

This is your **last** free member-only story this month.
[Sign up for Medium and get an extra one](#)



Prakhar Ganesh · Follow

Aug 12, 2019 · 4 min read ★

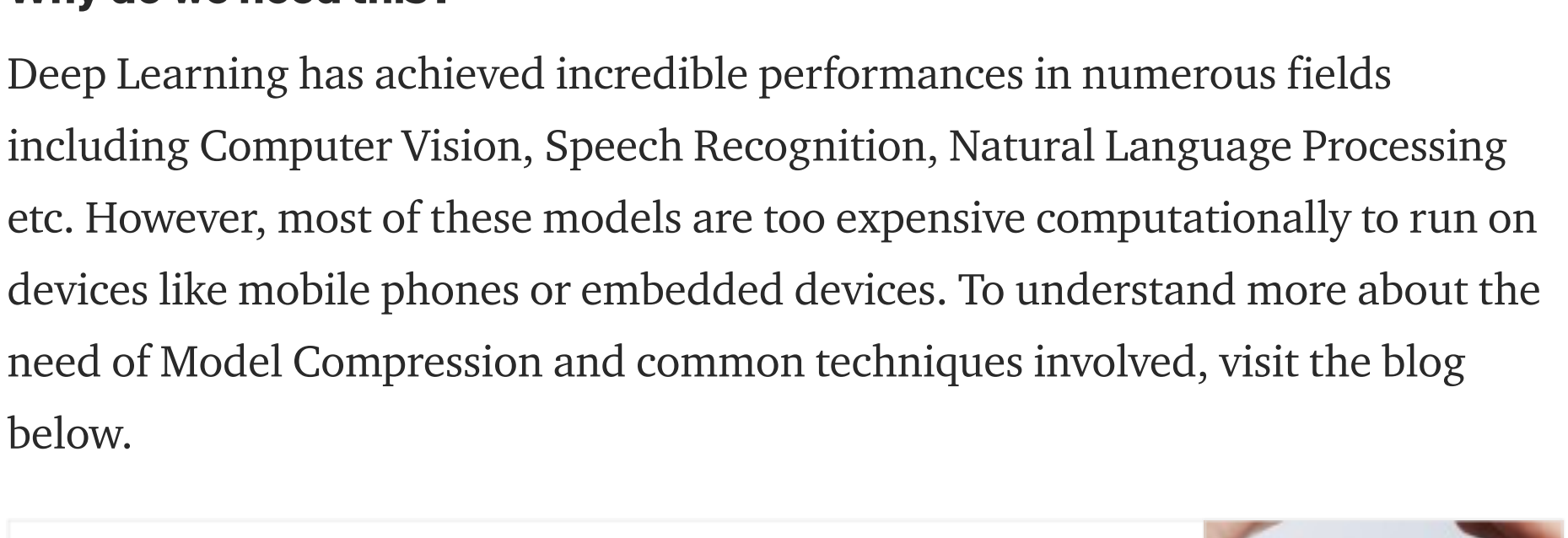
Knowledge Distillation : Simplified

Take a peek into the world of Teacher Student networks

What is Knowledge Distillation?

Neural models in recent years have been successful in almost every field including extremely complex problem statements. However, these models are huge in size, with millions (and billions) of parameters, and thus cannot be deployed on edge devices.

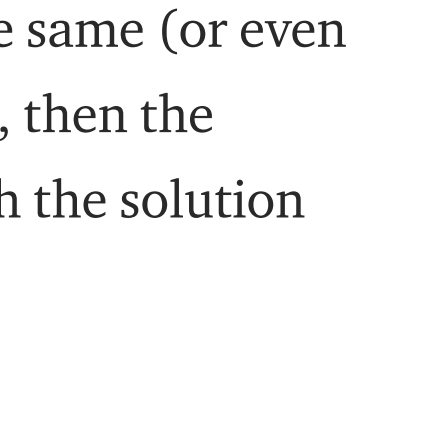
Knowledge distillation refers to the idea of model compression by teaching a smaller network, step by step, exactly what to do using a bigger already trained network. The ‘soft labels’ refer to the output feature maps by the bigger network after every convolution layer. The smaller network is then trained to learn the exact behavior of the bigger network by trying to replicate it’s outputs at every level (not just the final loss).



Why do we need this?

Deep Learning has achieved incredible performances in numerous fields including Computer Vision, Speech Recognition, Natural Language Processing etc. However, most of these models are too expensive computationally to run on devices like mobile phones or embedded devices. To understand more about the need of Model Compression and common techniques involved, visit the blog below.

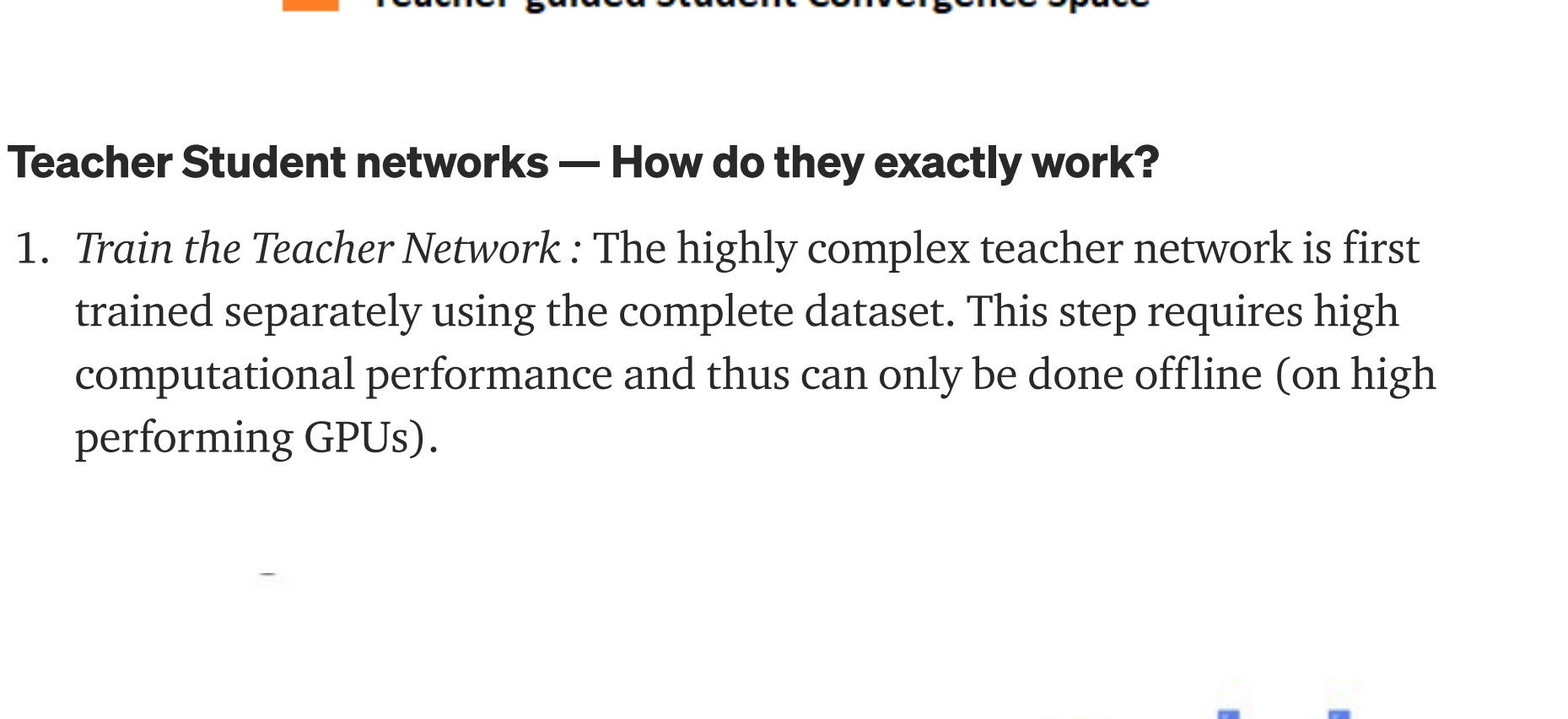
Deep Learning — Model Optimization and Compression: Simplified
Take a peek into the domain of compression, pruning and quantization of state-of-the-art Machine Learning models
towardsdatascience.com



How is this different from training a model from scratch?

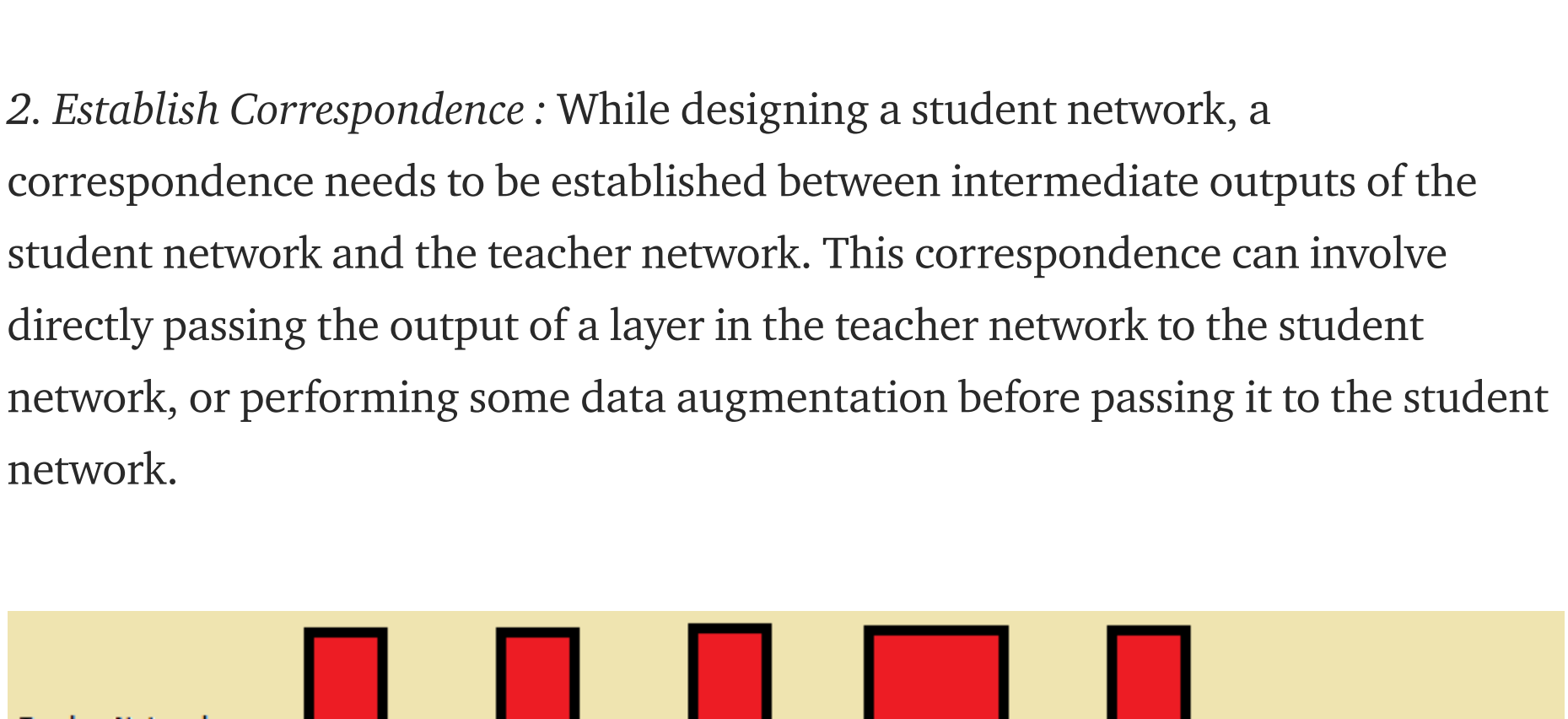
Obviously, with more complex models, the theoretical search space is larger than that of a smaller network. However, if we assume that the same (or even similar) convergence can be achieved using a smaller network, then the convergence space of the Teacher Network should overlap with the solution space of the student network.

Unfortunately, that alone does not guarantee converge for the student network at the same location. The student network can have a convergence which might be hugely different from that of the teacher network. However, if the student network is guided to replicate the behavior of the teacher network (which has already searched through a bigger solution space), it is expected to have its convergence space overlapping with the original Teacher Network convergence space.



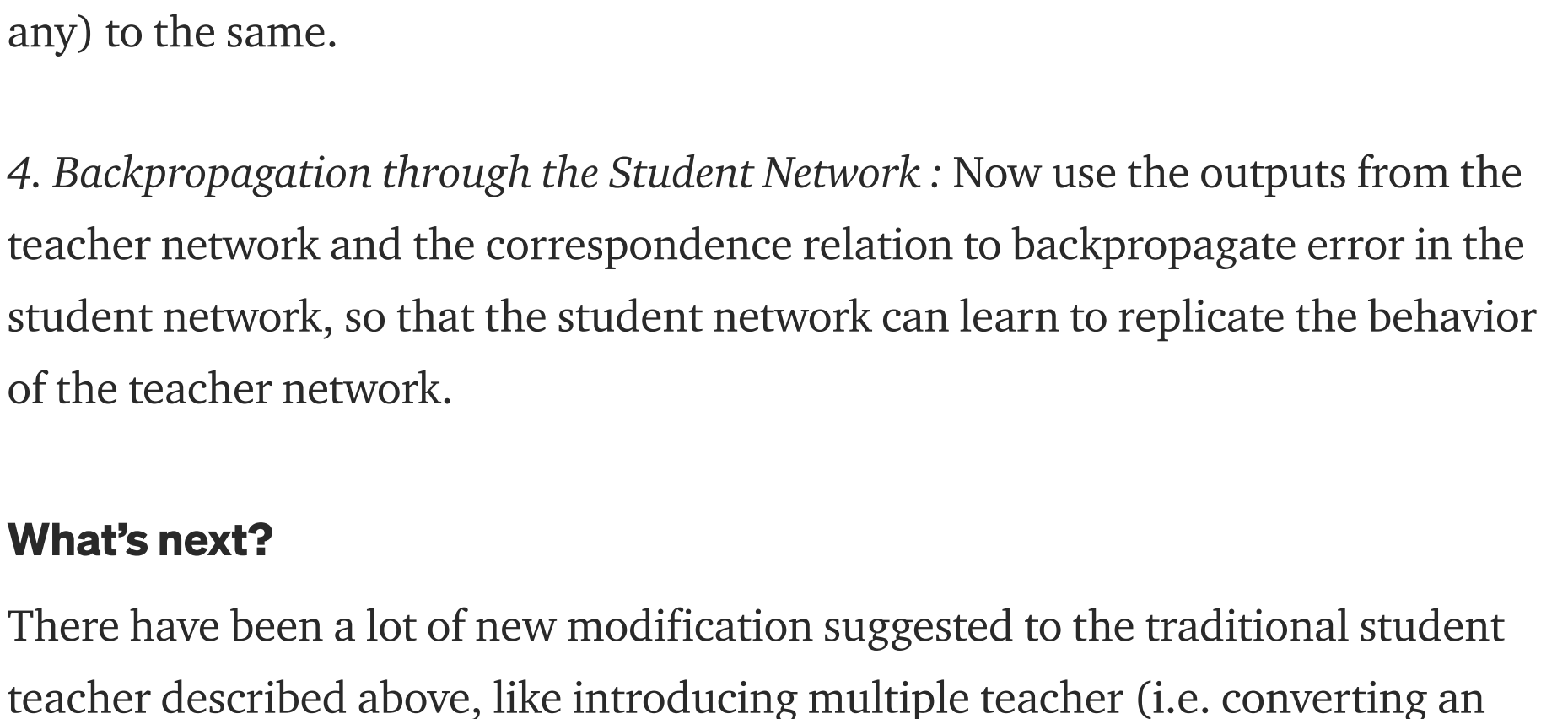
Teacher Student networks — How do they exactly work?

1. *Train the Teacher Network* : The highly complex teacher network is first trained separately using the complete dataset. This step requires high computational performance and thus can only be done offline (on high performing GPUs).



An example of a highly complex and Deep Network which can be used as a teacher network : GoogleNet

2. *Establish Correspondence* : While designing a student network, a correspondence needs to be established between intermediate outputs of the student network and the teacher network. This correspondence can involve directly passing the output of a layer in the teacher network to the student network, or performing some data augmentation before passing it to the student network.



An example of establishing correspondence

3. *Forward Pass through the Teacher network* : Pass the data through the teacher network to get all intermediate outputs and then apply data augmentation (if any) to the same.

4. *Backpropagation through the Student Network* : Now use the outputs from the teacher network and the correspondence relation to backpropagate error in the student network, so that the student network can learn to replicate the behavior of the teacher network.

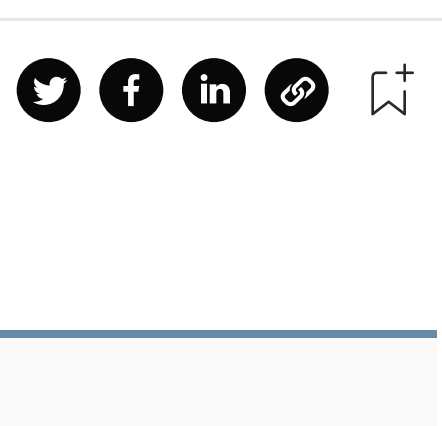
What's next?

There have been a lot of new modification suggested to the traditional student teacher described above, like introducing multiple teacher (i.e. converting an ensemble into a single network), introducing a teaching assistant (the teacher first teaches the TA, who then in turn teaches the student) etc. However, the field is still pretty young and is quite unexplored in many dimensions.

... ..

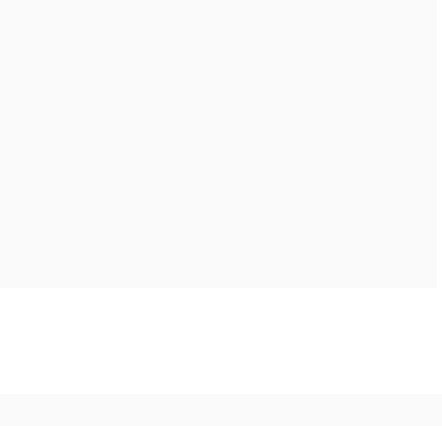
This blog is a part of an effort to create simplified introductions to the field of Machine Learning. Follow the complete series here

Machine Learning : Simplified
Know it before you dive in
towardsdatascience.com



Or simply read the next blog in the series

Growing your own RNN cell : Simplified
Take a peek into the 'deep' world of a single RNN cell
towardsdatascience.com



... ..


References


[1] Wang, Junpeng, et al. “DeepVID: Deep Visual Interpretation and Diagnosis for Image Classifiers via Knowledge Distillation.” *IEEE transactions on visualization and computer graphics* 25.6 (2019): 2168–2180.


[2] Mirzadeh, Seyed-Iman, et al. “Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher.” *arXiv preprint arXiv:1902.03393* (2019).


[3] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network.” *arXiv preprint arXiv:1503.02531* (2015).

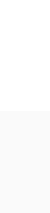
[4] Liu, Xiaodong, et al. “Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding.” *arXiv preprint arXiv:1904.09482* (2019).

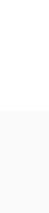
 442

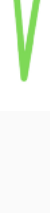
 4







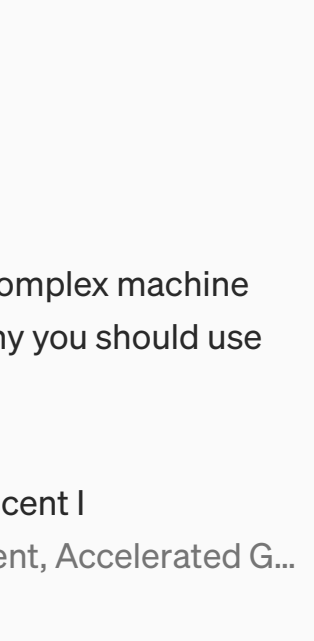




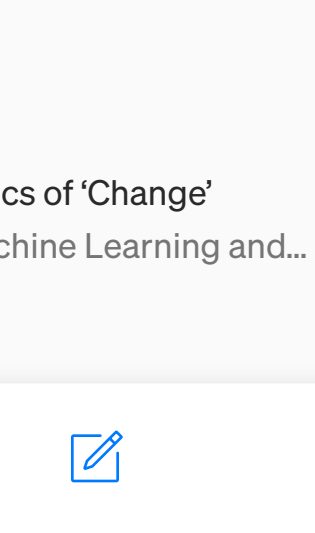
Sign up for The Variable
By Towards Data Science
Every Thursday, the Variable delivers the very best of Towards Data Science: from hands-on tutorials and cutting-edge research to original features you don't want to miss. [Take a look.](#)
[Get this newsletter](#)

More from Towards Data Science
Your home for data science. A Medium publication sharing concepts, ideas and codes.
Chintan Trivedi · Aug 12, 2019 ★

Proximal Policy Optimization Tutorial (Part 1/2: Actor-Critic Method)
Let's code from scratch a Reinforcement Learning football agent! — Welcome to the first part of a math and code tutorial series. I'll be showin...
Machine Learning · 7 min read



Serverless Recommendation System using PySpark and GCP
Behind the scenes of my online movie recommendation system and how it interacts with Google Cloud Platform. — “How Netflix predicts my taste?...
Machine Learning · 8 min read



Statistical Distributions
Breaking down discrete and continuous distributions and looking into how data scientists can apply statistics most efficiently. — What is a Probability Distribution? A probability distribution is a mathematical function that...
Data Science · 7 min read

More Data, More Sheets API
How the Google Sheets API uses magic (and code) to fill your Sheet with visual data. — As a huge fan of Yo-Yo Ma, I've been playing around with the Google Vision API by passing through an image of a cello. After obtaining...
Java Script · 6 min read

Sentiment analysis of the lead Characters on F.R.I.E.N.D.S
I am one of the biggest lovers of the American sitcom — FRIENDS ever since I started watching this show. Recently, I started looking into the...
Data Science · 6 min read

[Read more from Towards Data Science](#)

More from Medium

 RNN Simplified- A beginner's guide
Recurrent Neural Network(RNN) is a popular...
0.5959 · 1 · 0.1421 · world.com

 How to understand your complex machine learning algorithm, and why you should use SHAP.

 Know more about Machine Learning
This is a summarised introduction for machi...
0.5959 · 1 · 0.1421 · world.com

 Variations of Gradient Descent I
Stochastic Gradient Descent, Accelerated G...

 Five PyTorch Tensor Functions that are very useful and interesting for start

 Car costs prediction with PyTorch
Using Linear Regression

 Visualizing function approximation using dense neural networks in 1D, Part II

 Calculus—The Mathematics of 'Change'
How Calculus enables Machine Learning and...

Home

Search

Write