Sentiment Analysis with Transfer Learning in a Distributed Setting

Pramit Mallick, Jyotirmoy Mohapatra, Shucheng Yan

Agenda

1.Why Sentiment Analysis for Airline reviews

2.Data Understanding (Sentiment Treebank and Airline Reviews)

3.Modeling with Transfer Learning

4.Distributed Machine Learning Approach

5.Model evaluation

6.Conclusion and lessons learned

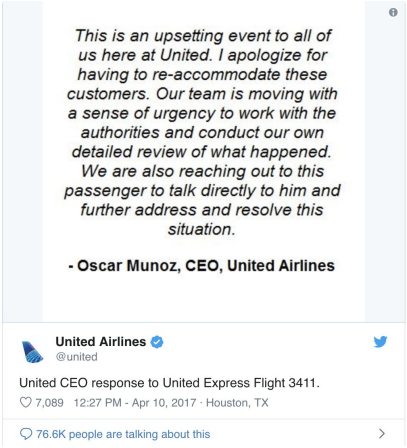
Why sentiment analysis in Airline Tweets

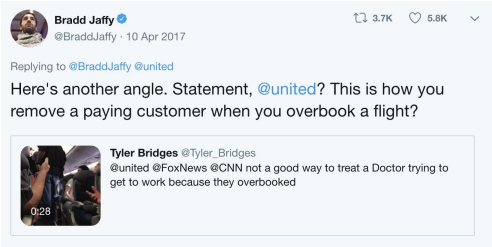
Public Relations

Marketing Analysis

Flying Experience feedback

Customer Service





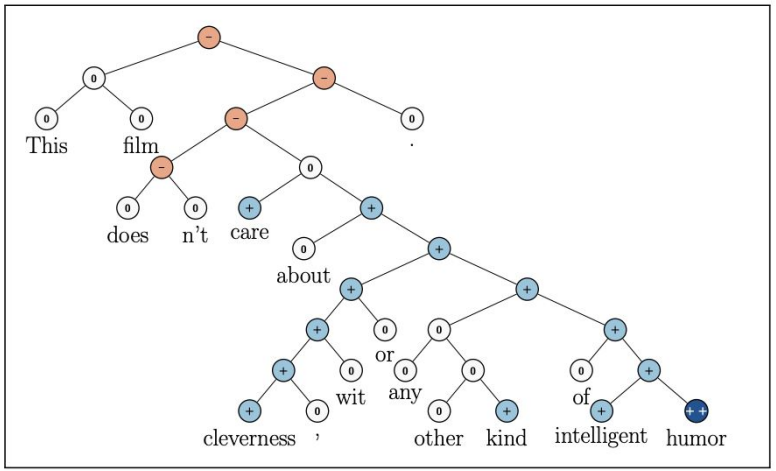
Data-Sentiment Tree (Stanford NLP Group)

Consists of 11,855 single sentences extracted from movie reviews

Each sentence is parsed with the Stanford parser

Includes fine grained sentiment labels for 215,154 phrases from those parse trees, each annotated by 3 human judges

First corpus with fully labeled parse trees that allows for a complete analysis of the compositional effects of sentiment in language



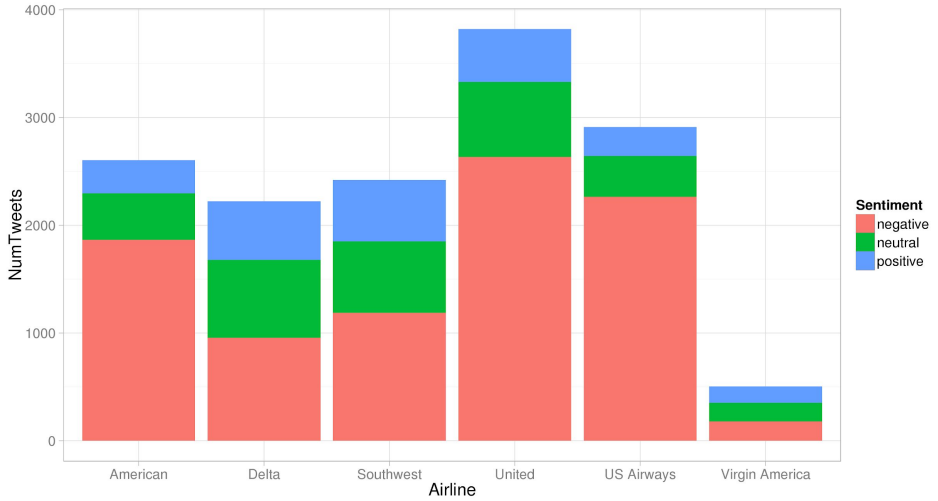
Data-Airline Tweets

Tweets tagged with airline names were scraped in February 2015

Labels: positive, negative, neural

14,641 labeled data points

@united thank you for getting our daughter home when @americanair Cancelled Flightled all their flights to Nashville
@united thanks for the re-upgrade to 1st class. It may be a 45 min flight, but it is appreciated.
@united makes total sense, except flight wasn't full :) I've got empty seats around me & overheads were more th
,Äú@united: @ItsAaronChriz Sorry to hear about your flight. Do you need help reFlight Booking Problems?,Äù
@united now we are trying to get to San Juan from Chicago O'Hare. Having lots of problems. May get a standby flight.
@united Alright, thank you. Is there a page that says the routes you have for each aircraft? Specifically the 787.
@united well sorta....we r trying to get to Aquadilla, PR but only 1 flight goes there a day. All are booked. UA Cancelle
@united why does it cost $547 to change the city of origin when the same flight on http://t.co/8FMZZOltv9 costs $165
@united I just sent a long note with some suggestions. Thanks for getting back to me.
@united there are at least 3 of us on UA1564 at ORD waiting to deplane to catch UA4232 to CLT. Any chance of waiti
@united @parryaftab done thnx
@united, I'm still frustrated I gave up my seat & the promised Travel Certificate was withheld w/o explanation. H
@united gotta love giving up 1st class upgrade b/c flight delayed, to get another flight (also delayed) just to ensure I m



Modeling with Transfer Learning

- What is Transfer Learning
- Types of Transfer Learning
- Model Architecture
- Hyperparameter tuning

Transfer Learning

> *Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned.*

— Chapter 11: Transfer Learning, Handbook of Research on Machine Learning Applications, 2009.

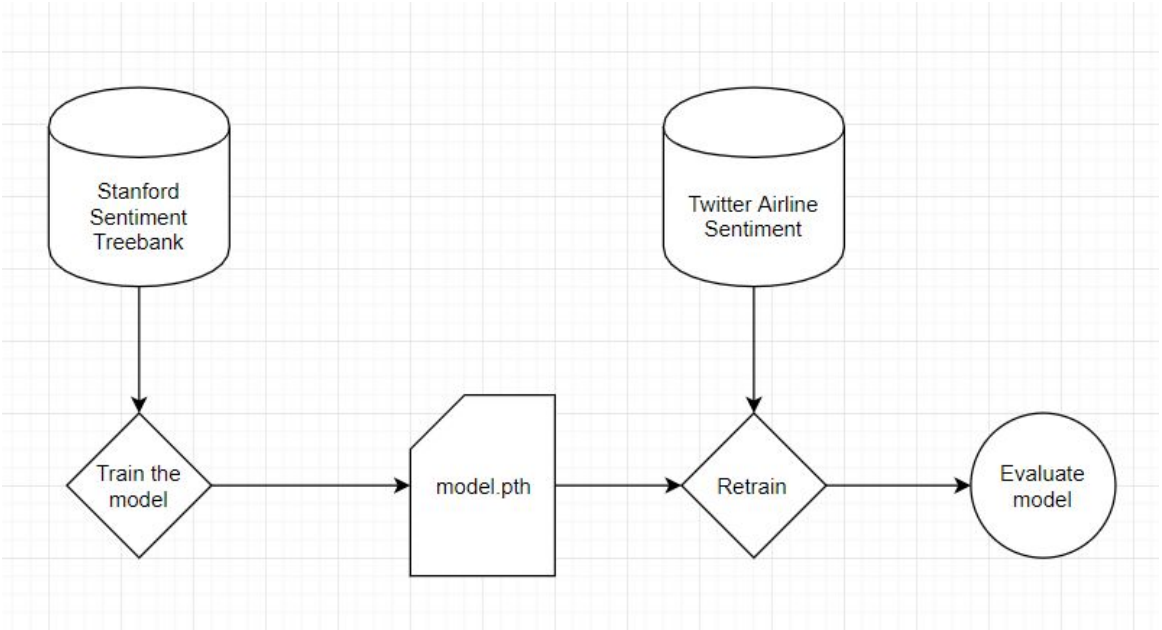
It is motivated by human learning. People can often transfer knowledge learnt previously to novel situations.

- Know how to ride a motorbike Learn how to ride a car
- Know how to play classic piano Learn how to play jazz piano
- Know math and statistics Learn machine learning

Types of Transfer Learning

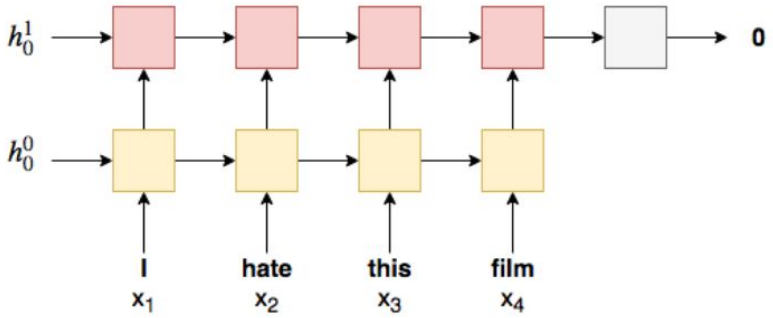
Transfer learning approaches	**Description**
Instance-transfer	To re-weight some labeled data in a source domain for use in the target domain
Feature-representation-transfer	Find a “good” feature representation that reduces difference between a source and a target domain or minimizes error of models
Model-transfer	Discover shared parameters or priors of models between a source domain and a target domain
Relational-knowledge-transfer	Build mapping of relational knowledge between a source domain and a target domain.

Our Approach



Model Architecture

- 2 Layer Bi-LSTM
- Word embeddings initialized with Glove 6B 100d
- Vocabulary size 25,000
- Hyperparameters Tuned:
  - Optimizer
  - Learning Rate
  - Hidden Dimension
  - Dropout



Hyperparameter Tuning - Optimizer and Learning Rate

Optimizer	Learning Rate	Train Accuracy	Validation Accuracy
Adam	0.1	51.33	62.67
Adam	0.01	90.95	76.24
Adam	0.001	90.51	78.92
Adam	0.0001	78.68	75.43
SGD	0.1	69.02	68.51
SGD	0.01	62.64	62.67
SGD	0.001	62.65	62.67
SGD	0.0001	62.65	62.67

Hyperparameter Tuning - Hidden Dimension

Hidden Dimension	Train Accuracy	Validation Accuracy	Number of Parameters
128	90.51	78.92	2,163,155
256	89.05	78.80	3,843,283
512	92.28	78.97	10,349,267

Hyperparameter Tuning - Dropout

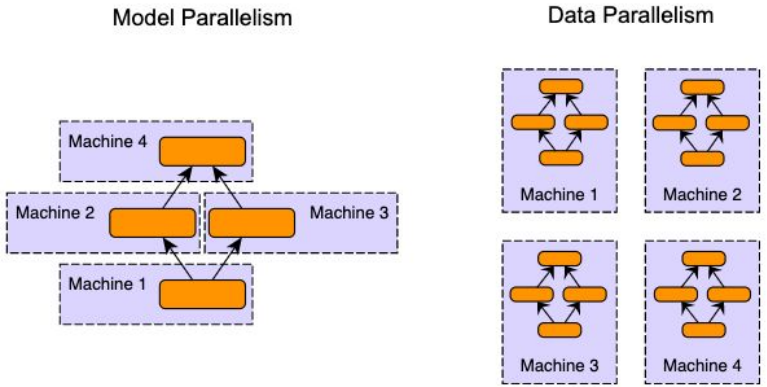
Dropout	Train Accuracy	Validation Accuracy
0.3	92.28	78.97
0.5	90.21	79.56
0.7	81.28	78.21

Distributed Machine Learning

Model vs Data Parallelism

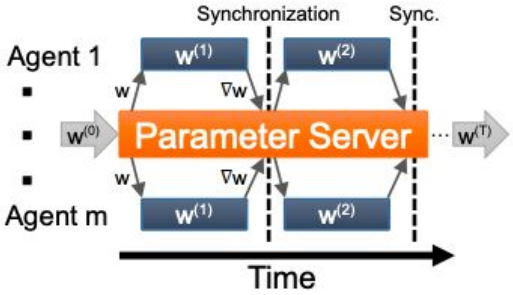
- Model - distribute the model, same data
- Data - distribute the data, same model

In both cases, we get some form of parallelism that helps us the decrease training time in the hopes of faster model convergence

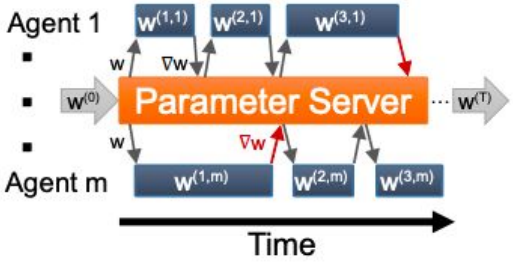


Distributed Machine Learning

- Updates to the model parameters can be handled synchronously, i.e the worker nodes run a mini-batch and send the gradients to the parameter server which aggregates and updates the model weights while the worker waits. The worker nodes then receive the latest parameters from the server.
- The worker nodes send the gradients and the server does an update on the model and sends back the weights without synchronizing or aggregating for others. This paradigm of update takes much longer to converge as it doesn't strictly follow the mini-batch stochastic gradient descent estimates.



(a) Synchronous, Parameter Server



(c) Asynchronous, Parameter Server

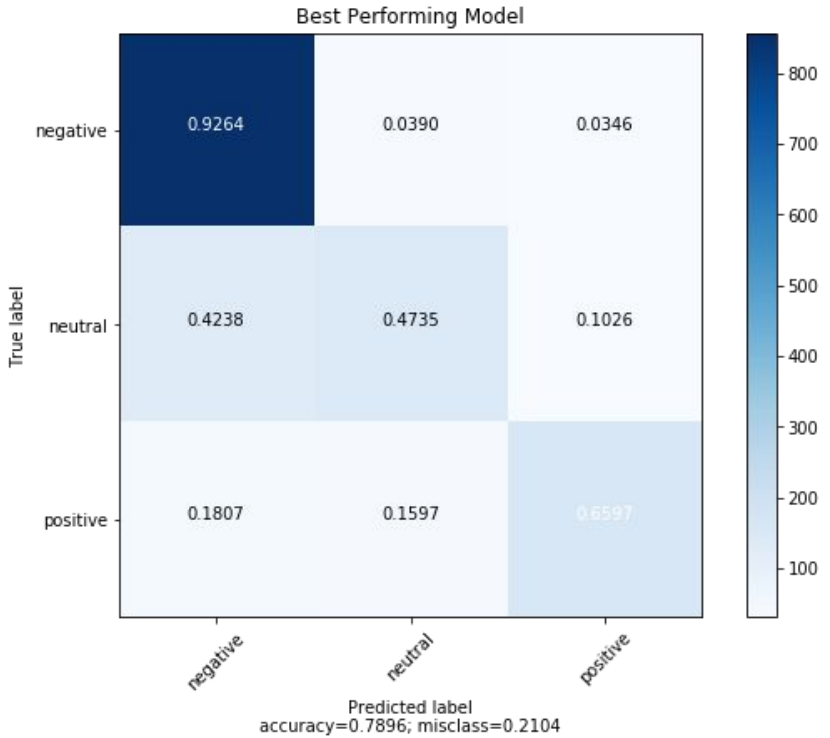
Best Performing Model

Optimizer = Adam

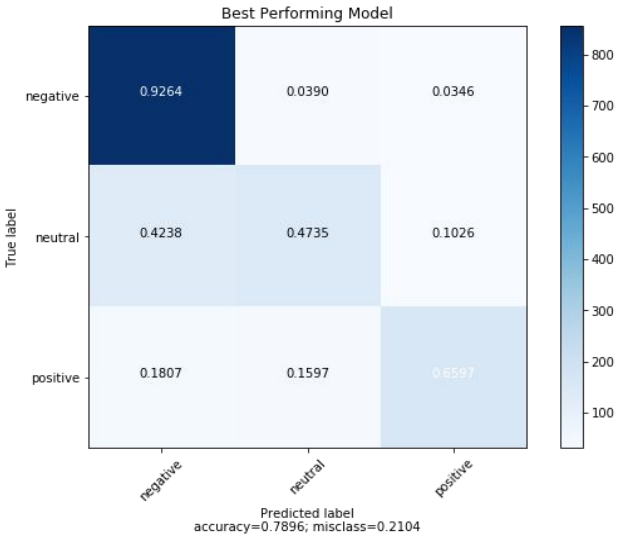
Learning Rate = 0.001

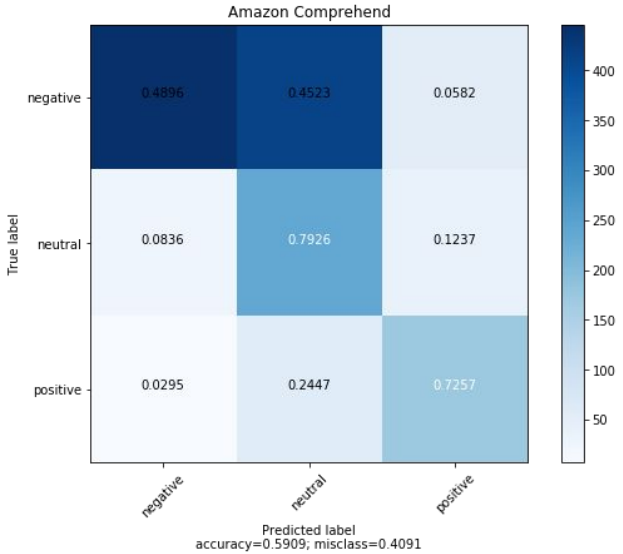
Dropout = 0.3

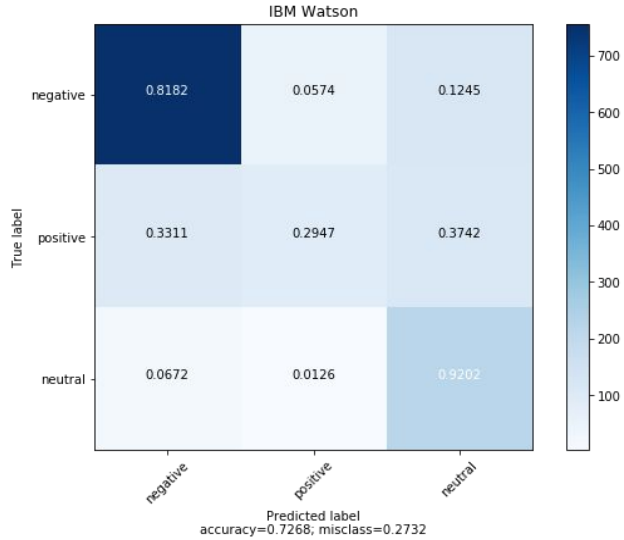
Hidden Dimension = 512



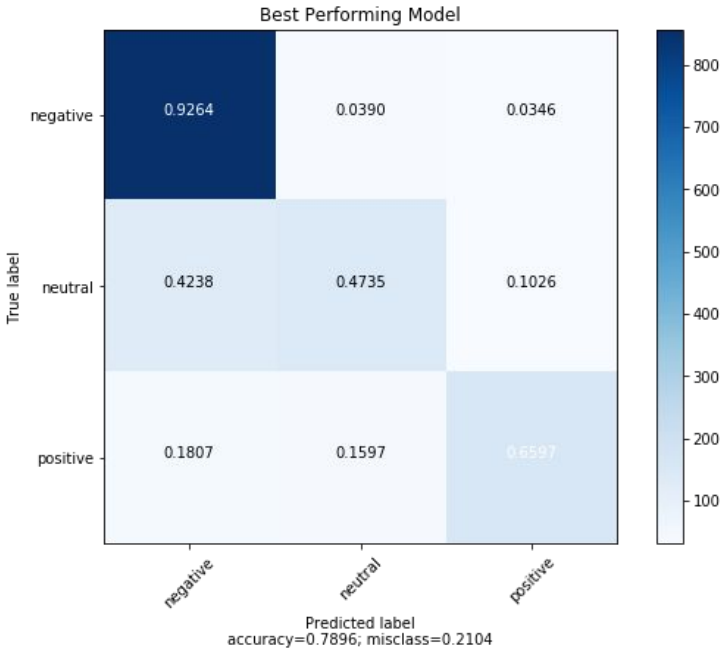
SaaS services

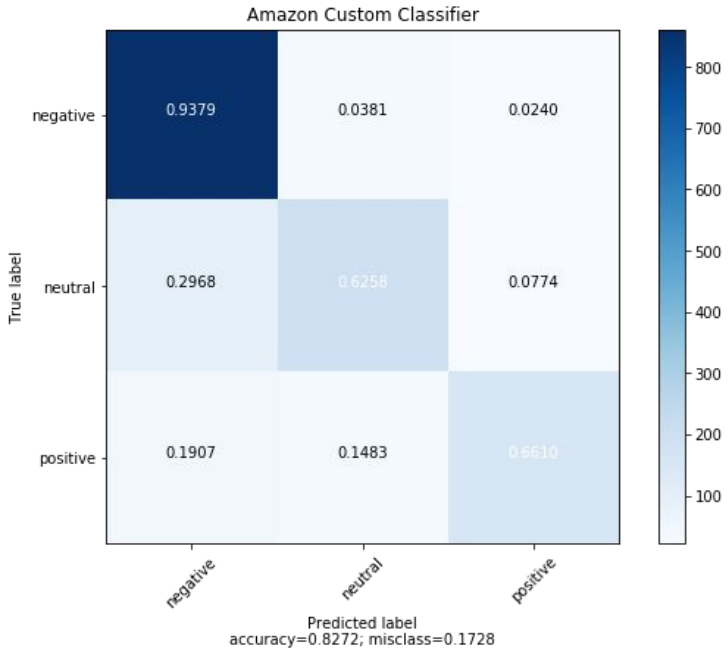






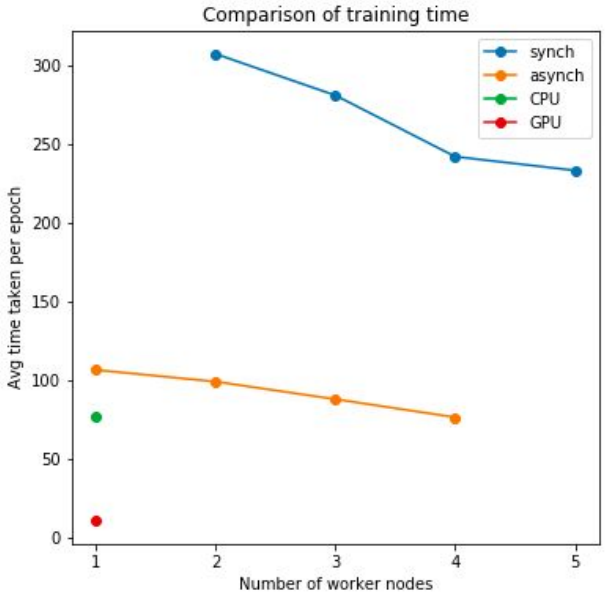
Custom Classifier using Amazon Comprehend





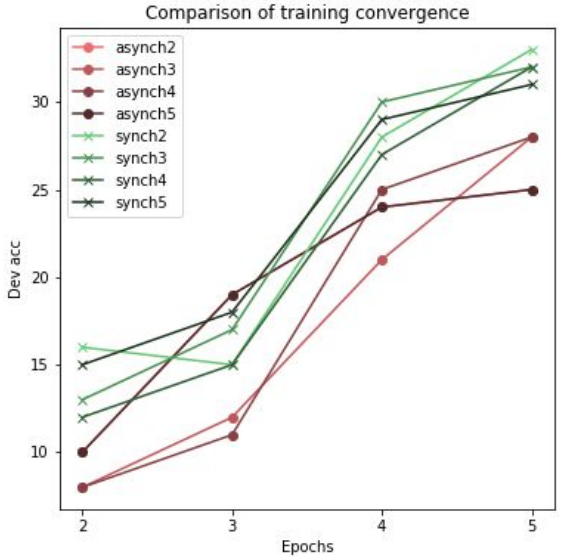
Distributed

- Avg time decreases with more workers (each worker works with lesser data)
- Synchronous SGD takes more time due to blocking waits when the parameter server aggregates the gradients
- We expect that as the number of nodes is increased, the time taken would stabilize and then increase. This is a natural behavior as, even though, the data size per node gets decreased, the entire system would mostly be occupied in communications with the parameter server and the latencies involved therewith would out benefit the gains from parallelization.



Distributed

- While we observe the obvious general rising trend in both cases, it is interesting to note that the synchronous SGD cases have higher accuracies.
- This is because asynchronous model parameters updates do not necessarily satisfy the mini-batch gradient descent estimates for convergence. Each update/step from every worker denotes a different descents due to stale parameters and thus convergence in the case of asynchronous SGD would take much longer than a synchronous one.



Conclusion

Domain specific data for training is important!

When domain specific data is not readily available or not labelled properly, transfer learning can be used to augment the performance of the model.

The synchronous data parallelism mode of distributed training can help improve the training time for faster model convergence.

Thank you!