Why do Transformers work so well?

Fabio James Fehr

18 March 2024

idiap RESEARCH INSTITUTE

nccr evolving language NATIONAL CENTRE OF COMPETENCE IN RESEARCH

EPFL

Outline

Who is this guy?

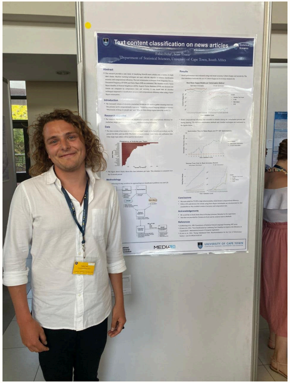
UCT

- **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

Why should we care?

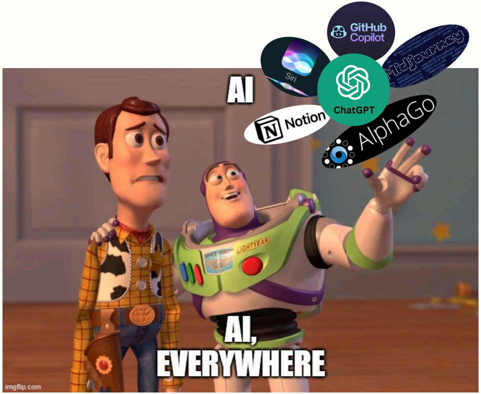
The "AI Revolution"

- 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 ...

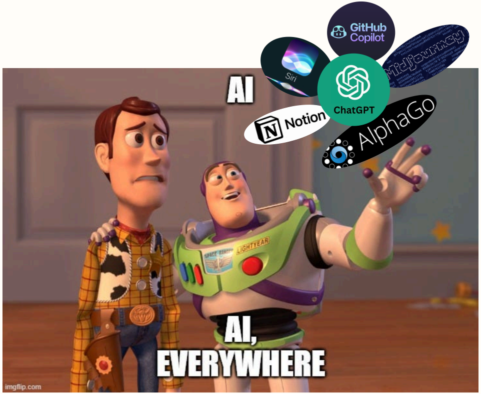


Why should we care?

The "AI Revolution"

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

What is the secret sauce?



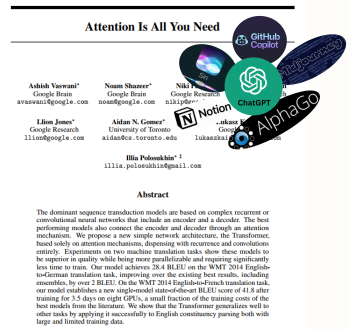
Why should we care?

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)



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Google Research

Llion Jones* Google Research llion@google.com

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

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.

Problem: Machine Translation

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

$$\underbrace{\textbf{Il m'a entarté}}_{\textit{French}} = \underbrace{???}_{\textit{English}}$$

How do we solve translation? (2000s)

Tokenization

- Break into smaller units (words)

$\underbrace{\text{Je}}\ \underbrace{\text{suis}}\ \underbrace{\text{un}}\ \underbrace{\text{chat.}}$ $\quad$ $\underbrace{\text{I}}\ \underbrace{\text{am}}\ \underbrace{\text{a}}\ \underbrace{\text{cat.}}$

How do we solve translation? (2000s)

Tokenization

- Break into smaller units (words)
- Build vocabulary

$$\underbrace{\text{Je}}_{x_1}\ \underbrace{\text{suis}}_{x_2}\ \underbrace{\text{un}}_{x_3}\ \underbrace{\text{chat.}}_{x_4} \qquad \underbrace{\text{I}}_{y_1}\ \underbrace{\text{am}}_{y_2}\ \underbrace{\text{a}}_{y_3}\ \underbrace{\text{cat.}}_{y_4}$$

How do we solve translation? (2000s)

Tokenization

- Break into smaller units (words)
- Build vocabulary

Vectorisation

- Make into numbers (0 or 1)

$$\underbrace{\text{Je}}_{x_1}\ \underbrace{\text{suis}}_{x_2}\ \underbrace{\text{un}}_{x_3}\ \underbrace{\text{chat.}}_{x_4} \qquad \underbrace{\text{I}}_{y_1}\ \underbrace{\text{am}}_{y_2}\ \underbrace{\text{a}}_{y_3}\ \underbrace{\text{cat.}}_{y_4}$$

$$\boxed{\begin{matrix}1\\0\\0\\0\\\vdots\end{matrix}}\ \boxed{\begin{matrix}0\\1\\0\\0\\\vdots\end{matrix}}\ \boxed{\begin{matrix}0\\0\\1\\0\\\vdots\end{matrix}}\ \boxed{\begin{matrix}0\\0\\0\\1\\\vdots\end{matrix}} \qquad \boxed{\begin{matrix}1\\0\\0\\0\\\vdots\end{matrix}}\ \boxed{\begin{matrix}0\\1\\0\\0\\\vdots\end{matrix}}\ \boxed{\begin{matrix}0\\0\\1\\0\\\vdots\end{matrix}}\ \boxed{\begin{matrix}0\\0\\0\\1\\\vdots\end{matrix}}$$

How do we solve translation? (2000s)

Tokenization

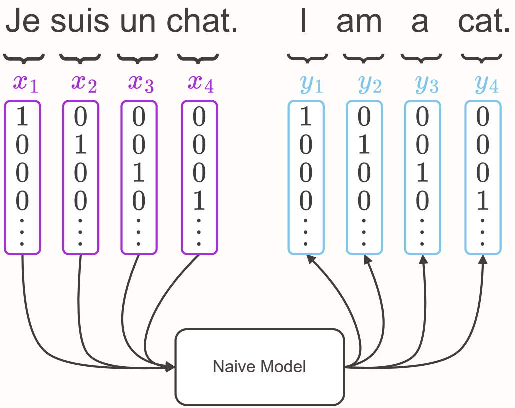
- Break into smaller units (words)
- Build vocabulary

Vectorisation

- Make into numbers (0 or 1)

Classification

- Multinomial logistic regression (per word) or Naive Bayes



How do we solve translation? (2000s)

Tokenization

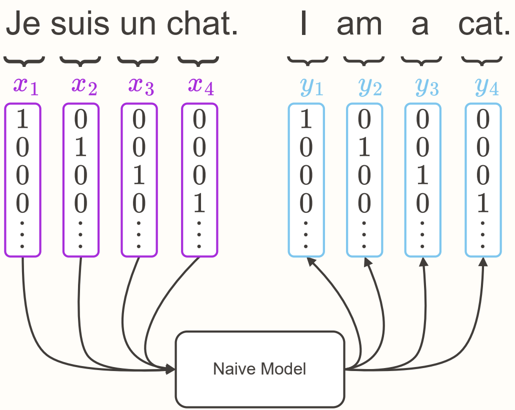
- Break into smaller units (words)
- Build vocabulary

Vectorisation

- Make into numbers (0 or 1)
  Semantics?

Classification

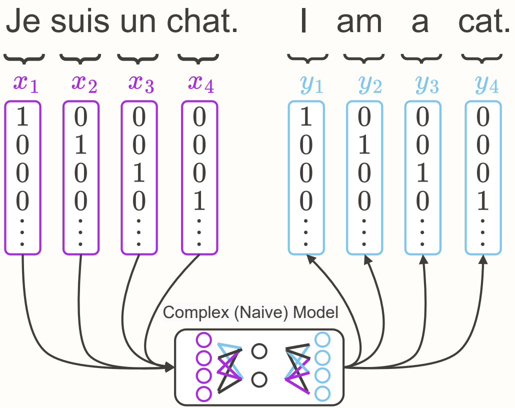
- Multinomial logistic regression (per word) or Naive Bayes
  Simplistic? Independence?



How do we solve translation? (<2014)

Models are too simple?

- Neural networks



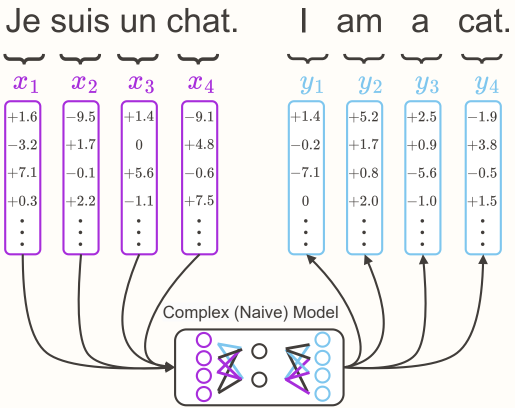
How do we solve translation? (<2014)

Models are too simple?

- Neural networks

Contextualised representations?

- Word2Vec [Mikolov et al., 2013]



How do we solve translation? (<2014)

Models are too simple?

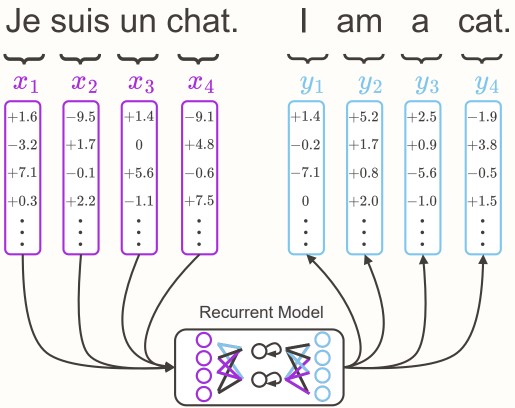
- Neural networks

Contextualised representations?

- Word2Vec [Mikolov et al., 2013]

Models assume independence?

- Recurrent Neural Networks



How do we solve translation? (<2014)

Models are too simple?

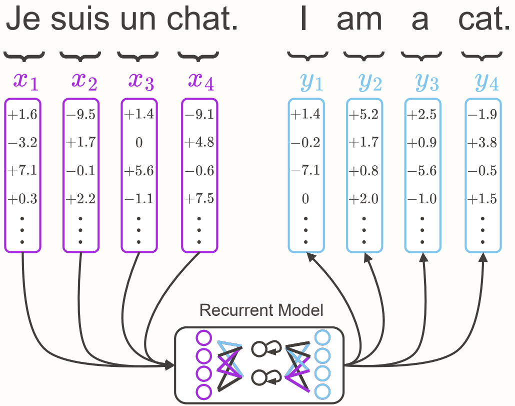
- Neural networks

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]

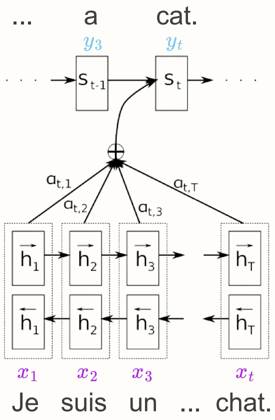


Figure: Bidirectional LSTM with attention

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]

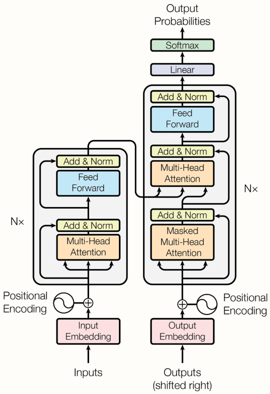


Figure: Transformer

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)

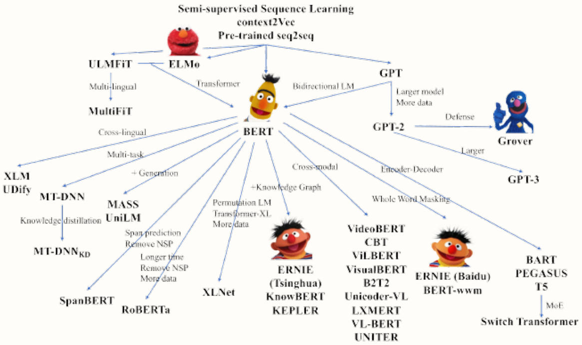


Figure: Pretrained models [Han et al., 2021]

How do we solve translation? (2014 and beyond)

Long term dependencies?

- Attention for translation [Bahdanau et al., 2014]

Recurrence?

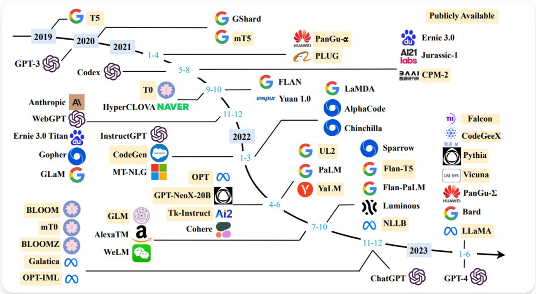
- Transformers - only attention [Vaswani et al., 2017]

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

$$\underbrace{\textbf{Il m'a entarté}}_{\textit{French}} = \underbrace{???}_{\textit{English}}$$

Takeaways:

The secret sauce:

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

$$\underbrace{\textbf{Il m'a entarté}}_{\textit{French}} = \underbrace{???}_{\textit{English}}$$

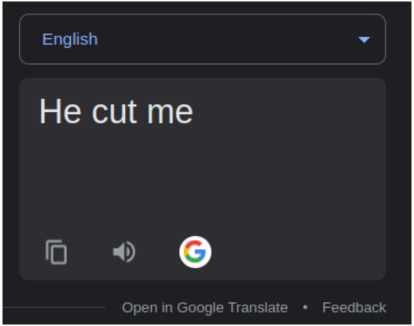


Figure: Google translate

Takeaways:

The secret sauce:

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

$$\underbrace{\textbf{Il m'a entarté}}_{\textit{French}} = \underbrace{???}_{\textit{English}}$$

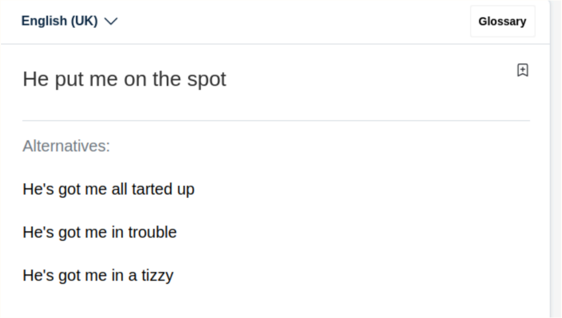


Figure: DeepL

Takeaways:

The secret sauce:

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

$$\underbrace{\textbf{Il m'a entarté}}_{\textit{French}} = \underbrace{???}_{\textit{English}}$$

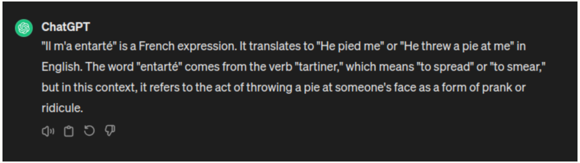


Figure: ChatGPT

Takeaways:

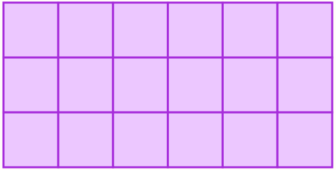
The secret sauce:

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

$$\underbrace{\textbf{Il m'a entarté}}_{\textit{French}} = \underbrace{???}_{\textit{English}}$$

Data

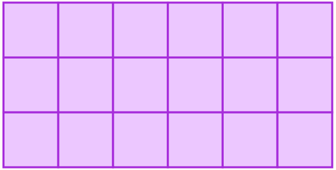
$\boldsymbol{Z} \in \mathbb{R}^{N \times P}$



- N observations, P dimensional

Data

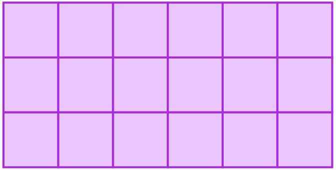
$\boldsymbol{Z} \in \mathbb{R}^{N\times P}$



- N observations, P dimensional
- No particular order

Data

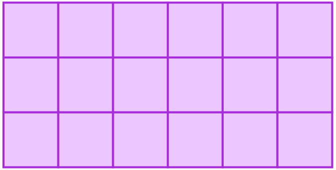
$\boldsymbol{Z} \in \mathbb{R}^{N\times P}$



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

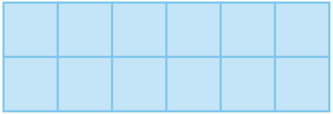
Data

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



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

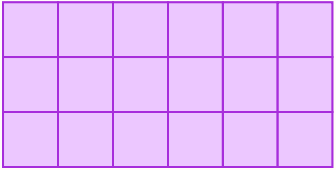
$\boldsymbol{U} \in \mathbb{R}^{M \times P}$



- M observations, P dimensional

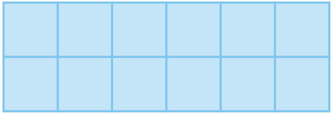
Data

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



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

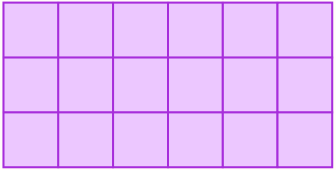
$\boldsymbol{U} \in \mathbb{R}^{M \times P}$



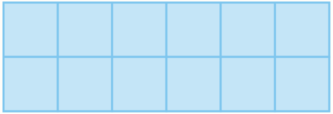
- M observations, P dimensional
- No particular order

Data

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



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

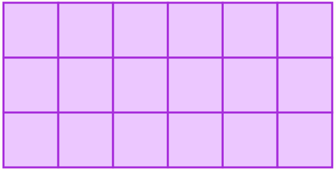


- N observations, P dimensional
- No particular order
- Common across domains

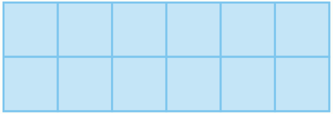
- M observations, P dimensional
- No particular order
- $\boldsymbol{U} = \boldsymbol{Z}$ or $\boldsymbol{U} \neq \boldsymbol{Z}$

How do we get Z and U to interact?

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

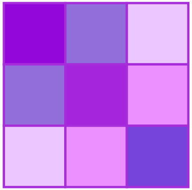


$\boldsymbol{U} \in \mathbb{R}^{M \times P}$

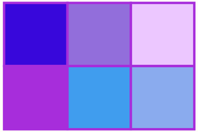


How do we get Z and U to interact?

$\boldsymbol{Z}\boldsymbol{Z}^{\boldsymbol{T}} \in \mathbb{R}_{+}^{N \times N}$



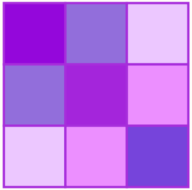
$\boldsymbol{U}\boldsymbol{Z}^{\boldsymbol{T}} \in \mathbb{R}^{M \times N}$



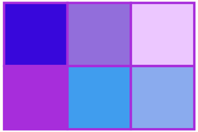
- $\boldsymbol{U} = \boldsymbol{Z}$: $Variance(\boldsymbol{Z}, \boldsymbol{Z})$
- $\boldsymbol{U} \neq \boldsymbol{Z}$: $Covariance(\boldsymbol{U}, \boldsymbol{Z})$

How do we project this interaction forward?

$\boldsymbol{Z}\boldsymbol{Z}^{\boldsymbol{T}} \in \mathbb{R}_{+}^{N \times N}$



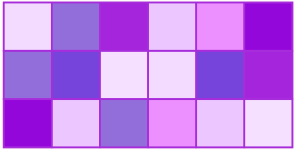
$\boldsymbol{U}\boldsymbol{Z}^{\boldsymbol{T}} \in \mathbb{R}^{M \times N}$



- $\boldsymbol{U} = \boldsymbol{Z}$: $Variance(\boldsymbol{Z}, \boldsymbol{Z})$
- $\boldsymbol{U} \neq \boldsymbol{Z}$: $Covariance(\boldsymbol{U}, \boldsymbol{Z})$

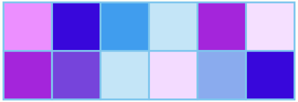
How do we project this interaction forward?

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



- $\boldsymbol{U} = \boldsymbol{Z}$: $Variance(\boldsymbol{Z}, \boldsymbol{Z})$
- Project by $\boldsymbol{Z}$

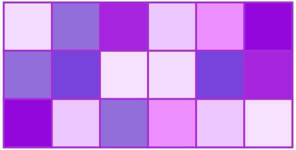
$\boldsymbol{U}\boldsymbol{Z}^{\mathbf{T}}\boldsymbol{Z} \in \mathbb{R}^{M \times P}$



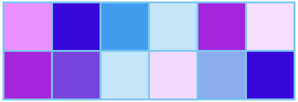
- $\boldsymbol{U} \neq \boldsymbol{Z}$: $Covariance(\boldsymbol{U}, \boldsymbol{Z})$
- Project by $\boldsymbol{Z}$

What are the problems with this?

$ZZ^TZ \in \mathbb{R}^{N\times P}$

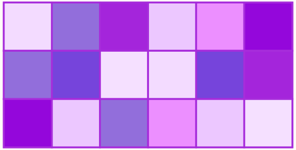


$UZ^TZ \in \mathbb{R}^{M\times P}$

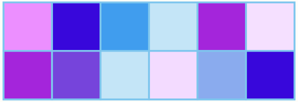


What are the problems with this?

$\boldsymbol{ZZ^TZ} \in \mathbb{R}^{N\times P}$



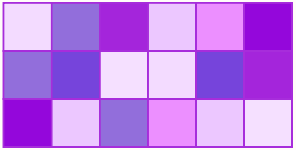
$\boldsymbol{UZ^TZ} \in \mathbb{R}^{M\times P}$



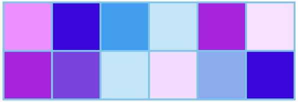
- Strictly positive multiplications?

What are the problems with this?

$\boldsymbol{ZZ^TZ} \in \mathbb{R}^{N \times P}$



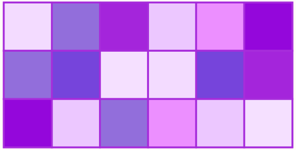
$\boldsymbol{UZ^TZ} \in \mathbb{R}^{M \times P}$



- Strictly positive multiplications?
- Normalisations and scaling?

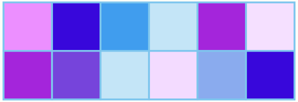
What are the problems with this?

$\boldsymbol{ZZ^TZ} \in \mathbb{R}^{N\times P}$



- Strictly positive multiplications?
- Normalisations and scaling?

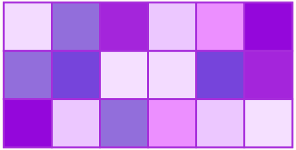
$\boldsymbol{UZ^TZ} \in \mathbb{R}^{M\times P}$



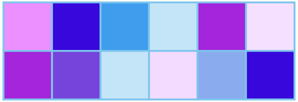
- Single interaction value?

What are the problems with this?

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



$$\boldsymbol{U}\boldsymbol{Z}^{\mathbf{T}}\boldsymbol{Z} \in \mathbb{R}^{M \times P}$$



- Strictly positive multiplications?
- Normalisations and scaling?

- Single interaction value?
- Quadratic?

The attention mechanism

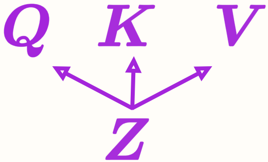


Figure: Self attention

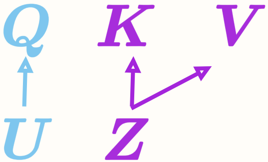


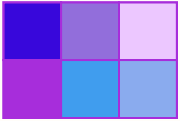
Figure: Cross attention

- $\boldsymbol{W}^{\boldsymbol{Q}}, \boldsymbol{W}^{\boldsymbol{K}}, \boldsymbol{W}^{\boldsymbol{V}} \in \mathbb{R}^{P \times d}$

The attention mechanism

- Cross attention

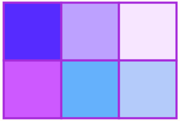
$$\boldsymbol{Q}\boldsymbol{K}^T \in \mathbb{R}^{M \times N}$$



The attention mechanism

- Cross attention
- Scaling

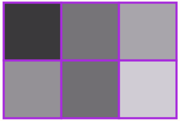
$$\left(\frac{Q\boldsymbol{K}^T}{\sqrt{d}}\right) \quad \in \mathbb{R}^{M\times N}$$



The attention mechanism

- Cross attention
- Scaling
- Normalisation

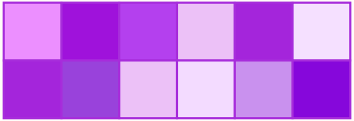
$$\text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \quad \in \mathbb{R}^{M\times N}$$



The attention mechanism

- Cross attention
- Scaling
- Normalisation
- Projection

$$\text{Softmax}\left(\frac{Q\boldsymbol{K}^T}{\sqrt{d}}\right)V \in \mathbb{R}^{M\times d}$$

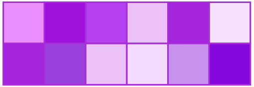


Why is attention so cool?

Advantages:

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

$$\mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \in \mathbb{R}^{M\times d}$$



Why is attention so cool?

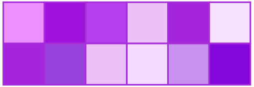
Advantages:

- 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?

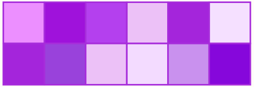
Advantages:

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

Problems solved:

- Strictly positive multiplications? ✓
- Normalisations and scaling? ✓

$$\text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \in \mathbb{R}^{M\times d}$$



Why is attention so cool?

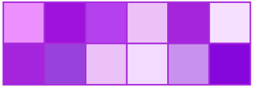
Advantages:

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

Problems solved:

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

$$\text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \in \mathbb{R}^{M \times d}$$



Why is attention so cool?

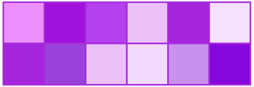
Advantages:

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

Problems solved:

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

$$\mathrm{Softmax}\left(\frac{Q\boldsymbol{K}^T}{\sqrt{d}}\right)V \in \mathbb{R}^{M\times d}$$

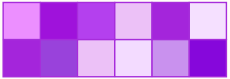


Intuition

Attention mechanism:

- No particular order
- Unbounded inputs
- Probabilistic interpretation

$$\text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \in \mathbb{R}^{M\times d}$$



Intuition

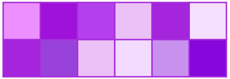
Attention mechanism:

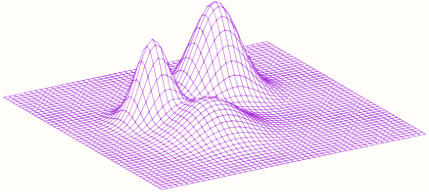
- No particular order
- Unbounded inputs
- Probabilistic interpretation

Nonparametric distributions

- Exchangable set
- Theoretically infinite
- Mixture of distributions

$$\mathrm{Softmax}\left(\frac{Q K^T}{\sqrt{d}}\right) V \in \mathbb{R}^{M \times d}$$

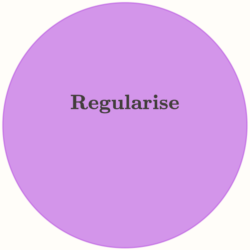




Nonparametric latent variable modelling

Regularisation:

- Generalisation
- Sparse representations



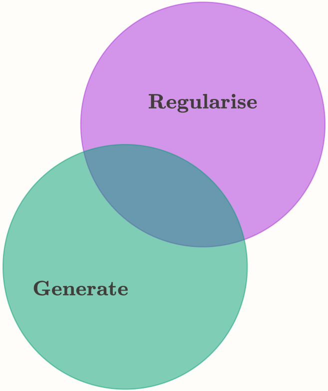
Nonparametric latent variable modelling

Regularisation:

- Generalisation
- Sparse representations

Generation:

- Generative modelling



Nonparametric latent variable modelling

Regularisation:

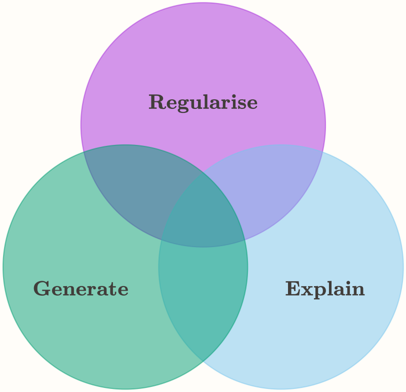
- Generalisation
- Sparse representations

Generation:

- Generative modelling

Explainability:

- Disentanglement



Nonparametric latent variable modelling

Regularisation:

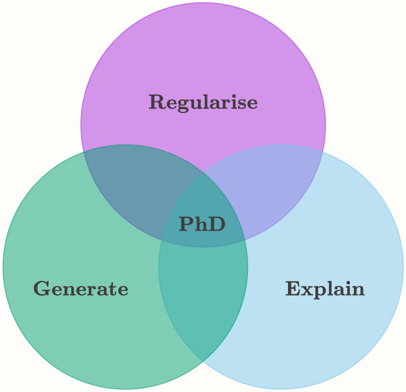
- Generalisation
- Sparse representations

Generation:

- Generative modelling

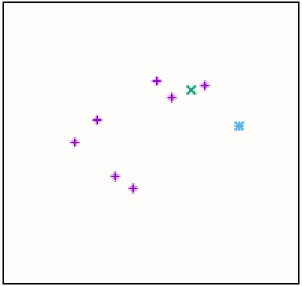
Explainability:

- Disentanglement



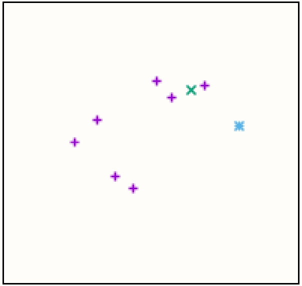
The denoising attention mechanism

$$\mathrm{Attn}(Q\,\boldsymbol{K}\,\boldsymbol{V}) = \mathrm{Softmax}\left(\frac{Q\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}$$



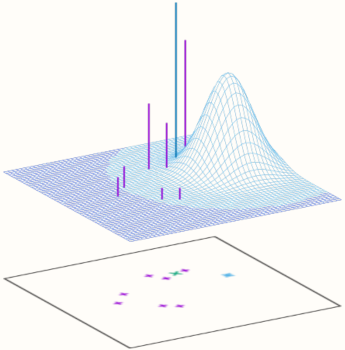
The denoising attention mechanism

$$\text{Attn}(\boldsymbol{U}\,\boldsymbol{Z}) = \text{Softmax}\left(\frac{\boldsymbol{U}\boldsymbol{Z}^T}{\sqrt{d}}\right)\boldsymbol{Z}$$



The denoising attention mechanism

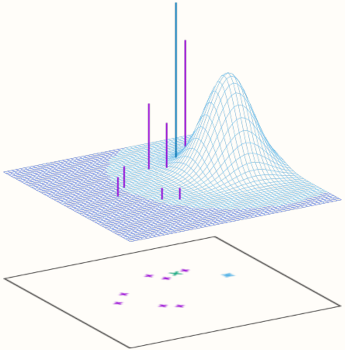
$$\mathrm{Attn}(\boldsymbol{U}\,\boldsymbol{Z}) = \mathrm{Softmax}\left(\frac{\boldsymbol{U}\boldsymbol{Z}^T}{\sqrt{d}}\right)\boldsymbol{Z}$$



The denoising attention mechanism

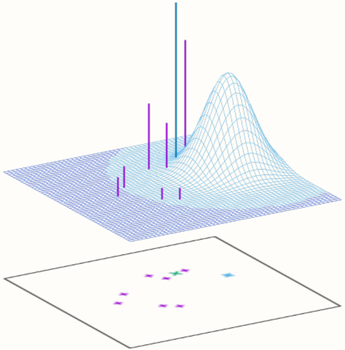
$$\mathrm{Attn}(\boldsymbol{U}\,\boldsymbol{Z}) = \mathrm{Softmax}\left(\frac{\boldsymbol{U}\boldsymbol{Z}^T}{\sqrt{d}}\right)\boldsymbol{Z}$$

- Prior information Z



The denoising attention mechanism

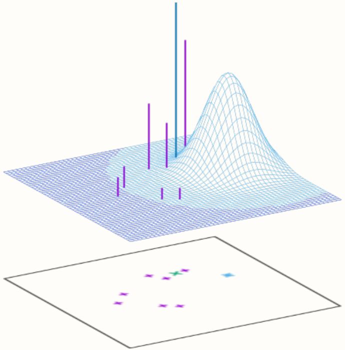
$$\mathrm{Attn}(\boldsymbol{U}\,\boldsymbol{Z}) = \mathrm{Softmax}\left(\frac{\boldsymbol{U}\boldsymbol{Z}^T}{\sqrt{d}}\right)\boldsymbol{Z}$$



- Prior information $\boldsymbol{Z}$
- Noisy query $\boldsymbol{U}$

The denoising attention mechanism

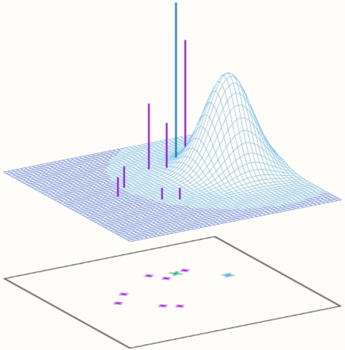
$$\mathrm{Attn}(\boldsymbol{U}\,\boldsymbol{Z}) = \mathrm{Softmax}\left(\frac{\boldsymbol{U}\boldsymbol{Z}^T}{\sqrt{d}}\right)\boldsymbol{Z}$$



- Prior information $\boldsymbol{Z}$
- Noisy query $\boldsymbol{U}$
- Posterior update $Attn(\boldsymbol{U}\,\boldsymbol{Z})$

The denoising attention mechanism

$$\mathrm{Attn}(\boldsymbol{U}\,\boldsymbol{Z}) = \mathrm{Softmax}\left(\frac{\boldsymbol{U}\boldsymbol{Z}^T}{\sqrt{d}}\right)\boldsymbol{Z}$$



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

Current research

Theory:

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

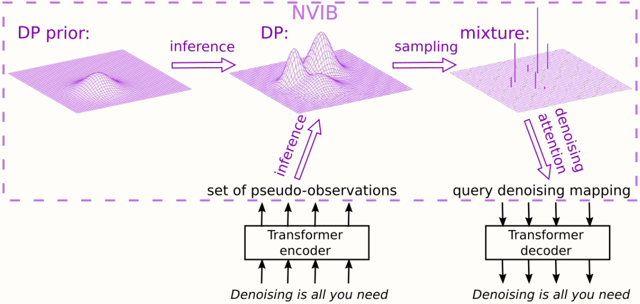


Figure: Nonparametric Variational Information Bottleneck (NVIB)

Current research

Theory:

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

Regularisation:

- Post-training regularisation [Fehr and Henderson, 2023]

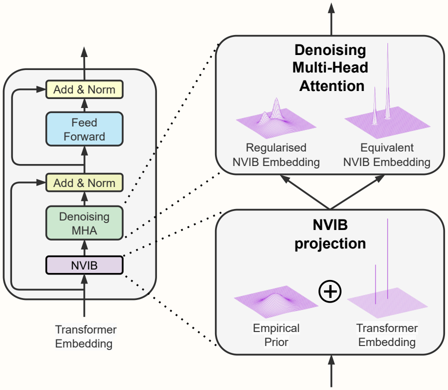


Figure: Nonparametric Variational Regularisation

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]

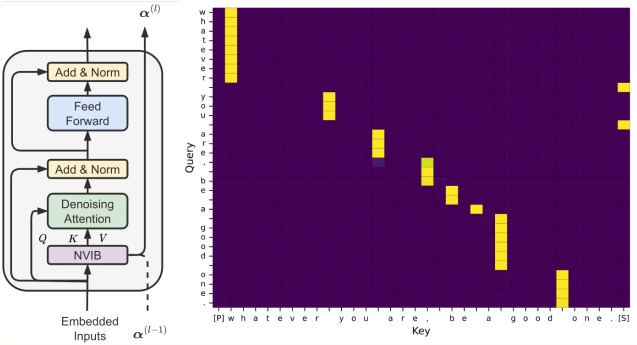


Figure: Learned layer-wise abstraction

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!

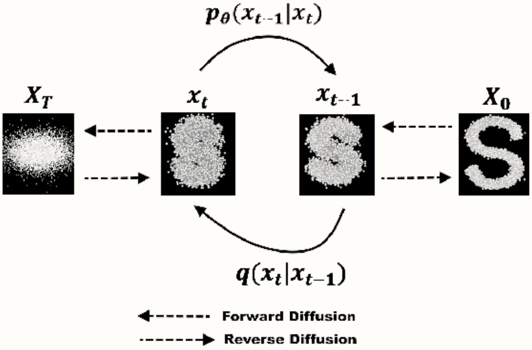
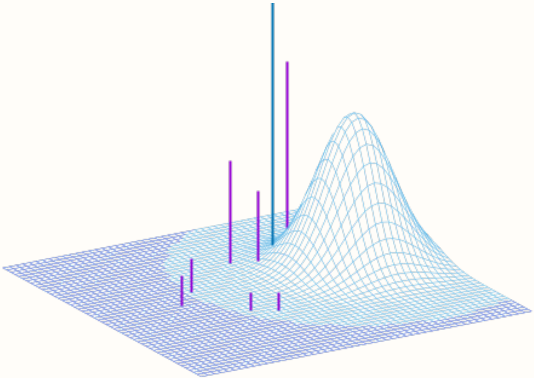


Figure: Latent attention-based diffusion for text

Fin

Fin



Thank you for your attention!

References I

Bahdanau, D., Cho, K., and Bengio, Y. (2014).
Neural machine translation by jointly learning to align and translate.
arXiv preprint arXiv:1409.0473.

Behjati, M., Fehr, F. J., and Henderson, J. (2023).
Learning to abstract with nonparametric variational information bottleneck.
In The 2023 Conference on Empirical Methods in Natural Language Processing.

Fehr, F. and Henderson, J. (2023).
Nonparametric variational regularisation of pretrained transformers.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., rong Wen, J., Yuan, J., Zhao, W. X., and Zhu, J. (2021).
Pre-trained models: Past, present and future.
AI Open, 2:225–250.

Henderson, J. and Fehr, F. (2023).
A VAE for Transformers with Nonparametric Variational Information Bottleneck.
In International Conference on Learning Representations.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013).
Efficient estimation of word representations in vector space.
In International Conference on Learning Representations.

References II

Sutton, R. (2019).
The bitter lesson.
March, 13:2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017).
Attention is all you need.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.