Learnings from productionizing a semi-supervised deep learning system model at the petabyte scale.

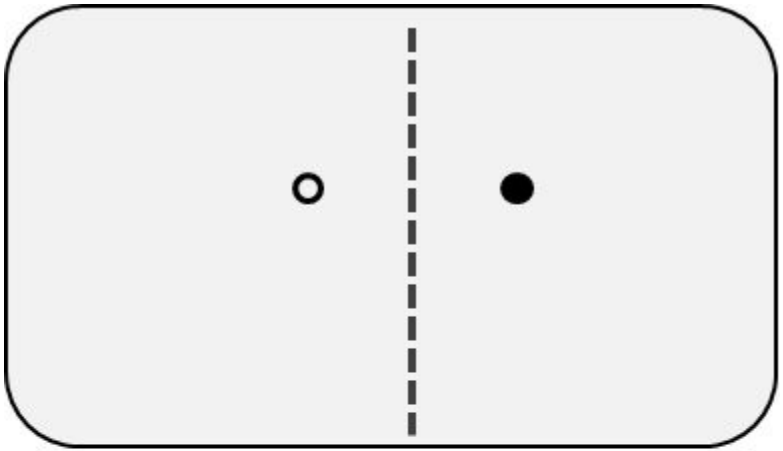
Samiran Roy

Senior Lead Data Scientist at Envestnet | Yodlee, Masters at IIT Bombay

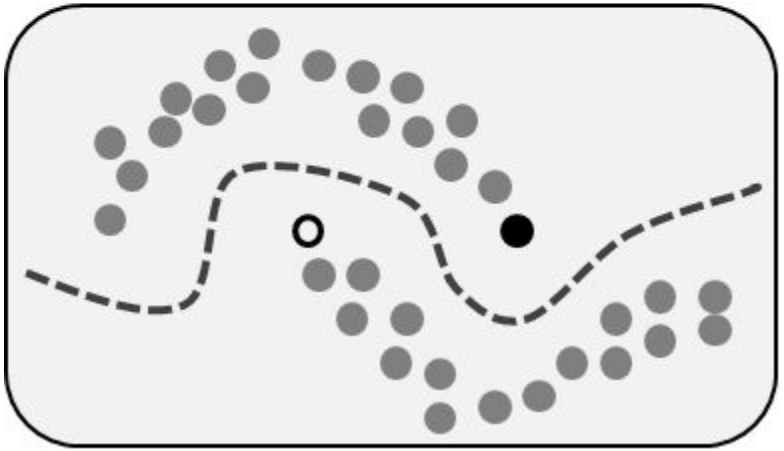
LinkedIn: samiranroy

Slides: http://bit.ly/pptsslprod

Supervised Setting



Semi-Supervised Learning



Yodlee Data Analytics - Scale

- 100 bn+ text datapoints
- 1 PB+ size
- 20 mn+ unique words
- 10+ classification challenges
- 3 mn+ classes*
- >90% precision/recall*

The Presentation

- Simple Ideas
- Common Sense - In hindsight

Tentative Outline

- Before
- During Model Creation
- After

The Power Law

debit purchase debit purchase mcdonalds debit purchase mcdonalds chicago il $40 on 19-10-2019

The Power Law

The Before Stage

Bunch of Data Points
Maybe a few labels

Where do you start?

Data Swimming

Data Swimming

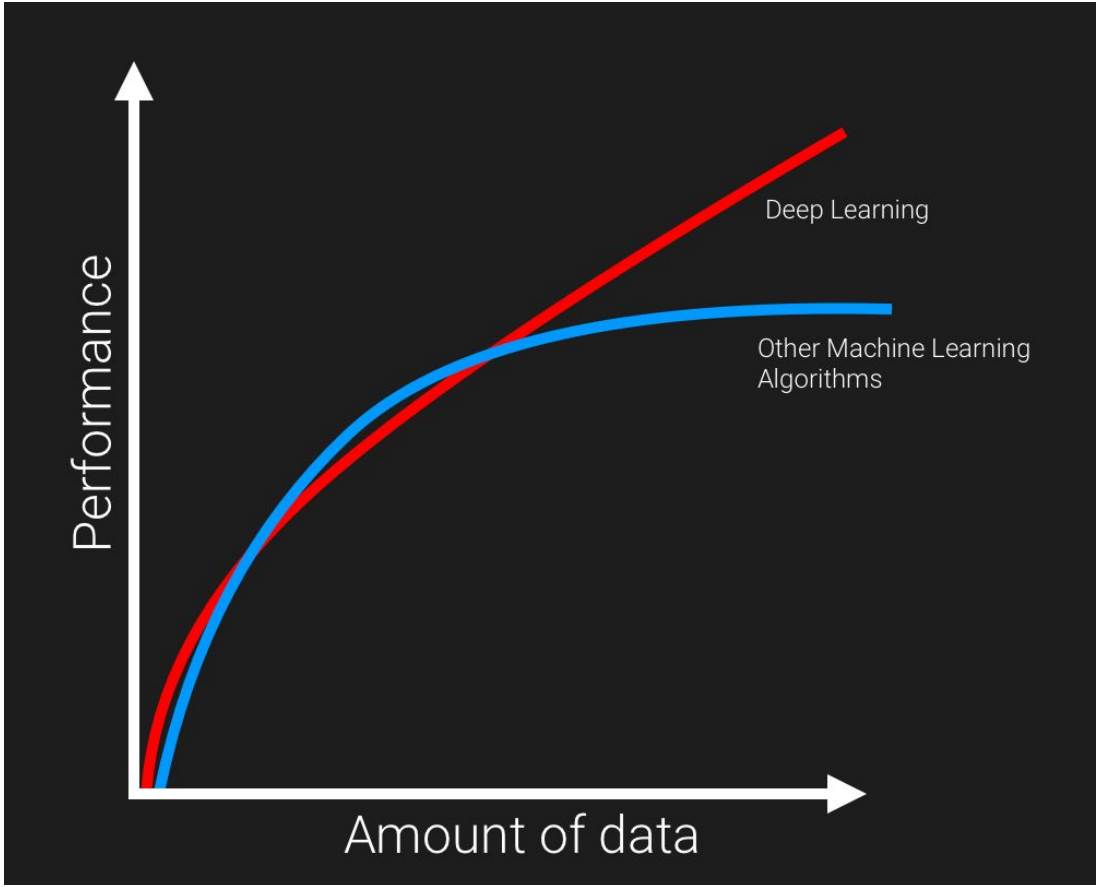
The **first step to training a neural net is to not touch any neural net code** at all and instead begin by thoroughly inspecting your data. **This step is critical**. I like to spend copious amount of time (measured in units of hours) scanning through thousands of examples, **understanding their distribution** and looking for patterns. Luckily, your brain is pretty good at this. One time I discovered that the data contained **duplicate examples**. Another time I found **corrupted images / labels**. I look for **data imbalances and biases**. I will typically also **pay attention to my own process for classifying the data**, which hints at the kinds of architectures we'll eventually explore. As an example - are very local features enough or do we need global context? How much variation is there and what form does it take? What variation is spurious and could be preprocessed out? Does spatial position matter or do we want to average pool it out? How much does detail matter and how far could we afford to downsample the images? How noisy are the labels?

~ Andrej Karpathy, (Senior Director of AI at Tesla)
(http://karpathy.github.io/2019/04/25/recipe/)

Accuracy?

Ref: https://gradientdescent.co/t/tesla-autonomy-day-watch-the-full-event/216/18

ML vs DL


Performance
Amount of data
Deep Learning
Other Machine Learning
Algorithms

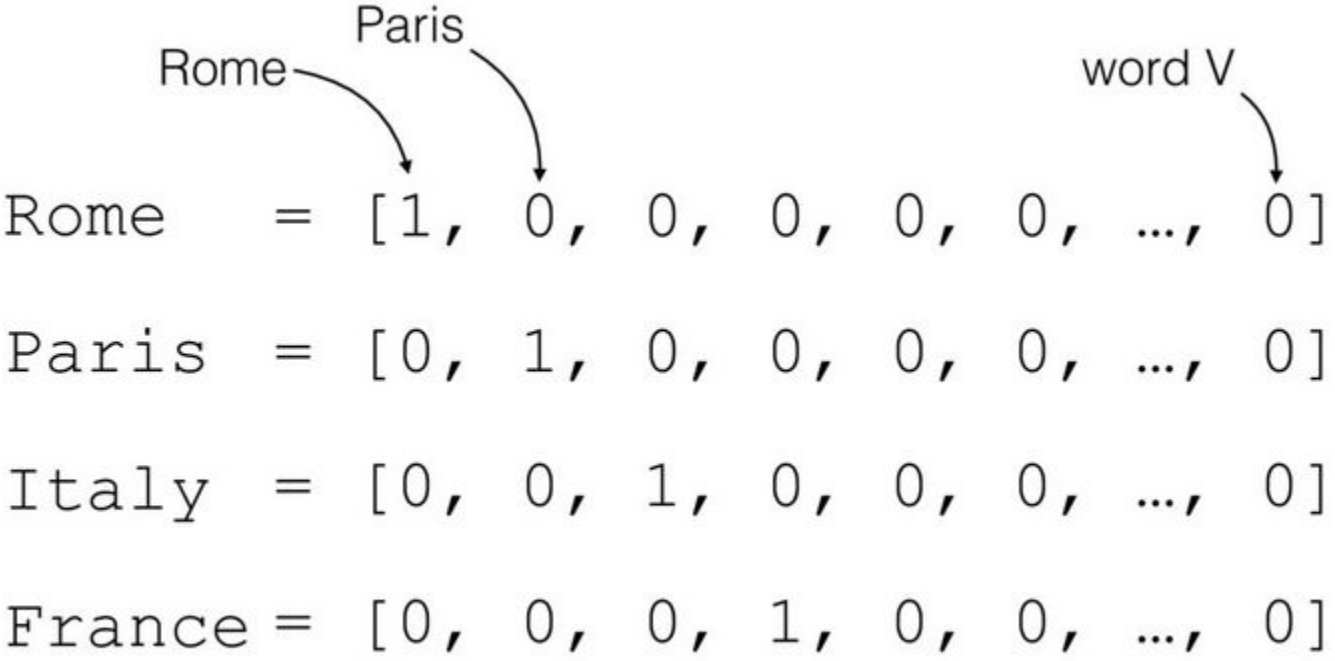
100 bn -> 1 bn (sample)

(Data driven - Domain Expertise)
Be very careful to avoid bias

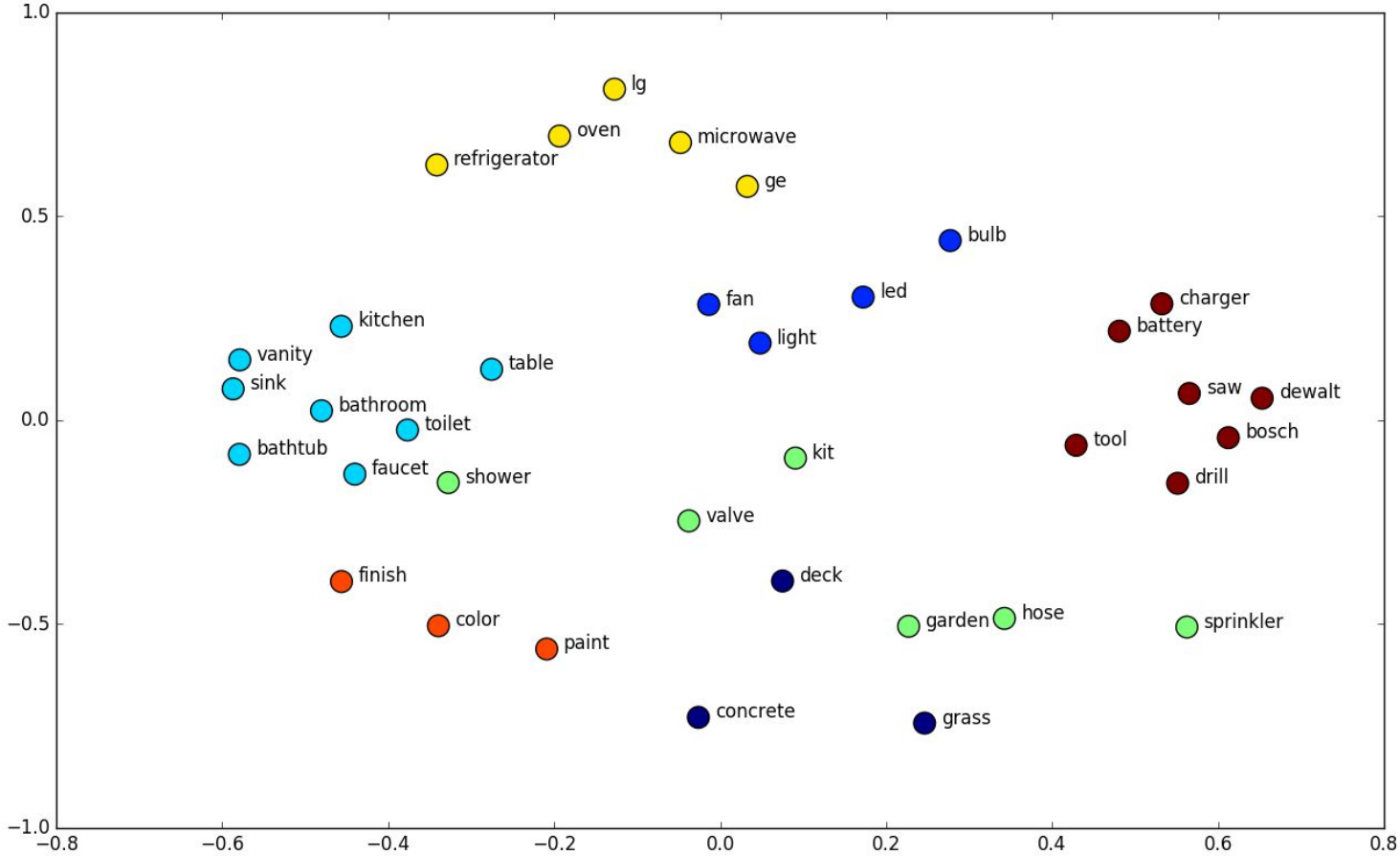
1bn -> 1 mn (labelled sample)

Machine learning based

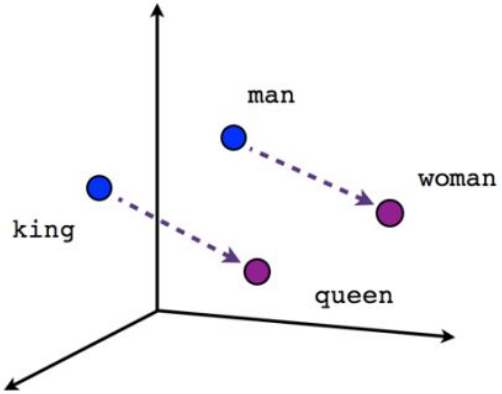
Representation is key



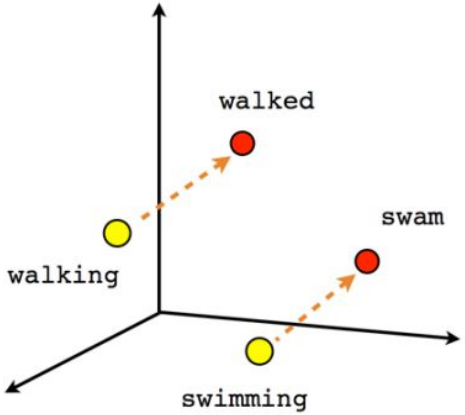
Word Embeddings



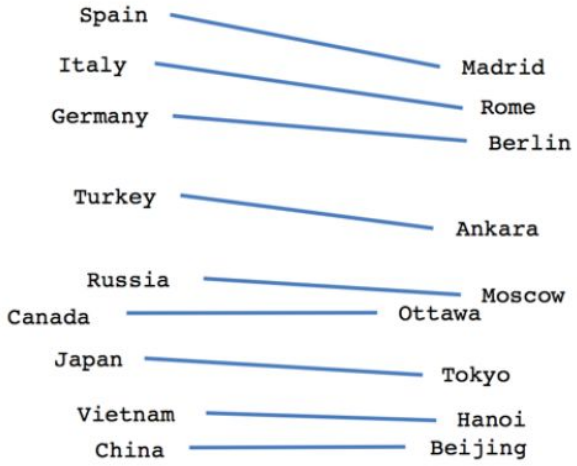
Vector Arithmetic



Male-Female

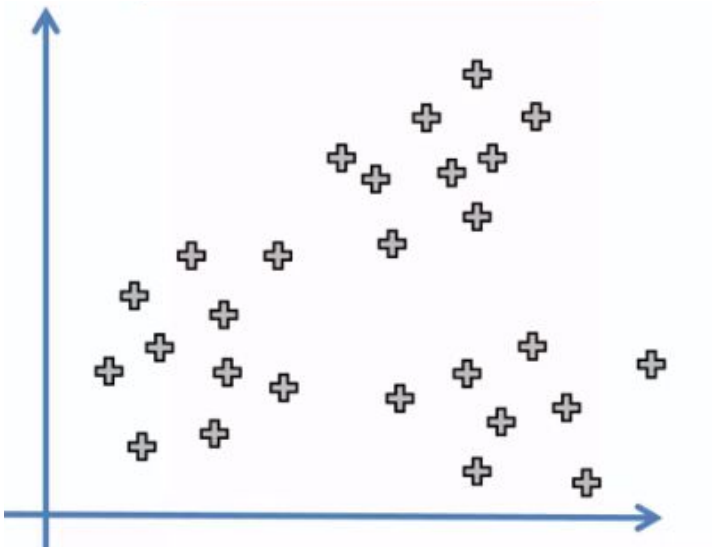


Verb tense

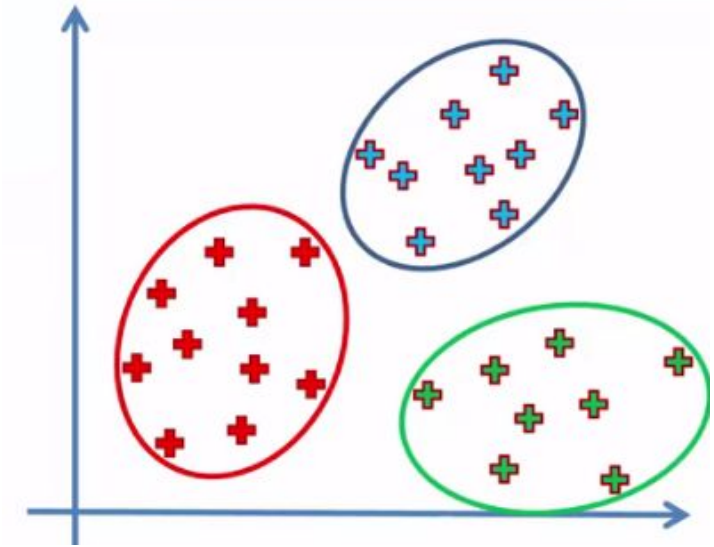


Country-Capital

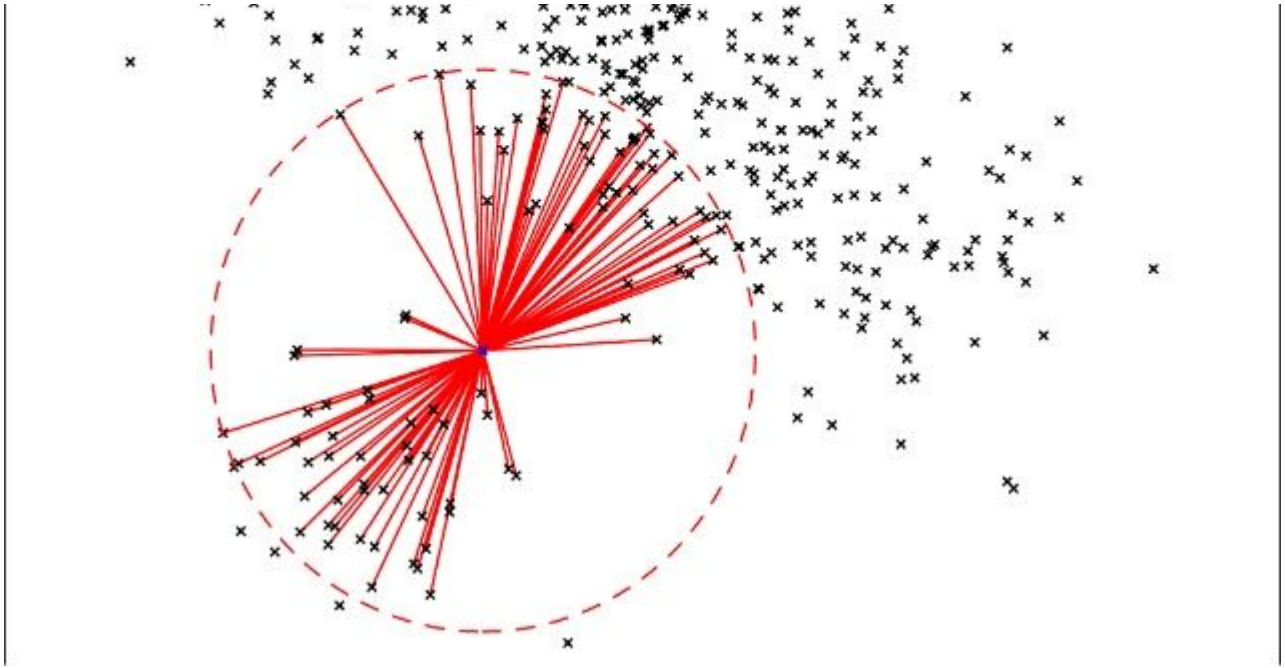
Visualizing embeddings



Unsupervised Learning (hdbscan, faiss)



Quantifying forgetfulness is important!
Nearest neighbor sampling



100 bn -> 1 bn -> 1 mn

Data Annotation

Semi Supervised Approach 1

Unsupervised word Embeddings (1bn) + Deep Learning Models (1mn)

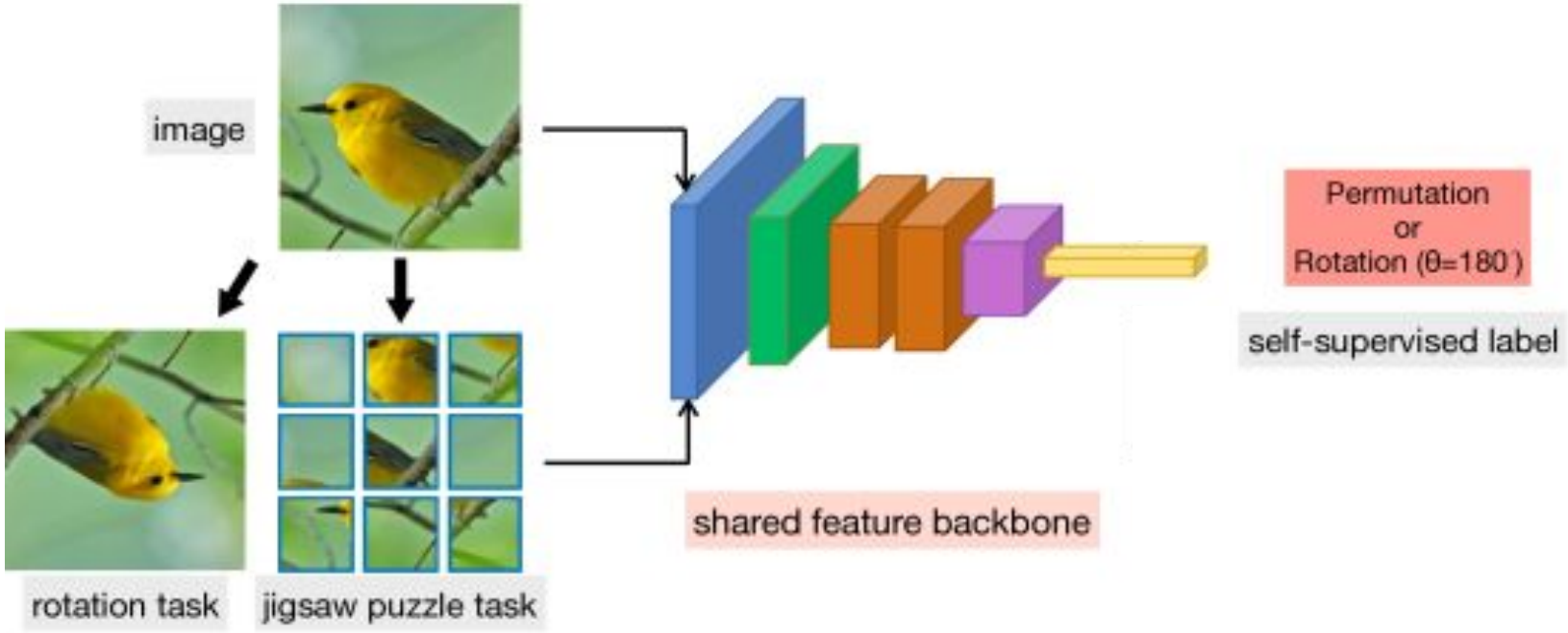
To read: https://arxiv.org/abs/1509.01626

Semi Supervised Approach 2

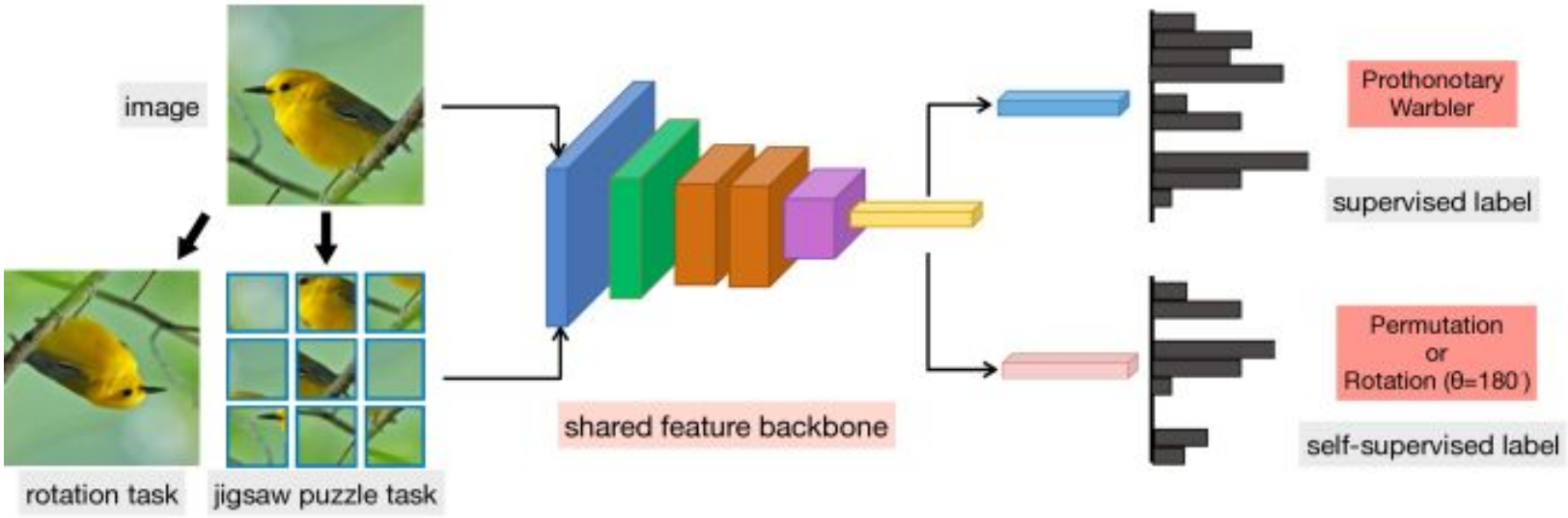
Jointly optimize Self-supervised + Supervised objective function

To read: https://arxiv.org/pdf/1905.03670v2.pdf

Self Supervision

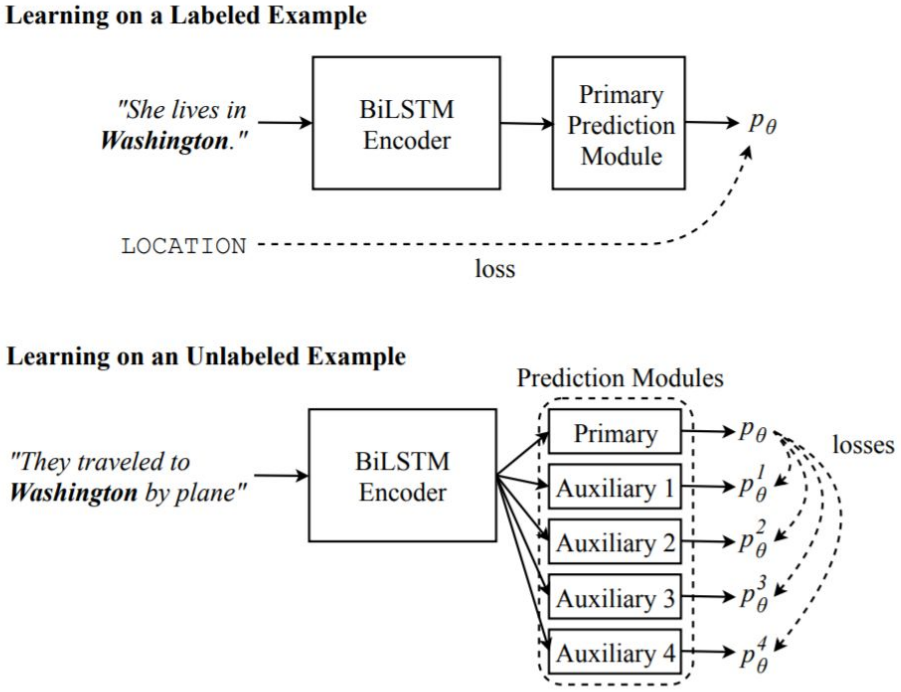


https://deepai.org/publication/boosting-supervision-with-self-supervision-for-few-shot-learning
https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf



https://deepai.org/publication/boosting-supervision-with-self-supervision-for-few-shot-learning

Cross View Training



Inputs Seen by Auxiliary Prediction Modules

Auxiliary 1: *They traveled to* _____

Auxiliary 2: *They traveled to* ***Washington*** _____

Auxiliary 3: _____ ***Washington*** *by plane*

Auxiliary 4: _____ *by plane*

Refer to: https://arxiv.org/pdf/1809.08370.pdf

(Semi-Supervised Sequence Modeling with Cross-View Training)

Practical Considerations

Cache frequently

Model from scratch

Consider *hidden* technical debt
(https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf)

Bonus: https://arxiv.org/abs/1801.09536

Tooling

1) Root Cause Analysis

Plug and play

Output metrics

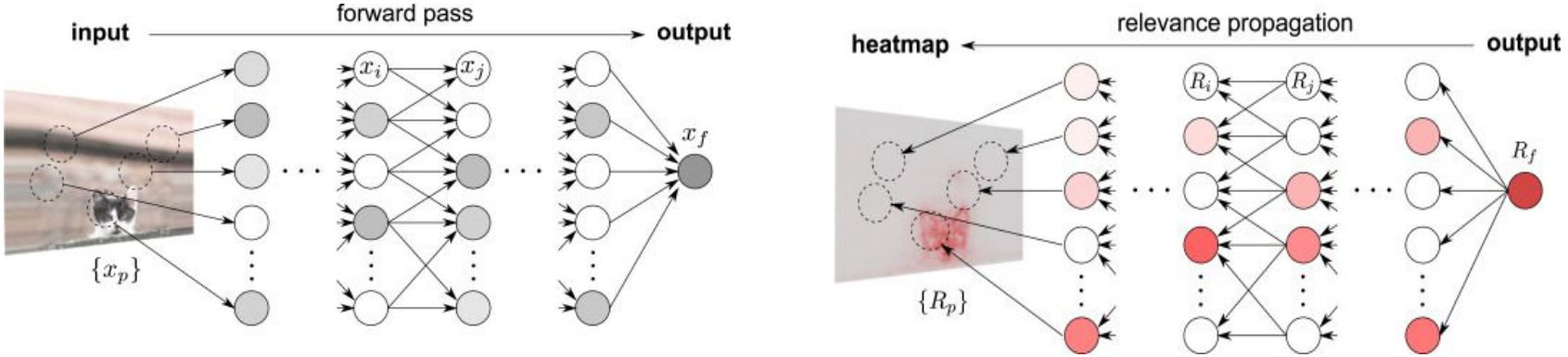
Confusion Matrices

2) Model Interpretation

Category: Sports

Alongside Tendulkar and Sehwag, several other retired international cricketers will be **seen in action**

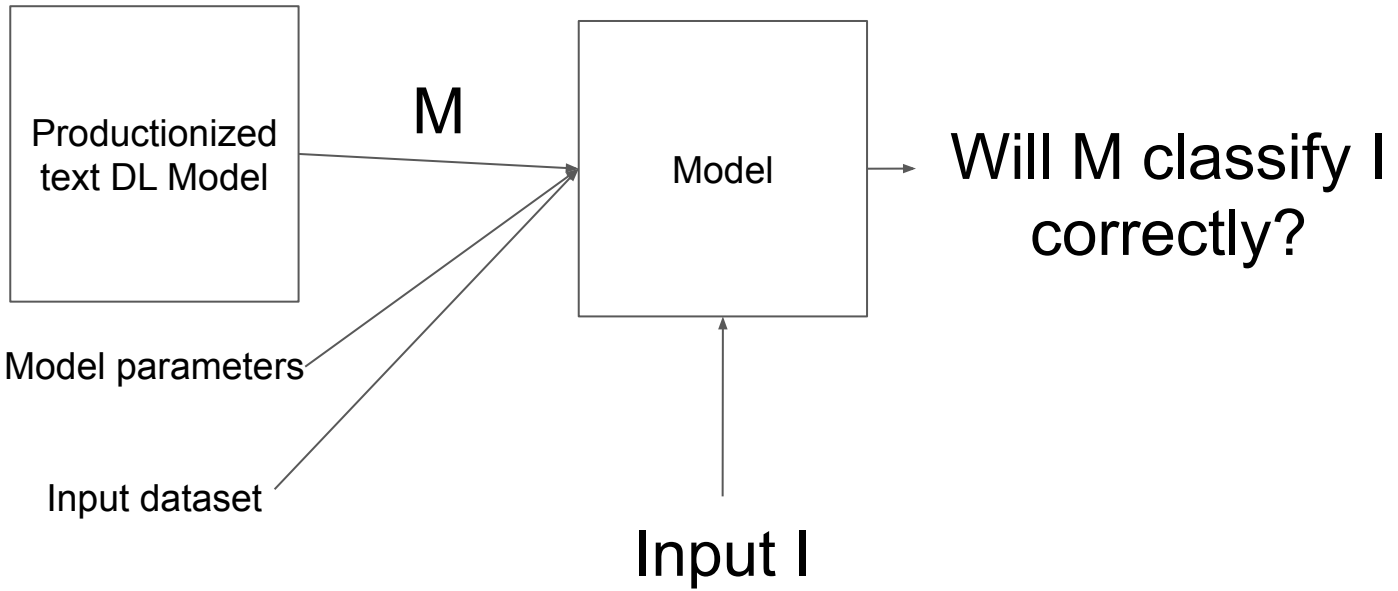
Alongside **Tendulkar** and **Sehwag**, several other retired international **cricketers** will be seen in action



https://indico.scc.kit.edu/event/344/contributions/2434/attachments/1258/1759/Talk_Samek.pdf

3) Automatic feedback loops

Ties in with Root Cause Analysis



3) Automatic feedback loops

1mn -> 1k problem

3) Automatic feedback loops

Ensemble of models:

Word based sampling

Embedding based sampling

Random Sampling

Automatic Validation

Model 1: Productionized, needs to meet SLAs.
~n millisecond per query

Model 2: High Precision Classifier, no compute constraints

https://arxiv.org/abs/1506.02142
https://arxiv.org/pdf/1709.01907.pdf

Compare the outputs of the two models

Thank you!

Slides: http://bit.ly/pptsslprod

Take away:

Power Law
Data Swimming
Semi Supervised Approaches
Detecting hidden technical debt