**Improving Question Answering Model with Data Augmentation**

**Members:**

Seunghwan Hyun U68035800

**Brief Description:**

This project intends to create augmented data to improve domain generalization and robustness of models. This research aligns with the growing interest in generative AI, as discussed in our class, and responds to the critical need for diverse and extensive datasets. Using DistilBERT model for Question Answering task, I will demonstrate that data augmentation can make the QA model more robust and perform better to untrained data.

**Problem Statement:**

The field of AI often encounters the issue of insufficient data, which can impede the development of robust and generalizable models. This project seeks to mitigate this problem by generating additional data through various augmentation techniques and analyzing the improvement on Question Answering task especially on a model using DistilBERT which as proven to be effective in Question Answering mission.

**Data Analysis:**

To reduce the computing cost and time, 80,000 rows of SQuAD was loaded. SQuAD data has different question-answer pairs under ‘title’. For 80,000 rows, both ‘train’ and ‘validation’ data had 405 titles such as 'University\_of\_Notre\_Dame', 'Buddhism', 'Kanye\_West', 'Alfred\_North\_Whitehead', 'Portugal', 'Prime\_minister', etc.

I divided data into two groups so different data would not have related context and questions. Group1 had 1735 and 466 QA pairs in training and validation data about American entertainment (‘Kanye\_West', 'Beyoncé', 'American\_Idol', 'PlayStation\_3'). Group2 had 1329 and 366 QA pairs in training and validation data about Science ('Antibiotics', 'Genome', 'Solar\_energy', 'Brain', 'Mammal', 'Diarrhea', 'Incandescent\_light\_bulb', 'Apollo', 'Neptune', 'On\_the\_Origin\_of\_Species')

**Data Augmentation:**

Earlier research by Wei et al.(2019)[1] discovered that simple data augmentation strategies, including synonym replacement, random deletion, random swap, and random insertion, are effective in enhancing text classification tasks, particularly for smaller datasets. So QA datasets will be augmented to compare the model’s performance before and after.

In NLP research, especially when dealing with QA datasets like SQuAD, choosing where to apply data augmentation is crucial. Typically, augmenting the context is the best route because it adds variation without changing the answers, which need to be accurate for proper evaluation. Tweaking questions is risky; it can easily alter their meaning and make them mismatched with the answers. So, by changing the context, we're aiming to teach the model to understand different ways of saying the same thing, which is super useful for making it better at dealing with various texts it might come across later on.

When working on data augmentation in NLP, it's typically best to augment just the training data and not the test data. Augmenting the training data is a great way to expose your model to a wider variety of examples. This can really help in preventing the model from just memorizing the training data (which we call overfitting) and instead improve its ability to handle new, unseen data. If you start augmenting the test data, you're essentially changing the very thing you're trying to measure. It would be like changing the questions in an exam after teaching the students — it wouldn't give you a clear picture of what they've actually learned.

**Improvements:**

I evaluated DistilBERT model with different train data and validation data to gain insight on the effectiveness of simple data augmentation techniques. ‘Enter’ means entertainment data, ‘Sci’ means science data and ‘AugEnter’ means augmented entertainment data.

|  |  |  |  |
| --- | --- | --- | --- |
| Train Data | Validation Data | Exact Match Score | F1 Score |
| Enter + Sci | Enter + Sci | 36.213 | 47.523 |
|  |  |  |  |
| Enter | Sci | 17.003 | 24.311 |
| Enter + AugEnter | Sci | 33.723 | 47.269 |
|  |  |  |  |
| Enter + Sci | Sci | 33.718 | 46.004 |
| Enter + AugEnter  + Sci + AugSci | Sci | 39.349 | 52.594 |

First output is using same data to split training and validation data. This outperforms zero-shot models where training was done on entertainment data and validation was done on science data. However, training the model with augmented data has resulted in performance improvement in zero-shot learning model and non-zero-shot learning model.

**Significance:**

Enhancing data diversity and volume is crucial for improving model performance, particularly in areas where data collection is challenging or limited. This endeavor is vital for advancing the capabilities of NLP models in real-world applications

**Resources:**

* “Evaluation of Text Generation: A Survey” Celikyilmaz et al.(2021) (<https://arxiv.org/pdf/2006.14799.pdf>)
* “An Empirical Survey of Data Augmentation for Limited Data Learning in NLP” Chen et al.(2021) (<https://arxiv.org/pdf/2106.07499.pdf>)
* “A Survey of Data Augmentation Approaches for NLP” Feng et al.(2021) (https://arxiv.org/pdf/2105.03075.pdf)
* “Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering” Yang et al.(2019) (<https://arxiv.org/pdf/1904.06652.pdf>)

“EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks” Wei el al. (2019) (<https://arxiv.org/pdf/1901.11196.pdf>)

* + Simple augmentation techniques

**Evaluation Strategy:**

EM and F1 score.

**References**

[1] Jason W. Wei and Kai Zou. EDA: easy data augmentation techniques for boosting performance on text classification tasks. CoRR, abs/1901.11196, 2019.