Less is More

Data Augmentation Through Summarization for Sentiment Analysis Prediction

Shehzad Shahbuddin
School of Information, UC Berkeley
shehzad@berkeley.edu

Abstract

Data Augmentation is primarily used in NLP to supplement low-resource domains like machine translation. Surprisingly, there is limited research investigating the benefits of abstractive summarization supplementing datasets to assist in other NLP tasks. In this paper, I explore the usefulness of abstractive summarization to augment datasets to improve prediction in sentiment analysis. First, this paper outlines some areas where data augmentation is used in NLP. Next, I review the use of pre-trained models and the motivation for fine-tuning on Amazon video reviews. Finally, the effect of summarization when used to replace and supplement the IMDB reviews dataset is evaluated in the prediction of sentiment analysis.

1. Introduction

Data Augmentation (DA) is a methodology that provides additional training data to improve model performance without actively collecting more data manually. In image processing, transformation of images to provide a more robust training set is widely studied and cited (Perez and Wang. 2017)[1]. In NLP, data augmentation techniques supplement low-resource domains, new tasks, and the popularity of large-scale neural networks that require large amounts of training data (Parida and Motlicek, 2019)[2]. However, there is very little research on the use of summarization as an augmentation strategy and how it can be used to benefit other NLP tasks such as Sentiment Analysis (SA).

This paper aims to provide motivation for the continued exploration leveraging abstractive summarization to build upon SOTA models to improve model performance across NLP tasks, specifically sentiment analysis. The structure of this paper is to first provide background and information and related works on summarization, DA, and SA in Section 2. Section 3 details the methodology used, specifically outlining the summarization techniques, the motivation behind fine-tuning the pre-trained models, and the framework of the various experiments used in the SA task are discussed. In Section 4, the results of the experiments are presented and the findings are discussed. Finally, Section 5 provides the conclusion and suggests recommendations to build on the experiment presented in this research to validate the effectiveness of summarization as a valid data augmentation task.

2. Background

Summarization is an NLP technique that creates a concise synopsis of a given source text. There are two common types of summarization, extractive and abstractive. While extractive captures important spans from the source text and presents them as they were in the original, abstractive summarization identifies the core message of the source material and generates a novel summary (Mehta, 2016)[3]. Several pretrained models stand out in abstractive summarization including Pegasus (Zhang et al., 2019)[4], T5 and ProphetNet (Puspitaningrum 2021)[5].

The goal of DA is increasing the diversity of the training data without collecting additional data.

DA is effectively used in Image Classification tasks by cropping, rotating, and flipping images and by using GANs to create additional training examples[1]. In NLP, a plethora of DA techniques ranging from rule-based, interpolation, and modeling (Feng et al., 2021)[6] have been used. While the augmentation techniques have enhanced many NLP tasks including summarization, the application of summarization for SA has limited research (Yang et al., 2020)[7].

Text classification, also known as SA, is a classic problem heavily studied in NLP. Text classification predicts the sentiment of a source text as positive, negative or neutral. The primary methodologies utilized in SA are neural models that identify text representations, pretrained models trained on a large corpus of text and fine-tuning of pretrained models to increase the accuracy of the classification task (Sun et al., 2020)[8].

3. Methods

This section delves into the summarization task, the models used for text classification and the framework of the research approach combining the two tasks.

3.1 Summarization

The first summarization model is Pegasus (Pre-training with Extracted Gap-sentences for Abstractive Summarization). A standard encoder-decoder Transformer is the base architecture of the model (Figure 1). Unlike T5 and BART, Pegasus masks multiple whole sentences instead of smaller spans of text.

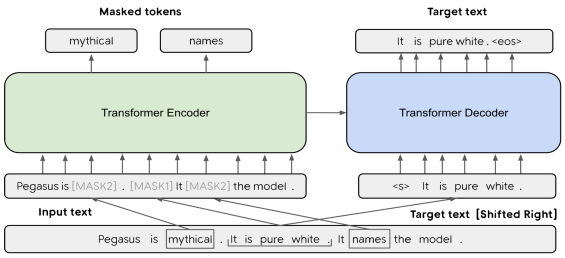

Figure 1: Base Pegasus architecture [4]

The Pegasus model used in the first summarization task was pretrained on the CNN/DailyMail dataset which is available on HuggingFace (Pegasus CNN). Since news articles and reviews are written very differently, I decided to fine-tune the pretrained Pegasus model on the Amazon US Video reviews dataset, which is also available on HugginFace. This data includes a full review and headline created by the author.

There were two fine-tuned models on the Amazon dataset, the first was trained for a single epoch (Pegasus Amazon) while the second was trained for 5 total epochs (Pegasus Amazon Fine Tuned).

The Amazon Videos dataset includes 301,232 reviews with fewer than 1024 characters (the cutoff chosen for the max length of the reviews). The dataset was split 70-30%, with 70% used as training and the remaining 30% equally split between validation and test. The models were trained and evaluated against the train and validation sets and evaluated against the test set. Each model's performance on the test set was evaluated by its ROUGE score on the summarization generated against the provided review headline.

3.2 Sentiment Analysis

The primary objective of this research was evaluating the benefit of summarization on sentiment analysis. Therefore, the model used for sentiment analysis was treated as the control and kept the same across each run of the experiment.

The model used was DistilBERT, a smaller, faster, and lighter version of BERT, while maintaining 95% of the performance as measured by the GLUE language understanding benchmark. The specific version of the model used was trained on the Stanford Sentiment Treebank (SST) dataset.

3.3 Experiment

Each Pegasus model generated summaries of the IMDB movie review train and test sets and were used in two forms during the experiment..

Before using the summaries, the DistilBERT model was trained on the full reviews of the training data and evaluated against the full reviews of the test data to set a baseline.

Next, the impact of each model's summaries were tested on the SA task individually by training their own DistilBert model. The first was trained on the summaries in isolation and the second used the summaries to augment the training data using a combination of the original reviews and the summaries. Two tests were run for the models, first the full reviews and then the summary of the test set. An outline of the methodology is captured in Figure 2, starting with training and creating the summaries, training the SA task, evaluating the results on the validation set, repeat for the next two summarization methodologies, and, finally, evaluate all the trained SA models on the test dataset.

As an additional experiment, the benefits of DA using summarization was tested for its effectiveness on an imbalanced dataset. To create an imbalanced dataset of the IMDB reviews, which is split 50/50 between negative and positive reviews, the training dataset was trimmed to the first 17,500 reviews making the split 71.4% to 28.6% in favor of negative reviews. The models tested were the baseline (no summaries), Pegasus Amazon Fine Tuned (summaries only), baseline with replacement on the positive reviews to resample, and the baseline with the Pegasus Amazon Fine Tuned summaries of only the positive reviews added. Both resampling and augmentation added 3,967 positive reviews making the training split 55.7% negative 44.3% positive.

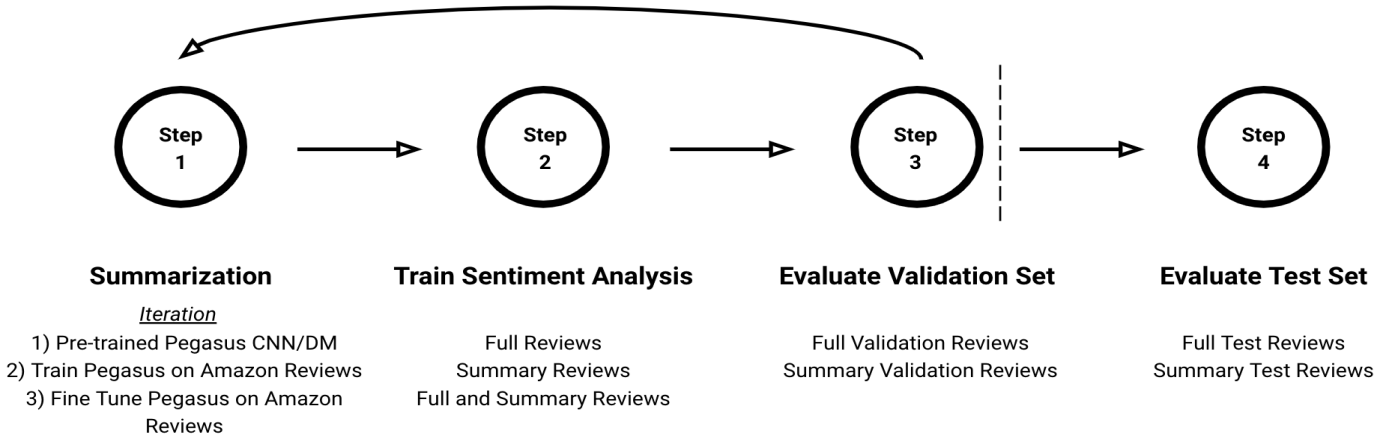

Figure 2: Sentiment Analysis Evaluation Methodology Workflow

4. Results and Discussion

4.1 Summarization

Training the Pegasus summarization task on the Amazon dataset did not drastically change the empirical results of the model performance.The fine tuning of the Pegasus model marginally improved the ROUGE-L score from 0.002983 to 0.005022 with reference to provided highlights. Additional metrics on the performance of the summarization task can be seen in Table 1. Although the Amazon reviews share a sense of positive or negative sentiment in their writing, the highlights are far shorter than what would generally be considered a summary of the

Review Body as exhibited by the sample summaries in Table 2. Therefore, the unimpressive performance of the ROUGE scores that were observed from the trained models is not surprising. Finding a sentiment based summarization dataset with longer reference highlights could be a worthwhile exercise to train the summarization model further.

Summarization	ROUGE-1	ROUGE-2	ROUGE-L
Pegasus CNN	0.002989	0.000083	0.002983
Pegasus Amazon	0.004957	0.000099	0.004926
Pegasus Amazon Fine Tuned	0.005035	0.000102	0.005022

Table 1: Pegasus Model ROUGE Score on Amazon Video Review Summarization Task

Original I bought this album around 1992 and consider it one of my best synth music albums (I have several of Tangerine Dream). The video was mostly shot in Europe and shows a number of types of architectures (including the Egyptian pyramids, French monistaries, german war fortifications, the Eifel Tower, etc.) Most of the film is time-lapse which is interesting. Michael Stearns composed the music and I consider this his best album. I have bought many of his other albums because I was looking for the same style or quality. The soundtrack to Chronos is excellent. The very end of the soundtrack (with riverboat in Paris) is similar to a Philip Glass recording.
Reference Highlight a masterpiece of video and music
Pegasus CNN Model Summary I bought this album around 1992 and consider it one of my best synth music albums .<n>The soundtrack to Chronos is excellent. The very end of the soundtrack is similar to a Philip Glass recording
Pegasus Trained on Amazon Summary The soundtrack to Chronos is excellent. The very end of the soundtrack (with riverboat in Paris) is similar to a Philip Glass recording. The video was mostly shot in Europe and shows a number of types of architectures (including the Egyptian pyramids, French monistaries, german war fortifications, the Eifel Tower, etc.)
Pegasus Trained on Amazon Summary The soundtrack to Chronos is excellent. The very end of the soundtrack (with riverboat in Paris) is similar to a Philip Glass recording. The video was mostly shot in Europe and shows a number of types of architectures (including the Egyptian pyramids, French monistaries, german war fortifications, the Eifel Tower, etc.) most of the film is time-lapse which is interesting. Most of the film is time-lapse which is interesting.

Table 2: Example Pegasus Summaries from Training Exercise

4.2 Sentiment Analysis

4.2.1 Balanced Dataset

The sentiment analysis tasks were evaluated on their prediction accuracy. The results of the different models and combinations of datasets are highlighted in Table 3. The baseline DistilBERT model on the original dataset outperformed all other models in its predictive accuracy. The additional training data was not able to improve the accuracy of the text classification model, thereby failing to reject the null hypothesis. However, by looking at only models trained by summary reviews, there is an improvement in accuracy from the pretrained model to the finetuned model across all test datasets except for on the evaluation on the test summaries alone.

Taking a closer look at the IMDB review summaries (Table 4), one can clearly observe the distinction between the output from the pretrained, partially trained, and fine tuned models. The pretrained model does a good job summarizing, but ultimately fails to clearly capture sentiment. The partially trained model appears to capture sentiment; however, the repetition within the summary fails to provide additional encodings that the model could learn from. The fine tuned model, the best of both worlds, captures the sentiment of the review and creates a (mostly) coherent summary. The improved summaries and the promising accuracy improvement observed in the model performance indicates that a model augmented with summarization data could potentially outperform baseline results given higher quality summaries. This researcher attempted to further fine tune the summarization model, but was faced with several insurmountable challenges during the summarization task retraining that prohibited additional exploration in this area. Primarily, attaining the necessary compute resources to train a summarization model beyond the 5 epochs was not possible in the Google Colab Pro environment as memory availability issues were continually encountered.

Summary Model	Training Data	Test Data	Accuracy
N/A (Baseline)	Full Reviews	Full Reviews	92.47%
		Pegasus Summaries	83.36%
		Pegasus Amazon Summaries	83.74%
		Pegasus Amazon Fine Tuned Summaries	84.24%
Pegasus CNN	Summary Reviews	Full Reviews	84.00%
		Pegasus Summaries	77.04%
	Full and Summary Reviews	Full Reviews	90.59%
		Pegasus Summaries	80.87%
Pegasus Amazon	Summary Reviews	Full Reviews	83.20%
		Pegasus Amazon Summaries	80.69%
	Full and Summary Reviews	Full Reviews	89.52%
		Pegasus Amazon Summaries	82.46%
Pegasus Amazon Fine Tuned	Summary Reviews	Full Reviews	83.61%
		Pegasus Amazon Fine Tuned Summaries	85.62%
	Full and Summary Reviews	Full Reviews	91.60%
		Pegasus Amazon Fine Tuned Summaries	87.00%

Table 3: Summary of SA results using DistilBERT

Original
"Holy crap. This was the worst film I have seen in a long time. All the performances are fine, but there is no plot. Really! No plot! A bunch of clowns talk about this and that and that's your film. Ug... Robert Duvall's character is senile and keeps asking the same people the same qestions over and over. This earns him the same responses over and over. I am pretty sure this film got upto a six because people think they should like it. Good performances with famous and well regarded actors, but the actual complete work is a steamy turd. Well, maybe that's a bit deceptive since steam rising from a fresh pile sounds a little like something happening and in this film NOTHING HAPPENS! Sack"
Pegasus CNN
The film is a remake of the 1970s classic, starring Robert Duvall as an ageing Hollywood star who moves to rural New Mexico with his wife (Susan Sarandon) and their two children.
Pegasus Amazon
The worst film I've seen in a long time. Horrible!
Pegasus Amazon Fine Tuned
This is the worst film I have seen in a long time. I am pretty sure this film got upto a six because people think they should like it. Well, maybe that's a bit deceptive since steam rising from a fresh pile sounds

Table 4: Generated Summaries from Pegasus on IMDB Reviews

4.2.2 Imbalanced Dataset

After the DA failed to show an improvement in the predictive accuracy of a sentiment analysis on a balanced dataset, this research aimed to determine if the same methodology could benefit an imbalanced dataset. This was motivated as data augmentation has been successfully used in other domains with SMOTE and other oversampling techniques improving a model's performance on imbalanced datasets.

Based on the results of the initial experiment, only the fine tuned model and the baseline were considered. Again, the baseline managed to outperform other models even with the imbalance with an accuracy of nearly 90%. This research experiment on an imbalanced dataset was surprising, but for the wrong reasons. The use of summaries heavily underperformed both the baseline and simple resampling in prediction. Looking at the two models which incorporated summaries, the full dataset with all summaries added, which still maintained the same label imbalance, did better than using only the positive summaries to balance the dataset. The accuracy of each model is seen in Table 5. Unlike the prior experiments on the balanced dataset, there is little promise shown from the use of summaries to augment imbalanced datasets.

Summary Model	Training Data	Test Data	Accuracy
N/A (no summaries)	Imbalanced Full Reviews	Reviews	89.64%
		Pegasus Amazon Fine Tuned Summaries	85.26%
Pegasus Amazon Fine Tuned	Imbalanced Summary Reviews	Full Reviews	66.76%
		Pegasus Amazon Fine Tuned Summaries	80.18%

Pegasus Amazon Fine Tuned	Full and Summary Reviews Together	Full Reviews	82.42%
		Pegasus Amazon Fine Tuned Summaries	81.79%
N/A (no summaries)	Full Reviews with Resampled Positive Reviews	Full Reviews	89.23%
		Pegasus Amazon Fine Tuned Summaries	82.40%
Pegasus Amazon Fine Tuned	Full Reviews Augmented with Positive Summaries	Full Reviews	77.98%
		Pegasus Amazon Fine Tuned Summaries	55.50%

Table 5: Summary of DistilBERT SA results from DA on imbalanced dataset

5. Conclusion

This work proposed that summarization can be used as a means of data augmentation, providing additional training data to an NLP model attempting to improve its performance at a given task. Specifically, this paper tested how using summaries of movie reviews in conjunction with the original training data can affect the performance of a text classification model. Beyond the presence of additional data, this paper highlighted that the quality of the summaries impacts the effectiveness of its ability to augment the data. While evaluating ROUGE scores of a summary on a similar dataset provides one method to gauge a summary, explicitly looking at the output of the summarization model and validating the quality with a spot check of the proposed DA can point out shortcomings that may negatively affect a model's predictive ability. Although this research failed to produce an evaluation metric on a model that used summarization for data augmentation on par or better than the baseline of the SA task on the full dataset alone, the improvement in the SA task between runs of fine tuning show promising results indicating further summarization improvements they can be used as an effective means of DA; which this research could not accomplish in this experiment due to insufficient compute resources to adequately perform the additional finetuning on the summarization training. While this paper primarily focused on evaluating Pegasus as an abstractive summarization model and DistilBERT as the SA model, there is room to investigate how other NLP tasks and SOTA models are impacted by incorporating summaries of the training data using Pegasus or other summarization models.

References

[1] Luis Perez and Jason Wang. 2017. The Effectiveness of Data Augmentation in Image Classification using Deep Learning

[2] Shantipriya Parida and Petr Motlicek. 2019. Abstract Text Summarization: A Low Resource Challenge

[3] Parth Mehta. 2016. From Extractive to Abstractive Summarization: A Journey

[4] Zhang, Zhao, Saleh, Liu. 2019. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization

[5] Diyah Puspitaningrum. 2021. A Survey of Recent Abstract Summarization Techniques

[6] Feng, Gangal, Wei, Chandar, Vosoughi, Mitamura, Hovy. 2021. A Survey of Data Augmentation Approaches for NLP

[7] Sen Yang, Leyang Cui, Jun Xie, Yue Zhang. 2020. Making the Best Use of Review Summary for Sentiment Analysis

[8] Chi Sun, Xipeng Qiu, Yige Xu, Xuanjing Huang. 2020. How to Fine-Tune BERT for Text Classification?