Transformers:
The Era of Rapid Scaling in NLP

Nikita Kitaev
February 22, 2022
The Era of Rapid Scaling in NLP

2017: Transformer is introduced

[Vaswani+17] Attention is All You Need

2022: Large-scale Transformer models are the dominant approach for many NLP tasks
Neural Machine Translation is in production at Google

Neural MT ca. 2016
There are computation paths through the RNN-based network that scale linearly with the sequence length, and can't be parallelized.
Maximum Path Length

RNN: \#tokens \times \#layers
Maximum Path Length

RNN: \#tokens \times \#layers

What about a Convolutional Neural Network?
Maximum Path Length

RNN: \#tokens \times \#layers

Convolutional: \#layers -- *but we need to connect all tokens*
Maximum Path Length

RNN: \( \#\text{tokens} \times \#\text{layers} \)

Convolutional: \( \log_{\text{kernel size}}(\#\text{tokens}) \)
Maximum Path Length

RNN: \( \#\text{tokens} \times \#\text{layers} \)

Convolutional: \( \log_{\text{kernel size}}(\#\text{tokens}) \)

Any other alternatives?
Maximum Path Length

RNN: \( \# \text{tokens} \times \# \text{layers} \)

Convolutional: \( \log_{\text{kernel size}}(\# \text{tokens}) \)

How about attention?
Maximum Path Length

RNN: \#tokens * \#layers

Convolutional: $\log_{\text{kernel size}}(\#tokens)$

How about attention?
Maximum Path Length

RNN: \( \#\text{tokens} \times \#\text{layers} \)

Convolutional: \( \log_{\text{kernel size}}(\#\text{tokens}) \)

Attention: \( \#\text{layers} \)
(1) Transformer Architecture
Transformer Architecture
Encoder

Diagram of the Encoder process, including layers such as Add & Norm, Feed Forward, Softmax, Linear, and Multi-Head Attention.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
Position-Based Attention

She enjoys playing tennis.
She enjoys playing tennis.
The word at $t=1$ is: 

enjoys

Query: $t'=1$

She

enjoys

playing

tennis

. 

$\checkmark$

$\times$

$\times$

$\times$

$t=0$

$t=1$

$t=2$

$t=3$

$t=4$
Position-Based Attention

The word at $t=1$ is: 

Query: $t'=1$

- $t=1$: enjoys
- $t=2$: $\times$
- $t=3$: $\times$
- $t=4$: $\times$

She enjoys playing tennis.

$t=0$  
$t=1$  
$t=2$  
$t=3$  
$t=4$
She enjoys playing tennis.
Encoder

She enjoys

Multi-Head Attention

Feed Forward

LayerNorm

LayerNorm

Multi-Head Attention

position

word

She 

enjoys

t=0 
t=1
Self-Attention

- Query: $q_t$
- Key: $k_t$
- Value: $v_t$

- Position
- Word

$S = S + G + H$
Self-Attention

- \(q_t\): query
- \(k_t\): key
- \(v_t\): value

- Position: \(p(t \rightarrow 1)\)
- Word: \(p(t \rightarrow T)\)

- \(k_1, v_1\)
- \(k_T, v_T\)
Self-Attention

\[
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}V\right)
\]
Self-Attention

Attention\((Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}V\right)\)

MultiHead\((X) = \sum_{i=0}^{h} \text{Attention}(XW_i^Q, XW_i^K, XW_i^V)W_i^O\)
Feed-Forward

FeedForward($x$) = $\max(0, xW_1 + b_1)W_2 + b_2$
Layer Normalization [Ba+16] improves stability of neuron activations

Residual Connections
useful across a variety of neural network architecture types, not just in NLP
Encoder
Decoder
She enjoys playing tennis.
Encoder vs. Decoder

Self-Attention

Masked Self-Attention
She enjoys playing tennis.
Decoder
Encoder-Decoder
## Transformer MT Results

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>Training Cost (FLOPs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EN-DE</td>
<td>EN-FR</td>
</tr>
<tr>
<td>ByteNet [18]</td>
<td>23.75</td>
<td></td>
</tr>
<tr>
<td>Deep-Att + PosUnk [39]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNMT + RL [38]</td>
<td>24.6</td>
<td>39.92</td>
</tr>
<tr>
<td>ConvS2S [9]</td>
<td>25.16</td>
<td>40.46</td>
</tr>
<tr>
<td>MoE [32]</td>
<td>26.03</td>
<td>40.56</td>
</tr>
<tr>
<td>Deep-Att + PosUnk Ensemble [39]</td>
<td></td>
<td>40.4</td>
</tr>
<tr>
<td>GNMT + RL Ensemble [38]</td>
<td>26.30</td>
<td>41.16</td>
</tr>
<tr>
<td>ConvS2S Ensemble [9]</td>
<td>26.36</td>
<td><strong>41.29</strong></td>
</tr>
<tr>
<td>Transformer (base model)</td>
<td>27.3</td>
<td>38.1</td>
</tr>
<tr>
<td>Transformer (big)</td>
<td><strong>28.4</strong></td>
<td><strong>41.8</strong></td>
</tr>
</tbody>
</table>
(2) Pre-Training
She enjoys playing tennis.
Pre-Training with LMs
Pre-Training with LMs

Representative Model: GPT

(GPT = Generative Pre-Training)
Pre-Training with LMs
Pre-Training with LMs

Help prince transfer huge inheritance

Important information about your final exam

Spam

Not Spam

pre-training

task-specific

fine-tuning
Fine-tuning with LMs

GPT

Classifier

SPAM

Help  prince  transfer  huge  inheritance
Fine-tuning with LMs
Fine-tuning with LMs
Fine-tuning with LMs
Fine-tuning with LMs
Summarization with LMs
## Summarization with LMs

### Training Dataset

<table>
<thead>
<tr>
<th>Article #1 tokens</th>
<th>&lt;summarize&gt;</th>
<th>Article #1 Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article #2 tokens</td>
<td>&lt;summarize&gt;</td>
<td>Article #2 Summary</td>
</tr>
<tr>
<td>Article #3 tokens</td>
<td>&lt;summarize&gt;</td>
<td>Article #3 Summary</td>
</tr>
</tbody>
</table>

### Transformer-Decoder

Output #1
- Position #114
- Time step #1

Output #2
- Position #115
- Time step #2

1 ... 113 114 256
## GLUE Benchmark


https://gluebenchmark.com

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Data example</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoLA</td>
<td>Is the sentence grammatical or ungrammatical?</td>
<td>&quot;This building is than that one.&quot; = Ungrammatical</td>
<td>Matthews</td>
</tr>
<tr>
<td>SST-2</td>
<td>Is the movie review positive, negative, or neutral?</td>
<td>&quot;The movie is funny , smart , visually inventive , and most of all , alive .&quot; = .93056 (Very Positive)</td>
<td>Accuracy</td>
</tr>
<tr>
<td>MRPC</td>
<td>Is the sentence B a paraphrase of sentence A?</td>
<td>A) &quot;Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 .&quot; B) &quot;The island reported another 35 probable cases yesterday , taking its total to 418 .&quot; = A Paraphrase</td>
<td>Accuracy / F1</td>
</tr>
<tr>
<td>STS-B</td>
<td>How similar are sentences A and B?</td>
<td>A) &quot;Elephants are walking down a trail .&quot; B) &quot;A herd of elephants are walking along a trail .&quot; = 4.6 (Very Similar)</td>
<td>Pearson / Spearman</td>
</tr>
<tr>
<td>QQP</td>
<td>Are the two questions similar?</td>
<td>A) &quot;How can I increase the speed of my internet connection while using a VPN?&quot; B) &quot;How can Internet speed be increased by hacking through DNS?&quot; = Not Similar</td>
<td>Accuracy / F1</td>
</tr>
<tr>
<td>MNLI-mm</td>
<td>Does sentence A entail or contradict sentence B?</td>
<td>A) &quot;Tourist Information offices can be very helpful.&quot; B) &quot;Tourist information offices are never of any help.&quot; = Contradiction</td>
<td>Accuracy</td>
</tr>
<tr>
<td>QNLI</td>
<td>Does sentence B contain the answer to the question in sentence A?</td>
<td>A) &quot;What is essential for the mating of the elements that create radio waves?&quot; B) &quot;Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field.&quot; = Answerable</td>
<td>Accuracy</td>
</tr>
<tr>
<td>RTE</td>
<td>Does sentence A entail sentence B?</td>
<td>A) &quot;In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members.&quot; B) &quot;Yunus supported more than 50,000 Struggling Members.&quot; = Entailed</td>
<td>Accuracy</td>
</tr>
<tr>
<td>WNLI</td>
<td>Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?</td>
<td>A) &quot;Lily spoke to Donna, breaking her concentration.&quot; B) &quot;Lily spoke to Donna, breaking Lily's concentration.&quot; = Incorrect Referent</td>
<td>Accuracy</td>
</tr>
</tbody>
</table>

[Figure by Chris McCormick and Nick Ryan]
GLUE Benchmark Results

Fig. 1: Language Model Size & GLUE Performance

[Figure from Ahmet & Abdullah, 2020]
She enjoys playing tennis.
Bi-directional Pre-Training

enjoys playing tennis . <eos>

She enjoys playing tennis .

this task is trivially solved with bi-directional self-attention
Masked Language Model

Mask out 15% of tokens, then predict the missing tokens

She enjoys playing tennis.
Masked Language Model

Mask out 15% of tokens, then predict the missing tokens

She [MASK] playing tennis [MASK]
Masked Language Model

Mask out 15% of tokens, then predict the missing tokens

enjoys

She [MASK] playing tennis [MASK]
She [MASK] playing tennis enjoys.
Pre-Training with Masked LMs
Fine-Tuning with Masked LMs

[Diagram of BERT architecture with masked tokens and class label]
Fine-Tuning with Masked LMs

Figure 4: Illustrations of Fine-tuning BERT on Different Tasks.
Summarization with Masked LMs?
Summarization with Masked LMs?

*Bi-directional Masked LMs are not ideal for sequence-to-sequence tasks*

positrionic

brain
GLUE Benchmark Results

Fig. 1: Language Model Size & GLUE Performance

[Figure from Ahmet & Abdullah, 2020]
Encoder-Decoder Pre-Training

Cross-Attention

Self-Attention

Masked Self-Attention
Encoder-Decoder Pre-Training

Representative Model: T5
(T5 = Text-To-Text Transfer Transformer)
Thank you for inviting me to your party last week.

Inputs:
Thank you \(<X>\) me to your party \(<Y>\) week.
Encoder-Decoder Pre-Training

Original text:
Thank you for inviting me to your party last week.

Inputs:
Thank you <X> me to your party <Y> week.

Targets:
<X> for inviting <Y> last <Z>

Self-Attention

Cross-Attention

Masked Self-Attention
Encoder-Decoder Pre-Training

Original text:
Thank you for inviting me to your party last week.

Inputs:
Thank you \(<X>\) me to your party \(<Y>\) week.

Targets:
\(<X>\) for inviting \(<Y>\) last \(<Z>\)
Encoder-Decoder Pre-Training

Original text:
Thank you for inviting me to your party last week.

Inputs:
Thank you \(<X>\) me to your party \(<Y>\) week.

Targets:
\(<X>\) for inviting \(<Y>\) last \(<Z>\)

for inviting \(<Y>\) last \(<Z>\) \(<s>\)
Encoder-Decoder Pre-Training

WWW → T5
Encoder-Decoder Fine-tuning

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"sts1 sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi."

"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."
GLUE Benchmark Results

Fig. 1: Language Model Size & GLUE Performance

[Figure from Ahmet & Abdullah, 2020]
# GLUE Benchmark Results

![Figure by Chris McCormick and Nick Ryan](https://example.com/figure.png)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model Description</th>
<th>URL Score</th>
<th>CoLA</th>
<th>SST-2</th>
<th>MRPC</th>
<th>STS-B</th>
<th>QQP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T5 Team - Google</td>
<td>89.7</td>
<td>70.8</td>
<td>97.1</td>
<td>91.9</td>
<td>89.2</td>
<td>92.5/92.1</td>
</tr>
<tr>
<td>2</td>
<td>ALBERT-Team Google LanguageALBERT (Ensemble)</td>
<td>89.4</td>
<td>69.1</td>
<td>97.1</td>
<td>93.4</td>
<td>91.2</td>
<td>92.5/92.0</td>
</tr>
<tr>
<td>3+</td>
<td>王玮 ALICE v2 large ensemble (Alibaba DAMO NLP)</td>
<td>89.0</td>
<td>69.2</td>
<td>97.1</td>
<td>93.6</td>
<td>91.5</td>
<td>92.7/92.3</td>
</tr>
<tr>
<td>4</td>
<td>Microsoft D365 AI &amp; UMD FreeLB-RoBERTa (ensemble)</td>
<td>88.8</td>
<td>68.0</td>
<td>96.8</td>
<td>93.1</td>
<td>90.8</td>
<td>92.4/92.2</td>
</tr>
<tr>
<td>5</td>
<td>Facebook AI RoBERTa</td>
<td>88.5</td>
<td>67.8</td>
<td>96.7</td>
<td>92.3</td>
<td>89.8</td>
<td>92.2/91.9</td>
</tr>
<tr>
<td>6</td>
<td>XLNet Team XLNet-Large (ensemble)</td>
<td>88.4</td>
<td>67.8</td>
<td>96.8</td>
<td>93.0</td>
<td>90.7</td>
<td>91.6/91.1</td>
</tr>
<tr>
<td>7</td>
<td>Microsoft D365 AI &amp; MSR AI MT-DNN-ensemble</td>
<td>87.6</td>
<td>68.4</td>
<td>96.5</td>
<td>92.7</td>
<td>90.3</td>
<td>91.1/90.7</td>
</tr>
<tr>
<td>8</td>
<td>GLUE Human Baselines GLUE Human Baselines</td>
<td>87.1</td>
<td>66.4</td>
<td>97.8</td>
<td>86.3</td>
<td>80.8</td>
<td>92.7/92.6</td>
</tr>
<tr>
<td>9</td>
<td>Stanford Hazy Research Snorkel MeTaL</td>
<td>83.2</td>
<td>63.8</td>
<td>96.2</td>
<td>91.5</td>
<td>88.5</td>
<td>90.1/89.7</td>
</tr>
<tr>
<td>10</td>
<td>XLM Systems XLM (English only)</td>
<td>83.1</td>
<td>62.9</td>
<td>95.6</td>
<td>90.7/87.1</td>
<td>88.8/88.2</td>
<td>73.2/89.8</td>
</tr>
</tbody>
</table>
Types of Transformer Models

- **Decoder only**
  - e.g. [GPT]

- **Encoder only**
  - e.g. [BERT]

- **Encoder-Decoder**
  - e.g. [T5]
Types of Task-Specific Adaptation

Fine-tuning: modify existing model parameters

Adapter modules: freeze existing parameters; insert and train new layers

Prompting: re-formulate a task in natural language (e.g. fill-in-the-blank)
Example from GPT-3 LM

System Prompt (Human-written)

In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Model Completion (Machine-written, 10 tries)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. “By the time we reached the top of one peak, the water looked blue, with some crystals on top,” said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, “We can see, for example, that they have a common ‘language,’ something like a dialect or dialectic.”

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, “In South America, such incidents seem to be quite common.”
Examples of Prompting

Prompt

I am a highly intelligent question answering bot. If you ask me a question that is rooted in truth, I will give you the answer. If you ask me a question that is nonsense, trickery, or has no clear answer, I will respond with “Unknown”.

Q: What is human life expectancy in the United States?
A: Human life expectancy in the United States is 78 years.

Q: Who was president of the United States in 1955?
A: Dwight D. Eisenhower was president of the United States in 1955.

Q: Which party did he belong to?
A: He belonged to the Republican Party.

Q: What is the square root of banana?
A: Unknown

Q: How does a telescope work?
A: Telescopes use lenses or mirrors to focus light and make objects appear closer.

Q: Where were the 1992 Olympics held?
A: The 1992 Olympics were held in Barcelona, Spain.

Q: How many squigs are in a bonk?
A: Unknown

Q: Where is the Valley of Kings?
A:

Sample response

The Valley of Kings is located in Luxor, Egypt.

Prompt

The following is a list of companies and the categories they fall into:

Apple, Facebook, FedEx

Apple
Category:

Sample response

Technology
Facebook
Category: Social Media
Fedex
Category: Delivery
The Era of Rapid Scaling

[Amodei, Hernandez, et al. / OpenAI]
The Era of Rapid Scaling

Two Distinct Eras of Compute Usage in Training AI Systems

[Amodei, Hernandez, et al. / OpenAI]
The Era of Rapid Scaling in NLP

ELMo is an RNN model; all others are Transformer models.