# Intro

Hi I’m Tony.

# Caveats

I am not a Data Scientist.

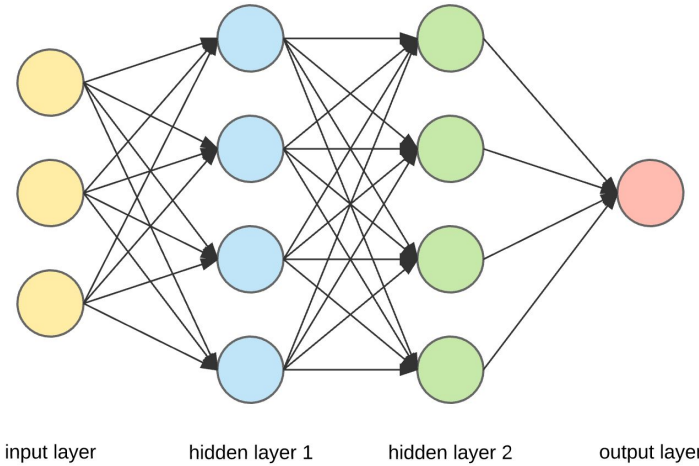
# Caveats

This stuff is super new.

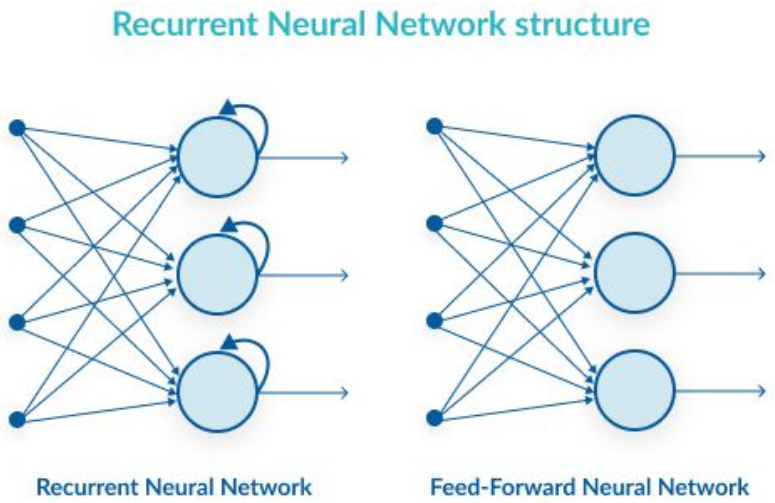
# What are doing here? LLMs

1. Brief History of Neural Networks
2. Attention / Transformer Model (the game changer)
3. Transformer Models at play now
4. SimplyPut

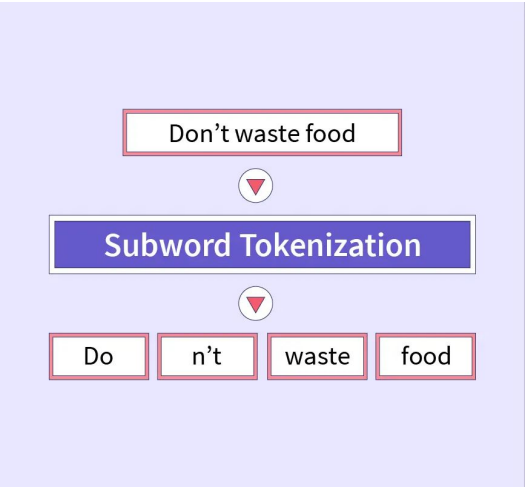
# Neural Networks



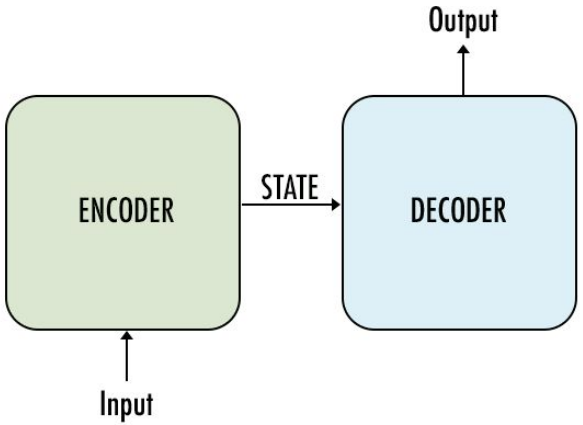
# Recurrent Neural Networks (RNNs)



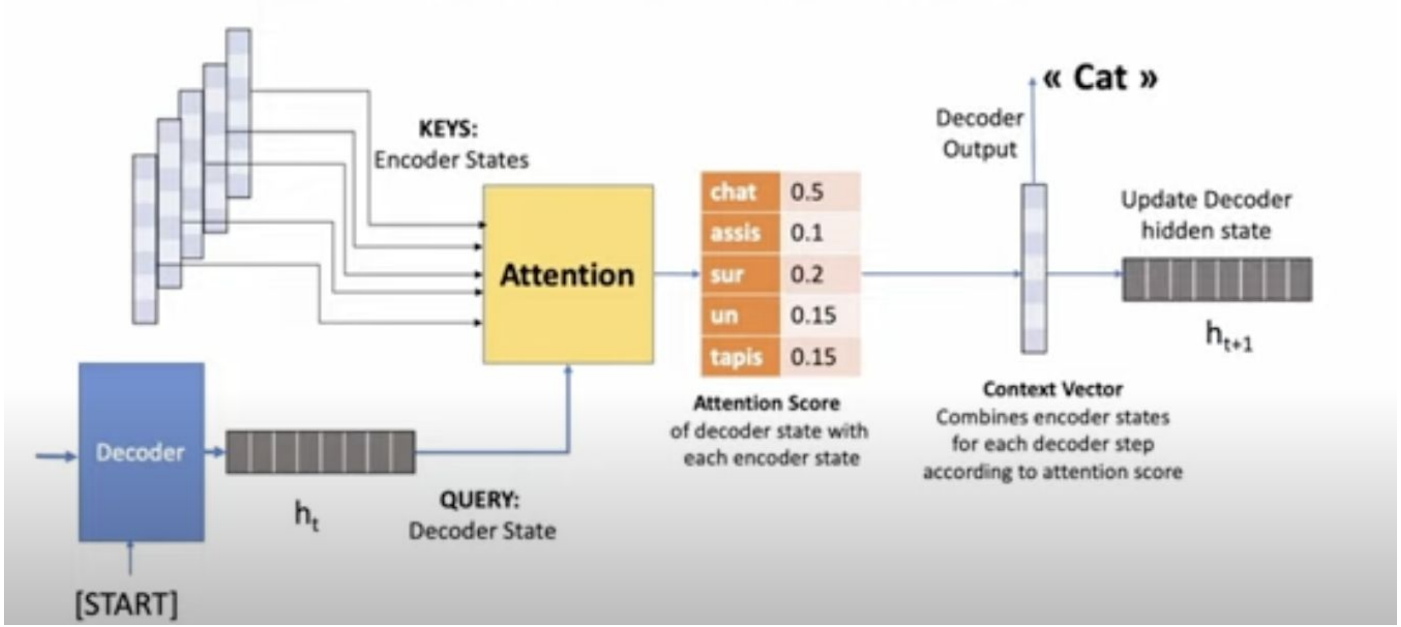
# Encoding (tokens) 2013 (we got word2vec)



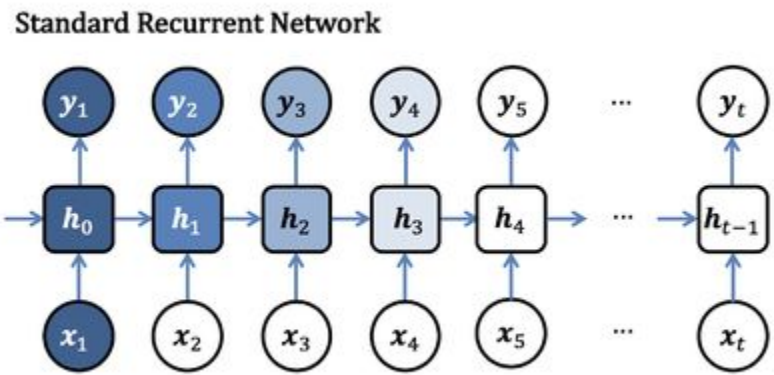
# Sequence-to-sequence models (2014)

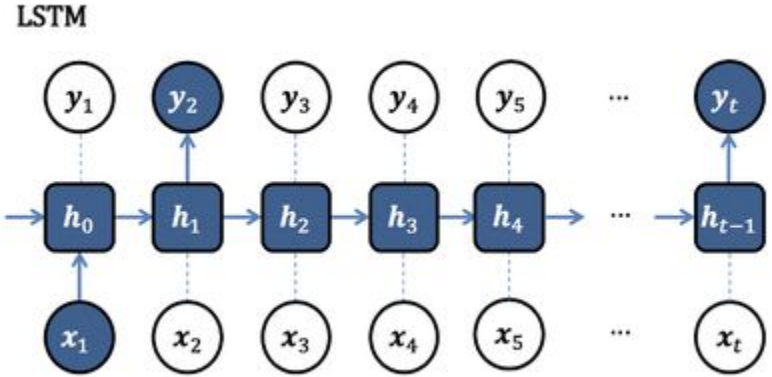


# Attention (2015)



# RNN, LSTM Problems



# Attention! (2017)



## Attention Is All You Need

**Ashish Vaswani***
Google Brain
avaswani@google.com

**Noam Shazeer***
Google Brain
noam@google.com

**Niki Parmar***
Google Research
nikip@google.com

**Jakob Uszkoreit***
Google Research
usz@google.com

**Llion Jones***
Google Research
llion@google.com

**Aidan N. Gomez*** †
University of Toronto
aidan@cs.toronto.edu

**Łukasz Kaiser***
Google Brain
lukaszkaiser@google.com

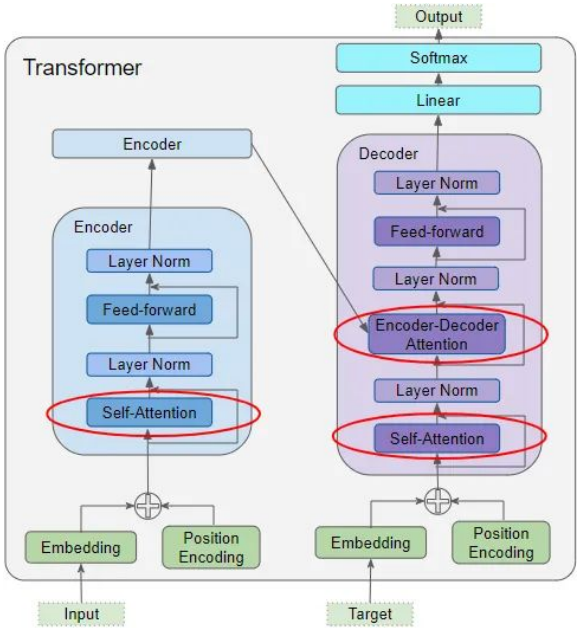
**Illia Polosukhin*** ‡
illia.polosukhin@gmail.com

### Abstract

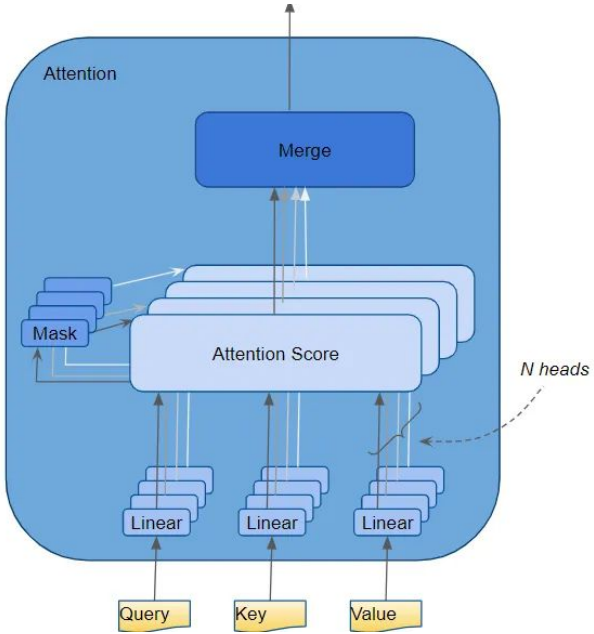
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

### 1 Introduction

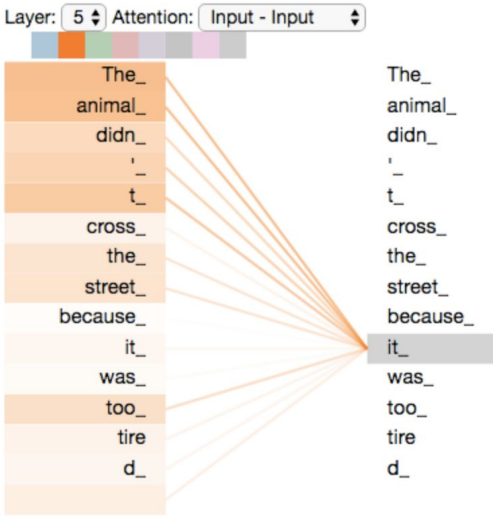
# Transformer

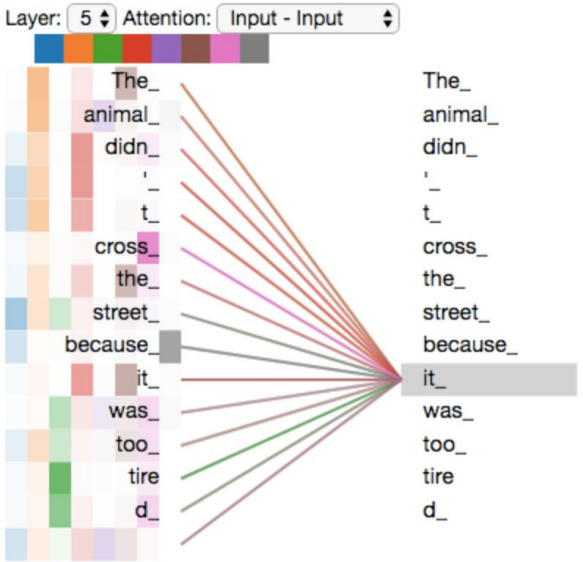


# Multi-Head Attention

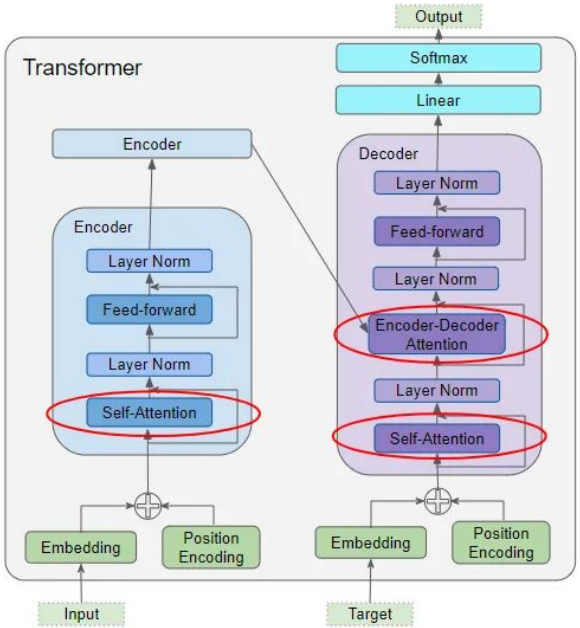


# Multi-Head Attention

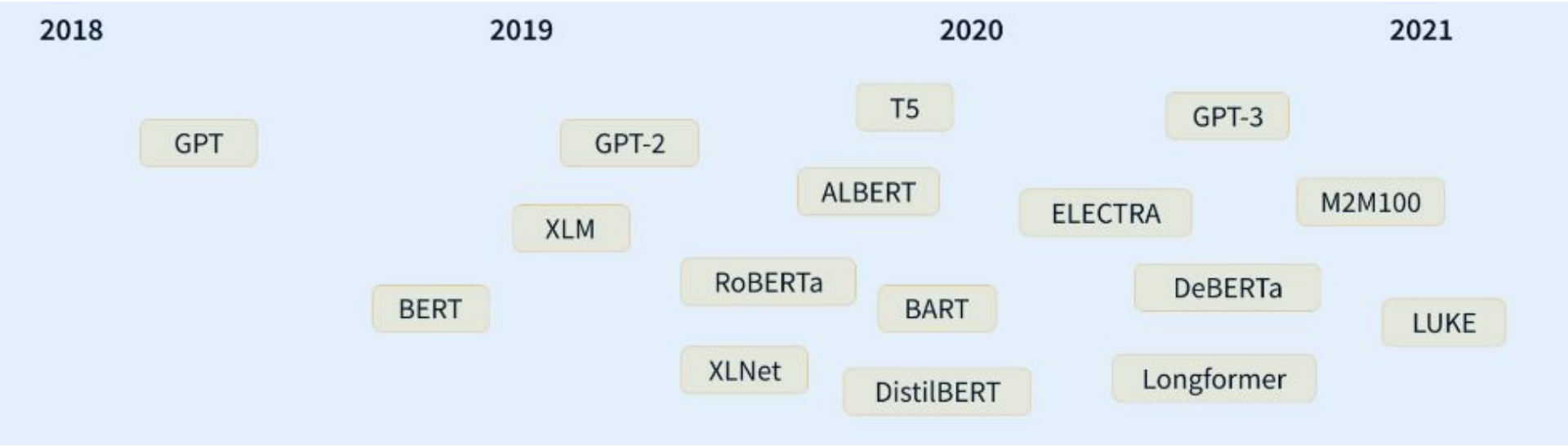


# What Does the Transformer Do?

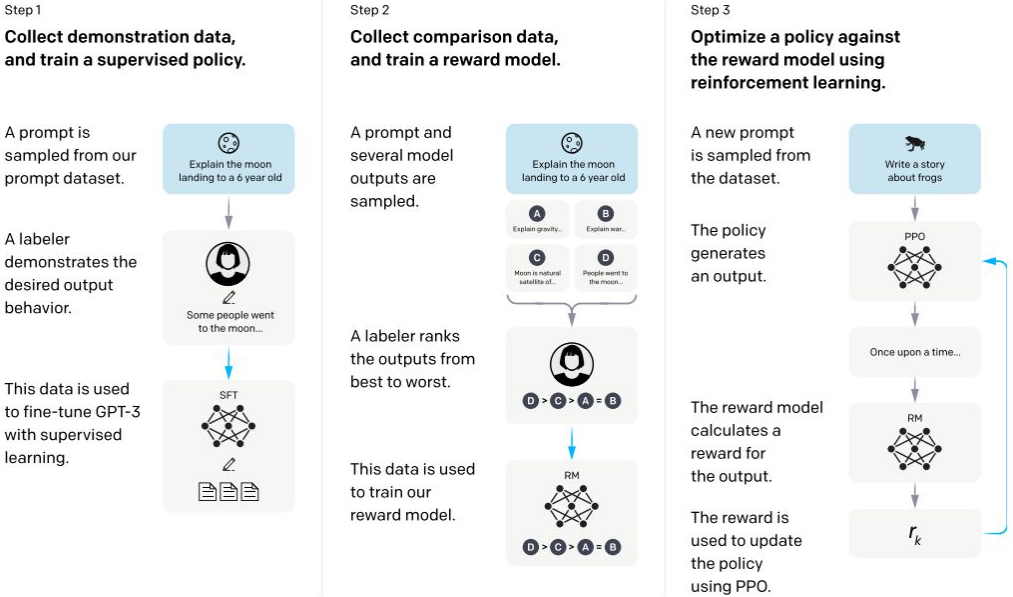


# Rise of the LLMS



# Zero Shot And Few Shot

# Hallucinations - Humans Needed (2022)



# LLMS Passing Tests

Med-PaLM 2 was the first LLM to perform at an “expert” test-taker level performance on the MedQA dataset of US Medical Licensing Examination (USMLE)-style questions, reaching 85%+ accuracy, and it was the first AI system to reach a passing score on the MedMCQA dataset comprising Indian AIIMS and NEET medical examination questions, scoring 72.3%.

# In healthcare

Thymia - Mental health

Med-Palm2 - Google LLM (asking medical questions)

RIKEN Center for Biosystems Dynamics Research - StemCells to help repair eyes?

Suki.ai - Physician Notes and other Admin tasks

# The Landscape Now

Make Better Models - Experts

Fine tuning existing models - Data Scientists (use HuggingFace, or google Vertex AI)

Make applications with Zero/Few Shot - Everyone else (use https://platform.openai.com/)