Machine Learning and Language Recognition

Ren He

ABSTRACT

Human languages use sounds to transmit information. Language has patterns. Words, sentences and paragraphs are built under rules and constraints. Language has diversity. Even in the same language, different individuals would have different preferences of phrasing. In this report, we aim to research for possibility of using machines to extract main idea of a passage using neural network algorithms.

Introduction

Human language is a tool originated since prehistorical time. It utilizes various combination of sounds to help an individual express their own feelings or communicate with other groups.

On one hand, languages have patterns. Consider modern English as an example. The smallest unit in the English language are letters, consonants and vowels. Letters make words, words make sentences and sentences make paragraphs. Consonants and vowels must be combined in certain ways such that sounds can be made. A sentence must consist of at least a noun and a verb. A paragraph should contain sentences that together convey a piece of logic.

On the other hand, languages have complexity. Often time, there exist multiple ways to express the same idea, and often time, an identical speech can have various meanings. "She is beautiful" and "she is pretty" both praises the appearance of subject "she". "I go to school" can either mean subject "I" attends school or they travels there based on different contexts.

Starting from the easier scenario. Consider that we are trying to extract sentiment inclination of a passage. For example, we want to analyze which candidate a passage supports during time of an election. In that case, by classical methods, we can either apply probit or logit regression models using parameters like frequency of key words. However, because the logic behind a passage would be too complex to be simply analyzed with one single regression. Consider the problem of context. For example, a key word associated with one party would imply different inclination from when it is assotiated with another party. If we want to incorporate with the effect of context, the amount of parameters to consider can at least double, not to mention that we are only doing binary analysis.

We therefore introduce the method of artificial neural networks (ANN). Inspired by the biological neural system, ANN is structured with nodes as neurons and links as synapses. Each neuron can be considered as a model. Links transmit outputs of the models usually in form of real numbers, and activate other models when thresholds are reached. Each neuron can be triggered multiple times in a run. By minimizing a cost function that quantifies deviation of the result from the desired output, the algorithm seek for a feasible solution based on optimization theory and statistical estimations. General neural networks have been used in various scenarios to look for algorithms to convert certain inputs to certain outputs. Compared to other methods, it is more accurate in recognizing and finding nonlinear relationships.

In this report, we explore the possibility for applying ANN in the task of language recognition. We will go over two existing methods, and potential possibility of future research directions.

Existing Works

Word Embedding for Sentiment Analysis

This algorithm is based on the nature that some words can be represented as vectors of attributes. The vectors can have multiple dimensions. For example, a Twitter post can be categorized to be happy, angry, excited or sad. Then in sentiment analysis of a Twitter post, we can decomposed words into four-dimensional vectors that represent their degrees of the four attributes.

Frequency of certain words of the same category can have correlation with inclination of a passage. Consider a financial report constantly mentioning words like "growing" and "thriving" without mentioning anything like "depression" or "recession", then the report is very likely positive about the subject they describes.

Consider an algorithm to give characteristics to words. In our case, we can view each word as an one dimensional vector pointing either to the negative side or positive side. For example, "fabulous" would have more positive meaning than "good". Meanwhile, context needs to be considered. Imagine a report about American political parties with a lot of positive words spotted. The inclination of the report cannot be determined unless we can associate the positive words with either of the parties.

Mr. Théo Szymkowiak conducted a project on classifying posts on the IMDB movie review data set. They used a package, Word2vec, in assistance of turning words into vectors. The package has following features:

1. Sub sampling: As words with frequency over a threshold would provide less and less information, extra words would be subsampled to increase training efficiency. This feature may not have too much use in this case, because movie reviews generally does not have a large input of texts.

2. Dimensionality: Performance of the algorithm increases with higher dimension it takes into account. But the effect of diminishing marginal gain exists. If too much dimensions are added, differences between the dimension would become smaller and it would be harder for the algorithm to differentiate between dimensions, also causing inefficiency in training time. But normally any dimensionality lower than 1000 can be acceptable, which is more than enough for the purpose of sentiment analysis.

3. Context Window: Context is certainly an important feature we need to consider when we analyze a passage. The size of the context window determines how many words before and after a given word would be included as context words of the given word. An adjective can imply opposite sentiment if it is used to modify different parties, while it can imply different intensity when it goes after different adverbs like "very" and "merely". According to the authors' note, the recommended value is 10.

This algorithm is useful in categorizing data. In Théo Szymkowiak's case, their team had been proven in actual tests that accuracy of their model can go up to 86.7% in binary classification. Currently, the algorithm is only tested in cases of shorter tests. As package Word2vec in fact supports larger amount of input, tests on longer texts would be interesting to see. However, in more daily life, people often time would be interested in actual contents of a passage instead of mere category it belong to.

Summary Based on Extraction

Longer passages sometimes contain information more than mere categorizable qualities. It is reasonable for users to want to extract a brief summary of main idea of the passage. In this case, we would need more complex algorithms.

Consider a long passage, an intuitive way to summarize it is to extract a list of topic sentences that can represent the whole passage. In a well-formulated passage, we can find such sentences at beginning or ending of a paragraph. However, for the algorithm to become more general, we can develop ways using ANN to locate a topic sentence in a passage.

In the work of K. Kaikhah, they considered three steps to achieve the goal. They assumed a topic sentence would have certain characteristics including phrasing structure, opinion strength and location. First, they apply algorithms of ANN to recognize those characteristics of a topic sentence. Secondly, they conduct the feature fusion process, that is, they prune the network and collapsed hidden layers to generalize important features and find trends of topic sentences. Eventually, they apply ANN again to locate and select through the sentences in the passage, leaving only the most relevant and non-repeating sentences, and build up a final summary.

This method showed their efficiency in their ability of identifying topic sentences. As a result of the project, their product can have an accuracy up to 99% when compared with human reader's summaries. Theoretically the method can prune a long passage to any length the reader requires, and individual readers can train the ANN according to their own preferences. Incorporating this algorithm with the word-embedding algorithm we discussed, we might be able to accurately extract sentiment of a longer passage.

However, because accuracy of the algorithm is calculated based on comparison with human reader's subjective summaries, while it might satisfy all subjective needs, its result may not be useful universally. Moreover, the algorithm assumes well-formatting of the passage. While the algorithm can be expected to work well when summarizing more rigid styles, it might face hardship when it is used on more artistic literatures, where main arguments are often time not explicitly stated.

Communicative Robots

Building up an algorithm that can enable the machine to effectively communicate with human beings is close to the final goal of machine language recognition. Recent years, large technology companies had worked on related topics. 2010, Apple Inc. purchased and reproduced Siri. 2014, Microsoft launched their first communicating bot Xiaoice in China. 2016, Microsoft launched the infamous bot Tay. Because the bots are developed by technology companies, algorithms behind them are not completely transparent, but guesses and hypothesis can still be made.

Most often time, a database can be built to make high frequency conversations more efficient. Greetings like "Hi" and "How are you" generally have very limited responses. Technical questions like "what is the weather now" and "how tall is the Pyramid of Khufu" has single correct answer. When it come to more creative interactions, the reacting algorithm would be more diverse. A commonly used technique is to search for similar topic on the internet and return a common response. Using Xiaoice as an example. There was history that they answer a question from a user with a short post under a relevant topic on Douban.com (a Chinese forum-structured website), only varying the punctuation, and the answer itself did not follow context of their discussion with the user.

Communicating robots can assist human beings in a more interactive way than regular machines. Very importantly, because those robots have access to the internet, they have a large data set of actual human talking styles, and therefore they have the potential to build more intimacy with common users than other form of machine learning algorithms with more restricted data sets.

However, existing communicating robots often time has the problem of loosing context. Due to limitation of context window, sometimes in a conversation, not all necessary context could be covered in the window, resulting in problem in communication. Some robots would not even have a context window, or their function is not obvious. Siri has related history. A user once asked them, "Could you sing?" They answered, "I would scare you." When the user continued to ask, "Why would I be scared?" They searched for that question on Google.

Future Possibilities

In none of the three cases we discussed above, the algorithm can actually "comprehend" the sentences they input or output. Reasonable interaction might be possible, but their potential is limited within the constraint of human history. Imagine an ape manages to find out rules in English pronunciation, and it knows it could be rewarded if it respond to certain sounds with certain sounds, it can fluently communicate with human beings, though their words are practically meaningless.

Not discussing the philosophical aspect of meanings, we focus on searching for ways to making language recognition of machines more efficient and creative, we would first want the machine to actually associate the words they speak with real world phenomena and concepts. If the machine knows that a banana is yellow, it cannot be misguided by any other sources saying that a banana is blue. Most importantly, the machine may even start to develop more creative combination of words, or more efficient ways of expressions.

However, it is also reasonable to argue that, given a significantly large amount of data, a machine can reach the same height without actually requiring any association with the physical world. Problem of determining color of a banana would not be a problem, as certainly opinions viewing bananas to be yellow is the majority, and the machine would not be misguided on the problem. But the machine is not always limited by following the majority. As there also exists examples in human history dataset that going against the majority can yield more beneficial outcomes, the machine may also learn from the rule, and bring up creative ideas. The debate is currently hard to settle, unless we eventually have the technology support to test the theories.

Relation with Class and Further Connections

First of all, this report went over a topic that is differ from the second presentation I had due to the immense mistakes and vanity in that presentation. I continued working with machine learning, but chose a topic that is more specific and practical.

Doing a project on machine generated music inspired me in conducting this project researching for machine learning and language recognition. There exists a significant amount of similarities between musics and language. Both are tools used by human beings utilizing sounds to express themselves and communicate, the former in a more artistic way, the latter in a more universally understandable way.

Direction of this project would be different from the machine composition project, that the main focus this time is not composing texts or inventing new ways of speaking, but more of understanding existing pieces of text and figuring out ways to respond. However, based on current technologies, programming machines to compose literature work is not an impossible task. In fact, the Microsoft AI, Xiaoice, we discussed above had written a book of poem, and a group had developed an AI, after analyzing seven books of Harry Potter, created a new Harry Potter story based on the analyzed data. In my opinion, the area still have a lot of potential. It would be interesting to continue my future researches on related topics.

Beyond that, algorithms of machine learning and ANN can be applied a various topics that our classmates had discussed in class. As mathematical as optimizing, training modeling, and as daily as sudoku solving, card counting and Google map routing. After attempts on all kinds of classical models, it is always good to take a try on ANNs. They might sometimes return some unexpectedly creative models that we are never awared of before.

References

https://medium.com/@thoszymkowiak/how-to-implement-sentiment-analysis-using-word-embedding-and-convolutional-neural-networks-on-keras-163197aef623

http://disi.unitn.it/ severyn/papers/sigir-2015-short.pdf

https://link.springer.com/article/10.1023/B:ORIG.0000016440.53346.dc

http://en.people.cn/n3/2017/0531/c90000-9222463.html

https://www.zhihu.com/question/24273642