

Making a Contextual Recommendation Engine using Python and Deep Learning



Muktabh Mayank Srivastava
CoFounder



@muktabh



Muktabh-Mayank



About ParallelDots

Machine Learning products company

- TIMELINE

• Arvind Subramanian

01 Jun 2015

Is India's growth overtaking China? India's road, rail drive could lay doubts to rest

30 May 2015

GDP figures indicate substantial improvement in economy: Arvind Subramanian

+ TWEETS

Powered by ParallelDots



"Janet Yellen (US Federal Reserve Chair) is saying there will be some rate hike s expressed concerns over the labour market. What that means is the timing will be something which we have to watch," Chief Economic Advisor Arvind Subramani

Related Articles

GDP figures indicate substantial improvement in economy: Arvind Subramanian

CEA Arvind Subramanian talks rain, rates, revenue

Indian economy is recovering: Arvind Subramanian

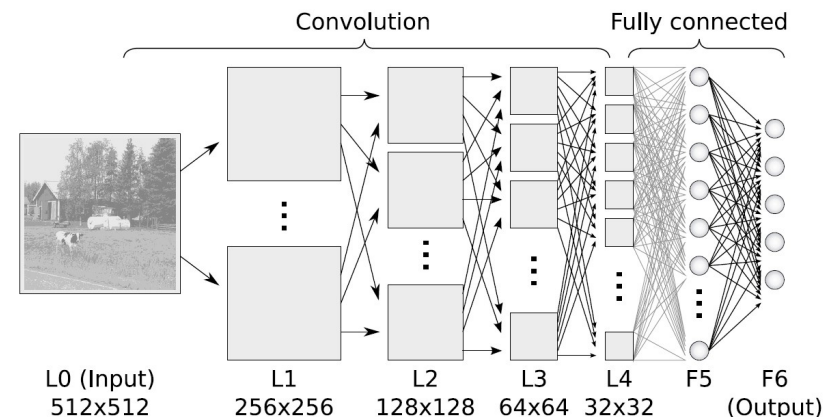
India will have to remain watchful about hike by the US Federal Reserve which this year, Chief Economic Advisor Arvi said.

A hike in interest rates by the US, it is the flight of capital from the emerging India.

"Janet Yellen (US Federal Reserve Chair) be some rate hike sometime this year, concerns over the labour market. What

timing will be carefully done. That is something which we have

- Deep Learning as a Service APIs. Semantic Similarity, Entity Extraction, Autotagging, Sentiment Analysis.
- Contextual Recommendation Engines for large publishers



Why make recommendation engine ?

- Solves problem of anomalous state of knowledge for the user.
- Helps in content discovery.
- Increases user engagement on website.
- Helps monetize indirectly.

Pre-existing Solutions

- CMS (say WordPress) provides a “Related Posts” plugin. This is TFIDF based search of article tags.
- Fails many a times, Related Posts for all articles not visible.
- Article tags often not contextual for SEO, produces garbage results many times.

Aims of new solution

- Should be more accurate than TFIDF tag search.
- Should be able to generate related posts for all articles.
- Should be cheap to deploy. (Its still “related posts” at the end of it)

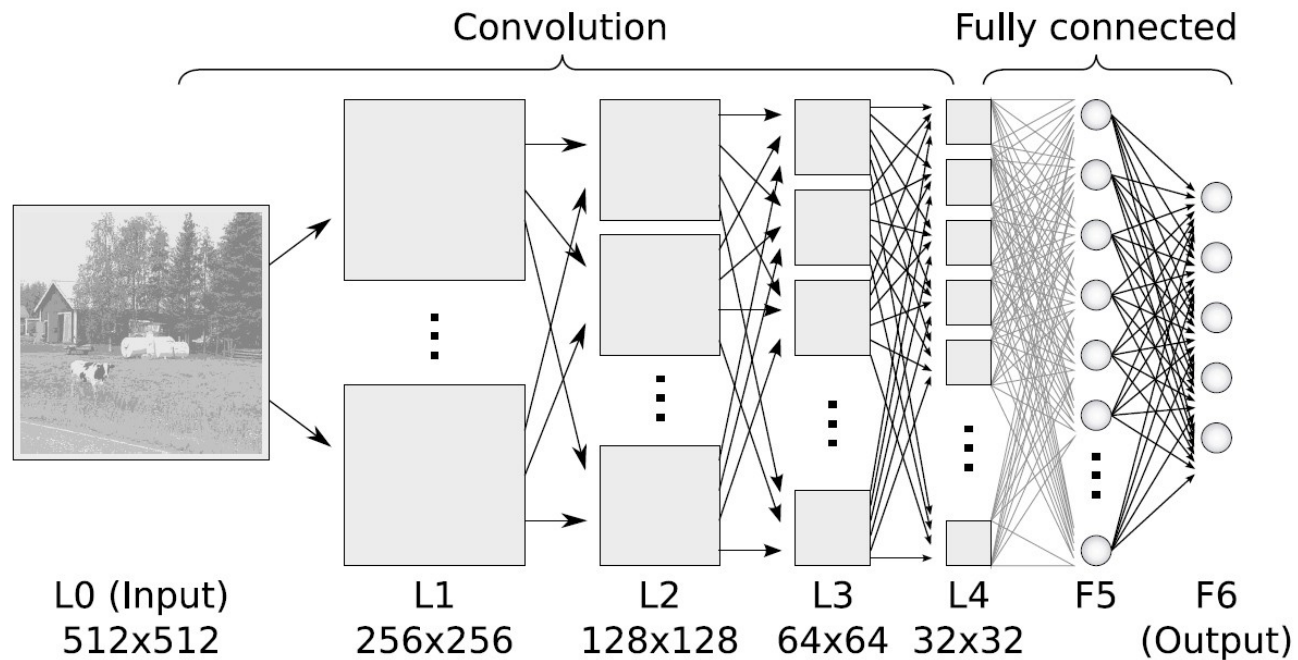
Why traditional methods wont work ?

- Made our MVP with Latent Semantic Analysis.
- Matrix Factorization methods like LSA were hard to scale on a vocabulary with size = (vocab(English) + vocab(Indian Names)).
- Switched to incremental LDA, but were not satisfied with accuracy.
- So decided to try out representation learning. This minimizes hardware costs at deployment. Can use spot instances for training, which is still heavy, but a batch job and periodic.

Current Solution Overview

- Generate low dimensional semantic document representations using Deep Neural Networks and heuristics.
- Create a Space-Partitioning Tree of documents.
- Refresh Trees at a higher frequency to add new documents.
- Refresh Representation Learning algorithms at low frequency to avoid topic drift.

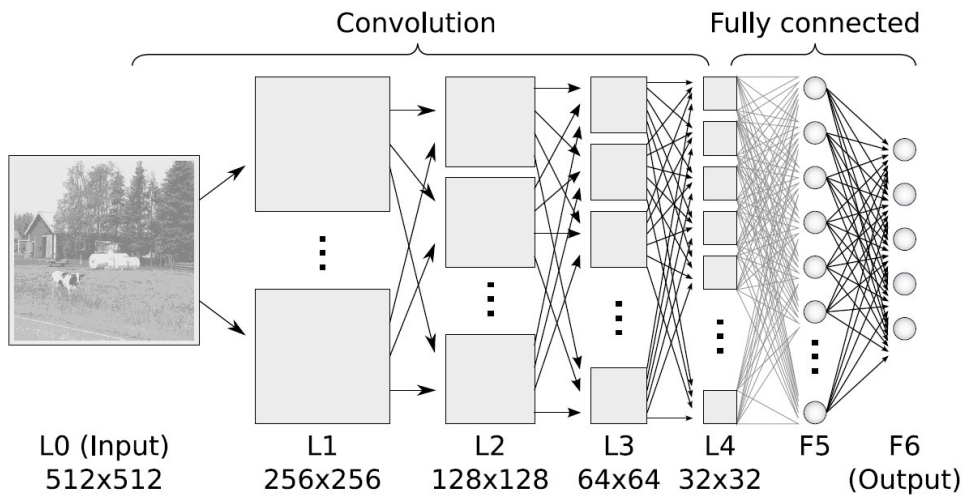
Basics of Deep Learning



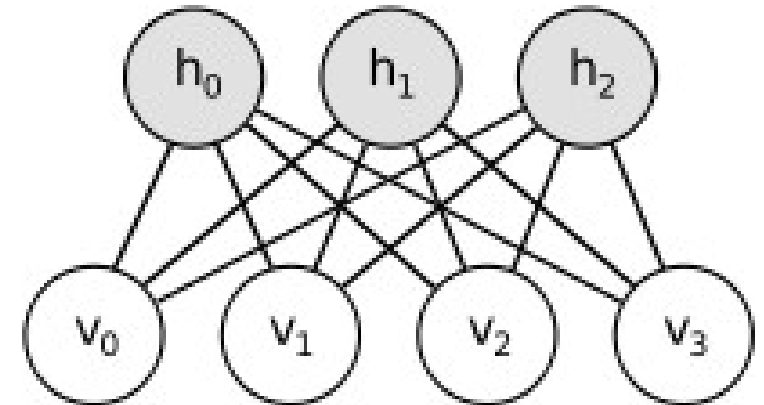
- Deep Learning is name given to multi layered Neural Networks.
- Layer(s) of weight are stacked on top of each other separated by layers of activation functions. Activation functions bring non-linearity into the learning, else multiple layers of weights would be same as one layer.
- They are trained by backpropagation of errors generally by Gradient Descent.

Popular Deep Neural Nets

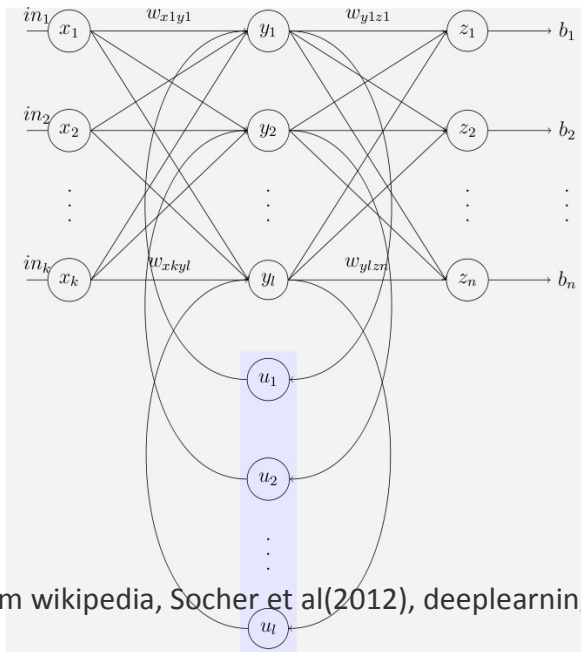
Convolutional



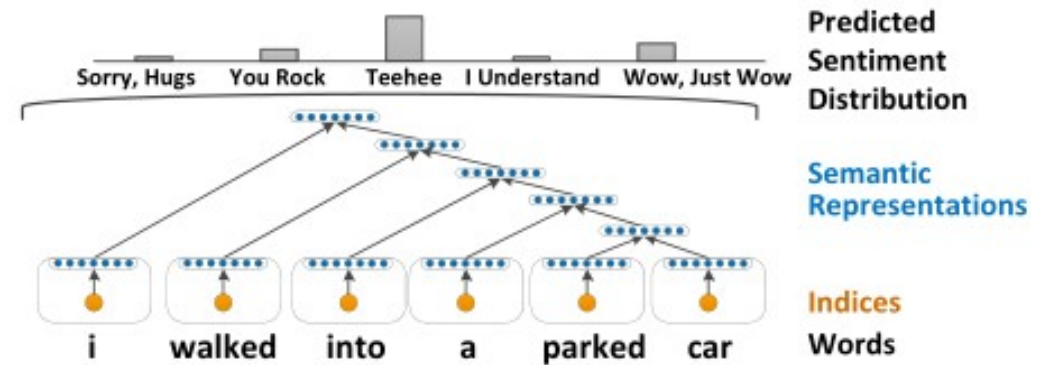
Boltzmann Machines (RBMs more common)



Recurrent



Recursive



More Neural Network Vocab

Architecture

Recurrent

Recursive

Convo NN

DBN

RBM

Dropout

LSTM

Sigmoid

ReLU

tanh

Neural Net Units

Clipping

Hessian

MSGD

Simulated
Annealing

SGD

Momentum

Adadelata

Adagrad

RMSprop

Optimization

Deep Learning for NLP

Basic Representation (Inputs to Neural Networks)

Bag of Words like

Word Embeddings

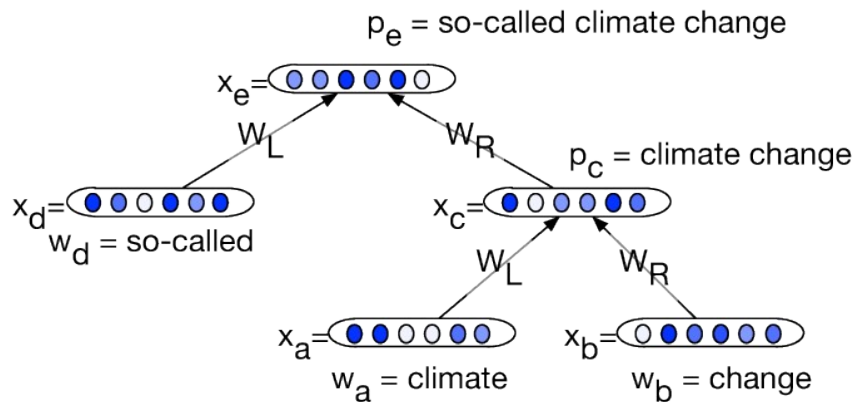
- Dense Low dimensional Representation of each word as a “thought vector”
- Are points in vector spaces corresponding to real world meanings, s.t., $R(\text{“king”}) - R(\text{“Man”}) \sim R(\text{“Queen”})$
- Two approaches: predicting next word given last n words (Word2Vec) or based on Co-occurrence Matrix (Glove). Both seem to be theoretically related.
- New approach: Use Autoencoder to predict co-occurrence matrix:
<https://github.com/ParallelDots/WordEmbeddingAutoencoder>

Character Level

- Newest (and possibly coolest) Trend. Already writing music and poetry. Based on a 2011 paper.
- <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Models we use at ParallelDots

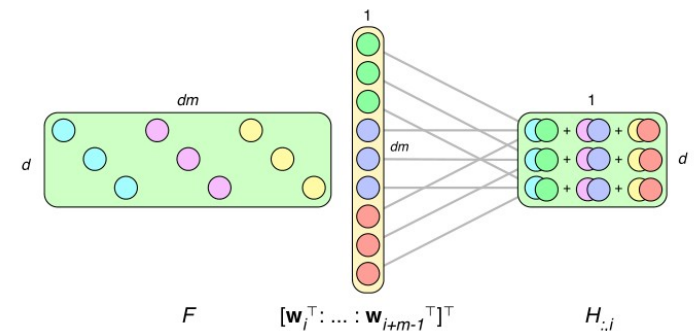
In Search Engine



- Recursive Neural Networks to combine Word Vectors into phrase Vectors for semantic closeness.
- Heuristically combining Word/Phrase Vectors for similar entity based near neighbors.

Others

- Convolutional Neural Networks to capture Sentiments in text.
- Recursive Neural Net based entity extraction.



Implementing Models

theano

Production Workhorse @ ParallelDots

Light CPU tryouts

Kayak: Library for Deep Neural Networks


Pylearn2

Lasagne

Lasagne is a lightweight library to build and train neural networks in Theano.

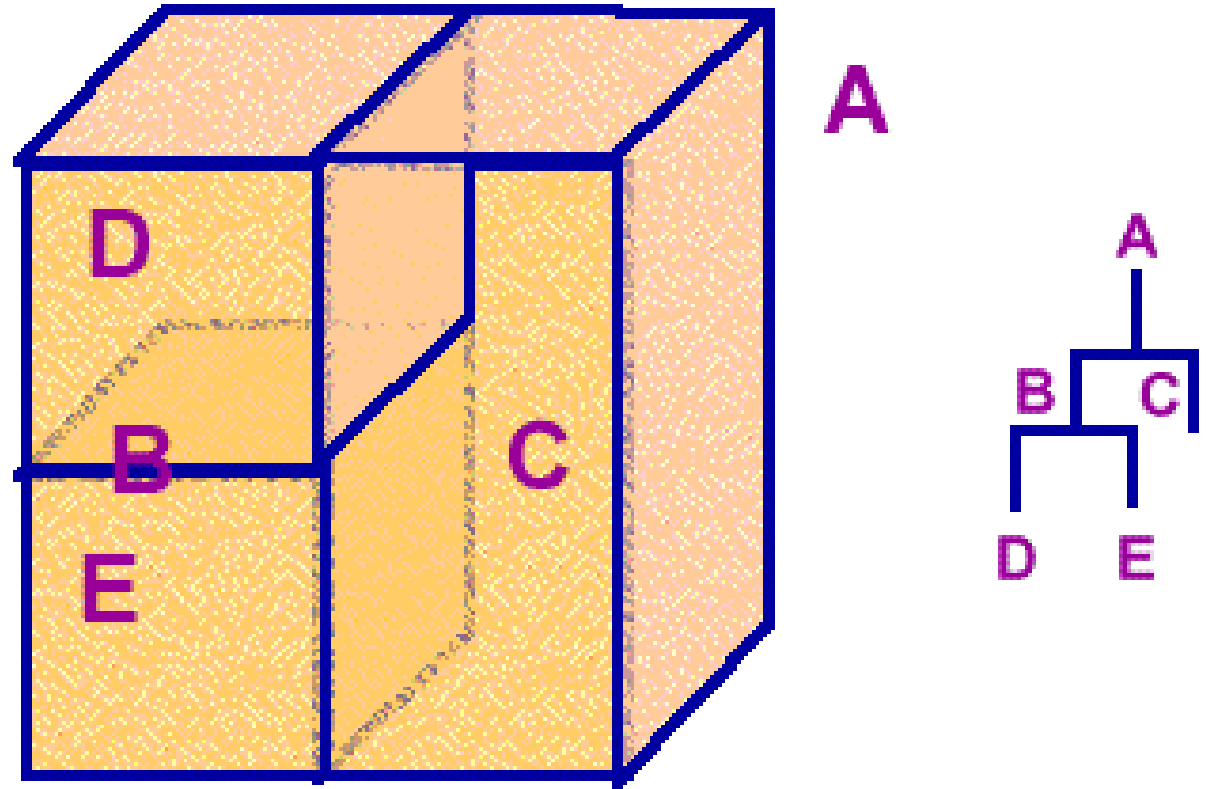


Caffe

 **cuda-convnet2**
Fast convolutional neural networks in C++/CUDA

Searching for Similar Documents

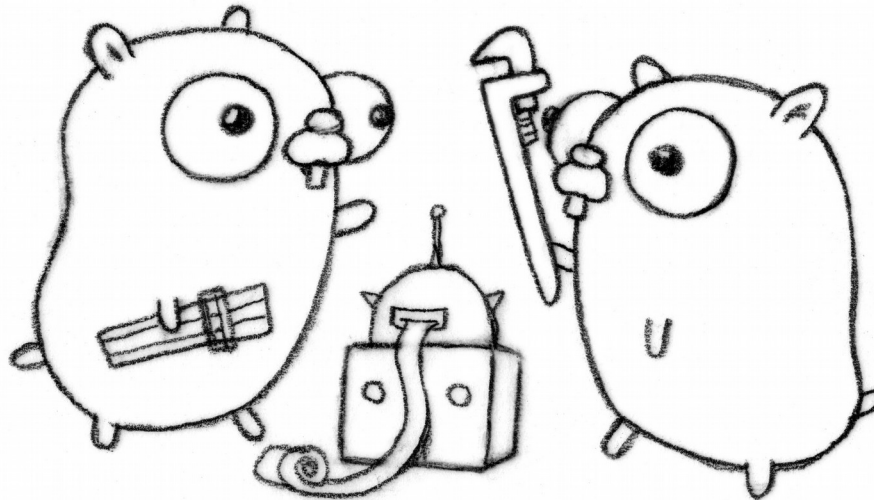
- Arrange Document Representation on a Space partitioning Tree.
- Uses VP Tree because of minimal requirements.
- Makes query $O(\log n) < \text{querytime} < O(N)$.



- Fast implementation in Numpy.
- Right now documents divided into buckets and hosted one bucket/core using Python's multiprocessing.
- Future Work: Shared memory model to make it work as true $O(\log(n))$.

Handling News Website Traffic

- We thought that we could just put Redis with Python during MVP, cache recommendations and serve. It did not work out.
- The most similar physical phenomenon to News Traffic is an Earthquake. One might have 1000s of concurrent users on a webpage in a second as soon it is shared on Social Media. This puts insane load on server before results are cached.
- We now use Golang channels to group these requests and deduplicate hits on Machine Learning infrastructure. Redis is used for caching. Handles upto 3000 concurrent users on a single box.



Thank You :) .

Please try our demo at ParallelDots.com