Adversarial Text Generation

NLP and Deep Learning — Final Project

Andreas Holck Høeg-Petersen
anhh@itu.dk

Mathias Bastholm
mbas@itu.dk

August 13, 2020

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1 Introduction

In recent years, Generative Adversarial Networks (GANs) have gained a lot of traction in the Deep Learning community because of their impressive results in image generation. The general idea is that a generator and a discriminator are jointly trained to produce an image output that is seemingly indistinguishable from non-generated images. This model were first described in Goodfellow et al. 2014.

We want to attempt to apply this strategy for text generation. The main difficulty for this task is that whereas image outputs can be considered a continuous value, a sentence is inherently discrete as it is a sequence of words each of which is chosen by the model using the non-differentiable $argmax$ function. To remedy this, we propose a model where the discriminator is trained to distinguish between the continuous outputs of a pre-trained encoder given a 'true' sentence from the generated, 'fake' output stemming from our generator.

In our project, we will construct and train an autoencoder model that can encode and decode a sentence from English to English. The encoded sentences are then used as labelled training data for the discriminator, representing 'true' values. The job of the generator is to produce similar encodings but doing this from random noise in a way that makes the discriminator unable to distinguish between the encodings stemming from the autoencoder and the encodings stemming from the generator.

Ideally, this would train the generator to produce sentence encodings that can be fed to the decoder of the Transformer model which would then produce meaningful sentences from this artificially generated input. See Figure 1 for an overview of the complete model.

This project thus have two objectives: one is to construct a working autoencoder that can map an English sentence to some hidden state $\mathbf{X}$ with a corresponding decoder that can extract the original sentence from $\mathbf{X}$. For convenience, we will refer to the encoder part of this model as the 'Teacher'.

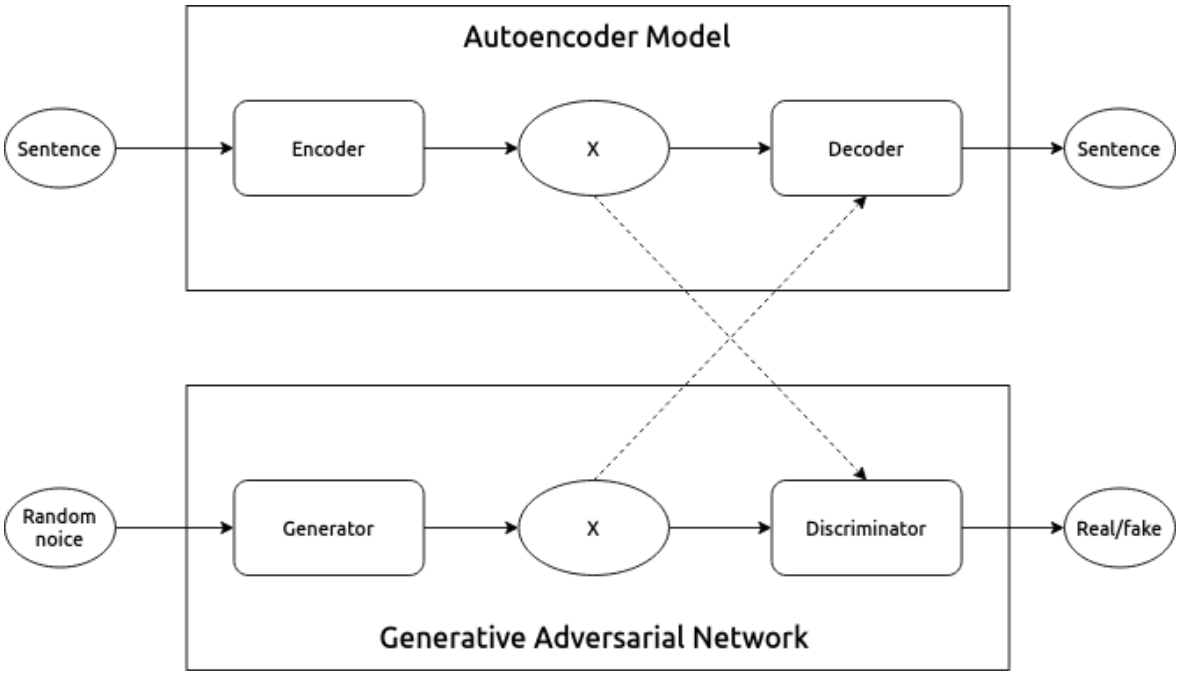

Figure 1: Overview of the model architecture. The dotted lines from the **X**s represents that the encoded and generated **X**s will be fed to the discriminator and the decoder during training and evaluation, respectively.

The second objective is to build a GAN network, where a generator — the 'Student' — must learn to produce approximations of **X**.

The second objective is highly experimental as explained in Section 2, where we will also describe other approaches at using the GAN architecture for NLP problems. In Section 3 we will describe how we have build the different parts of the model and how we utilize our dataset. Then in Section 4 we will present our results and discuss the shortcomings of the model*s*, and in Section 5 we will proceed to suggest improvements and ideas for further research. Lastly, in Section 6 we conclude on our project.

2 Background

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

3 Method

3.1 Baseline

3.2 Dataset

3.3 Models

3.4 Training

4 Analysis

4.1 Results

4.2 Discussion

5 Further research

6 Conclusion

References

Goodfellow, Ian J., Jean Pouget-Abadie, M. Mirza, B. Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio (2014). "Generative Adversarial Networks". In: *ArXiv* abs/1406.2661.