Project

**Generating text using LSTM GAN in PyTorch**

**Introduction**

Text generation is one of the common tasks in text mining and natural language processing (NLP). This project focuses on usage of generative adversarial networks (GANs) in the text generation. GANs are usually used for image data, in that case their components are CNNs, for working with text, RNN architecture should be used.

The second assumed model is a Markov chain model, a simple model that can perform satisfactory on a sufficiently big training dataset. Predictions of the model are then classified by the discriminator to directly measure its performance and thus indirectly evaluate the quality of the text generated by the GAN. Absolute quality of the generated text is also evaluated manually by a human using a labeling of fakes amongst the real comments.

Each training sample contains a flag, marking binary presence of negative sentiment in the text, such as racism or attack. Even though this information is not utilized in the models in any way, its possible usage as an additional input to the generator network is discussed and analyzed.

**Theory**

*Briefly present relevant theoretical background for your project, in particular machine learning models that were not covered in the course.*

In the statistical approach the generated text is prediction of a certain model, fitted to train data. Based on a kind and quality of the training set and the nature of the model we might

the generator and the discriminator, two networks reciprocally training each other,

The second assumed model for comparison is a Markov chain model building a *next-word* distribution for each word in training dataset.

RNN, LSTM

GANS

Embeddings

Sentiment

**Data**

The dataset was published on [figshare.com](https://mendeley.figshare.com/articles/dataset/Cyberbullying_datasets/12423407) as part of online *datathlon* [Data Sprint #13: Cyberbullying](https://dphi.tech/practice/challenge/42#data) held by *DPhi* community. The data collection consists of 8 csv files of overall size 171 MB, containing user comments from various social network platforms like Kaggle, Twitter, YouTube and Wikipedia Talk. The data contains text and binary flag marking the presence of cyberbullying or sentiment in terms of text mining. Cyberbullying aspects include behavior such as hate speech, aggression, insults and toxicity. (Elsafoury 2020)

The data files are downloaded and merged as a single format. The samples are deduplicated with text equality since the files have heavy overlap. All sentiment flags are unified as a single binary cyberbully-present flag. Then each text is tokenized and stored in the file data/words.txt. This data-engineering step takes about 45 minutes and is being done separately in src/fetch.py and executed out of the training pipeline.

File data/words.txt contains 196292 distinct comments and 3 attributes: text (stringified list of words), label (0/1 sentiment flag) and source (twitter/youtube/kaggle/unknown). An example line of the file is

|  |
| --- |
| "['able', 'to', 'do', 'better', 'than', 'that']",0,unknown |

The word is mapped to numeric value as an index of word2vec in gensim package. This makes every different word distinct including case-differences and a word is represented as a scalar, integer. On the other hand, the mapping removes all the semantic connections such as synonyms capturable by more complex mappings. The word2vec model is trained on the data, the vocabulary size is 307967.

(case-sensitive) a separate entity and ignoring all the s

The dataset comes from a datath is part of

Select the text dataset

The dataset – why? Description.

Fetching and processing.

Mapping to numeric vector.

Padding.

**Method**

Generator

Discriminator

GAN – executed on Google Colab

Contra-technique - HMM generator

(Liu 2020)(Mosquera 2018)

**Results**

Evaluation of the results

**Discussion**

*Most of work should be here. Present your analysis of the results that you obtained, discus the possibilities and limitations of your approach and compare your study to related work.*

**Conclusion**

*Built on the discussion – answer the question what new knowledge you take away from your project.*

**Resources**

Elsafoury, Fatma. 2020. „Cyberbullying datasets.“ 6. doi:10.17632/jf4pzyvnpj.1.

Liu, Kanghui. 2020. „Pytorch 20: Implement RNN(LSTM) Generate Random Text.“ 28. 6. https://www.bigrabbitdata.com/pytorch-20-implement-rnnlstm-generate-random-text/.

Mosquera, Diego Gomez. 2018. „GANs from Scratch 1: A deep introduction. With code in PyTorch and TensorFlow.“ 1. 2. https://medium.com/ai-society/gans-from-scratch-1-a-deep-introduction-with-code-in-pytorch-and-tensorflow-cb03cdcdba0f.