Lecture Notes for

Machine Learning in Python

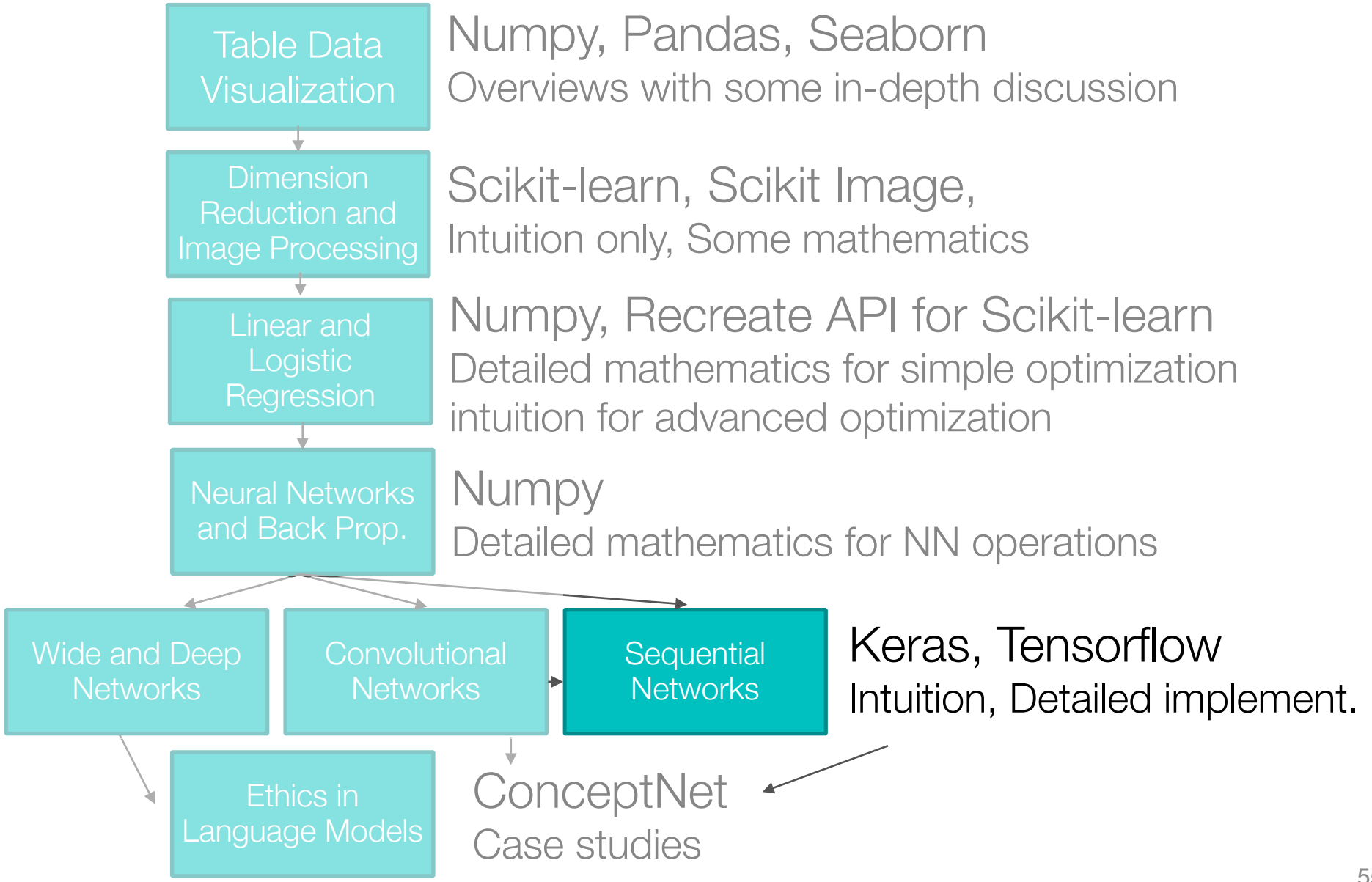
Professor Eric Larson

Sequential Networks Demonstration

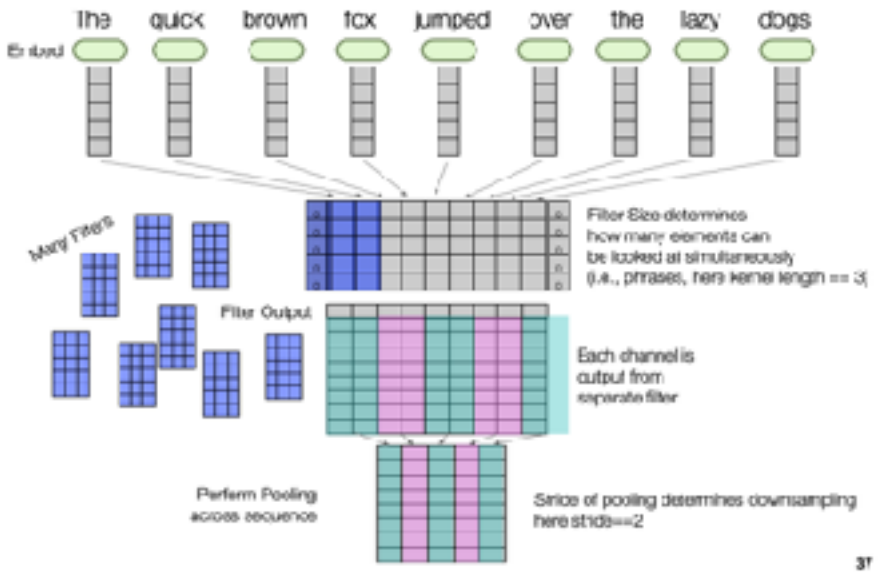
Lecture Agenda

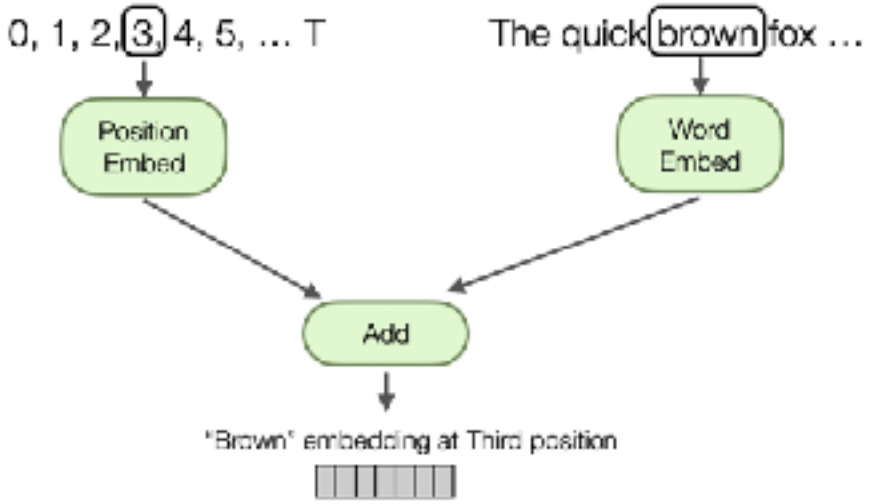
- Logistics
  - Grading Update
  - Sequential Networks due **Last Day of Finals**
    - **Around NOON on December 13**
- Agenda
  - Sequential Networks Demo
    - Extended Demo
  - Final Town Hall
  - Next time:
    - Ethical Principles
    - Retrospective and Evaluations

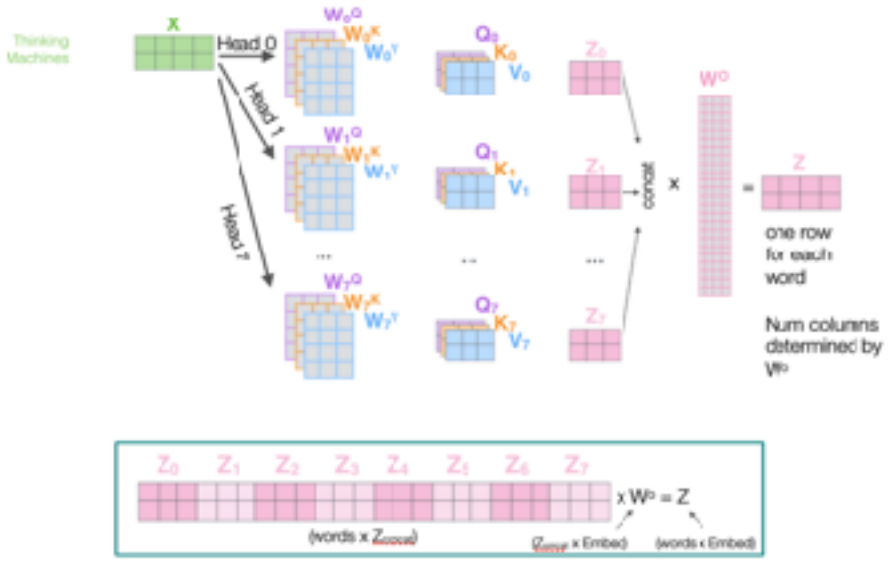
Class Overview, by topic

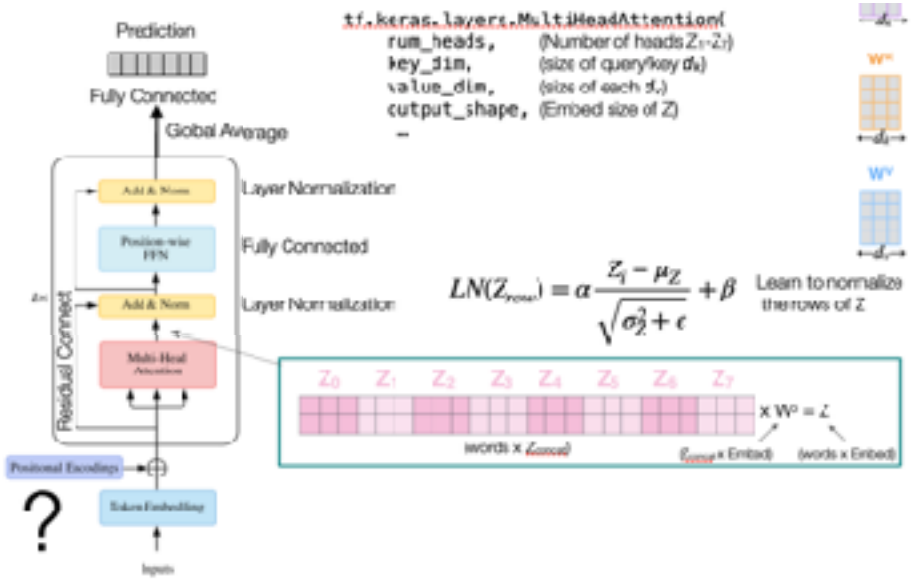


Last Time

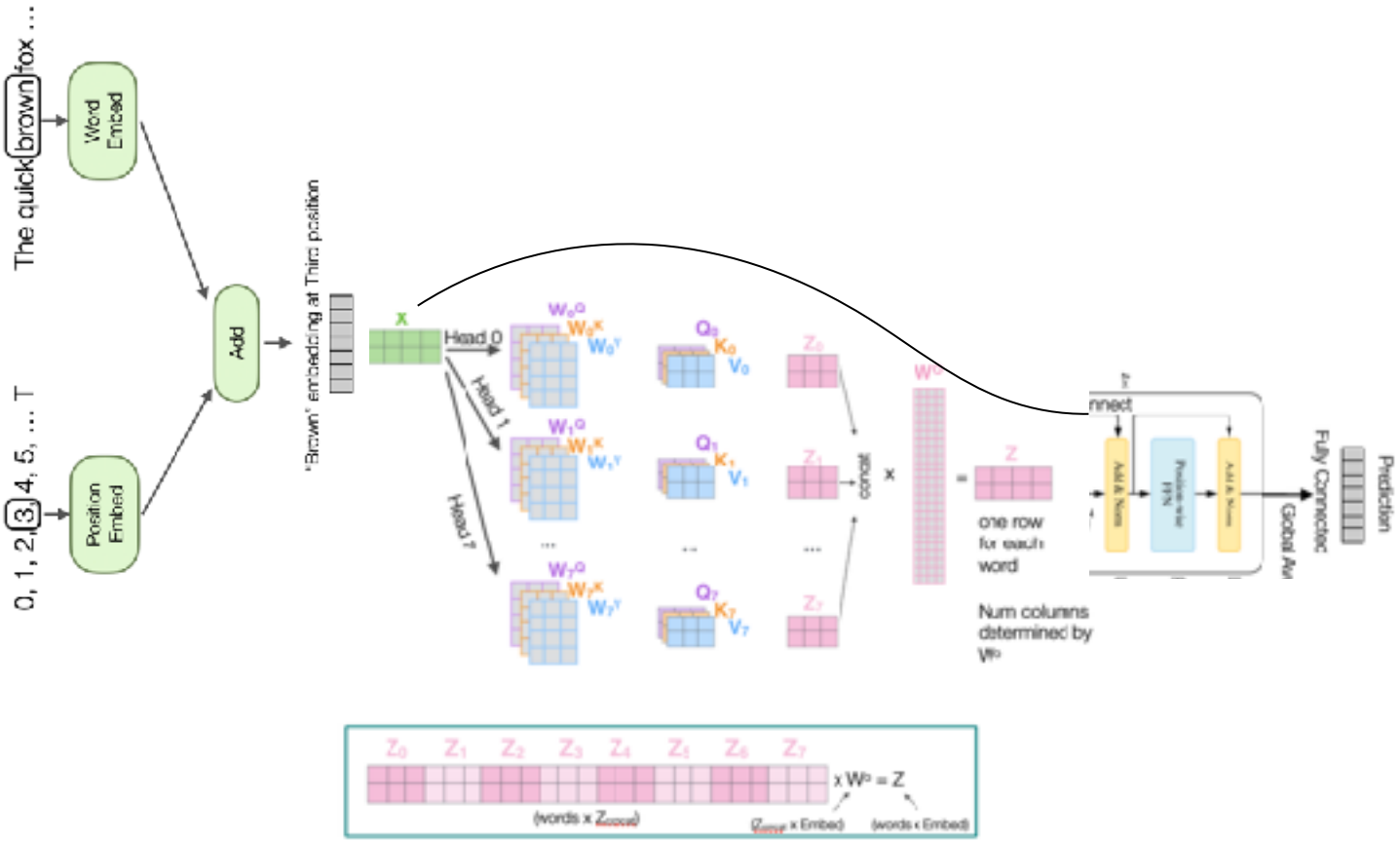

The quick brown fox jumped over the lazy dogs
Many Filters
Filter Size determines how many elements can be looked at simultaneously (i.e., phrases, here kernel length == 3)
Filter Output
Each channel is output from separate filter
Perform Pooling across sequence
Stride of pooling determines downsampling here stride==2
37


0, 1, 2, 3, 4, 5, ... T
The quick brown fox ...
Position Embed
Word Embed
Add
"Brown" embedding at Third position


Thinking Machines
X
Head 0
Head 1
Head 7
Q0 K0 V0
Q1 K1 V1
Q7 K7 V7
Z0
Z1
Z7
concat
x
W^O
=
Z
one row for each word
Num columns determined by W^O
Z0 Z1 Z2 Z3 Z4 Z5 Z6 Z7
x W^O = Z
(words x Z_concat)
Z_concat x Embed
(words x Embed)


Prediction
Fully Connected
Global Average
Layer Normalization
Fully Connected
Layer Normalization
Residual Connect
Positional Encodings
Inputs
?
tf.keras.layers.MultiHeadAttention(
num_heads, (Number of heads Z1-Z7)
key_dim, (size of query/key d_k)
value_dim, (size of each d_v)
output_shape, (Embed size of Z)
)
W^O
W^V
LN(Z_row) = α (Z_i − μ_Z) / √(σ_Z^2 + ε) + β
Learn to normalize the rows of Z
Z0 Z1 Z2 Z3 Z4 Z5 Z6 Z7
x W^O = Z
words x Z_concat
Z_concat x Embed
(words x Embed)

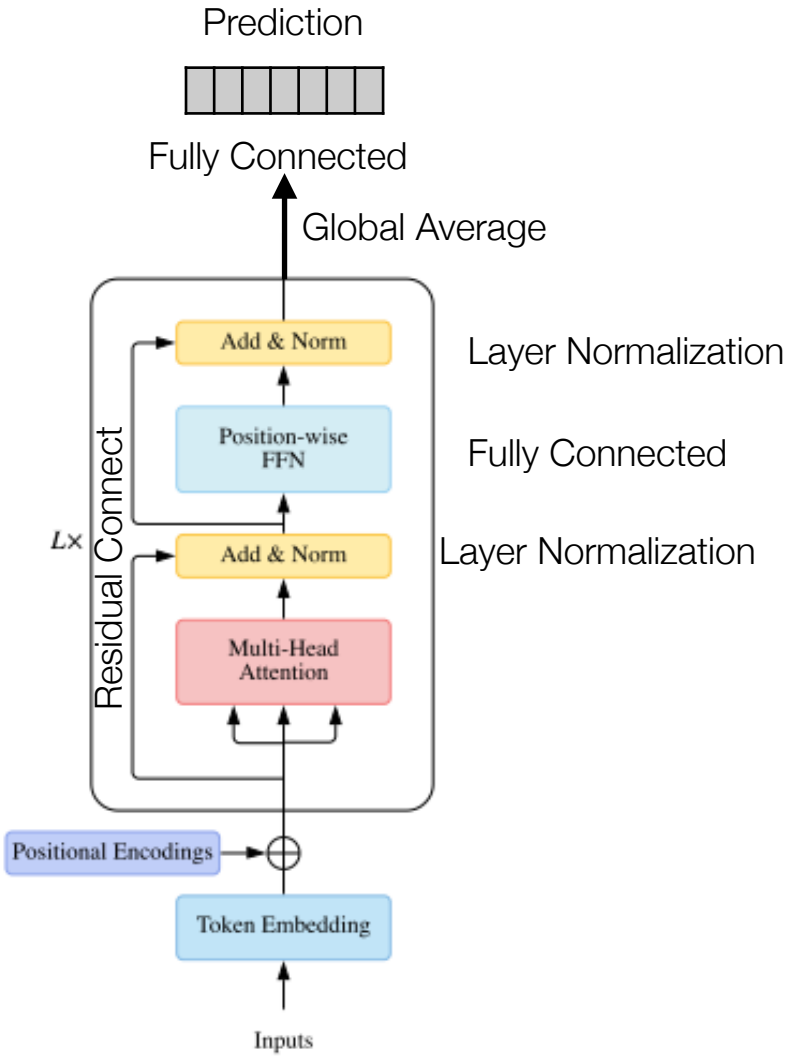
Transformers

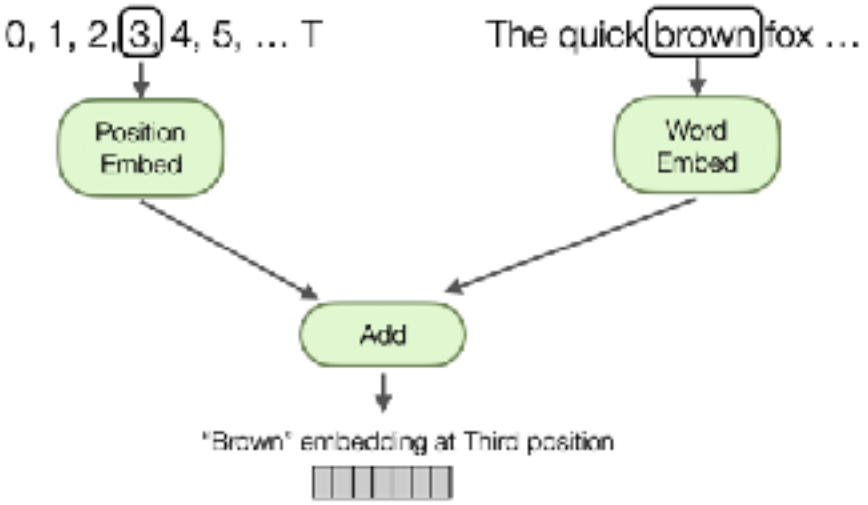


Do transformers have a lot of trainable parameters?

Do transformers have a relatively large memory footprint?

"Finish" Demo





- The Transformer and 20 news groups with GloVe

`13a. Sequence Basics [Experimental].ipynb`

Sequential Networks Town Hall