Machine Learning Engineer Nanodegree

Capstone Proposal

Elvyna Tunggawan
March 3rd, 2018

Proposal

Domain Background

Emergence of web 2.0 has increased variety, veracity, velocity and volume of textual data. Demand of text analytics researches also increases, such as sentiment analysis [ref1], [ref2], topic modelling, text summarization, and text duplication detector. One of the challenges on text analytics, especially on text duplication detector, is identifying semantically equivalent sentences. Different phrases could have similar meaning, while a word also could have different meanings (homonyms). Moreover, there are thousands of language in the world ([ref]) with various grammars and vocabularies, which add the complexity.

Number of past researches were conducted to identify text similarity. Achananuparp et. al. ([ref]) evaluated performance of different sentence similarity measures and found that linguistic measures perform better than word overlap and TF-IDF measures. However, it couldn't produce satisfactory result on text which requires more specific textual entailment. Bogdanova et al. ([ref]) represented documents from Ask Ubuntu Community Questions and Answers site using word embeddings and fed them into convolutional neural network (CNN). CNN with pre-trained in-domain word embeddings resulted in higher accuracy than using English Wikipedia's word2vec. Based on experiments on different domains, neural network could achieve high accuracy regardless of the domain and size of training data.

In this section, provide brief details on the background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited in this section, including why that research is relevant. Additionally, a discussion of your personal motivation for investigating a particular problem in the domain is encouraged but not required.

Problem Statement

As a question-and-answer site, Quora enables people to submit questions, give high quality answers, and learn from each other. Having more than 100 million users, submitted questions could have similar words, and multiple questions might have similar intent. Quora emphasizes on providing canonical questions, in order to ease user experience on finding the best answers of similar questions and answering on similar questions. This problem also arises on other question-and-answer sites, such as Stack Overflow. To provide canonical questions, identification of duplicated questions could be done using a classification model. Each question pair could be represented as numerical features and be used as input on the model.

In this section, clearly describe the problem that is to be solved. The problem described should be well defined and should have at least one relevant potential solution. Additionally, describe the problem thoroughly such that it is clear that the problem is quantifiable (the problem can be expressed in mathematical or logical terms) , measurable (the problem can be measured by some metric and clearly observed), and replicable (the problem can be reproduced and occurs more than once).

Datasets and Inputs

The dataset used on this work is retrieved from Quora Question Pairs Kaggle Competition (Kaggle link), which contains full text of the question pair. The dataset contains 404,290 question pairs, which will be split into training, validation, and test set. Each question pair is labeled manually by human experts, which might not be 100% accurate. However, the labels are considered to represent reasonable consensus and will be used as the ground truth. The dataset structure is described on the table below.

Field Name	Description	Data Type
id	id of a question pair on training set	int64
qid1	unique ids of each question	int64
qid2	unique ids of each question	int64
question1	full text of each question	string
question2	full text of each question	string
is_duplicate	target variable (boolean): 1 if question1 and question2 have essentially same meeting, otherwise 0	int64

Features on `question1` and `question2` will be extracted and used as input variables to the model; while `is_duplicate` resembles the target variable.

In this section, the dataset(s) and/or input(s) being considered for the project should be thoroughly described, such as how they relate to the problem and why they should be used. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included with relevant references and citations as necessary It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.

Solution Statement

As a solution of this problem, I plan to implement Siamese neural network model, since it is popular among tasks that involve finding similarities or relationships between objects (ref). Prior to model training, data preprocessing and feature engineering will be done to extract important features of the dataset. I consider using word2vec to extract semantic representation of each word in the question pairs.

In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).

Benchmark Model

Quora developed random forest model as initial solution to this classification task (ref), then built another LSTM-based model. Their best model achieved 87% accuracy and 88% F1 score, which will be an aspirational target. Another research done by Shankar and Shenoy (ref) achieved 85% accuracy from Support Vector Machine (SVM) model. Homma et al. (ref) also achieved 85% accuracy with 84% F1 score from Siamese Gated Recurrent Unit (GRU) with 2-layer similarity network trained on an augmented dataset. Thus, 85% accuracy should be a more attainable target.

In this section, provide the details for a benchmark model or result that relates to the domain, problem statement, and intended solution. Ideally, the benchmark model or result contextualizes existing methods or known information in the domain and problem given, which could then be objectively compared to the solution. Describe how the benchmark model or result is measurable (can be measured by some metric and clearly observed) with thorough detail.

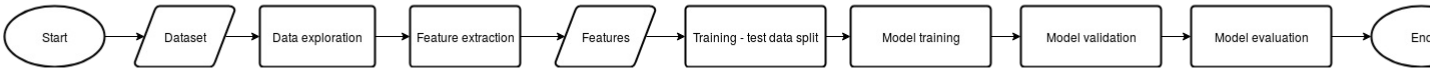
Evaluation Metrics

Model performance will be measured based on its accuracy, which is defined as number of correctly predicted class (duplicate or not) divided by number of question pairs on test set. In addition, F1 score will be calculated as measure of model performance on identifying the duplicated pairs. Model training duration will be reported as comparison for further works.

In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).

Project Design

(approx. 1 page)



The first stage of this project is data acquisition, which is done by downloading the dataset via Kaggle. Initial data exploration will be done to identify characteristics of the dataset, as well as taking proper treatment on missing values (if any). During data exploration, observation of characteristics on each record will be done, such as number of unique words, sentence length, and number of common words between question pairs.

After having a glance of the dataset characteristics, we will extract features of each question pair. Part-of-speech (POS) tagging will be performed on each question, as well as named entity recognitions. Punctuations will be removed from the dataset. We will also extract similarity measures from each question pair, such as cosine and Jaccard distances. We consider using TF-IDF measures and some linguistic measures proposed by Achananuparp et. al. (ref), as well as word embedding (e.g. word2vec) to enrich the input features. Word embedding and other features will be combined and be used as input to the classification model.

The handcrafted features will then be used as input on the Siamese neural network. As initial approach, we will use similar model architecture with Cohen (ref), which contains embedding matrix as input to Long Short Term Memory (LSTM) network. Each question on the question pair will be represented as an embedding matrix, and both questions will be processed on separated LSTM network and result in two outputs. Similarity of the outputs will be calculated using exponent of negative Manhattan distance, which will result in 0 or 1:

$$ \exp(-||h^\text{(left)} - h^\text{(right)}||_1) $$

The trained model will be cross validated and tested on the test set. If its accuracy is lower than the benchmark model's, another architecture will be tested on the model until model with best accuracy is gained. F1 score will also be reported on the final result.

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

References

[1] ....
[2] ....

Note:
Sample architecture by Bradley Allen - without Siamese (ref).

Before submitting your proposal, ask yourself. . .

- Does the proposal you have written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Solution Statement** and **Project Design**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your proposal?
- Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?