Indian Institute of Technology, Kanpur

SUMMER '17 PROJECT

PROGRAMMING CLUB

Science and Technology Council

Summarizing News Articles

End-term Evaluation

Project Members

Vipul Bajaj Hunar Preet Singh
Ritesh Kumar Richeek Awasthi

Project Mentor

Prannay Khosla

Contents

1. About the project

- As the project name suggests our aim is to teach computer enough so that it can give summary of the lengthy news articles or as a matter of fact any articles (given proper data set to train and required processing units).
- So basically we are using various concepts of machine learning and deep learning to train our model for generating summary of a given article.
- In our case we have used 'reuters' corpus from NLTK as training dataset and tensorflow module for training the model.

2. Decisions made

We have decided to use tensorflow instead of torch to implement deep learning due to its easinesss since it handles them very well, and extension of code to GPUs is quite minimal And the speed ups are tremendous.

3. Things learnt

3.1. Linux

Getting acquainted with ubuntu and terminal.

3.2. Python

All of the team members according to their pre proficiency in python spent time learning numpy.
One of the sources used was this.

3.3. Machine Learning

Did CS229,An introduction to machine learning by Andrew Ng from Coursera till Neural Nets.

- Supervised learning (parametric/non-parametric algorithms, support vector machines, kernels, neural networks)
- Best practices in machine learning (bias/variance theory; innovation process in machine learning and AI)
- Linear regression,Logistic Regression
- Gradient descent,Stochastic Gradient Descent

3.4. Deep Learning

Udacity course on Deep Learning

- Deep Neural Networks
- Convolutional Neural Networks
- Deep Models for text and sequences

3.5. Information retrieval

We read a book on information retrieval using NLP techniques. From the book we went through Chapter 1, 2, 6, 7, 11, 12

irbook_manning.pdf

3.6. Natural Language Processing

Natural language processing (NLP) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages, and, in particular, concerned with programming computers to fruitfully process large natural language corpora.

Natural language processing is not very distant from creating a robot like Jarvis (Just A Rather Very Intelligent System)from the movie Iron Man. There have been numerous examples over the last two decades of how Natural Language Processing, or NLP, is being used by companies to provide an intelligent voice to gadgets and searches. Think, for instance, how the world of search engines—from Yahoo, Microsoft and Google—have changed the Internet with text-based search algorithms driving and augmenting the World Wide Web. NLP, though, does much more than just that and text analytics. NLP exploration on our current digital planet includes voice searches on automobiles and then, of course, the dictation mechanics of the software world.

3.7. Word2vec

Word2vec is a two-layer neural net that processes text. Its input is a text corpus and its output is a set of vectors: feature vectors for words in that corpus. While Word2vec is not a deep neural network, it turns text into a numerical form that deep nets can understand.

The purpose and usefulness of Word2vec is to group the vectors of similar words together in vectorspace. That is, it detects similarities mathematically. Word2vec creates vectors that are distributed numerical representations of word features, features such as the context of individual words. It does so without human intervention.

Given enough data, usage and contexts, Word2vec can make highly accurate guesses about a word's meaning based on past appearances. Those guesses can be used to establish a word's association with other words (e.g. "man" is to "boy" what "woman" is to "girl"), or cluster documents and classify them by topic. Those clusters can form the basis of search, sentiment analysis and recommendations in such diverse fields as scientific research, legal discovery, e-commerce and customer relationship management.

The output of the Word2vec neural net is a vocabulary in which each item has a vector attached to it, which can be fed into a deep-learning net or simply queried to detect relationships between words.

4. Accomplishments

- First task
    1. Download MNIST dataset
    2. Create a dataset of 100,000 images using Numpy where there is a presence of 2 digits in the same image(i.e. 100 labels)
    3. Tag the dataset while creating it. If the image contains a number 3,4 then mark it as 34. Note this would be as to which number is left to what number in the image.
    4. The size of the images have to 64x64.
    5. Train a 5 layer convolutional network using this dataset and report accuracy. Use tensorflow and GPUs.The data should be completely random and the task should be completely automated.

    Link to the task implemented

- We have implemented the basic version of summarizer i.e. extractive summarizer which extracts sentences from articles which are most likely candidates to be included in summary.
  Github Link

Brief explanation :

- Did some basic language processing e.g. word tokenization , sentence tokenization ,removing numeric characters etc
- Used skip gram word2vec model to train on processed dataset and get word vectors
- Calculated sentence vectors (from word vectors)
- Clustering sentence vectors by k means algorithm
- Choose the central vectors from heavily populated clusters and that is our summary (most likely!!)

5. Plans ahead

- We will focus on abstractive summarization i.e. make our computer intelligent enough so that it can create a suitable headline for a news article.
- In order to generate a summary it will search the space of all possible summaries to find the most likely sequence of words for the given article.
- Next we'll focus on text generation using Generative Adversarial Networks and hence use it in summarization.

6. Study resources

- A paper on neural text generation, using Adversarial models
- A paper on GAN
  GAN_ian_goodfellow.pdf
- http://colah.github.io/
- http://www.wildml.com/
- Deep learning by Udacity
- Machine Learning By Coursera(Andrew Ng)
- A Book on IR by Manning