Data Augmentation for NLP using Transformers

Manu Suryavansh

09/23/2021 - BayPiggies Meetup

NLP Challenges

- Lack of large labelled datasets
- Cost of labelling data is high
- Expertise in specific domain is needed for labelling

Augmentation Methods

- Back Translation
- Random word replacement
- Random word insertion
- Adding synthetic generated text to existing text

Example tasks

- **Hate Speech Classification -** Used Jigsaw-Wiki dataset for training binary identity_hate classifier with ~1000 samples
- **Consumer complaints dataset -** Multiclass classifier with less than 1000 samples for each of the five classes.
- **Sentiment Classifier -** Binary classifier using IMDB reviews dataset.

Links

- Github for Streamlit Demo - https://github.com/suryavanshi/nlp_augment_streamlit
- Blog Post - https://towardsdatascience.com/nlp-data-augmentation-using-transformers-89a44a993bab
- Streamlit Examples - https://streamlit.io/gallery
- Transformers - https://huggingface.co/transformers/

Demo and Questions