Attention & Transformers for Language Translation

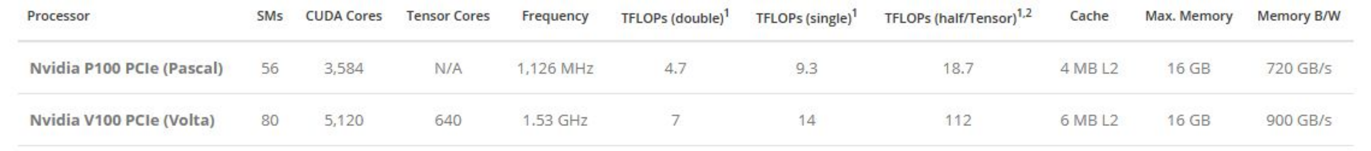
By: Ben Paulson & John Cisler

Why Is This Important?

- **Attention**
  - Aid in increasing “focus” on input
  - Allow longer text sequences
- **Transformers**
  - Extremely effective sequence model
  - Use “self-attention”
  - GPT, BERT, Google NMT (Neural Machine Translation)

Our Task – Language Translation

- **Why is this difficult?**
  - Non-direct syntax mapping
  - Nuanced semantics
- **What will this look like?**
  - Following "Attention is All You Need"
    - Best English-to-French WMT 2014
    - ~1 day on 8 P100 GPUs
  - We'll tackle different language from WMT



Processor	SMs	CUDA Cores	Tensor Cores	Frequency	TFLOPs (double)[1]	TFLOPs (single)[1]	TFLOPs (half/Tensor)[1,2]	Cache	Max. Memory	Memory B/W
Nvidia P100 PCIe (Pascal)	56	3,584	N/A	1,126 MHz	4.7	9.3	18.7	4 MB L2	16 GB	720 GB/s
Nvidia V100 PCIe (Volta)	80	5,120	640	1.53 GHz	7	14	112	6 MB L2	16 GB	900 GB/s

1

Background Info

Crucial Terms & How You Can Learn as Well!

Natural Language Processing (NLP)

- Language Translation, Sentiment Analysis, Question Answering, Text Summarization, etc…
- **Token:** String of values representing text unit
- **Encoding:** Deterministic mapping of token
- **Embedding Space:** All vectorized tokens; closer means similar meaning

↳ **What about "right" or "cool"?**

Semantic, not syntactic!!

Attention – Background

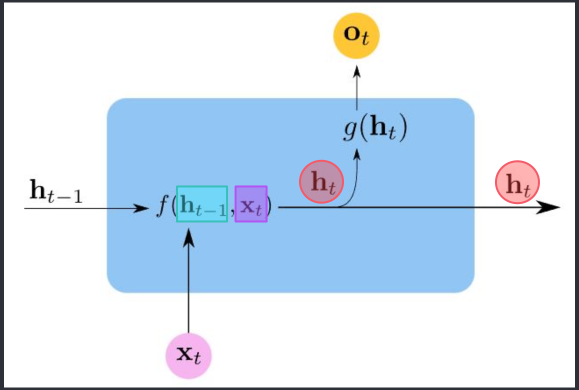
- Sequence Model Basics (RNNs)
  - MLP not effective w/ sequence data
  - Strung-together element-level models
  - Require sequence context and input

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \theta)$$

$$\mathbf{o}^{(t)} = g(\mathbf{h}^{(t)}; \theta')$$

o^t = output (token)

h^t = Sequence Encoding



$$f(\dots f(\mathbf{h}^0, \mathbf{x}^1; \theta), \dots \mathbf{x}^{(t)}; \theta)$$

In-Class

$$a_j = g_j(WX + B)$$

New

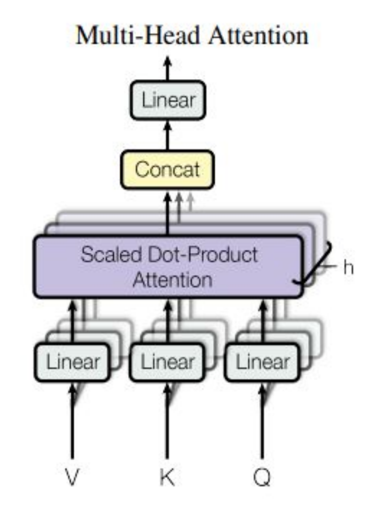
$$h_t = g_t(Wh^{(t-1)} + Ux^{(t)} + B)$$

$$o_t = Wh^t + B$$

Attention – How it Works

- **Problem with RNN:**
  - Vanishing/exploding gradient as you build large sequences…

- **Solution = Attention!**
  - **Pay attention** to portions of the input sequence
  - From **<100** to **>1000** tokens
  - Self-Attention = transformers



Parallel Self-Attention in Transformers 😵‍💫

Transformers - Background

- **Transformers are seq2seq**
  - **Encoder:** Input Sequence **->** Context Vector
  - **Decoder:** Context Vector **->** Output Sequence
  - Both are often **RNNs!**

- **What's Learned?**
  - Encoding/Decoding Functions **(Fe() / Gd())**

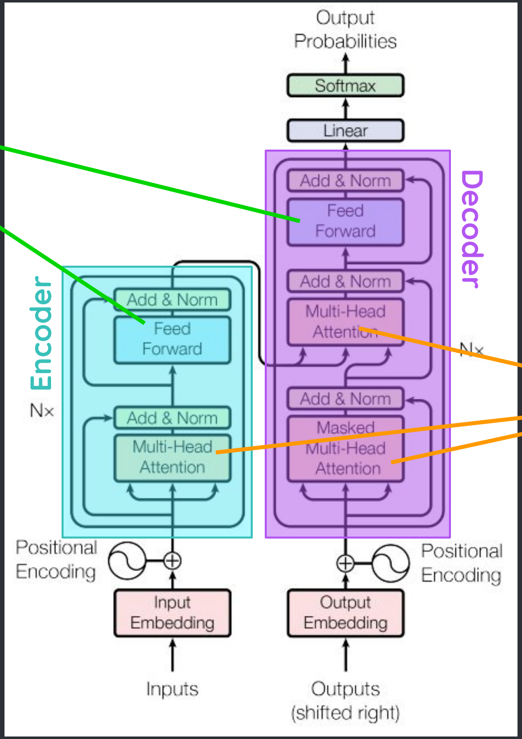
Note: A "hidden state" is an encoded sequence token

Transformers - How It Works

Feed Forward Networks

Simple network to ensure non-linearity

Captures more complex relationships between input and output



Some Attention Sprinkled Around

Can consider all portions of input sequence

Can parallelize computation

Transformer Architecture from “Attention is All You Need”

Existing Use Cases

Encoder-Only

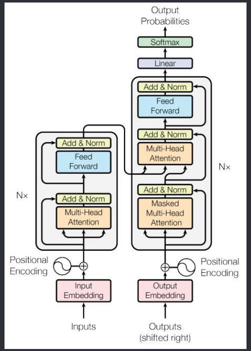
BERT

Google Search

Google ALBERT

FB RoBERTa

Transformer



Decoder-Only

GPT

GPT-2,3,4...

Google T5

Stanford Llama

Remember... we're focusing on "Attention Is All You Need"

2

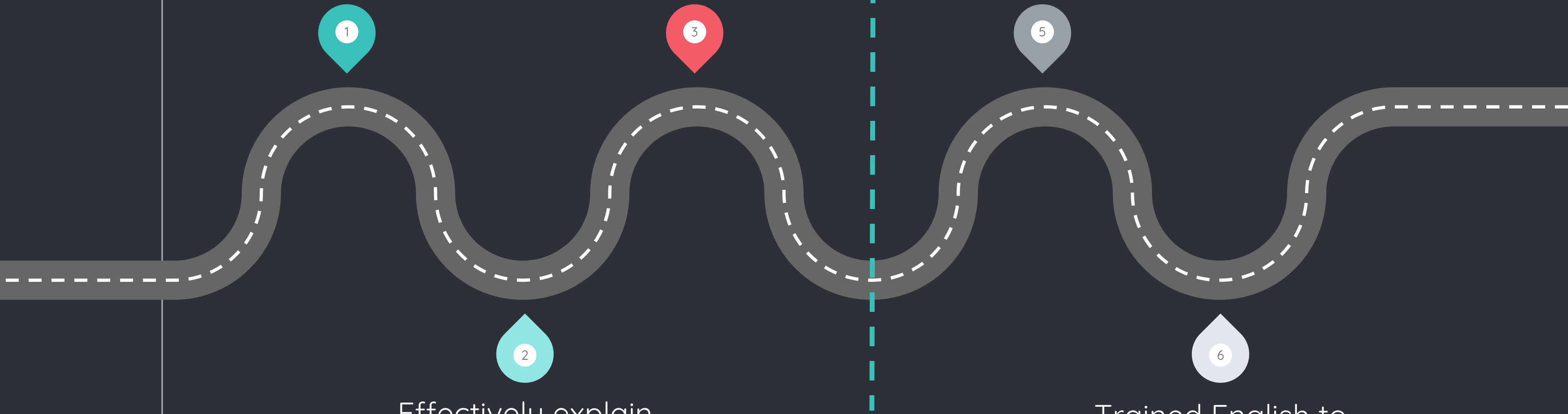
Our Project Proposal

Reporting & Development Goals

The Project

- Walk through “Attention is All You Need”
  - Explain attention
  - Explain transformers
  - Their English -> French/German
- Build a language translation model
  - English -> Spanish
  - Translation accuracy metrics

Roadmap



“Attention is All You Need” Analysis

Effectively explain attention & transformers to a general AI audience

English -> French implementation in Tensorflow

Week 9 Preso

Trained English to French Model

Trained English to Portuguese Model (w/ metrics)

Expected Challenges

- **Time Constraints**
  - At minimum, will require 20 hrs training
- **Knowledge Compilation**
  - Fitting complex topics into easy-to-consume format
  - Relating to concepts to code implementations

BIG QUESTION

Why are transformers, and attention, used in some of the most effective deep-learning models today for NLP?

CONCLUSION

“Attention Is All You Need” Analysis

Conceptual Exploration

Language Translation Implementation

Motivation: Learn about how some of the most state-of-the-art models work today!