Data Science for Retail

Retail Analytics using Advanced Machine Learning

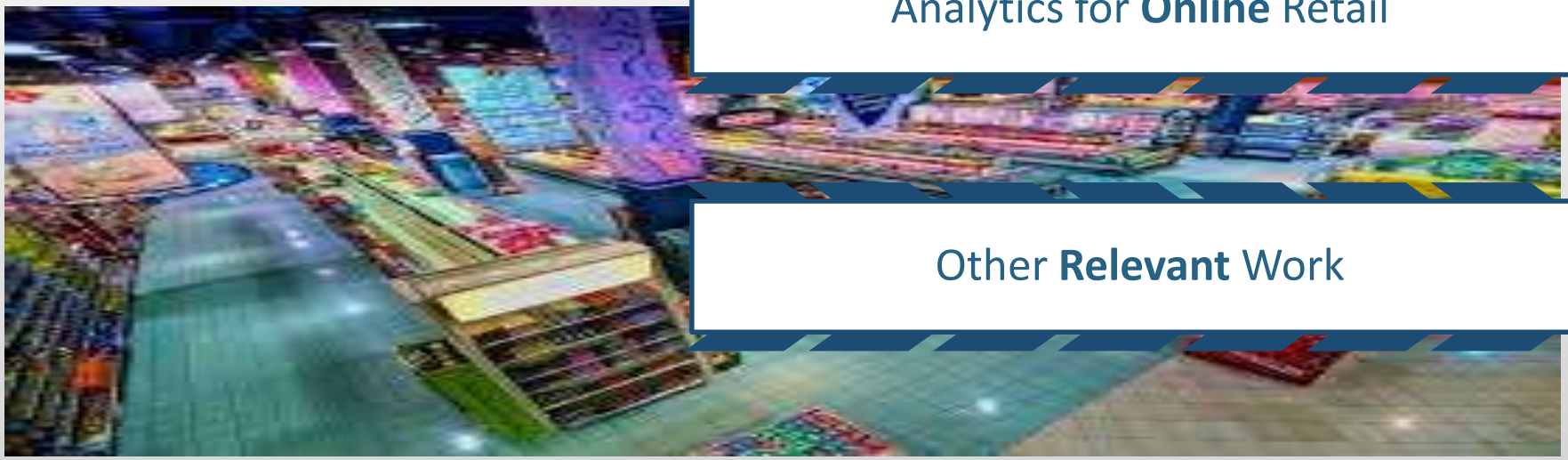
AlgoAnalytics

Outline

About **AlgoAnalytics**

Analytics for **Online** Retail

Other **Relevant** Work



AlgoAnalytics

CEO and Company Profile

About AlgoAnalytics

Analytics Consultancy

- Work at the intersection of mathematics and other domains
- Harness data to provide insight and solutions to our clients

Led by Aniruddha Pant

- +30 data scientists with experience in mathematics and engineering
- Team strengths include ability to deal with structured/ unstructured data, classical ML as well as deep learning using cutting edge methodologies

Expertise in Mathematics and Computer Science

- Develop advanced mathematical models or solutions for a wide range of industries:
- Financial services, Retail, economics, healthcare, BFSI, telecom, …

Working with Domain Specialists

- Work closely with domain experts – either from the clients side or our own – to effectively model the problem to be solved

Aniruddha Pant

CEO and Founder of AlgoAnalytics

PhD, Control systems, University of California at Berkeley, USA 2001

Highlights

- 20+ years in application of advanced mathematical techniques to academic and enterprise problems.
- Experience in application of machine learning to various business problems.
- Experience in financial markets trading; Indian as well as global markets.

Expertise

- Experience in cross-domain application of **basic scientific process**.
- Research in areas ranging **from biology to financial markets to military applications**.
- Close collaboration with premier educational institutes in India, USA & Europe.
- Active involvement in startup ecosystem in India.

Prior Experience

- Vice President, Capital Metrics and Risk Solutions
- Head of Analytics Competency Center, Persistent Systems
- Scientist and Group Leader, Tata Consultancy Services

AlgoAnalytics

AlgoAnalytics - One Stop AI Shop

Financial Services

- Dormancy prediction
- Recommender system
- News summarization – automated 60 words news summary

Healthcare

- Medical Image Diagnostics
- Work flow optimization
- Cash flow forecasting

Legal

- Contracts Management
- Structured Document decomposition
- Document similarity in text analytics

Internet of Things

- Assisted Living
- Predictive in ovens
- Air leakage detection
- Engine/compressor fault detection

Others

- Algorithmic trading strategies
- Risk sensing – network theory
- Network failure model
- Multilanguage sentiment analytics

- We use structured data to design our predictive analytics solutions like churn, recommender sys
- We use techniques like clustering, Recurrent Neural Networks,

Structured Data

- We used text data analytics for designing solutions like sentiment analysis, news summarization and many more
- We use techniques like natural language processing, word2vec, deep learning, TF-IDF

Text Data

- Image data is used for predicting existence of particular pathology, image recognition and many others
- We use techniques like deep learning – convolutional neural network, artificial neural networks and technologies like TensorFlow

Image Data

- We use sound data to design factory solutions like air leakage detection, identification of empty and loaded strokes from press data, engine-compressor fault detection
- We use techniques like deep learning

Sound Data

AlgoAnalytics

Technologies

NLTK

R

TensorFlow

theano

Pandas

python

Microsoft Azure

cassandra

mongoDB

php

PostgreSQL

C++

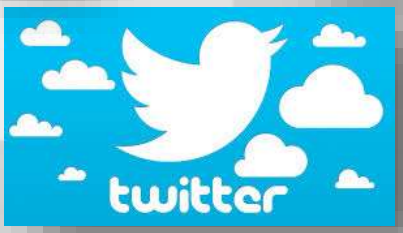
Microsoft CNTK

Java

Spark

Google news

H_2O.ai



AlgoAnalytics

Analytics in Online Retail

Recommender system

- Analysis of user behavior for personalized shopping experience
- Product recommendations for upselling and cross selling

Demand Prediction

- Demand modeling based on price or brand, price of competing products, etc.
- Useful in price optimization and sales event planning

Image Analytics in Retail

- Image recognition – item tagging, differentiating between original and duplicate, substitute product
- Generating image descriptions

Marketing

- Customer segmentation for focused marketing
- Brand marketing – customizable ad placement

Customer churn preventions

- Improved customer engagement
- Loss prevention through customer retention

AlgoAnalytics

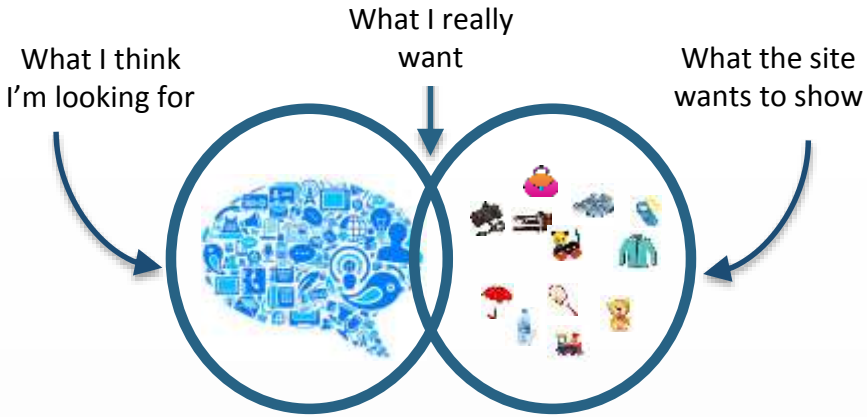
Recommender System

What is RecSys?

- Aims to predict user preferences based on historical activity and implicit / explicit feedback

- Helps in presenting the most relevant information (e.g. list of products / services)

Value of Recommendation



RecSys Modeling and Applications

Collaborative filtering: User's behavior, similar users

Content-based filtering: using discrete characteristic of items

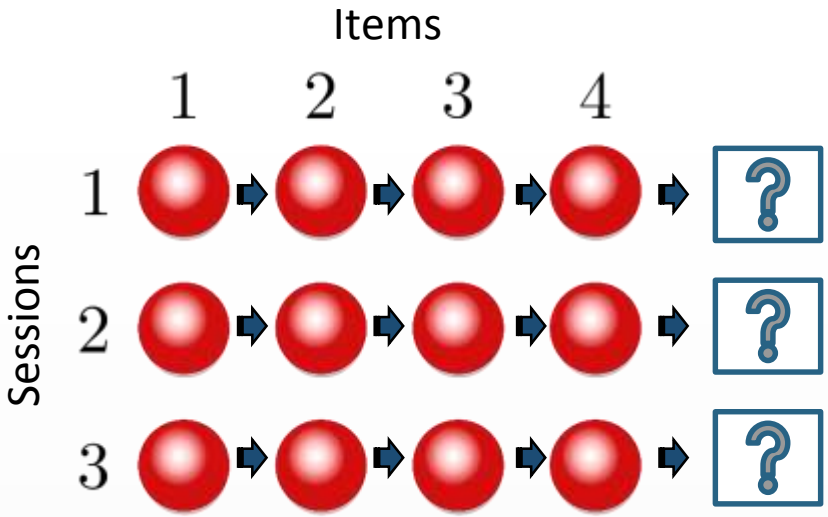
- Nearest Neighbor modeling
- Matrix factorization and factorization machines
- Classification learning model

* Movies, music, news, books, search queries, social tags, etc.

* Financial services, insurance Intel business unites (BUs), sales and marketing

Session Based Recommender System

Use historical sessions data from all customers to recommend products



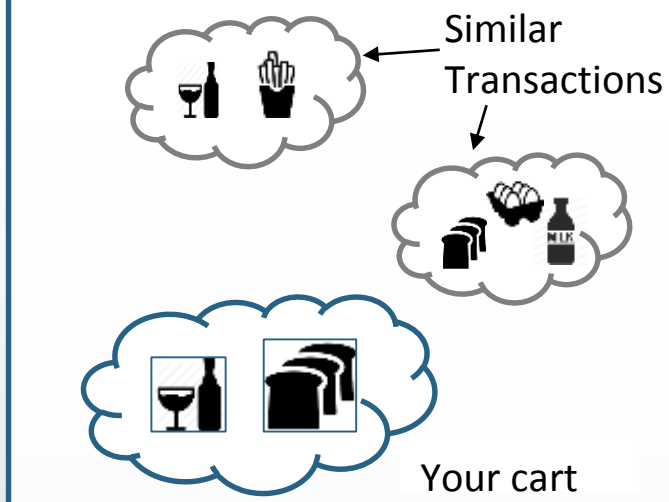
Deep Learning

- Recommendations are independent of customer profile
- Tracking user is not required
- Ideal recommendations for guest users

AlgoAnalytics

Session Based Recommender System

Product Recommendation



Dataset Description

- 3 Months worth of raw click-stream data
- ~800K products for RecSys
- ~2 million user sessions for building a model

AlgoAnalytics

Performance of Our Session-Based RecSys (Recall@N)

Recall@N represents % of times the desired item appeared in top-N recommendations

Higher the recall, better the RecSys, increase in cross-sell and up-sell

53% recall@20 has been achieved using Session Based Recommendation System

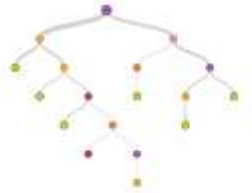
Thus it is more likely that a customer will view one of the recommended products!

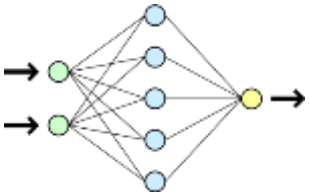
AlgoAnalytics

Image Analytics in Retail

Methods

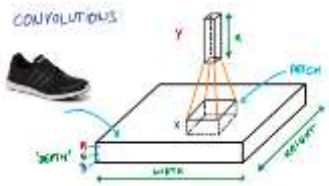
Statistical Learning





Artificial Neural Network

Convolutional Neural Network



Transfer Learning (Deep Learning)

Technologies

R Programming for statistical models: using pixel values as features and applying models such as logistic regression, random-forest

TensorFlow-Python for neural networks (feed-forward and CNN)

Google's Inception model (pre-trained TensorFlow model on Imagenet dataset)

AlgoAnalytics

Identification of unseen images

Input Data: Images of different products along with category label.

Trained Model

Test Image : Image without category label

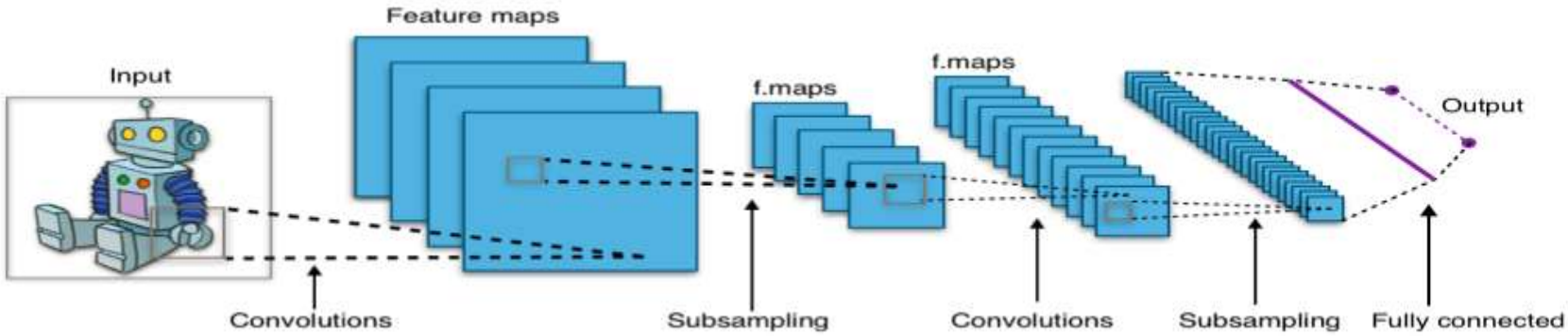
Classification Engine

Output: Chance of input item being floaters, boots, etc

Floaters (9%)
Athletic Shoes (32%)
⋮
Boots (6%)

Convolutional Neural Network

In **convolutional** (filtering and encoding by transformation) **neural networks** (CNN) every **network** layer acts as a detection filter for the presence of specific features or patterns present in the original data.

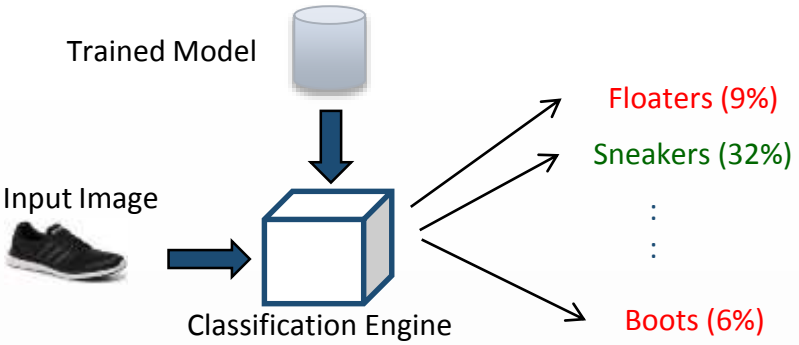


AlgoAnalytics

Image analytics in Retail – predictive model identifying class of unseen images with high accuracy

Product Category Identification

Input Data: Images of different products along with category label



Brand Logo Classification

Input Data:

- Images of different various brand logos, such as Adidas, Google, Coca Cola, etc. (total 32 brands)
- Masks for logo location in an image

Approach:

- Extract logos from input images using masks
- Reshape to 64x64 size
- Statistical models, Neural Networks, transfer learning using Google's Inception model

❑ **Results (best so far)**

1. Product Category Identification
   - Identifying sneakers vs. others: **91% accuracy**
   - Classifying types of shoes in 10 different classes: **76% accuracy**
2. Brand Logo Classification
   - Classifying brand logo correctly in one of 32 classes (image size of 64x64 pixels): **88%**

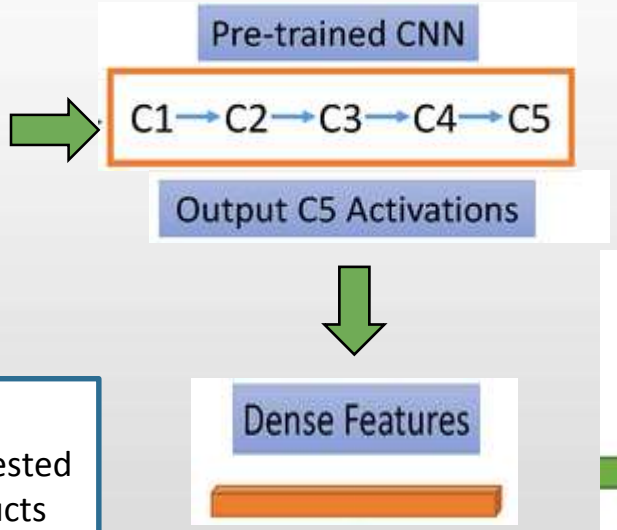
❑ **Business Value**

1. Product Category Identification: automating task of categorizing millions of untagged product catalog images for e-commerce websites
2. Brand Logo Classification: brand tracking on social media
3. Others: detection of pathologies in medical images (healthcare domain), OCR (optical character recognition), face recognition (biometrics), etc.

AlgoAnalytics

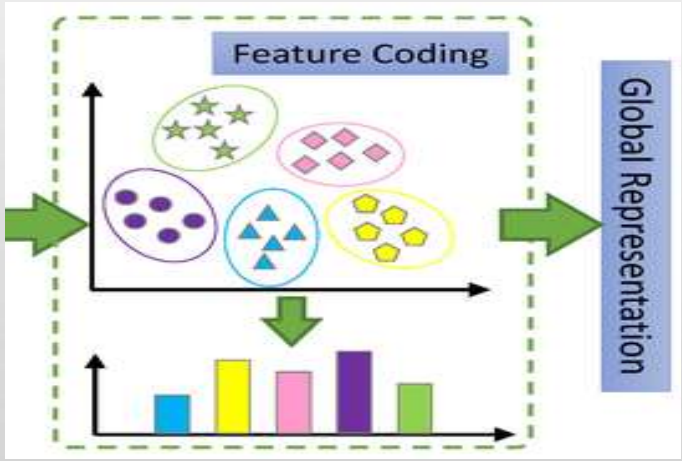
Image Classification Results

Our model assigned correct labels to given untagged images **94%** of the time
Advanced methodology of **Transfer Learning** is used to get the best classification model.



Transfer Learning : Improve learning in new task through transferring the knowledge from already learned related task

Testing : Final model is tested on ~5k images of products from 10 different categories

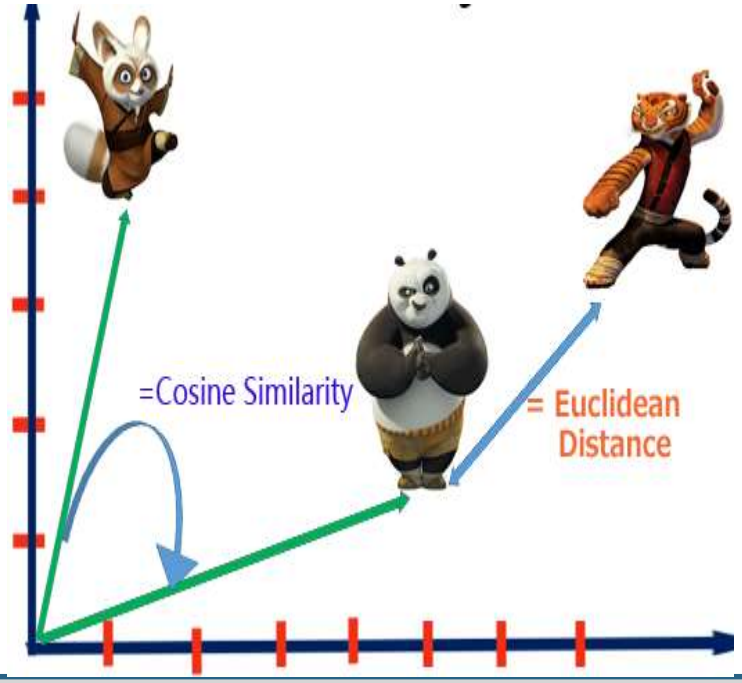


AlgoAnalytics

Image Similarity Based Recommendations

Similarity Measure:

- **Cosine Similarity :** A measure of **similarity** between two non zero vectors of an inner product space that measures the **cosine** of the angle between them.
- **Euclidean Distance :** The distance between two points defined as the square root of the sum of the squares of the differences between the corresponding coordinates of the points.
- **Nearest Neighbour :** Finding the item in a given set that is **closest** (or most similar) to an input item.



AlgoAnalytics

Performance of Our Image-Based RecSys

Total number of products for analysis: ~700K

More than 30 product categories (electronics, clothing, etc.) each with 100+ subcategories

Real World Example (from click-stream dataset of a retail client)

Input Product Image

Recommendations by our Image based RecSys

- Time taken for generating real-time recommendation for input product: **< 100 milliseconds**
- Practical benefit: only **product image** is required to build a recommendation system
- Other information (such as customer's data and product description) can help to improve results further

AlgoAnalytics

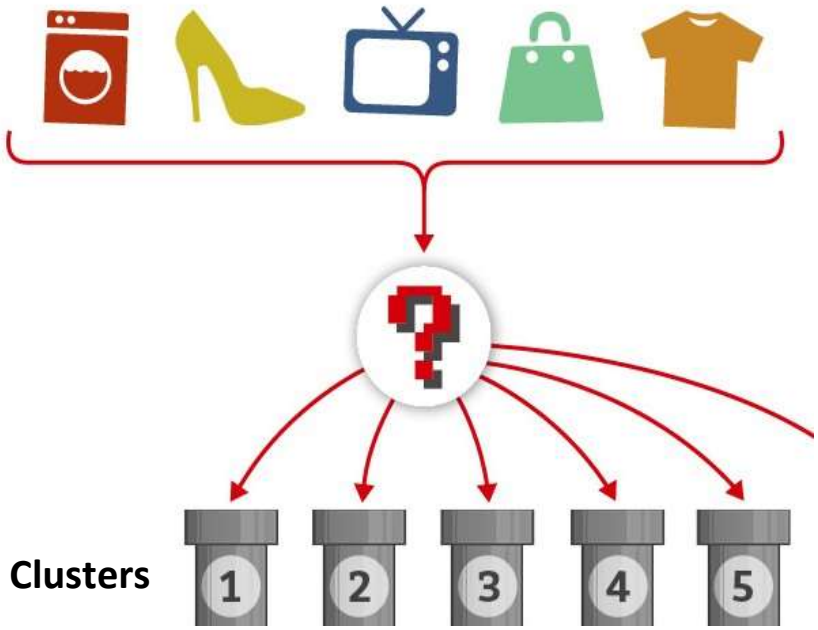
Image Based Clustering

Problem Statement

Creating clusters of products using images of products as input to create high level categories for unlabeled products.

Use Cases

- Organising huge amount of unlabeled products
- Processing and Analyzing the data.
- Extracting knowledge,insights from the data and preparing data for supervised learning



AlgoAnalytics

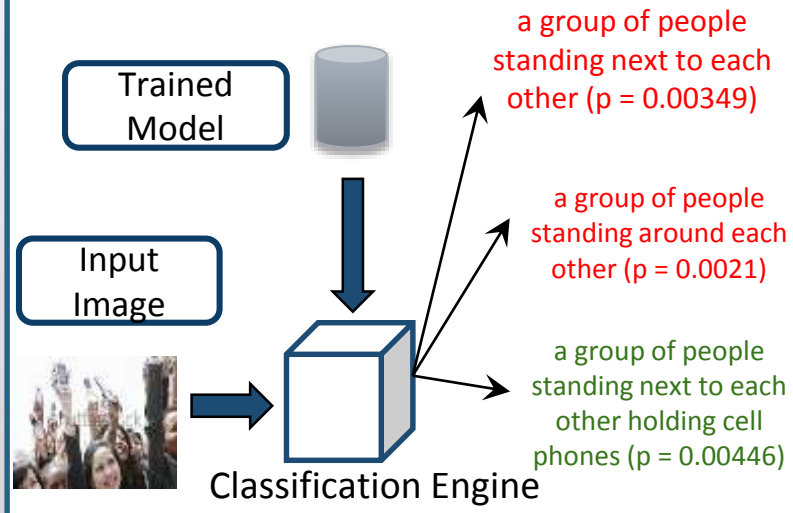
Deep Visual-Semantic Alignments for Generating Image Descriptions

Problem Statement

Given a set of images, with its caption, create a predictive model which generates relevant caption for the unseen images.

Automatic caption generation

Input Data: Some Image



Dataset: MS COCO dataset of images annotated with captions

Model: Convolutional Neural Network followed by Recurrent Neural Network

Result:

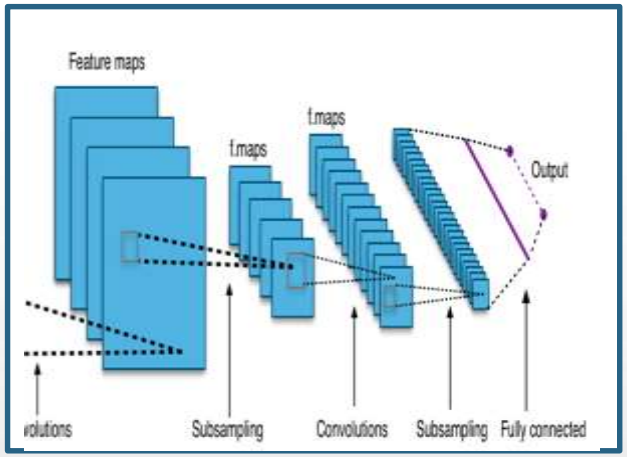
- **Accuracy = ~67%**
- **%of times correct caption was one of top four predicted captions = ~92 %**

AlgoAnalytics

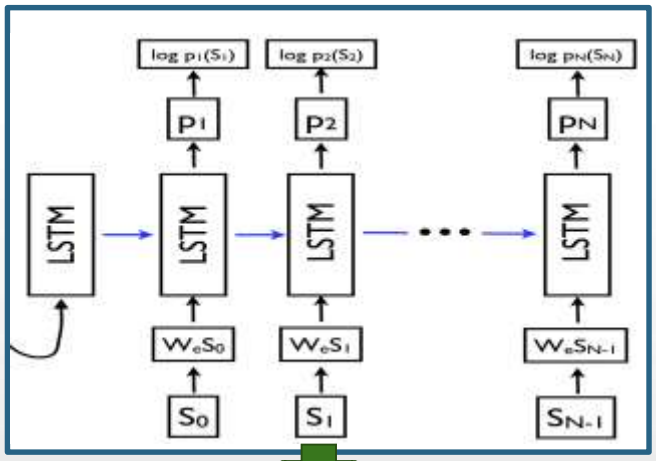
Methodology

Input Image

Deep Convolutional Neural Network



LSTM Language Modelling (RNN)



Group of People around
Next to each other
Cell phone holding
Crowd
Girl
Standing

word by word generation

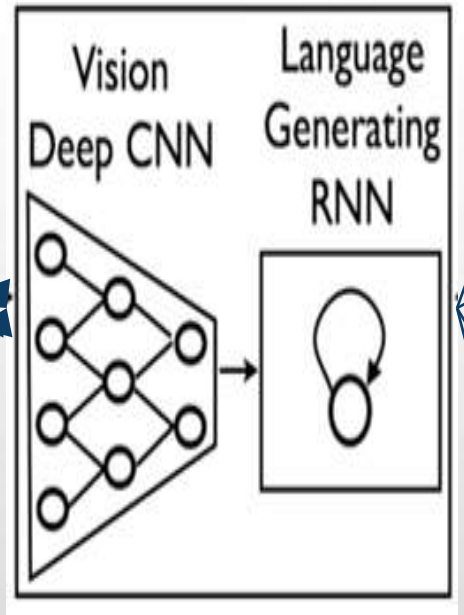
1. Group of people around each other holding cell phones
2. a group of people standing around each other
3. a group of people standing next to each other holding cell phones

Caption

AlgoAnalytics

Examples of tested images

Input Images



Captions with probability

- a man riding a wave on top of a surfboard . (p=0.036320)
- a person riding a surf board on a wave (p=0.016302)
- a man on a surfboard riding a wave. (p=0.010878)

- a group of men standing next to each other . p=0.004558)
- a group of people standing next to each other. (p=0.003918)
- a group of people standing in a room . (p=0.001977)

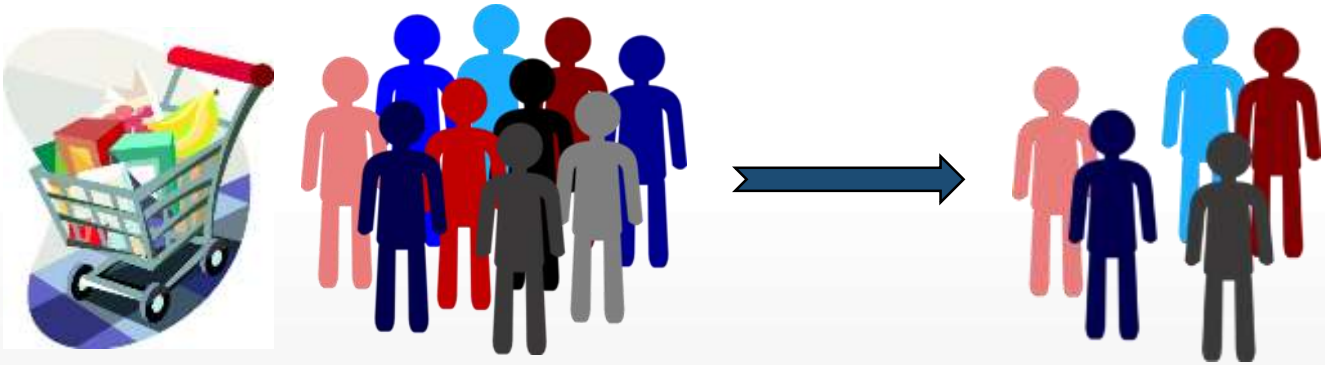
- a baseball player pitching a ball on top of a field . (p=0.003140)
- a baseball player pitching a ball on a field. (p=0.002312)
- a baseball player pitching a baseball on a field . (p=0.001413)

- a man in a hat and sunglasses is talking on a cell phone . (p=0.000018)
- a man with a hat and a hat on . (p=0.000016)
- a man with a hat and a hat on (p=0.000008)

AlgoAnalytics

Customer Churn Prediction

Take customers' past activities clickstream data to predict the customers' retention



Process Data → Compute Features → Train the classification model → Predict/Score

Predicted Returning Customers	• Target with loyalty programs
Predicted 'Not Returning' Customers	• Target with other offers and discounts

AlgoAnalytics

Customer Churn Prediction: Case study

The capability of **predicting a churning risk for important customers** leads to huge revenue benefits for every business.

Terabytes of Clickstream data

Handling large dataset?

Distributed File Systems
Cluster Computing

Results

66% Accuracy	51% Sensitivity	79% Specificity
• % of times model predicted customer churning activity correctly	• % of actually churned customers identified	• % of active customers identified

AlgoAnalytics

Interested in knowing more:

Contact us: info@algoanalytics.com

AlgoAnalytics