Domain Background

For capstone - Machine Learning Nanodegree project, proposing a project under Natural Language Processing (NLP). To be specific, under open domain question answering system.

Natural Language Processing is field in computer science which deals with interaction between humans and computers, by so making computer understand human languages. History of NLP goes as back as 1950. In the year, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence. For a truly intelligent system, one of the important aspects is to understand human language and reply back as humans do. With this in mind, historically many work has been done till date, like, English to Russian translation in 1954 to today's IBM Watson, Apple Siri, Microsoft Cortana etc.

One of the sub problems in NLP is open domain question answering system. Given textual information to computer/agent, how accurately it can answer by understanding the context of the question. In this proposal, under open domain question answering system, proposing a project to build a supervised machine learning model to predict the answer in given paragraph/text.

I am looking forward to take up this particular project because of my personal interest in NLP which is one of the key aspects to develop human like AI and huge progress made in recent days. Another motivation for taking up this project is, deep learning foundation nanodegree which I am currently enrolled. Interested in solving the problem using techniques in deep learning like RNN, LSTM which are specifically tuned for this kinda problem.

Problem Statement

The ability to read text (Reading Comprehension) and then answer questions, is a challenging task for machines, requiring both understanding of natural language and knowledge about the world. There have been many attempts to build better models using different techniques. Here I will be trying to build a supervised model to extract exact answer from given reading comprehension text. Dataset I choose is **S**tanford **Qu**estion **A**nswering **D**ataset (SQuAD) which consists good amount of question answer set with corresponding text for training and testing.

Datasets and Inputs

Dataset: **S**tanford **Qu**estion **A**nswering **D**ataset (SQuAD)

Dataset link: https://rajpurkar.github.io/SQuAD-explorer/

Input: A reading comprehension (RC) dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, With 100,000+ question-answer pairs on 500+ articles.

Dataset is divided into 2 sets, training and testing in json format. Each of it contain paragraphs (RC), question and corresponding answer.

Solution Statement

Given a question on a paragraph, solution not only should identify the sentence of the answer, but also identify exact text or span in the sentence as answer.

Few algorithms which I want to implement are Recurrent Neural Networks (RNN) and Long Short Term Memory networks (LSTM) deep learning techniques which are specifically modeled for sequence-to-sequence processing. Specialty of RNN is that a neuron can pass on information it learnt to its successor like a loop.

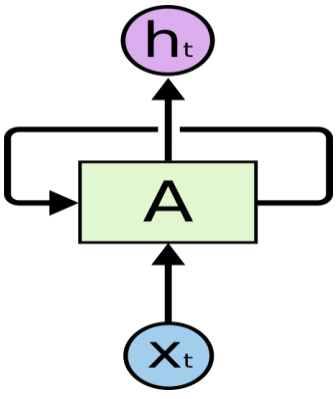

Fig 1: A typical RNN neuron.

Because of sequence-to-sequence property, RNN has huge impact in solving problem like speech recognition, language modeling, translation etc.

LSTM is a special kind of RNN, which specifically designed for retaining for information for long term which is very helpful in language processing.

Benchmark Model

Here want to define 3 Benchmark model,

a. Simple baseline model using sliding window: This model predicts answers with an accuracy of 20%.
b. Dataset author's logistic regression model: This model predicts answers with an accuracy of 51%.
c. Human performance: Human's answer prediction accuracy is of 86.8%

Will be aiming to outperform simple baseline model and near or excide author's logistic regression model.

Please refer following link for author's paper on dataset for benchmark model,

https://arxiv.org/pdf/1606.05250.pdf

Evaluation Metrics

Metrics for the model are exact match (EM) and F1 score.

Exact match (EM): these metric measures the percentage of predictions that match any one of the ground truth answers exactly.

(Macro-averaged) F1 score: This metric measures the average overlap between the prediction and ground truth answer. Here the prediction and ground truth are treated as bags of tokens, and compute their F1. Then the maximum F1 over all of the ground truth answers are taken for a given question, and then average over all of the questions.

To evaluate the solution, author of the dataset has provided two different test dataset. Dev set for developer to test the accuracy after training and a test dataset, which is not released for public to keep the integrity of the solutions. Will be evaluation on dev set provided.

Project Design

Project design is divided into following steps,

Step 1: Data Preprocessing

- Convert data from json to data frame format for better handling
- Check the integrity of the dataset.

Step 2: Data analysis

- Dataset summary
- Analysis paragraphs in data
- Analysis question and answers and verities in it

Step 3: Advance data processing

- Co reference resolution: Replace words like he, she , it etc with proper nouns
- Word embedding: vector representation of words

Step 4: Training

- Build model using LSTM technique to predict the exact answer (text or span) with the input being training data.
- Tune parameters like number of hidden layers, number of hidden units per layer, learning rate, dropout rate etc. to improve the model
- Test the model with dev set.