Why do Transformers work so well?

Fabio James Fehr

18 March 2024
Outline

Personal Background

A Brief History of NLP and Deep Learning

The Attention Mechanism

My PhD Research
Personal Background

A Brief History of NLP and Deep Learning

The Attention Mechanism

My PhD Research
Who is this guy?

UCT

- **BBusSci**: Analytics (2015-2018)
  Thesis: Natural Language Processing
  - Stefan Britz
Who is this guy?

UCT

- **BBusSci**: Analytics (2015-2018)
  Thesis: Natural Language Processing
  - Stefan Britz

- **MSc**: Statistics (2019-2020)
  Thesis: Nonparametric methods vs deep learning - Allan Clark
Who is this guy?

**UCT**

- **BBusSci**: Analytics (2015-2018)
  Thesis: Natural Language Processing
  - Stefan Britz
- **MSc**: Statistics (2019-2020)
  Thesis: Nonparametric methods vs deep learning - Allan Clark

**EPFL & Idiap - Switzerland**

- **PhD** Electrical Engineering (2021-2025)
  Nonparametric methods for NLP
Personal Background

A Brief History of NLP and Deep Learning

The Attention Mechanism

My PhD Research
Why should we care?

The “AI Revolution”
- ChatGPT (Text)
- MidJourney (Images)
- AlphaGo (Games)
- Siri (Audio)
- etc ...
Why should we care?

The “AI Revolution”
- ChatGPT (Text)
- MidJourney (Images)
- AlphaGo (Games)
- Siri (Audio)
- etc ...
Why should we care?

The “AI Revolution”
- ChatGPT (Text)
- MidJourney (Images)
- AlphaGo (Games)
- Siri (Audio)
- etc ...

What is the secret sauce?
Why should we care?

The “AI Revolution”

- ChatGPT (Text)
- MidJourney (Images)
- AlphaGo (Games)
- Siri (Audio)
- etc ...

What is the secret sauce?

- The attention mechanism (NLP)
- Large-scale pretraining (Deep Learning)
Problem: Machine Translation

Given text examples of French and English language pairs, translate the following:

\[ \text{Il m’a entarté} = \text{???} \]

\text{French} \quad \text{English}
How do we solve translation? (2000s)

Tokenization
- Break into smaller units (words)

Je suis un chat.  I am a cat.

I am a cat.  Je suis un chat.

UCT Seminar  A Brief History of NLP and Deep Learning
How do we solve translation? (2000s)

Tokenization
- Break into smaller units (words)
- Build vocabulary

Je suis un chat. I am a cat.

\[
x_1 \quad x_2 \quad x_3 \quad x_4 \\
y_1 \quad y_2 \quad y_3 \quad y_4
\]
How do we solve translation? (2000s)

Tokenization
- Break into smaller units (words)
- Build vocabulary

Vectorisation
- Make into numbers (0 or 1)

<table>
<thead>
<tr>
<th>I am a cat.</th>
<th>Je suis un chat.</th>
<th>I am a cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$ 1</td>
<td>$x_2$ 0</td>
<td>$x_3$ 0</td>
</tr>
<tr>
<td>$y_1$ 1</td>
<td>$y_2$ 0</td>
<td>$y_3$ 0</td>
</tr>
<tr>
<td>0 1 0 0</td>
<td>0 0 1 0</td>
<td>0 0 1 0</td>
</tr>
<tr>
<td>0 0 0 1</td>
<td>0 0 0 1</td>
<td>0 0 0 1</td>
</tr>
</tbody>
</table>
How do we solve translation? (2000s)

**Tokenization**
- Break into smaller units (words)
- Build vocabulary

**Vectorisation**
- Make into numbers (0 or 1)

**Classification**
- Multinomial logistic regression (per word) or Naive Bayes

![Diagram of translation model]

**Example:**

I am a cat.

Je suis un chat.

Naive Model
How do we solve translation? (2000s)

Tokenization
- Break into smaller units (words)
- Build vocabulary

Vectorisation
- Make into numbers (0 or 1)

Semantics?

Classification
- Multinomial logistic regression (per word) or Naive Bayes
  Simplistic? Independence?

Naive Model

I am a cat.
Je suis un chat.

Naive Model

UCT Seminar
A Brief History of NLP and Deep Learning
How do we solve translation? (<2014)

Models are too simple?
• Neural networks

Complex (Naive) Model

I am a cat.
Je suis un chat.

 Complex (Naive) Model

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>y1</th>
<th>y2</th>
<th>y3</th>
<th>y4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
How do we solve translation? (<2014)

Models are too simple?
- Neural networks

Contextualised representations?
- Word2Vec [Mikolov et al., 2013]

I am a cat.
Je suis un chat.

![Diagram showing translation models and word vectors](attachment://diagram.png)
How do we solve translation? (<2014)

Models are too simple?
- Neural networks

Contextualised representations?
- Word2Vec [Mikolov et al., 2013]

Models assume independence?
- Recurrent Neural Networks

The image shows a comparison of the English phrase “I am a cat.” to the French translation “Je suis un chat.” with corresponding vector representations for each word in both languages. The model takes into account the context of each word to make translations more accurate.
How do we solve translation? (<2014)

Models are too simple?
- Neural networks

Contextualised representations?
- Word2Vec [Mikolov et al., 2013]

Models assume independence?
- Recurrent Neural Networks
  Long term dependencies?
  Recurrence on GPUs?
  Vanishing gradients?
How do we solve translation? (2014 and beyond)

Long term dependencies?
- Attention for translation [Bahdanau et al., 2014]

Figure: Bidirectional LSTM with attention

Jean suis un chat.
... a cat.
How do we solve translation? (2014 and beyond)

Long term dependencies?
- Attention for translation
  [Bahdanau et al., 2014]

Recurrence?
- Transformers - only attention
  [Vaswani et al., 2017]

**Figure: Transformer**
How do we solve translation? (2014 and beyond)

Long term dependencies?
- Attention for translation
  [Bahdanau et al., 2014]

Recurrence?
- Transformers - only attention
  [Vaswani et al., 2017]

Scale
- The bitter lesson [Sutton, 2019]
How do we solve translation? (2014 and beyond)

Long term dependencies?
- Attention for translation [Bahdanau et al., 2014]

Recurrence?
- Transformers - only attention [Vaswani et al., 2017]

Scale
- The bitter lesson [Sutton, 2019]

Pretraining
- “Large” foundation models (2017-2021)

Figure: Pretrained models [Han et al., 2021]
How do we solve translation? (2014 and beyond)

Long term dependencies?
- Attention for translation
  [Bahdanau et al., 2014]

Recurrence?
- Transformers - only attention
  [Vaswani et al., 2017]

Scale
- The bitter lesson [Sutton, 2019]

Pretraining
- “Large” foundation models (2017-2021)
- Large Language Models (2023)
Takeaways:

The secret sauce:

1. The attention mechanism
2. Scaling data and model size
Takeaways:

The secret sauce:
1. The attention mechanism
2. Scaling data and model size

Il m’a entarté = ???

French  English
Takeaways:

The secret sauce:

1. The attention mechanism
2. Scaling data and model size

\[ \text{Il m’a entarté} = \text{???} \]

\text{French} \quad \text{English}

Figure: Google translate
Takeaways:

The secret sauce:
1. The attention mechanism
2. Scaling data and model size

Il m’a entarté = ???

French

English

Figure: DeepL
Takeaways:

The secret sauce:

1. The attention mechanism
2. Scaling data and model size

Il m’a entarté = ???

French

English

Figure: ChatGPT
Takeaways:

The secret sauce:
1. The attention mechanism
2. Scaling data and model size

Il m’a entarté = ???

French

English
Personal Background

A Brief History of NLP and Deep Learning

The Attention Mechanism

My PhD Research
Data

- \( N \) observations, \( P \) dimensional
- No particular order
- Common across domains

\[ Z \in \mathbb{R}^{N \times P} \]

- \( N \) observations, \( P \) dimensional
Data

- $N$ observations, $P$ dimensional
- No particular order

$Z \in \mathbb{R}^{N \times P}$
Data

• $N$ observations, $P$ dimensional
• No particular order
• Common across domains

$Z \in \mathbb{R}^{N \times P}$
Data

\[ Z \in \mathbb{R}^{N \times P} \]

\[ U \in \mathbb{R}^{M \times P} \]

- \( N \) observations, \( P \) dimensional
- No particular order
- Common across domains

- \( M \) observations, \( P \) dimensional
Data

- $N$ observations, $P$ dimensional
- No particular order
- Common across domains

$Z \in \mathbb{R}^{N \times P}$

- $M$ observations, $P$ dimensional
- No particular order

$U \in \mathbb{R}^{M \times P}$
Data

- $N$ observations, $P$ dimensional
- No particular order
- Common across domains

$$Z \in \mathbb{R}^{N \times P}$$

- $M$ observations, $P$ dimensional
- No particular order
- $U = Z$ or $U \neq Z$

$$U \in \mathbb{R}^{M \times P}$$
How do we get $Z$ and $U$ to interact?

- $U = Z$: Variance ($\text{Var}(Z, Z)$)
- $U \neq Z$: Covariance ($\text{Cov}(U, Z)$)
How do we get $Z$ and $U$ to interact?

- $U = Z$: Variance($Z, Z$)
- $U \neq Z$: Covariance($U, Z$)
How do we project this interaction forward?

- $U = Z$: Variance($Z$, $Z$)
- $U \neq Z$: Covariance($U$, $Z$)
How do we project this interaction forward?

- \( U = Z \): Variance \((Z, Z)\)
- Project by \( Z \)

- \( U \neq Z \): Covariance \((U, Z)\)
- Project by \( Z \)
What are the problems with this?

- Strictly positive multiplications?
- Normalisations and scaling?
- Single interaction value?
- Quadratic?

\[ ZZ^T Z \in \mathbb{R}^{N \times P} \quad \text{and} \quad UZ^T Z \in \mathbb{R}^{M \times P} \]
What are the problems with this?

- Strictly positive multiplications?
- Normalisations and scaling?
- Single interaction value?
- Quadratic?

\[
ZZ^T Z \in \mathbb{R}^{N \times P}
\]

\[
UZ^T Z \in \mathbb{R}^{M \times P}
\]
What are the problems with this?

- Strictly positive multiplications?
- Normalisations and scaling?

\[ ZZ^T Z \in \mathbb{R}^{N \times P} \]

\[ U Z^T Z \in \mathbb{R}^{M \times P} \]
What are the problems with this?

- Strictly positive multiplications?
- Normalisations and scaling?
- Single interaction value?
What are the problems with this?

- Strictly positive multiplications?
- Normalisations and scaling?
- Single interaction value?
- Quadratic?
The attention mechanism

\[
Q \quad K \quad V
\]

\[
Z
\]

\text{Figure: Self attention}

\[
Q \quad K \quad V
\]

\[
U \quad Z
\]

\text{Figure: Cross attention}

\begin{itemize}
  \item \(W^Q, W^K, W^V \in \mathbb{R}^{P \times d}\)
\end{itemize}
The attention mechanism

- Cross attention

\[ QK^T \in \mathbb{R}^{M \times N} \]
The attention mechanism

- Cross attention
- Scaling

\[
\left( \frac{QK^T}{\sqrt{d}} \right) \in \mathbb{R}^{M \times N}
\]
The attention mechanism

- Cross attention
- Scaling
- Normalisation

\[
\text{Softmax} \left( \frac{QK^T}{\sqrt{d}} \right) \in \mathbb{R}^{M \times N}
\]
The attention mechanism

- Cross attention
- Scaling
- Normalisation
- Projection

\[
\text{Softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V \in \mathbb{R}^{M \times d}
\]
Why is attention so cool?

Advantages:
• Simple layer
• No particular order
• Unbounded inputs
• Probabilistic interpretation

\[
\text{Softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V \in \mathbb{R}^{M \times d}
\]
Why is attention so cool?

Advantages:
- Simple layer
- No particular order
- Unbounded inputs
- Probabilistic interpretation

Problems solved:
- Strictly positive multiplications? ✓

\[
\text{Softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V \in \mathbb{R}^{M \times d}
\]
Why is attention so cool?

Advantages:
- Simple layer
- No particular order
- Unbounded inputs
- Probabilistic interpretation

Problems solved:
- Strictly positive multiplications? ✓
- Normalisations and scaling? ✓
Why is attention so cool?

**Advantages:**
- Simple layer
- No particular order
- Unbounded inputs
- Probabilistic interpretation

**Problems solved:**
- Strictly positive multiplications? ✓
- Normalisations and scaling? ✓
- Single interaction value? ✓

\[
\text{Softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V \in \mathbb{R}^{M \times d}
\]
Why is attention so cool?

Advantages:
- Simple layer
- No particular order
- Unbounded inputs
- Probabilistic interpretation

Problems solved:
- Strictly positive multiplications? ✓
- Normalisations and scaling? ✓
- Single interaction value? ✓
- Quadratic? ✗
Personal Background

A Brief History of NLP and Deep Learning

The Attention Mechanism

My PhD Research
Intuition

Attention mechanism:

- No particular order
- Unbounded inputs
- Probabilistic interpretation
Intuition

Attention mechanism:
• No particular order
• Unbounded inputs
• Probabilistic interpretation

Nonparametric distributions
• Exchangable set
• Theoretically infinite
• Mixture of distributions

\[ \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V \in \mathbb{R}^{M \times d} \]
Nonparametric latent variable modelling

Regularisation:
- Generalisation
- Sparse representations
Regularisation:
- Generalisation
- Sparse representations

Generation:
- Generative modelling
Nonparametric latent variable modelling

Regularisation:
- Generalisation
- Sparse representations

Generation:
- Generative modelling

Explainability:
- Disentanglement
Nonparametric latent variable modelling

Regularisation:
• Generalisation
• Sparse representations

Generation:
• Generative modelling

Explainability:
• Disentanglement
The denoising attention mechanism

$$\text{Attn}(QKV) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} \right)V$$
The denoising attention mechanism

\[ \text{Attn}(UZ) = \text{Softmax} \left( \frac{UZ^T}{\sqrt{d}} \right) Z \]
The denoising attention mechanism

\[ \text{Attn}(UZ) = \text{Softmax} \left( \frac{UZ^T}{\sqrt{d}} \right) Z \]
The denoising attention mechanism

\[ \text{Attn}(UZ) = \text{Softmax} \left( \frac{UZ^T}{\sqrt{d}} \right) Z \]

- Prior information \( Z \)
The denoising attention mechanism

\[
\text{Attn}(UZ) = \text{Softmax} \left( \frac{UZ^T}{\sqrt{d}} \right) Z
\]

- Prior information \( Z \)
- Noisy query \( U \)
The denoising attention mechanism

\[ \text{Attn}(UZ) = \text{Softmax} \left( \frac{UZ^T}{\sqrt{d}} \right) Z \]

- Prior information \( Z \)
- Noisy query \( U \)
- Posterior update \( \text{Attn}(UZ) \)
The denoising attention mechanism

\[ \text{Attn}(UZ) = \text{Softmax} \left( \frac{UZ^T}{\sqrt{d}} \right) Z \]

- Prior information \( Z \)
- Noisy query \( U \)
- Posterior update \( \text{Attn}(UZ) \)
- Proof of concept [Henderson and Fehr, 2023]
Current research

Theory:
- A VAE for Transformers [Henderson and Fehr, 2023]

Figure: Nonparametric Variational Information Bottleneck (NVIB)
Current research

Theory:
• A VAE for Transformers [Henderson and Fehr, 2023]

Regularisation:
• Post-training regularisation [Fehr and Henderson, 2023]

Figure: Nonparametric Variational Regularisation
Current research

Theory:
• A VAE for Transformers
  [Henderson and Fehr, 2023]

Regularisation:
• Post-training regularisation
  [Fehr and Henderson, 2023]

Explainability:
• Abstraction
  [Behjati et al., 2023]

Figure: Learned layer-wise abstraction
Current research

Theory:
• A VAE for Transformers
  [Henderson and Fehr, 2023]

Regularisation:
• Post-training regularisation
  [Fehr and Henderson, 2023]

Explainability:
• Abstraction
  [Behjati et al., 2023]

Generation:
• Coming soon 2024!

Figure: Latent attention-based diffusion for text
Fin

Personal Background
A Brief History of NLP and Deep Learning
The Attention Mechanism
My PhD Research
Fin

Personal Background
A Brief History of NLP and Deep Learning
The Attention Mechanism
My PhD Research

Thank you for your attention!
Neural machine translation by jointly learning to align and translate.

Learning to abstract with nonparametric variational information bottleneck.

Nonparametric variational regularisation of pretrained transformers.

Pre-trained models: Past, present and future.
AI Open, 2:225–250.

A VAE for Transformers with Nonparametric Variational Information Bottleneck.
In International Conference on Learning Representations.

Efficient estimation of word representations in vector space.
In International Conference on Learning Representations.
The bitter lesson.

Attention is all you need.