Data Augmentation using GPT-2

Statistics 98

Masaoud Haidar

Harvard University

May 2022

Outline

Imbalanced Data

- Imbalanced data can result in biased models. One solution to this is to over-sample from minority data points. Simple Bootstrapping can lead to over-fitting, so we want to create new synthetic data.

Text Data Augmentation Methods

- **EDA: Easy Data Augmentation**:
  Presented by Wei and Zou [2019]. Combines 4 methods:
  1. Synonym Replacement
  2. Random Insertion
  3. Random Swap
  4. Random Deletion

  Improves CNNs and RNNs. Works best on small datasets. On specific tasks, training with 50% of data with EDA performs as well as training with all the data.

Text Data Augmentation Methods

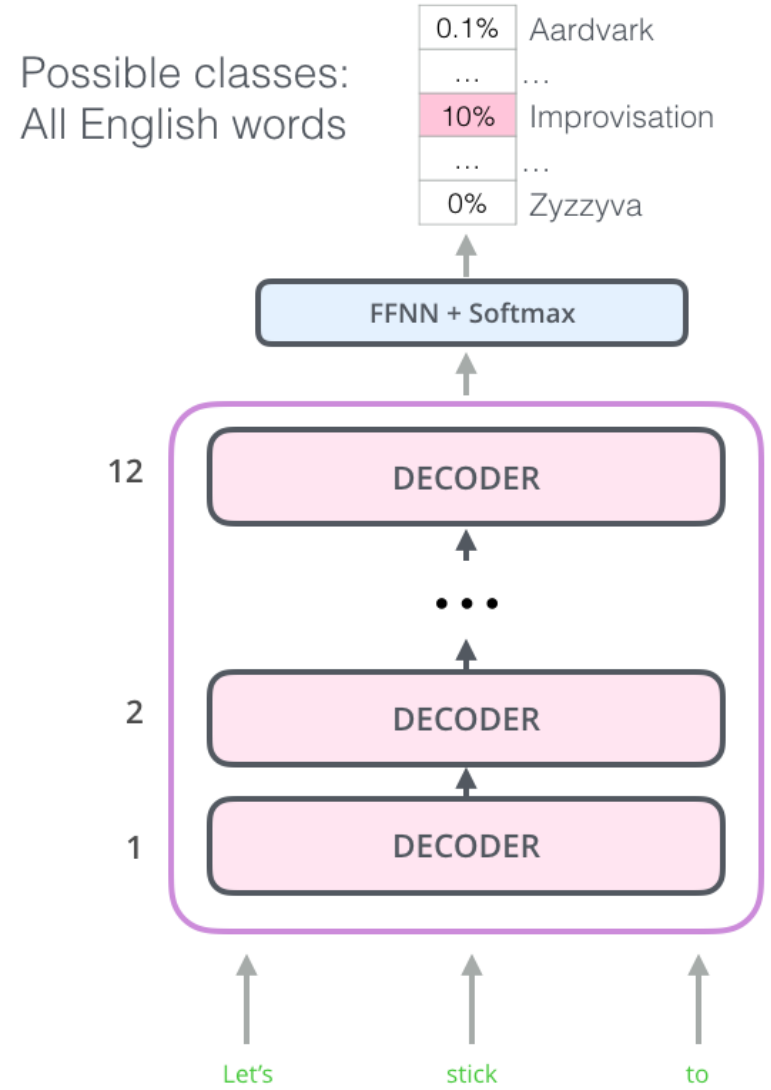
- **EDA: Easy Data Augmentation**:
  Presented by Wei and Zou [2019]. Combines 4 methods:
  1. Synonym Replacement
  2. Random Insertion
  3. Random Swap
  4. Random Deletion

  Improves CNNs and RNNs. Works best on small datasets. On specific tasks, training with 50% of data with EDA performs as well as training with all the data.

- **GPT-2**
  Presented by Radford et al. [2018] and [2019]. Can be used to generate new text data. Shown to improve F1 score on some tasks. Might not preserve the label When conditioned to generate data from multiple classes.

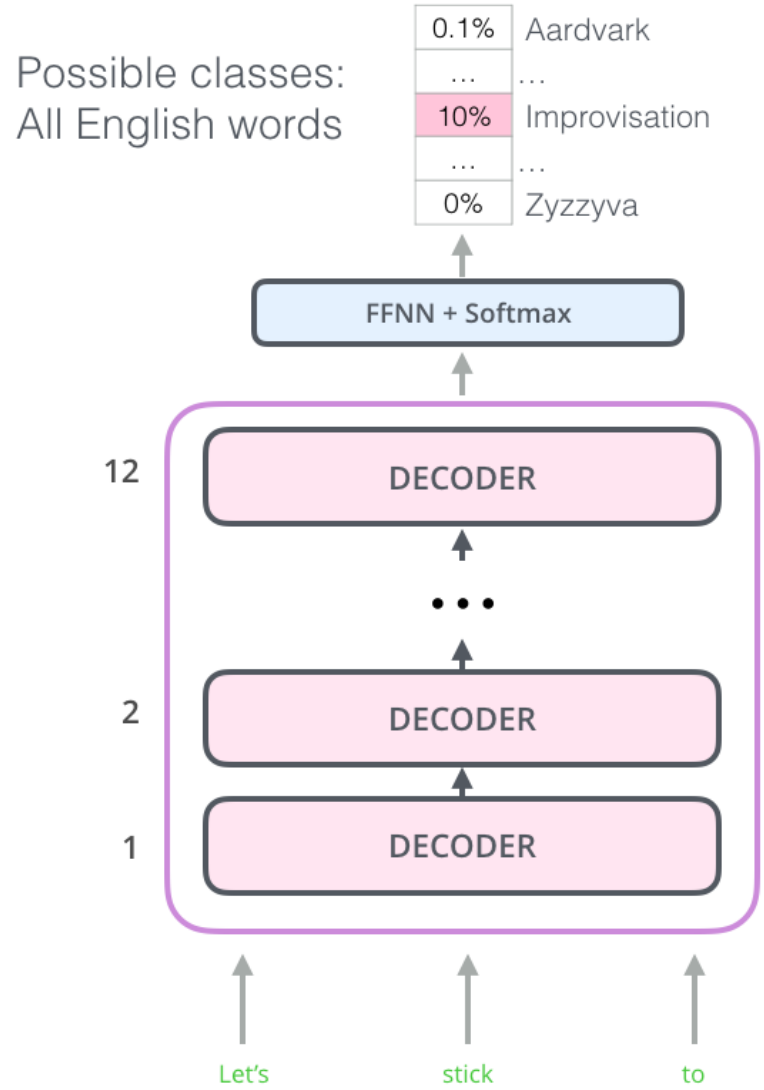
GPT-2 Structure

- Generative Pre-trained Transformer 2.
  Pre-trained on a large amount of data, and can be tuned for specific tasks.



GPT-2 Structure

- Generative Pre-trained Transformer 2. Pre-trained on a large amount of data, and can be tuned for specific tasks.
- Uses multiple layers of decoder-only transformers that use the self attention mechanism to transform the data. The output of the last layer is passed to a FFNN to predict the next word.



Research Question

- *How does Data Augmentation using GPT-2 affect the distribution of the data? And how does it compare with EDA when used with word-based models or context-based models?*

Experimental Setup

- Create 5 main Training Datasets:

  (A) Main training set of Amazon Reviews. 80:20 rate.

Experimental Setup

- Create 5 main Training Datasets:
  - (A) Main training set of Amazon Reviews. 80:20 rate.
  - (B) Artificially Imbalanced training set. 95:5 rate

Experimental Setup

- Create 5 main Training Datasets:
  - (A) Main training set of Amazon Reviews. 80:20 rate.
  - (B) Artificially Imbalanced training set. 95:5 rate
  - (C) Augment (B) using GPT-2 to 80:20 rate

Experimental Setup

- Create 5 main Training Datasets:
  - (A) Main training set of Amazon Reviews. 80:20 rate.
  - (B) Artificially Imbalanced training set. 95:5 rate
  - (C) Augment (B) using GPT-2 to 80:20 rate
  - (D) Augment (B) using GPT-2 to 90:10 rate

Experimental Setup

- Create 5 main Training Datasets:
  - (A) Main training set of Amazon Reviews. 80:20 rate.
  - (B) Artificially Imbalanced training set. 95:5 rate
  - (C) Augment (B) using GPT-2 to 80:20 rate
  - (D) Augment (B) using GPT-2 to 90:10 rate
  - (E) Augment (B) using EDA to 80:20 rate.

Experimental Setup

- Create 5 main Training Datasets:
  - (A) Main training set of Amazon Reviews. 80:20 rate.
  - (B) Artificially Imbalanced training set. 95:5 rate
  - (C) Augment (B) using GPT-2 to 80:20 rate
  - (D) Augment (B) using GPT-2 to 90:10 rate
  - (E) Augment (B) using EDA to 80:20 rate.
- We use Distilled GPT-2 for (C) and (D).
  We use EDA with $\alpha = 0.1$ for (E)

Experimental Setup

- Record the following:
  (1) **Word-based Model Performance:** Train TF-IDF plus Logistical Regression classifier. Record the accuracy and F1 scores.

Experimental Setup

- Record the following:
  (1) **Word-based Model Performance:** Train TF-IDF plus Logistical Regression classifier. Record the accuracy and F1 scores.

  (2) **Context-based Model Performance:** Train a BERT Classifier. Record the accuracy and F1 scores.
  We use tiny-bert with only two layers of encoder transformers.

Experimental Setup

- Record the following:
  (1) **Word-based Model Performance:** Train TF-IDF plus Logistical Regression classifier. Record the accuracy and F1 scores.

  (2) **Context-based Model Performance:** Train a BERT Classifier. Record the accuracy and F1 scores.
  We use tiny-bert with only two layers of encoder transformers.

  (3) **Distribution of the Data:** We compare the distribution of the data in (A), (D), and (E).

Experimental Setup

- Record the following:
  (1) **Word-based Model Performance:** Train TF-IDF plus Logistical Regression classifier. Record the accuracy and F1 scores.

  (2) **Context-based Model Performance:** Train a BERT Classifier. Record the accuracy and F1 scores.
  We use tiny-bert with only two layers of encoder transformers.

  (3) **Distribution of the Data:** We compare the distribution of the data in (A), (D), and (E).

  (4) **Preserving Labels:** Which is better at preserving labels, (D) or (E)?

Word-Based Model Performance

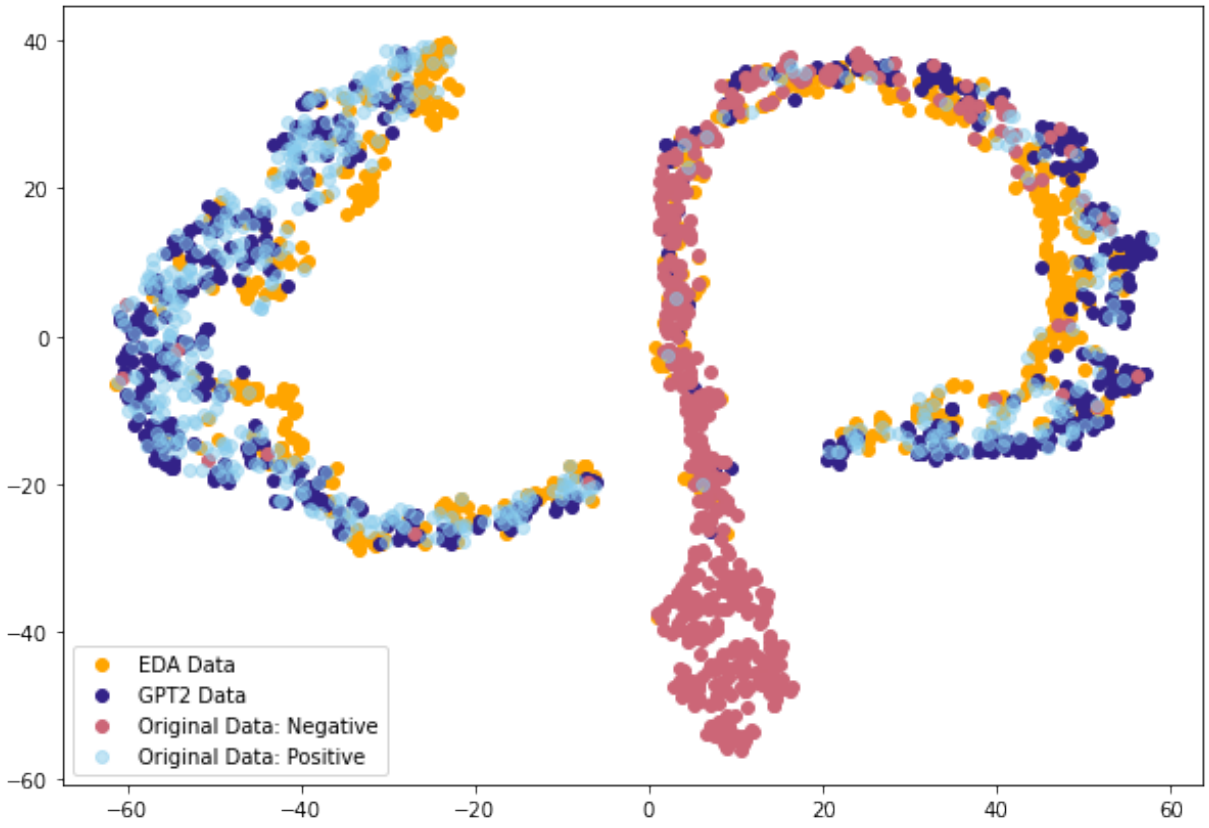
Model	Accuracy	F1 Score
A. Original	0.8799	0.662
B. Imbalanced	0.821	0.246
C. GPT2	0.8366	0.4209
E. EDA	0.8189	0.2715

Context-Based Model Performance

Model	Accuracy	F1 Score
A. Original	0.8886	0.7002
B. Imbalanced	0.8443	0.4198
C. GPT2	0.8403	0.4289
E. EDA	0.7931	0.0038

Distribution of the Data

- Train a Base BERT (the original large model) on (A). Record the outputs of the last layer of the model (before classification). Make TSNE plot of 500 samples from each group.



Preserving Labels

- Editing too much in a data point can change its meaning, and thus change its label. We want to measure that.

Preserving Labels

- Editing too much in a data point can change its meaning, and thus change its label. We want to measure that.
- Train a Base BERT (the original large model) on (A). Record the accuracy on the generated samples.

Preserving Labels

- Editing too much in a data point can change its meaning, and thus change its label. We want to measure that.
- Train a Base BERT (the original large model) on (A). Record the accuracy on the generated samples.
- The model achieves 90.9% accuracy on the positive data from its training set.
  Achieves 76.32% accuracy on the GPT-2 data.
  Achieves 64.16% accuracy on the EDA data.

Conclusions

- Models trained on GPT2-Augmented Datasets do better than EDA-Augmented datasets in all tested settings. GPT2 is also better at preserving the minority labels.

Conclusions

- Models trained on GPT2-Augmented Datasets do better than EDA-Augmented datasets in all tested settings. GPT2 is also better at preserving the minority labels.
- On TF-IDF with logistical regression, GPT-2 improves the F1 score by 18 points.

Conclusions

- Models trained on GPT2-Augmented Datasets do better than EDA-Augmented datasets in all tested settings. GPT2 is also better at preserving the minority labels.
- On TF-IDF with logistical regression, GPT-2 improves the F1 score by 18 points.
- With tiny-bert, it only improves F1 score by 0.9%. Might be because BERT was already not very prone to Imbalanced Data.

Conclusions

- Models trained on GPT2-Augmented Datasets do better than EDA-Augmented datasets in all tested settings. GPT2 is also better at preserving the minority labels.
- On TF-IDF with logistical regression, GPT-2 improves the F1 score by 18 points.
- With tiny-bert, it only improves F1 score by 0.9%. Might be because BERT was already not very prone to Imbalanced Data.
- Models trained on (C) and (E) don't come close to the models trained on original data. There is a lot of work to be done on imbalanced data.