

Anatomy and Interpretability of Neural Networks

Leon Yin ~ Data Scientist | Research Engineer

SMaPP and CDS

PRG 2017-11-15

Today's talking points:

How do Neural Networks work?

How can we see what they're learning?

Discussion about training data and policy.

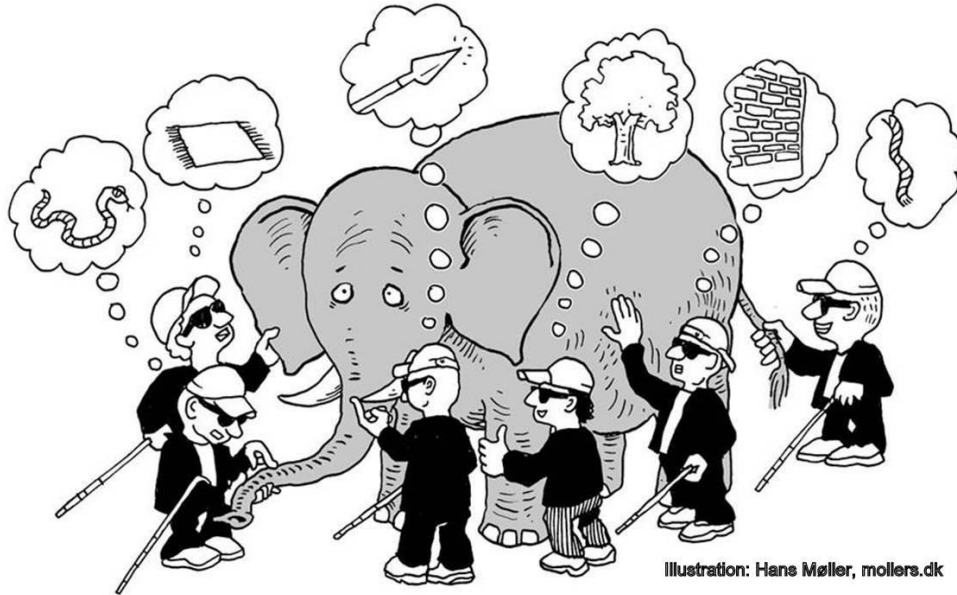
First of all

All models are wrong, but some are useful!

Neural Networks:

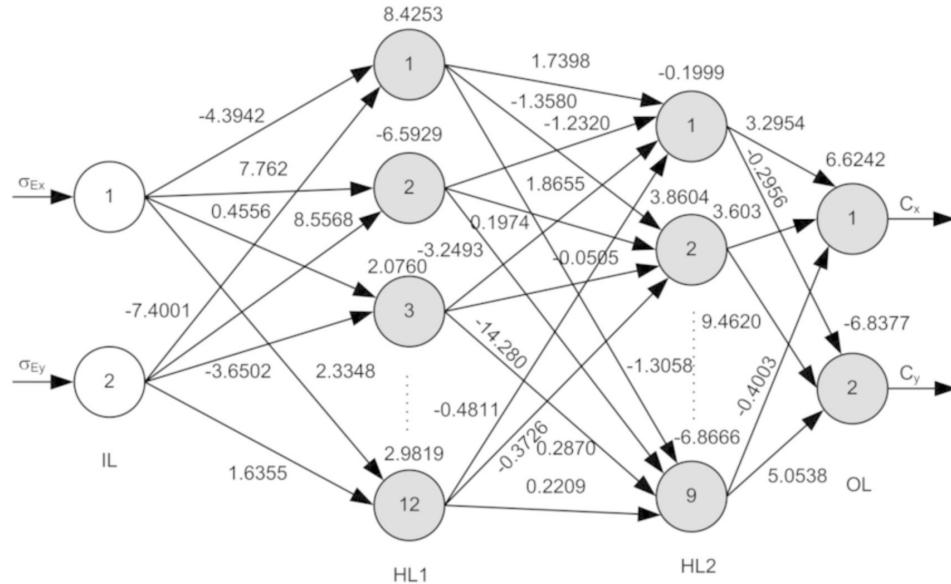
Transforms one dataset (D) into another dataset (D').

The D' is optimized for discrimination.



Basic Functions

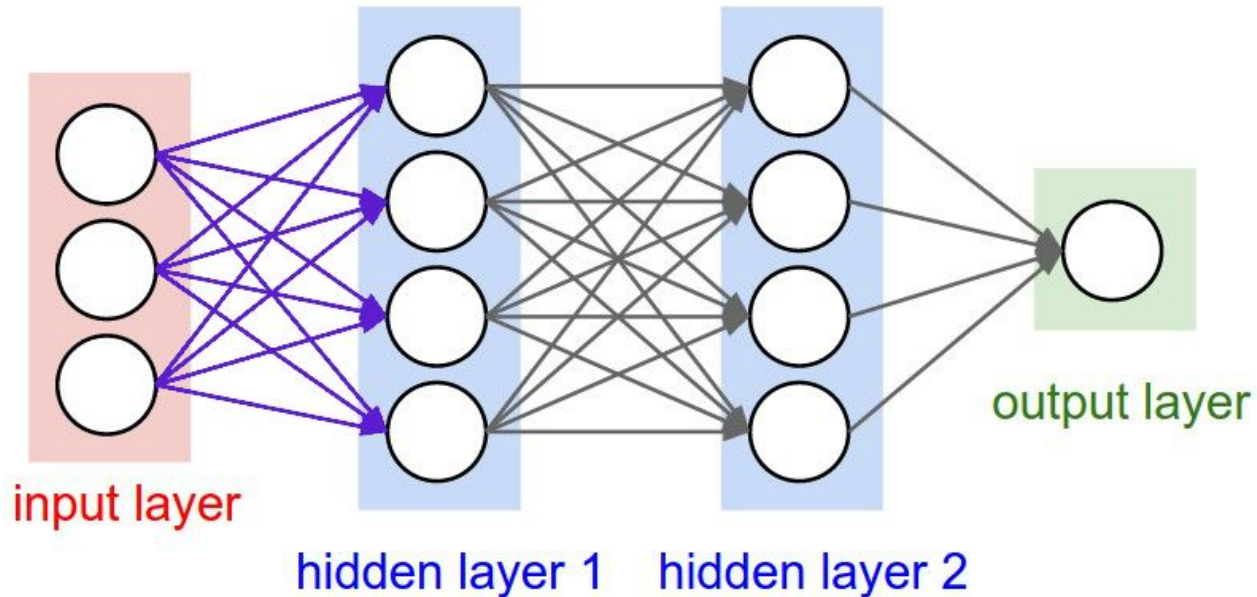
1. Matrix multiplication
2. Thresholding



Matrix Multiplication

Input gets multiplied by N randomly initialized **weights**,

Where N is equal to the number of nodes (neurons) in the next layer.



Convolutions

1	0	1
0	1	0
1	0	1

Kernel or Filter

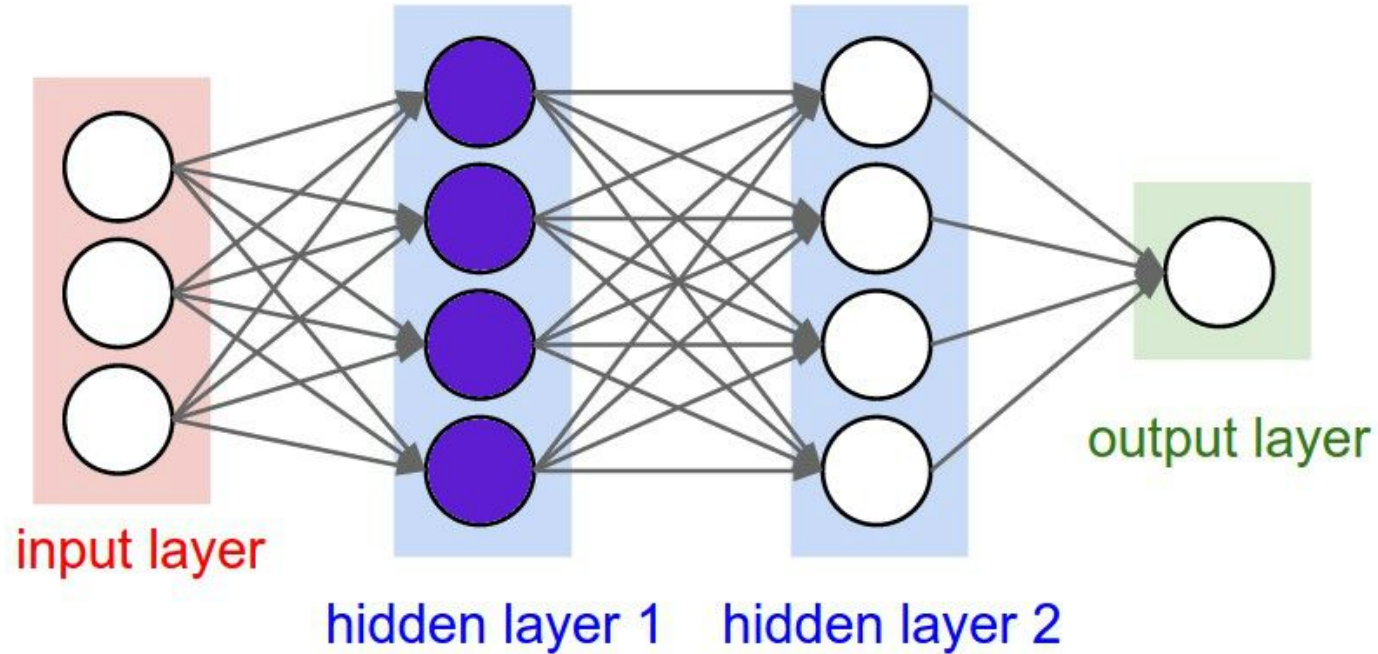
1 _{x1}	1 _{x0}	1 _{x1}	0	0
0 _{x0}	1 _{x1}	1 _{x0}	1	0
0 _{x1}	0 _{x0}	1 _{x1}	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

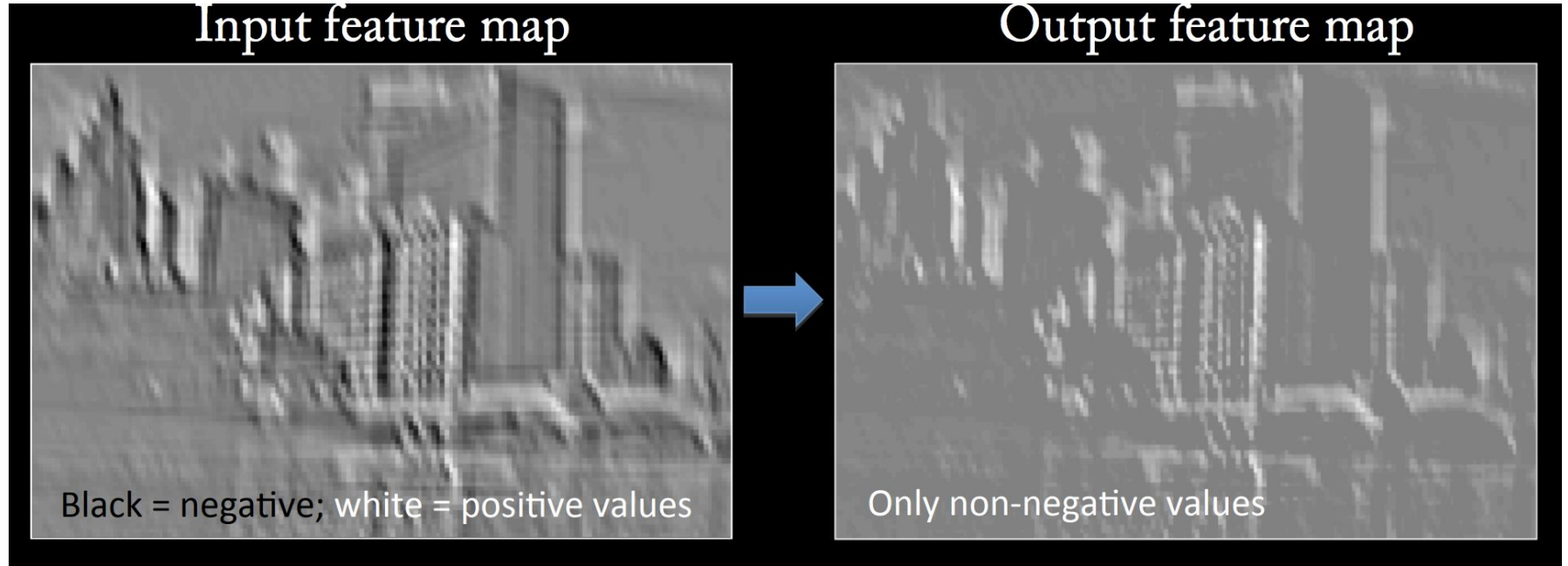
Convolved
Feature

Thresholding



Thresholding or Activation Functions

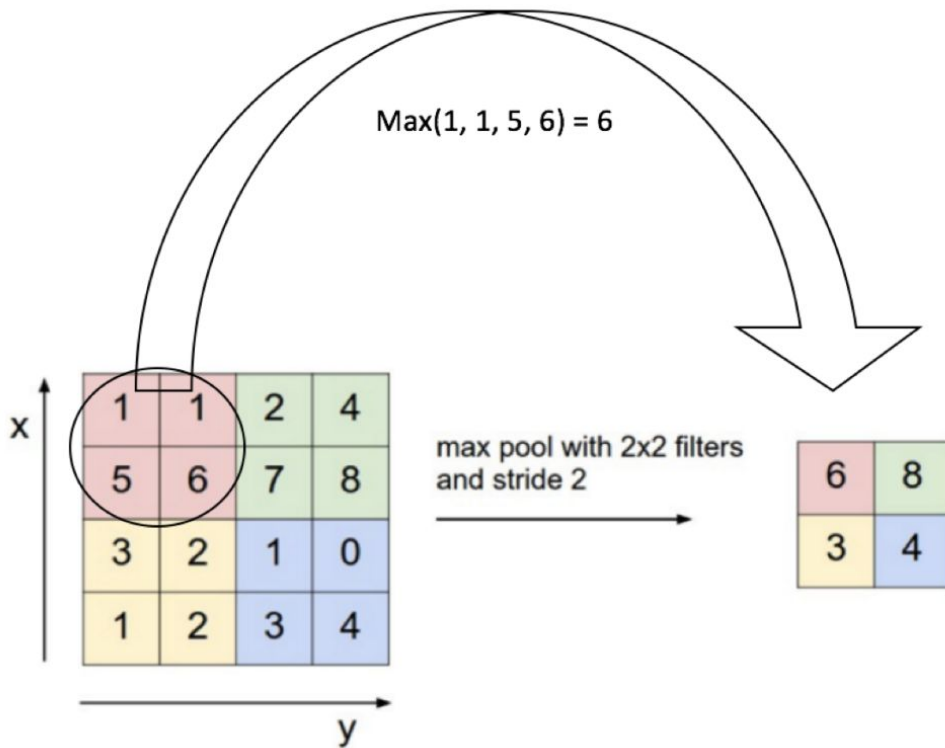
Rectified Linear Units (ReLU) remove negative values.



Downsampling

Use pooling function either Max,
Avg, Sum

Also for simplification and
amplification



Recap:

Matrix multiplication creates new features.

Thresholding and downsampling simplify the math and amplify signals.

This is repeated and combined to identify patterns with increasing complexity.

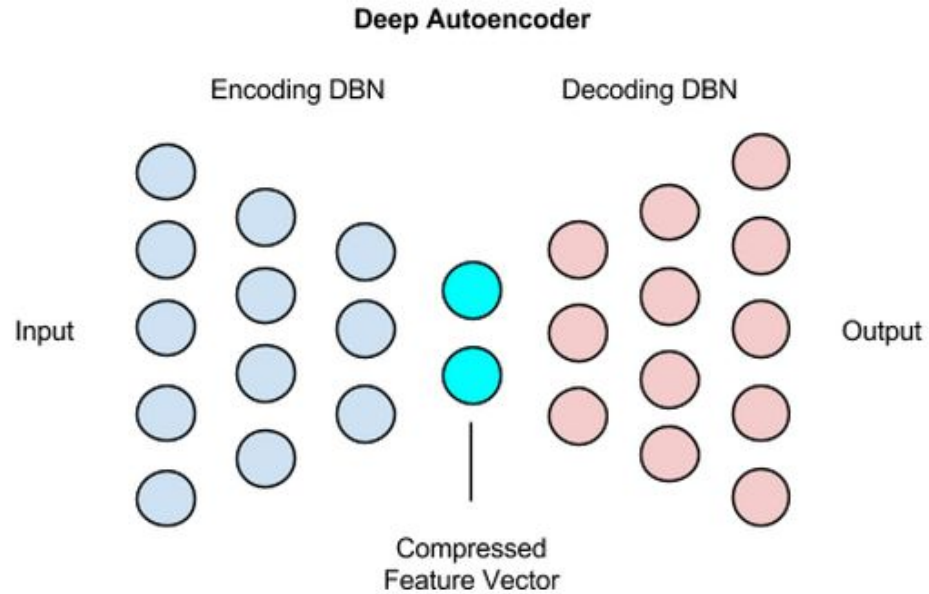
Feature Visualization

<https://distill.pub/2017/feature-visualization/>

Let's Look at Logits:

https://nbviewer.jupyter.org/github/yinleon/interpreting_neural_networks/blob/master/null_features/model_conv_feature_evaluation.ipynb

What about Text?



Bias on Yelp

Different tasks have the same outcomes:

Mexican food is associated with negative reviews and negative connotations!

Sentiment transfer from negative to positive

I would recommend find another place.

I would recommend this place again!

Do not like it at all!

All in all, it's great!

I regret not having the time to shop around.

I have a great experience here.

Average Mexican food.

Authentic Italian food.

Training Data

We build infrastructure around availability

What are we feeding models?

Cool paper about reducing training data gender bias:

<https://homes.cs.washington.edu/~my89/publications/bias.pdf>

Looking for Context

NLP community standardizing metadata RE: origin, app and audience.

Thoughts about Interpretability?

Thanks!

@leonyin