Generating Robust Question Answering Model with Data Augmentation

Boston University CS505 Final Project

Seunghwan Hyun

shhyun@bu.edu

Abstract

This comprehensive study delves into the application of data augmentation to enhance domain generalization and robustness in Question Answering (QA) models, specifically focusing on DistilBERT. Amidst the burgeoning interest in generative AI, this research addresses the acute need for diversified datasets in Natural Language Processing (NLP). We demonstrated strategic data augmentation bolstered the QA model's resilience and adaptability to untrained data, a critical challenge in AI. Comparing the performance of model trained with augmented data, F1 score increased by 6% and for zero-shot model, F1 score increased by 23%.

1. Introduction

In the evolving landscape of Artificial Intelligence, the quest for robust and adaptable models is often hampered by the scarcity of comprehensive data. This study aims to address this limitation by employing a series of data augmentation techniques, thereby enriching the training data and enhancing the model's ability to generalize across varied domains. This investigation is particularly pertinent in the context of QA tasks, where the ability to interpret and respond to a wide array of queries is paramount.

2. Background

In the realm of NLP, the generalization of models across diverse domains remains a pivotal challenge, often due to the limited scope and variety of training data. This issue is particularly pronounced in QA tasks where models must understand and respond accurately to a vast array of queries. We aim to address this by exploring how data augmentation can artificially expand the diversity and volume of training data, potentially leading to more robust and adaptable models.

3. Methodology

This study employs a systematic approach to augmenting the Stanford Question Answering Dataset (SQuAD) and subsequently training and evaluating a DistilBERT-based QA model. The methodology is divided into several key phases:

3.1 Data Preparation

Informed by the seminal work of Wei et al. (2019)[1], this study embraces a multifaceted data augmentation approach to enhance the robustness of QA models trained on the Stanford Question Answering Dataset (SQuAD).

To ensure a comprehensive examination across disparate domains, the dataset was strategically segregated into two distinct groups from a curated subset of 80,000entries. This separation was intended to prevent any overlap in context and questions between the groups. The first group, focusing on American entertainment, comprised 1,735 and 466 QA pairs in the training and validation sets, respectively, featuring titles like 'Kanye_West', 'Beyoncé', and 'American_Idol'. The second group was centered around scientific topics, containing 1,329 and 366 QA pairs in the training and validation sets, respectively, with titles including 'Antibiotics', 'Genome', and 'Solar_energy'. This meticulous division allows for an in-depth analysis of the model's adaptability and performance across varied thematic areas.

3.2 Data Augmentation

Four augmentation techniques are applied to the 'context' portion of the training data:

- Synonym Replacement: Introduces lexical diversity, enabling the model to recognize various expressions of the same concept.
- Random Deletion: Simulates scenarios of missing information, preparing the model for incomplete or noisy inputs.
- Random Swap: Enhances the model's understanding of context and sentence structure by presenting rearranged word orders.
- Random Insertion: Adds unexpected elements to the context, promoting the model's resilience to unfamiliar or extraneous information.

Each technique is applied separately, and the augmented data is stored alongside the original data,

These techniques were exclusively applied to the 'context' segments of the training data. Altering the 'context' rather than the 'questions' or 'answers' preserves the integrity of the QA pairs while introducing sufficient variability to challenge and train the model effectively and increasing the size of the training dataset by multi-fold.

The augmentation process was meticulously designed to balance variety with coherence. A controlled application ensured that the augmented data remained relevant and meaningful, avoiding excessive distortion that could lead to counterproductive training. Moreover, augmentation was confined to the training data to maintain the purity and reliability of the validation data. This approach upholds the validity of the evaluation metrics by testing the model against unaltered, real-world-like scenarios.

3.3 DistilBERT

The DistilBERT model, a lighter version of BERT maintaining competitive performance, is fine-tuned on the augmented training data. DistilBERT is chosen for its efficiency and effectiveness in understanding and generating language. The training process involves adjusting the pre-trained model to better fit the QA task based on the SQuAD dataset.

4. Improvements and Observations

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

The DistilBERT model's performance was evaluated using distinct combinations of training and validation datasets to ascertain the impact of data augmentation on model efficacy. The datasets were labeled for clarity: 'Enter' denoting entertainment data, 'Sci' indicating science data, and 'AugEnter' representing augmented entertainment data. This nomenclature facilitated a structured and clear comparison of results, shedding light on the augmentation techniques' effectiveness in enhancing the model's capabilities.

The DistilBERT model's performance was rigorously evaluated across different datasets: original and augmented, segmented by entertainment and science domains. The evaluation focused on two primary metrics: Exact Match (EM) and F1 Score, reflecting the model's accuracy and its ability to generalize. Notably, the model trained on augmented datasets demonstrated significant improvements in both zero-shot and non-zero-shot learning scenarios. This indicates the augmented data's effectiveness in enhancing the model's comprehension and adaptability to unseen contexts.

5. Significance and Implications

The observed improvements underscore the critical role of data diversity and volume in NLP. By expanding the dataset artificially, the study contributes to overcoming the inherent limitations of data scarcity and homogeneity. This has profound implications, particularly in resource-constrained environments where collecting extensive and varied datasets is impractical. The enhanced model performance signifies a step forward in developing more robust, versatile, and reliable NLP systems capable of real-world applications.

In summary, this study provides compelling evidence that strategic data augmentation can substantially improve the performance and generalization of QA models. By judiciously expanding the training data and rigorously evaluating the outcomes, the research contributes valuable insights and methodologies to the NLP community, driving the field towards more sophisticated and practical solutions.

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.

[2] Yang et al. Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering. (2019) https://arxiv.org/pdf/1904.06652.pdf.

[3] Celikyilmaz, Asli, et al. "Evaluation of Text Generation: A Survey." arXiv, 2021, https://arxiv.org/pdf/2006.14799.pdf.

[4] Chen, Jason, et al. "An Empirical Survey of Data Augmentation for Limited Data Learning in NLP." arXiv, 2021, arxiv.org/pdf/2106.07499.pdf.

[5] Feng, Steven, et al. "A Survey of Data Augmentation Approaches for NLP." arXiv, 2021, arxiv.org/pdf/2105.03075.pdf.

[6] "Data Augmentation Library for Text." Towards Data Science, Towards Data Science Inc., Apr 20 2019. https://towardsdatascience.com/data-augmentation-library-for-text-9661736b13ff.

[7] "Hugging Face's Transformers: Question Answering." Google Colab, Hugging Face, colab.research.google.com/github/huggingface/notebooks/blob/main/examples/question_answering.ipynb#scrollTo=glBa-_rUaibO&uniqifier=1.