Transformers and Self attention

Shubham Gupta

March 18, 2020

- This is a guest lecture given by Ashish Vaswani, author of "Attention is all you need".
- Representation learning is the core of all DL.
- Used for learning variable-length representations

1 CNN

- Easy to parallelize
- Exploits local deps
- Interaction distance between positions linear or log
- Long distance deps require many layers