“Deep Faking” Political Twitter using Transfer Learning and Transformers

Ryan Ressemeyer (ryanress), Sam Masling (smasling), Madeline Liao (mmliao)
Stanford University, Department of Computer Science

Abstract

In the modern political climate, Twitter has become one of, if not the most impactful mediums for both current and aspiring politicians to communicate with their constituents.

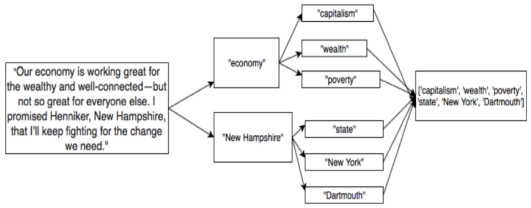
Our group attempted to replicate a variety of politicians’ twitter accounts using a variety of deep learning methods. We used the recently released GPT-2 model from OpenAI as a starting point for our text generation. Using transfer learning, we used GPT-2 to determine context, and trained the last few layers of our model on a large corpus of political tweets.

Data

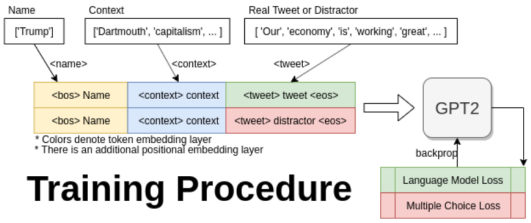
- ~4000 recent tweets from a variety of politician accounts
- Filtered out retweets and images
- Augmented each tweet with additional context via keyword extraction and similar word generation
- Total: 80,000 tweets, each labeled with its original account

Model (Transformer)

We used Rapid Automatic Keyword Extraction (RAKE) to extract keywords from each tweet and used GloVe vectors to generate similar words as context for each tweet..



We then used transfer learning to adapt OpenAI’s pretrained GPT-2 transformer-based model to generate tweets using a given context in the style of a specific twitter user.

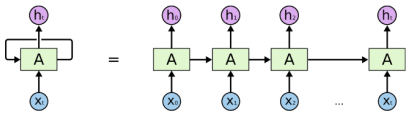


Training Procedure

Transformer models are similar to LSTMs in that they attempt to predict the next word given a list of prior words. However, they have dynamic “attention” and therefore take cues from the most relevant words in a sentence regardless of position.

Model (LSTM)

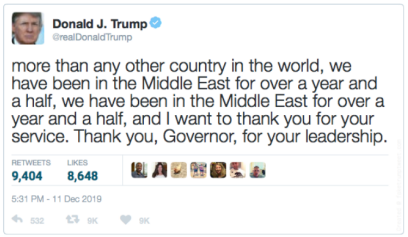
Used a LSTM RNN to repeatedly generate letters to form words based on previous context using hidden memory states.



$$\text{Loss} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

Results

LSTM generated example:



Transformer generated example:

Discussion

- Based on a human qualitative analysis, it’s clear that both the content and style of the GPT-2 generated tweet is more similar to a Trump tweet
- For further analysis, we plan to analyze readability scores of the generated tweets and compare
- May request more human analyses as well

Future Work

- Training on more accounts
- Better evaluation of quality of tweet-generation
- Creating a model that could also receive an input topic (e.g. “foreign policy” or “immigration”) and generate tweets from a specified account about that subject
- Fine-tuning hyperparameters

References

[1] Radford, A., Wu, J., Language Models are Unsupervised Multitask Learners. 2018.
[2] Rose, S., Engel, D., Cramer, N., Cowley, W. Automatic keyword extraction from individual documents. 2010.
[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.
[4] Wolf, T How to build a State-of-the-Art Conversational AI with Transfer Learning, Medium.com, 2019