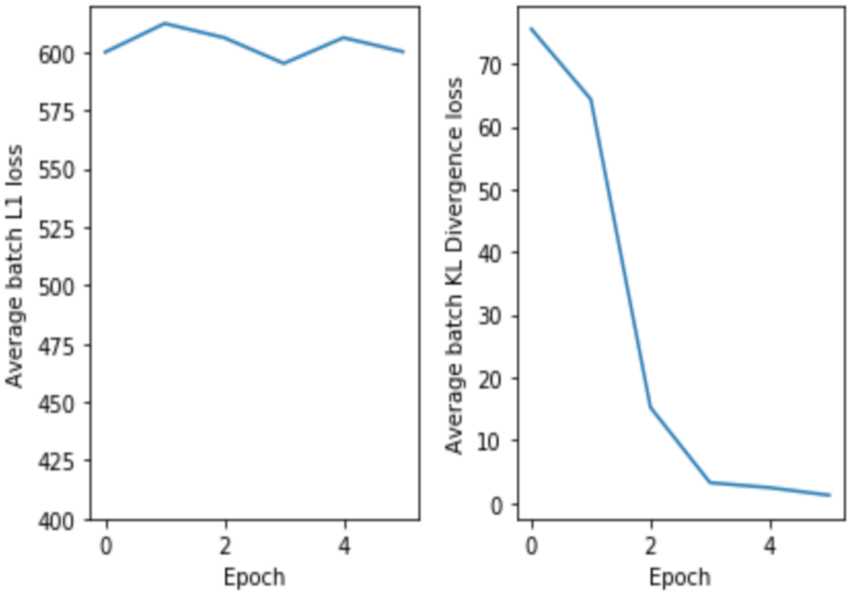


Figure 8: L1 and KL Divergence losses when training with KL Divergence

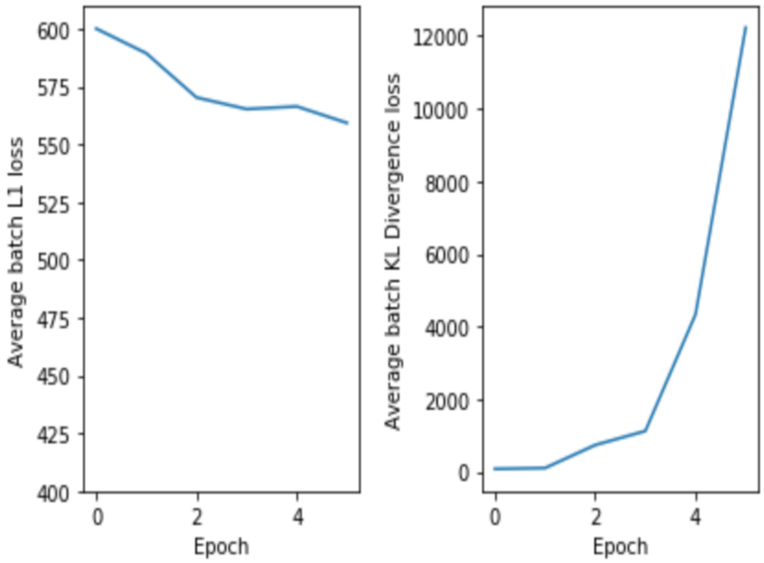


Figure 9: L1 and KL Divergence losses when training without KL Divergence

References

[1] Khan, S.H., et al. (2018) Adversarial Training of Variational Auto-encoders for High Fidelity Image Generation. *arxiv preprint* https://arxiv.org/pdf/1804.10323.pdf

[2] Karpathy, A. & Fei Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

[3] Kingma, D.P. & Welling, M. & (2014) Auto-Encoding Variational Bayes. https://arxiv.org/abs/1312.6114

[4] Vaswani, A., et al. (2017). Attention is all you need. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[5] Xu, T., et al. (2017). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. https://arxiv.org/pdf/1711.10485.pdf