Text Augmentation Literature Report

Joshua S Raju | 22BAI1213 | Augmentation

Report Date: 28.02.24

2021, Survey : Text Data Augmentation for Deep Learning

Shorten, C., Khoshgoftaar, T.M. & Furht, B. Text Data Augmentation for Deep Learning. J Big Data 8, 101 (2021). https://doi.org/10.1186/s40537-021-00492-0
https://rdcu.be/dzR4b

ABSTRACT

- key motifs of Text Data Augmentation as Strengthening Decision Boundaries, Brute Force Training, Causality and Counterfactual Examples, and the distinction between Meaning versus Form

METHOD

- key motifs of Text Data Augmentation as Strengthening Decision Boundaries, Brute Force Training, Causality and Counterfactual Examples, and the distinction between Meaning versus Form
- This is in reference to small changes to the original data points. In NLP this could be deleting or adding words, synonym swaps, or well controlled paraphrases.
- What Deep Learning struggles with, as we will unpack in Generalization Testing with Data Augmentation, is extrapolating outside of data points provided during training. A potential solution to this is to brute force the data space with the training data. Te upper bound solution to many problems in Computer Science is to simply enumerate all candidate solutions. Tis way, even the most extreme edge cases will have been covered in the training set
- e. In this example, two people are stranded on separate islands, communicating through an underwater cable. Tis underwater cable is intercepted by an intelligent octopus who learns to mimic the speaking patterns of each person. Te octopus does this well enough that it can substitute for either person, as in the Turing test. However, when one of the stranded islanders encounters a bear and seeks advice, the octopus is unable to help. Tis is because the octopus has learned the form of their communication, but it has not learned the underlying meaning of the world in which their language describes.
- Types of Augmentation
- Symbolic Augmentation
- The drawbacks of symbolic augmentation include its limited ability to perform global transformations on longer inputs, which restricts its usefulness in information-heavy applications like question answering or summarization.
- Overall, while symbolic augmentation offers interpretability and works well with short transformations, it may not be suitable for all types of data or tasks that require global transformations or high accuracy.
- Rule Based
- Rule-based augmentations involve constructing rules to generate augmented examples by using ifelse programs and symbolic templates to insert and rearrange existing data

- four augmentations: random swapping, random deletion, random insertion, and random synonym replacement
- Opportunities for improvement in rule-based augmentations include considering word classification for semantic invariances and designing token vocabularies with structured patterns
- Graph Structured
- MixUp
- Feature Space
- Neural
- Back Translation
- Back-translation is an efficient text augmentation technique that involves translating text from one language to another and then back to the original language.
- Style
- Generative
- Label
- Among these: the top 3 are Symbolic, Rule Based, Back Translation

SHORTCOMING

many details of implementing Text Data Augmentation

- use of consistency training on a large amount of unlabeled data to constrain model predictions to be invariant to input noise. It is a technique used in semi-supervised learning to improve deep learning models when labeled data is scarce
- Consistency regularization Consistency regularization is a strong compliment to the priors introduced via Data Augmentation. A consistency loss requires a model to minimize the distance in representations of an instance and the augmented example derived from it.
- Rather than minimizing the distance between representations of original and augmented examples, the framework requires that the model outputs the exact same answer when predicting from context, question inputs as when a separate model generates the question from context, answer inputs
- Contrastive Learning
- However, a key diference is that contrastive learning generally uses other data points as the negatives, whereas Negative Data Augmentation entails applying aggressive augmentations.

Report Date: *11.03.24*

EDA: Easy Data Augmentation

Key Points:

- Paper proposes 4 simple random operations: insertion, deletion, swap and synonym replacement.
- Demonstrated strong results, especially on smaller datasets. Avg. accuracy w/o augmentation noted was 88.3% at using 100% of the training dataset. Avg. accuracy of 88.6% was noted for EDA models using 50% of the same dataset.
- Little is given about if EDA conserves the true labels. But according to the paper 'for the most part, sentences augmented with EDA conserved the labels of their original sentences.'
- However, on large and complete datasets, the performance gain is marginal (less than 1%).
- Doesn't use LM (language model) or ext. datasets
- Besides synonym replacement, other 3 EDA operations haven't been fully explored in depth yet. However, upon conducting tests, the paper has concluded that all 4 operations contribute significantly.

AEDA: An Easier Data Augmentation

Key Points:

- While EDA proposes 4 different random operations, AEDA proposes to use only random insertion of punctuation marks {".", ";", "?", ":", "!", ","}.
- Punctuation marks maintains order, but changes positions of the words; claimed to lead to better generalized performance.
- EDA's deletion operation leads to loss of info, while AEDA preserves all input info.
- No. of insertions = A number randomly chosen between 1 and 1/3rd of sequence length. This ensures that at least, or not too many punctuation marks are inserted which can lead to noise.
- The position where the mark is inserted is also decided randomly.
- While EDA wasn't boosting performance on large datasets, AEDA was consistent in boosting the performance in all various dataset size; outperforming EDA.

- An absolute improvement of 1.5-2.5% is noted for all dataset sizes by a single augmentation. Increasing the number of augmentations created a significant impact only on the smaller datasets, but remained stagnant for the full datasets (<1% gain).

Thoughts on EDA & AEDA:

Clearly the two augmentation techniques, EDA and AEDA are simple and in their own right effective. They are low-resourced yet efficient in delivering an improved performance for the models. In case we are interested in using a fast data augmentation technique, these could be our go-tos.

CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP

Key Points

-