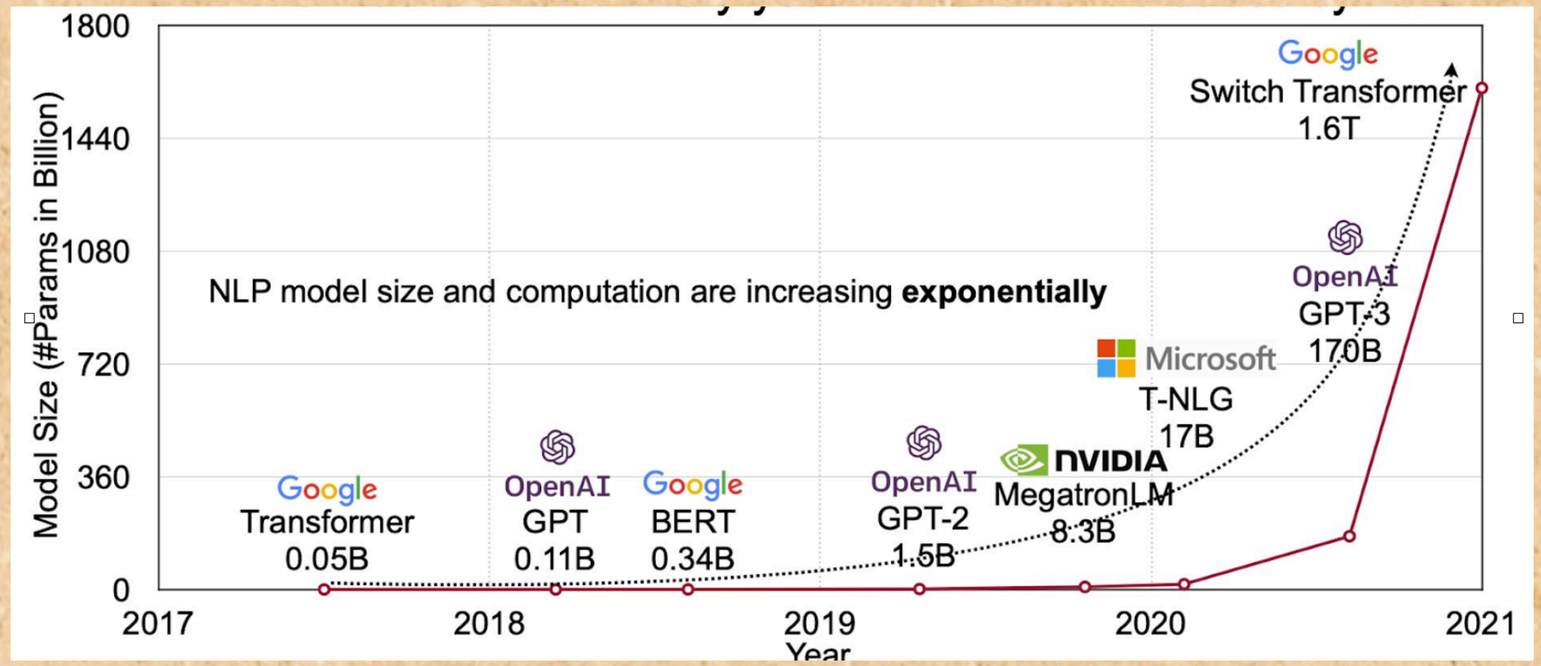
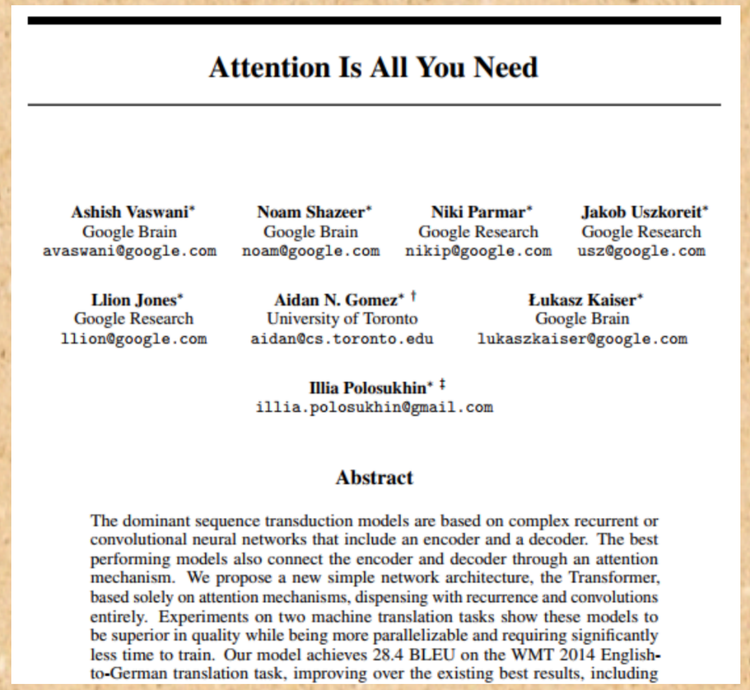
# Module 1 - Topic 1

# 1.3 Understanding Transformers

# Understanding Transformers



**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

Let’s understand

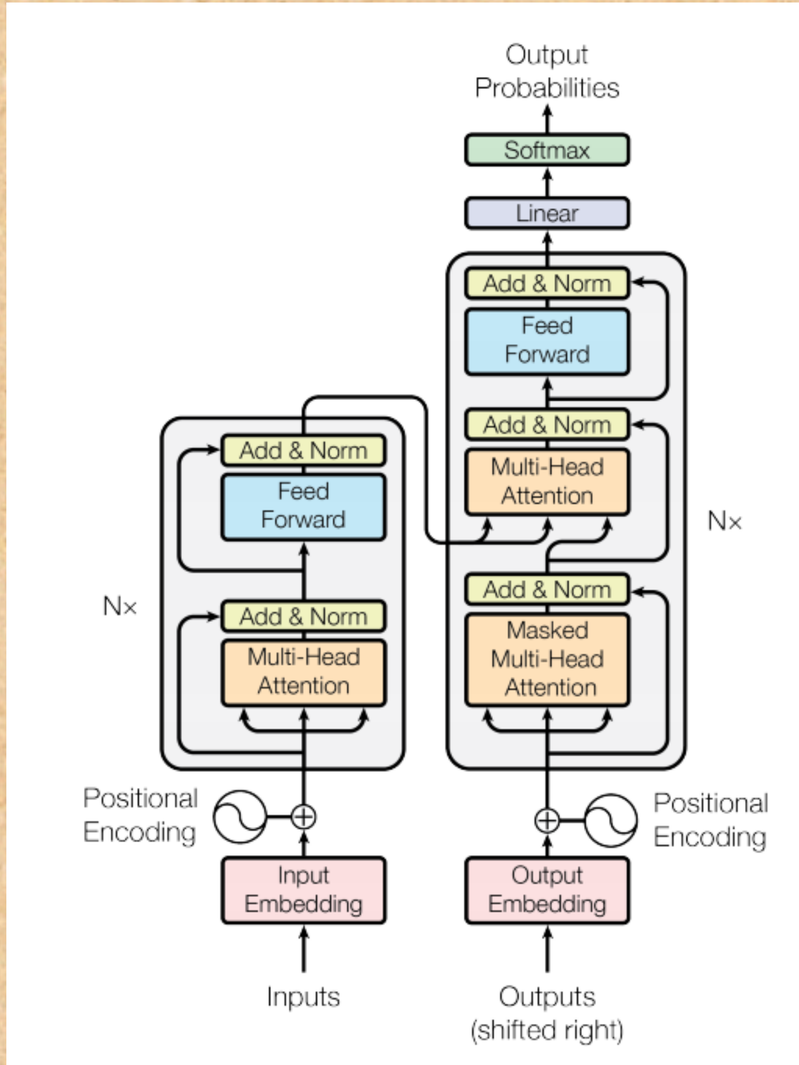
# Transformers

# How transformers work?

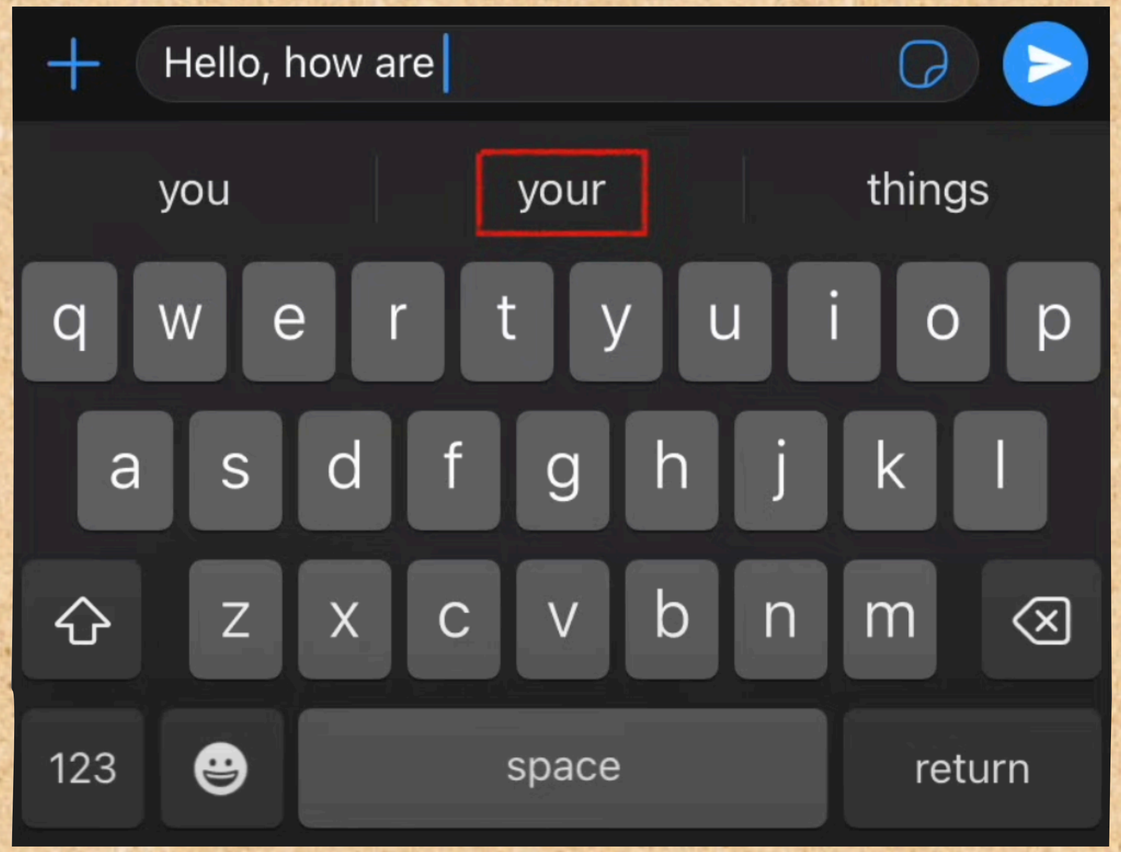
## Generating one word at a time

Once upon a ___ →  → time

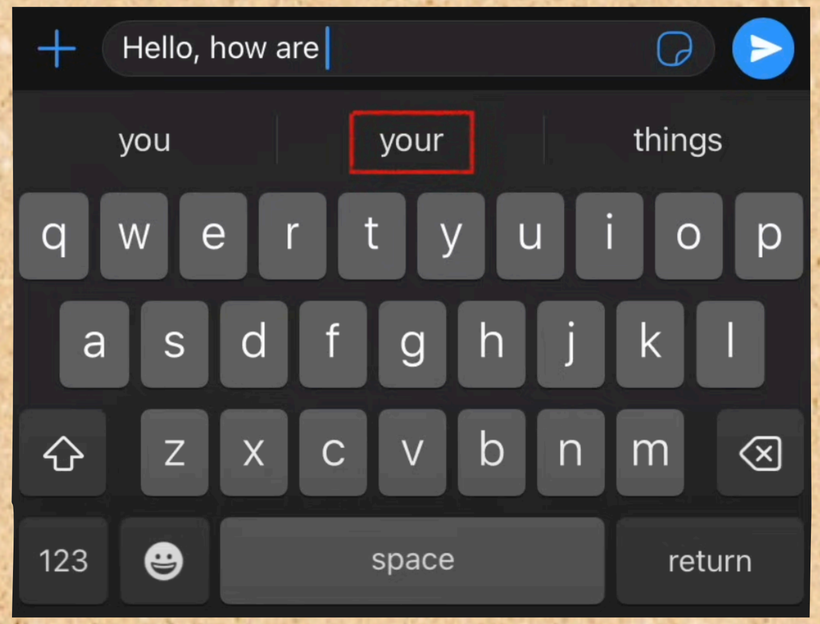
# But why do we need this complex architecture?



# Good old keypad



# N-gram models




DATA

… are **you** …

… are **sad** …

… are**?**…

… are **happy** …

… are **ready** …

… are **happy** …

… are **free** …

1-gram

Hello, how are **you**?

Hello, how are **things** going?

Hello, how are **things** today?

Hello, how are **the** kids?

Hello, how are **the** others?

Hello, how are **they** doing?

Hello, how are **things** happening?

3-gram

# N-gram models - Limitation

DATA

Hello, how are **you**?

Hello, how are **things** going?

Hello, how are **things** today?

Hello, how are **the** kids?

Hello, how are **the** others?

Hello, how are **they** doing?

Hello, how are **things** happening?

3-gram

The other day I was walking and five unicorns said ____

10 words

We need something more complicated

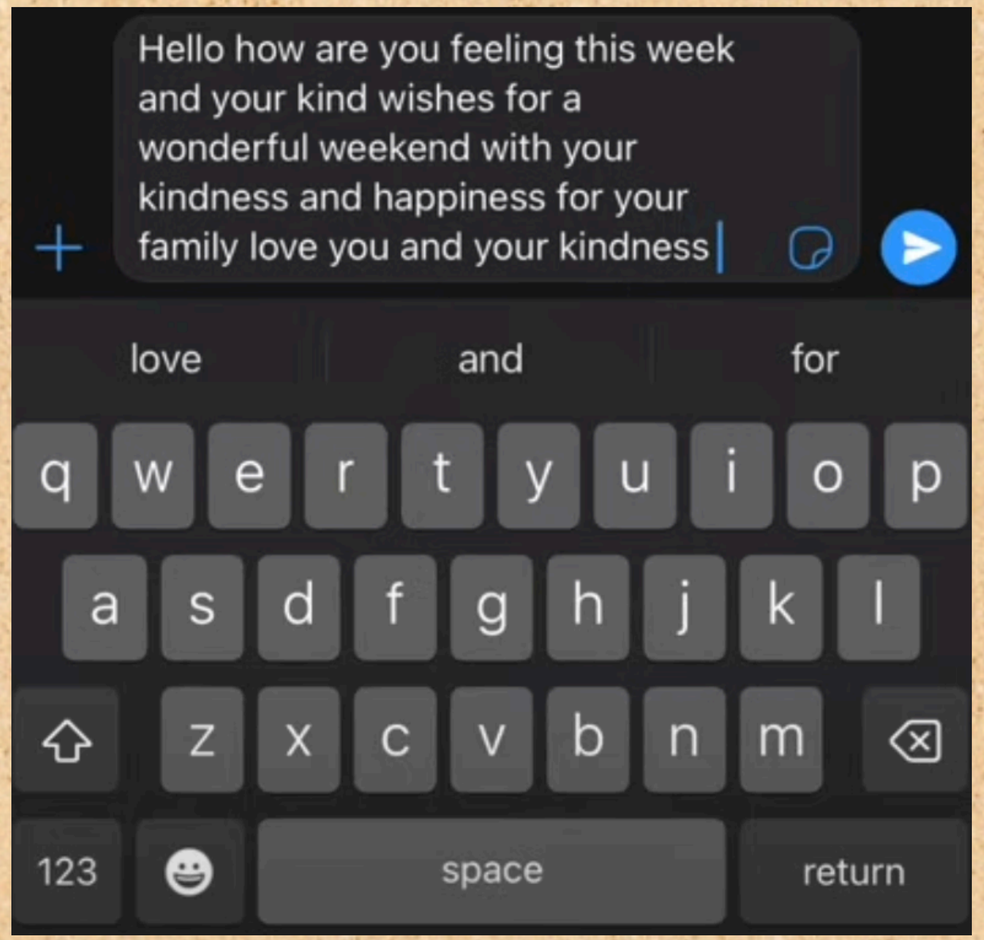
Neural Networks

# Neural Network (RNNs) - Limitations

- Sequential Processing: process inputs sequentially, one element at a time.

- Struggle to learn and main long context

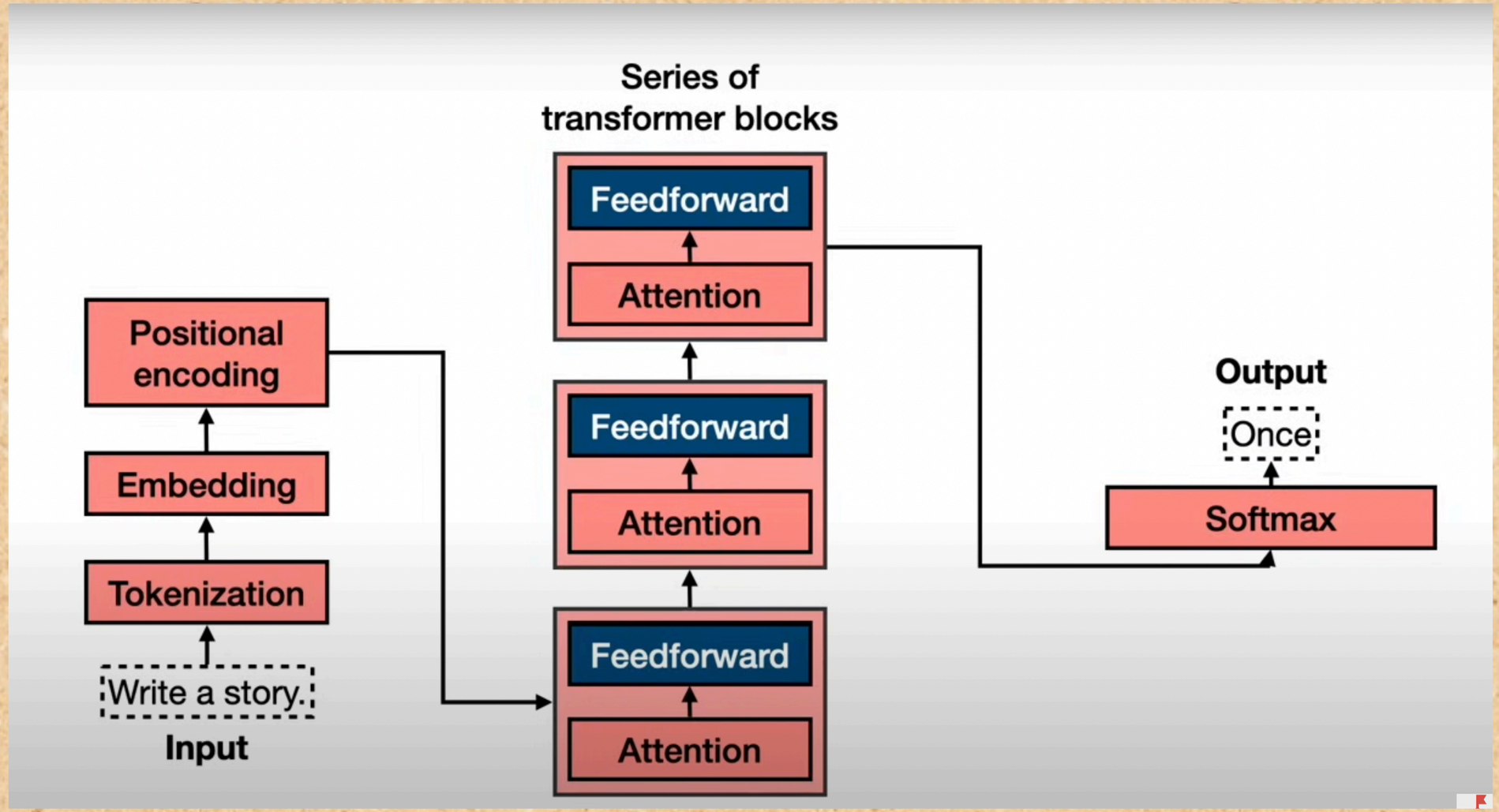
- Slow Training/Inference



# Transformers

- **Parallelization**: Can process all sequence elements in parallel for faster training/inference
- **Self-attention:** This allows directly modeling long-range dependencies
- Better Performance: Often outperforms RNNs on sequence modeling tasks

# Architecture



Let’s get an intuition!

# How do LLMs work?

# I want to play _____

a. Bag

b. Rainbow

c. Cricket

d. Cloud

# I want to play _____

a. Chess

b. Music

c. Cricket

d. Cards

## I love playing outdoor games. I want to play _____

a. Chess

b. Music

c. Cricket

d. Cards

# How do LLMs work?

- Trained on huge amounts of data.

- Understand to recognise patterns in language.

- Predict words/phrases based on their probability.