Machine Learning Engineer Nanodegree

Capstone Proposal

Chenyang Y.
March 6th, 2020

Proposal

Domain Background

Natural language processing (NLP) is the technology used to aid computers to understand the human's natural language. The objective of NLP is to read, decipher, understand, and even generate human languages in a manner that is valuable. It started in the 1950s, although works can be traced as early as Turning Test. It is one of the most important fields in machine learning and AI. NLP is commonly seen in Interactive Voice Response (IVR), language translation, grammar checker, and etc.

Machine learning techniques are used to be considered for NLP tasks, while deep learning is traditionally used for image processing, classification and etc. Since the research on Long short-term memory (LSTM) and transfer learning were published, we are able to deploy pre-trained deep learning models to solve any NLP problems. The motivation of NLP projects using LSTM and related techniques is to classify topics, tag sequences, analyze sentiments, and generate natural language in the future.

Problem Statement

The project is aiming to analyze how travelers in February 2015 expressed their feelings on Twitter, through creating and tuning an ULMFiT model to correctly classify Twitter airline sentiment data. The result should produce 1 of 3 sentiments (positive, negative, and neutral). The tweets are quantifiable through tokenization to create a numeric representation of words. I will be using a language model provided from the fast.ai library and applying the pre-calculated weights. It will provide a word embedding scheme that aligns with the corpus of airline tweets. In order to capture the meaning in each word, the language model hyperparameters will be tuned and it will lead to a sentiment result. If the model is accurate, my result will match the pre-labeled sentiment. The same methodology can be applied to any text processing.

Datasets and Inputs

Twitter airline sentiment data contains 14640 labeled tweets directed at six major US airline companies (Virgin America, United, Southwest, Delta, US Airways and American Airlines), originally came from Crowdflower's Data for Everyone library. Tweets were scraped from February of 2015 and it includes labeling of positive, negative, and neutral sentiments. Twitter dataset came in both .csv and database format since it is obtained from a Kaggle challenge. It is categorized as a natural language processing problem.

Columns include **tweet_id**, **airline_sentiment**, **airline_sentiment** confidence, **negativereason**, **negativereason_confidence**, **airline**, **airline_sentiment_gold**, **name**, **negativereason_gold**, **retweet_count**, **text**, **tweet_coord**, **tweet_created**, **tweet_location**, and **user_timezone**. The key columns are "text", which is the content of the tweet, "**airline_sentiment"** and **airline_sentiment**_confidence, which will be used to validate my model.

Solution Statement

LSTMs is one of the many solutions to NLP problems. LSTM is a subset of Recurrent Neural Network, but takes time and sequences into account. It stores the information outside the normal RNN flow in a cell gated by sigmoid activation function. The gates act on the information they receive to let the stored information in or out. The gates will learn to keep or forget the relevant information. The gated cells are the basic ideas of LSTM, while it is more complicated with different roles to addition and multiplication during the transformation of input. Similarly, some benefits of LSTM application on NLP are the ability to store certain texts in the memory to learn from, as well as to resolve the limitation of vanishing gradients.

Benchmark Model

Universal Language Model Fine-tuning (ULMFiT), a novel approach towards applying transfer learning to natural language processing tasks, was originally introduced by Jeremy Howard and Sebastian Ruder in 2018. Transfer learning takes advantage of models that have already been pre-trained on large datasets and re-purposes them to learn a related task. The base architecture of ULMFiT is ASGD Weight-Dropped LSTM (AWD-LSTM), which implements dropouts, BPTT, and multi-batch encoder. In addition, ULMFiT process includes LM (Language Model) pre-training, LM fine-tuning using novel techniques Discriminative fine-tuning and Slanted triangular learning rates, and finally classifier fine-tuning using Gradual unfreezing, BPTT, and bi-directional LM. The optimization techniques mentioned above were added to the benchmark model one by one. The final model implemented on IMDb dataset resulted in a 5% benchmark validation error rate.

Evaluation Metrics

1. F1 Error: Evaluation will be using the F1 score (F1μ) for the three sentiment classes - Positive, Negative and Neutral on the submissions made with predicted class of each sample in the evaluation data set. The formula of F1 metric is as following:

   The traditional F-measure or balanced F-score (**F_1 score**) is the harmonic mean of precision and recall:

   $$F_1 = \left(\frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} \right) = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

2. The error rate = 1 - Accuracy
   a. This is the error rate used in ULMFiT paper. I will be able to compare to their validation error benchmark on the IMDb sentiment classification task.

Project Design

My goal is to translate texts into numbers and then studies the “meaning” of the number sequences to produce sentiment classification. For this process, transfer learning and ULMFiT are good fits for the dataset due to limited data size and small word counts per tweet (limited to 140 words count in 2015). My approach is outlined in following steps:

- Data preparation and exploration.
    - cleaning and train-test split.
    - correlation and distributions of sentiment and other variables.
- Tokenize texts.
- Langue Model (LM) Fine-tuning:
    - We'll be using an AWD-LSTM language model trained on the Wikitext 103 corpus, provided from the fast.ai library and applying the pre-calculated weights to the tokenized airline sentiment dataset.
    - Discriminative fine-tuning allows me to tune each layer with different learning rates.
    - Slanted triangular learning rates allows me to tune learning rate through linearly increasing and decaying.
- Fine-tune classifiers is critical for transfer learning:
    - Gradually unfreezing first allows me to unfreeze last layer, then fine-tune all frozen layers for one epoch. The fine-tuning will repeat as we unfreeze the lower layer, until the convergence at last iteration.
    - Bi-directional language model will allow me to use two separate text classifiers on both the forward and backward language model encoders to create an ensemble classification model.
- Evaluate and compare performance on test set through evaluation metrics.
- Analysis and document my results.

References:

1. Kaggle Dataset https://www.kaggle.com/crowdflower/twitter-airline-sentiment
2. Universal Language Model Fine-tuning for Text Classification by Jeremy Howard and Sebastian Ruder https://arxiv.org/abs/1801.06146
3. Understanding building blocks of ULMFIT by Kerem Turgutlu https://medium.com/mlreview/understanding-building-blocks-of-ulmfit-818d3775325b
4. F1 Score Wikipedia page https://en.wikipedia.org/wiki/F1_score
5. Pathmind A Beginner's Guide to LSTMs and Recurrent Neural Networks https://pathmind.com/wiki/lstm