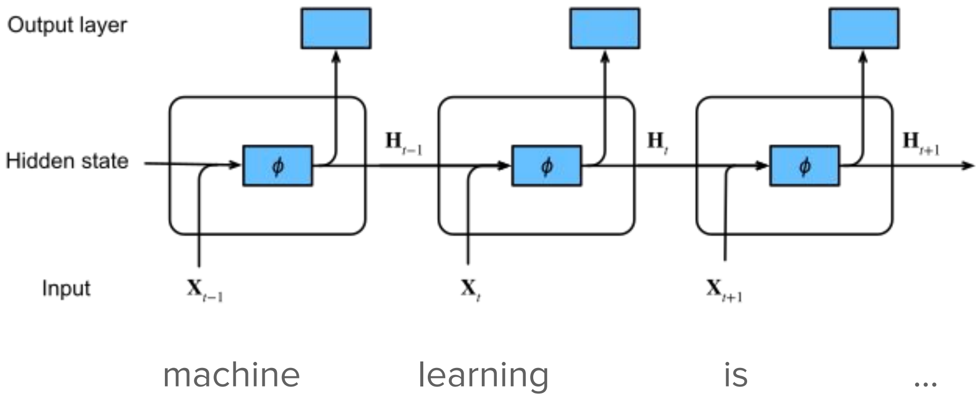
Attention and Transformers

He He

neuromatch academy

Recap: recurrent neural networks

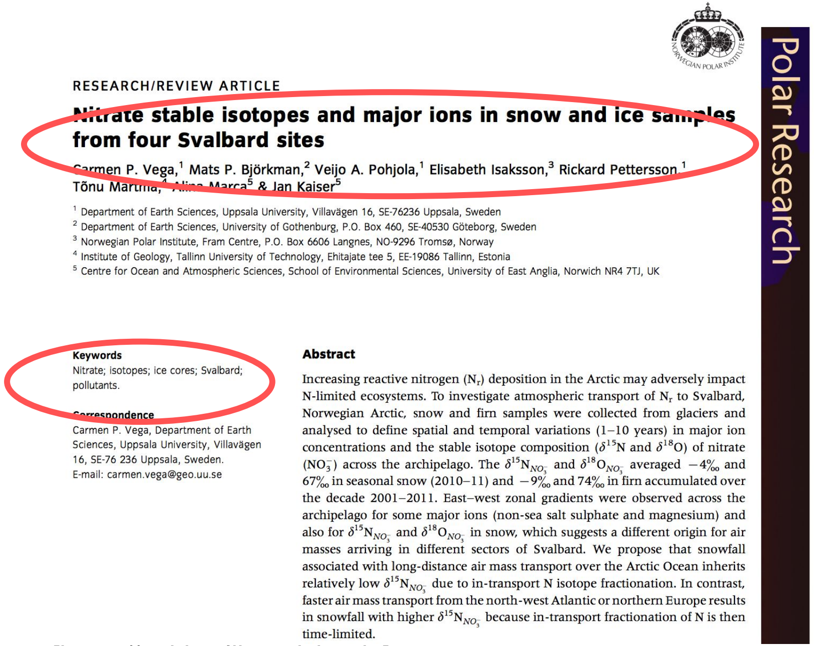


- Pros: handle long-range dependency
- Cons: inefficient, gradient vanishing/exploding

Can we handle dependencies in a more efficient way?

[https://www.d2l.ai]

Human attention



Polar Research

RESEARCH/REVIEW ARTICLE

Nitrate stable isotopes and major ions in snow and ice samples from four Svalbard sites

Carmen P. Vega,[1] Mats P. Björkman,[2] Veijo A. Pohjola,[1] Elisabeth Isaksson,[3] Rickard Pettersson,[1] Tõnu Martma,[4] Alina Marca[5] & Jan Kaiser[5]

[1] Department of Earth Sciences, Uppsala University, Villavägen 16, SE-76236 Uppsala, Sweden
[2] Department of Earth Sciences, University of Gothenburg, P.O. Box 460, SE-40530 Göteborg, Sweden
[3] Norwegian Polar Institute, Fram Centre, P.O. Box 6606 Langnes, NO-9296 Tromsø, Norway
[4] Institute of Geology, Tallinn University of Technology, Ehitajate tee 5, EE-19086 Tallinn, Estonia
[5] Centre for Ocean and Atmospheric Sciences, School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK

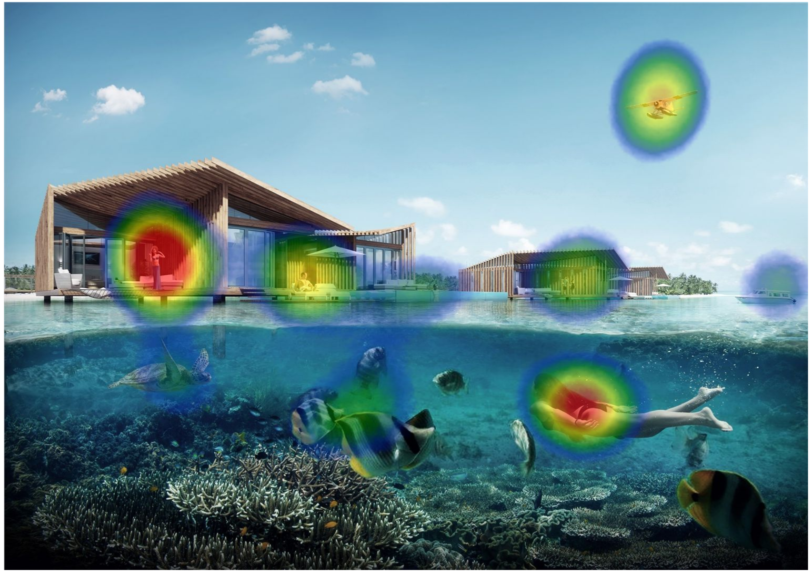
Keywords
Nitrate; isotopes; ice cores; Svalbard; pollutants.

Correspondence
Carmen P. Vega, Department of Earth Sciences, Uppsala University, Villavägen 16, SE-76 236 Uppsala, Sweden.
E-mail: carmen.vega@geo.uu.se

Abstract

Increasing reactive nitrogen (N_r) deposition in the Arctic may adversely impact N-limited ecosystems. To investigate atmospheric transport of N_r to Svalbard, Norwegian Arctic, snow and firn samples were collected from glaciers and analysed to define spatial and temporal variations (1–10 years) in major ion concentrations and the stable isotope composition ($\delta^{15}N$ and $\delta^{18}O$) of nitrate (NO_3^-) across the archipelago. The $\delta^{15}N_{NO_3^-}$ and $\delta^{18}O_{NO_3^-}$ averaged $-4‰$ and $67‰$ in seasonal snow (2010–11) and $-9‰$ and $74‰$ in firn accumulated over the decade 2001–2011. East–west zonal gradients were observed across the archipelago for some major ions (non-sea salt sulphate and magnesium) and also for $\delta^{15}N_{NO_3^-}$ and $\delta^{18}O_{NO_3^-}$ in snow, which suggests a different origin for air masses arriving in different sectors of Svalbard. We propose that snowfall associated with long-distance air mass transport over the Arctic Ocean inherits relatively low $\delta^{15}N_{NO_3^-}$ due to in-transport N isotope fractionation. In contrast, faster air mass transport from the north-west Atlantic or northern Europe results in snowfall with higher $\delta^{15}N_{NO_3^-}$ because in-transport fractionation of N is then time-limited.

[https://guides.lib.umich.edu]



[https://brickvisual.com]

Attention mechanism

Select content relevant to a query

Machine translation:

Time flies like an arrow

光阴 似 箭

Question answering:

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail...

What causes precipitation to fall?

Image captioning:

The dog is lying on the beach

[https://portcitydaily.com/]

Today: attention and transformers

Part I: transformer architecture

- Use ~~recurrence~~ attention to capture dependencies among inputs
- Computational efficiency → large-scale models
- Highly versatile: classification, sequence generation/labeling

Part II: pre-training and fine-tuning

- Language modeling as unsupervised pre-training
- Good representation → better performance with less data

Queries, Keys, and Values

Queries Keys and Values in CS

Select a **values** (referenced by a **key**) relevant to a **query**

```
select value where key is major
```

Value	Key
Lisa	name
linguistics	major
chess	hobby
2020	class

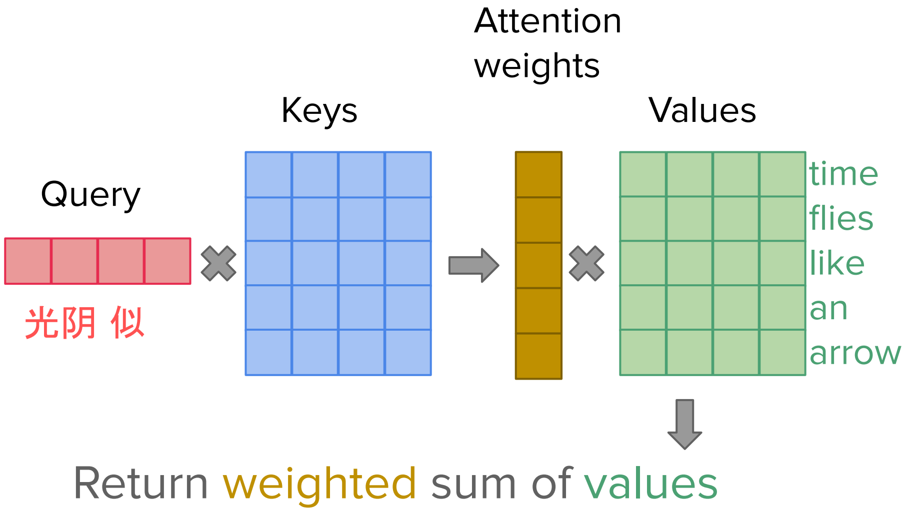
linguistics

Queries Keys Values in DL

Query: Embeddings without context (input)

Keys: The surrounding embeddings (context)

Values : Embeddings with context (Output)



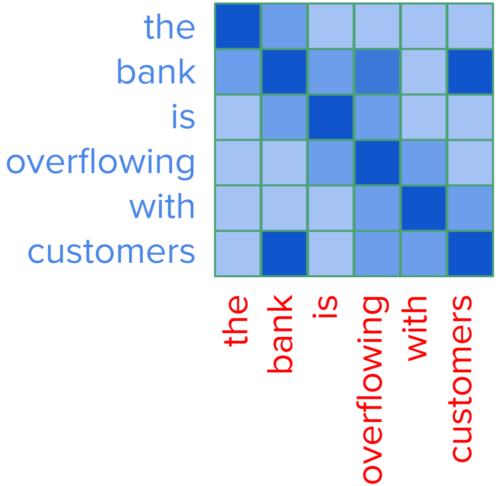
An example

Keys

Query: what is the context of "bank"?

the
bank
is
overflowing
with
customers

attention updates embeddings with context



Compute attention weights

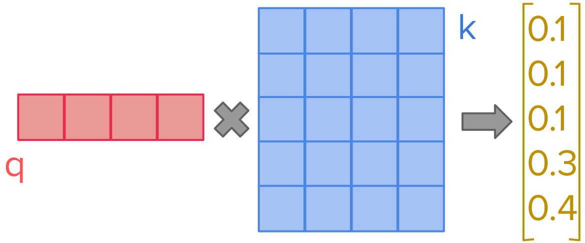
Weights: how strongly is the query matched to a key?

softmax(score(q, k))

Scaled dot-product attention

$$\text{score}(q, k) = \frac{q^T k}{\sqrt{d}}$$

d: dimension of the query/key vector



Compute attention weights

Weights: how likely is the query matched to a key?

softmax(score(q, k))

Scaled dot-product attention

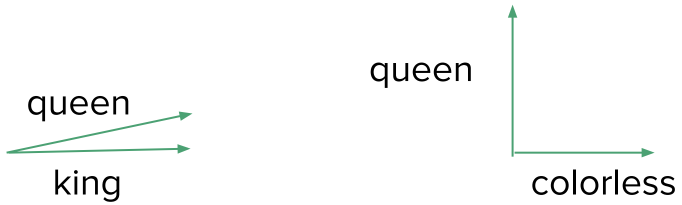
$$\text{score}(q, k) = \frac{q^T k}{\sqrt{d}}$$

d: dimension of the query/key vector

Word embeddings dot-product similarity

High similarity

Low Similarity



Compute attention weights

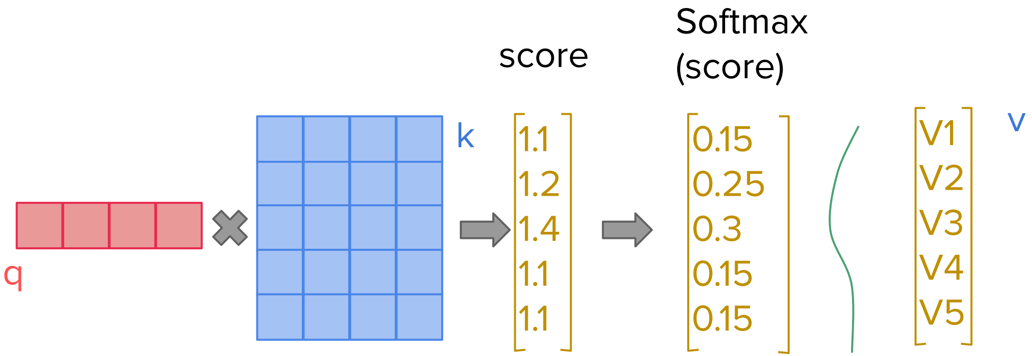
Weights: how likely is the query matched to a key?

softmax(score(q, k))

Scaled dot-product attention

$$\text{score}(q, k) = \frac{q^T k}{\sqrt{d}}$$

d: dimension of the query/key vector



Compute attention weights

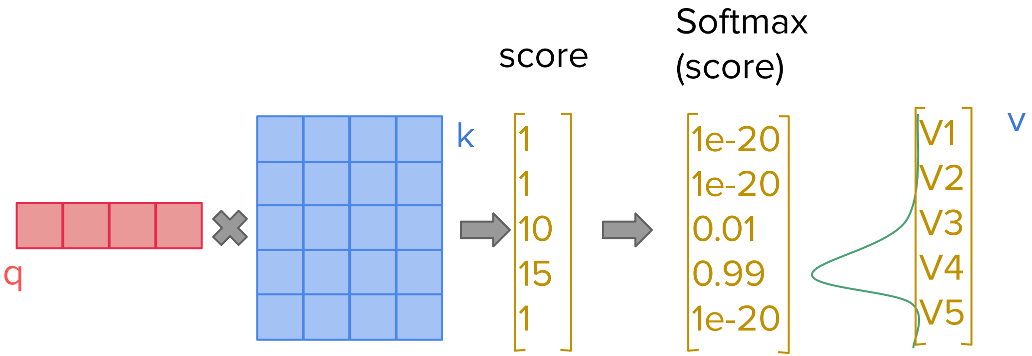
Weights: how likely is the query matched to a key?

softmax(score(q, k))

Scaled dot-product attention

$$\text{score}(q, k) = \frac{q^T k}{\sqrt{d}}$$

d: dimension of the query/key vector

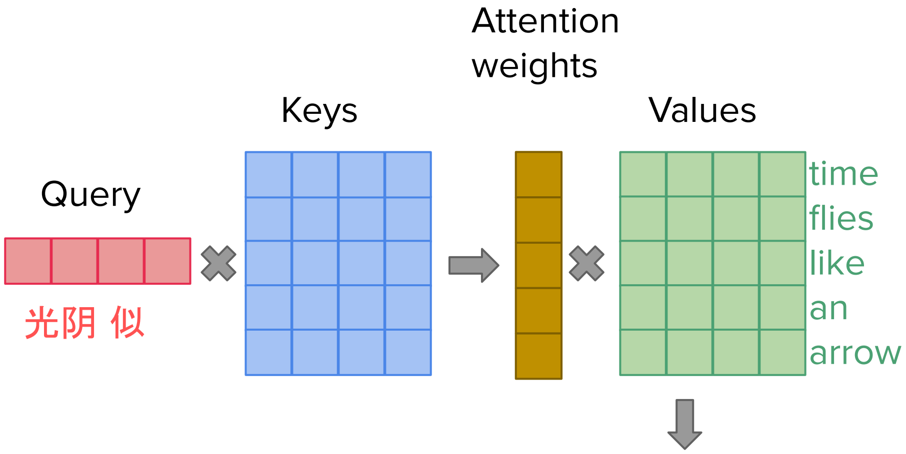


Queries Keys Values in DL

Query: Embeddings without context (input)

Keys: The surrounding embeddings (context)

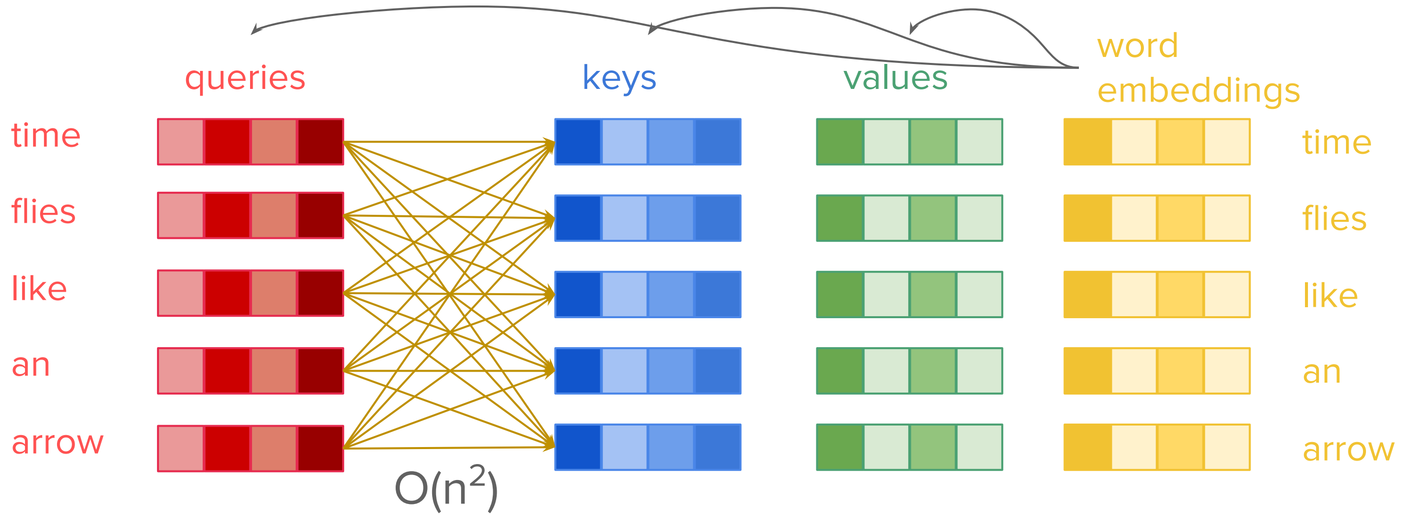
Values : Embeddings with context (Output)



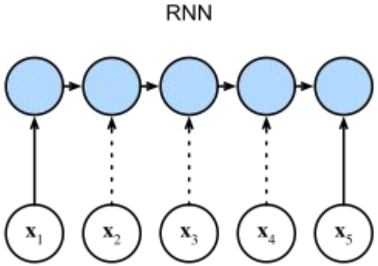
Return weighted sum of values

Self-attention

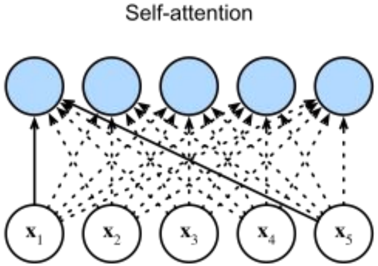
- Each word attends to all words in the sentence
- Word embeddings projected to values, keys, and queries



Comparison of RNN and self-attention



- Sequential O(n)
- Uni-directional and may forget past context
- Handle long sequence trivially



[https://www.d2l.ai]

- Parallelizable $O(n^2)$
- Direct interaction between any word pair
- Maximum sequence length is fixed

Multi-head Attention

Motivation

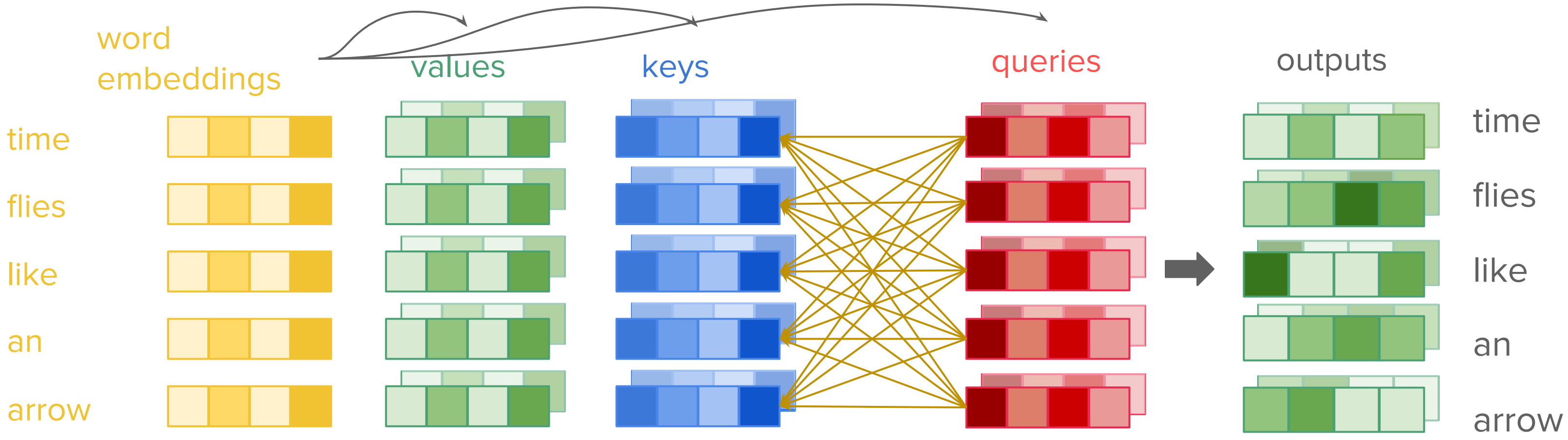
- "Time flies like an arrow"

- Which words should "like" attend to?
    - Semantics: "time", "arrow" (a simile)
    - Syntax: "flies", "arrow" (a preposition)

- Need to create multiple embeddings
    - Each embedding could attend to different things

Multi-head attention

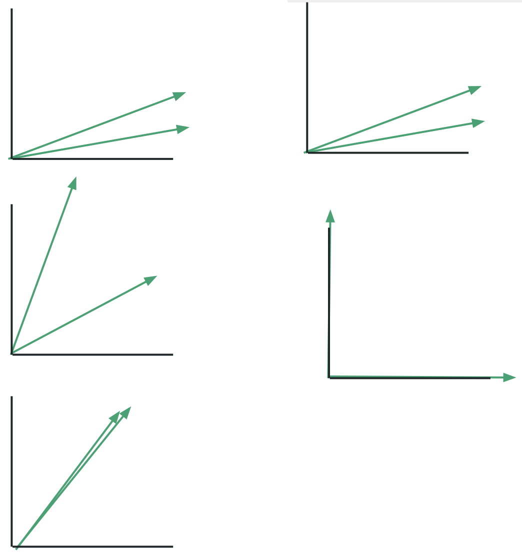
- Produce k sets of queries/keys/values
- Word embeddings ➔ k sets of attention outputs



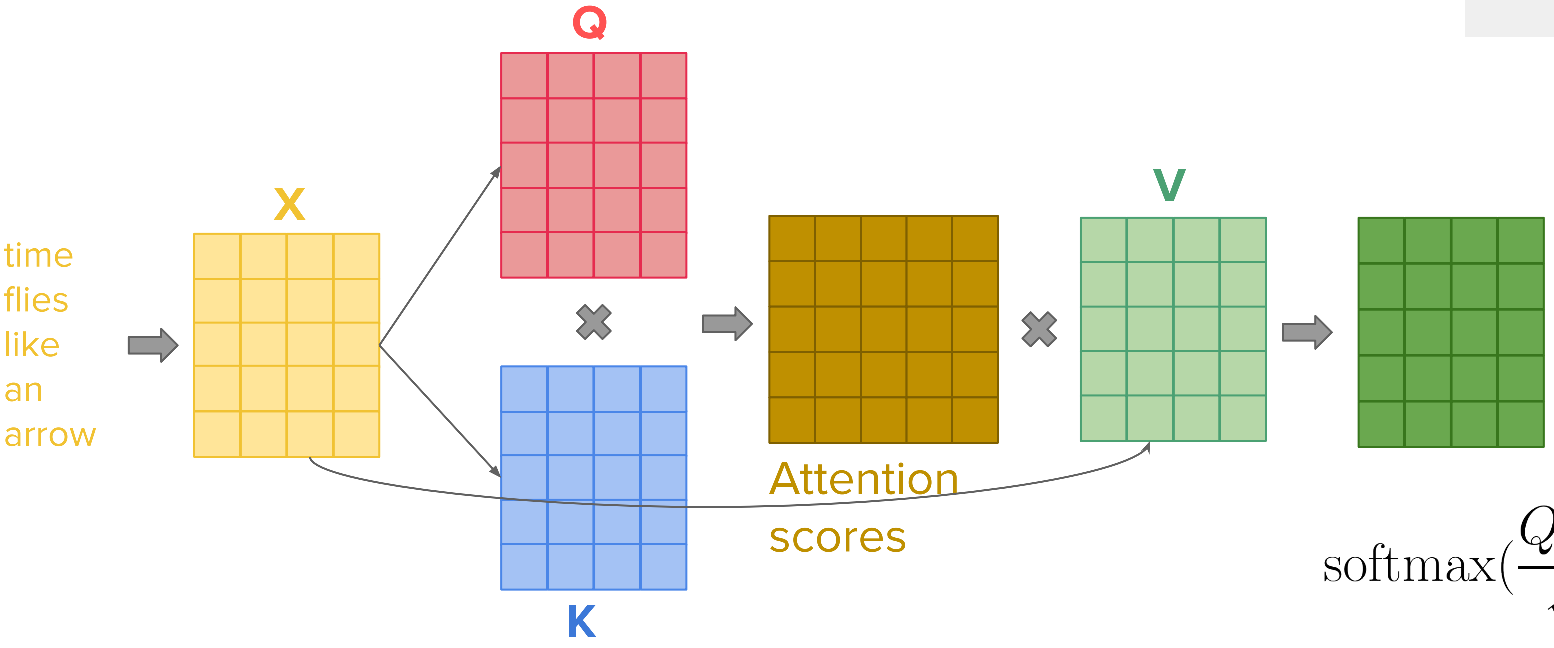
How do we create these multiple copies?

- Linear Projection!
  - Create N projections of queries values and keys to and perform self attention N times

- Where to project?
  - Let model decide!
  - Each projection randomly initiated and learned by the model

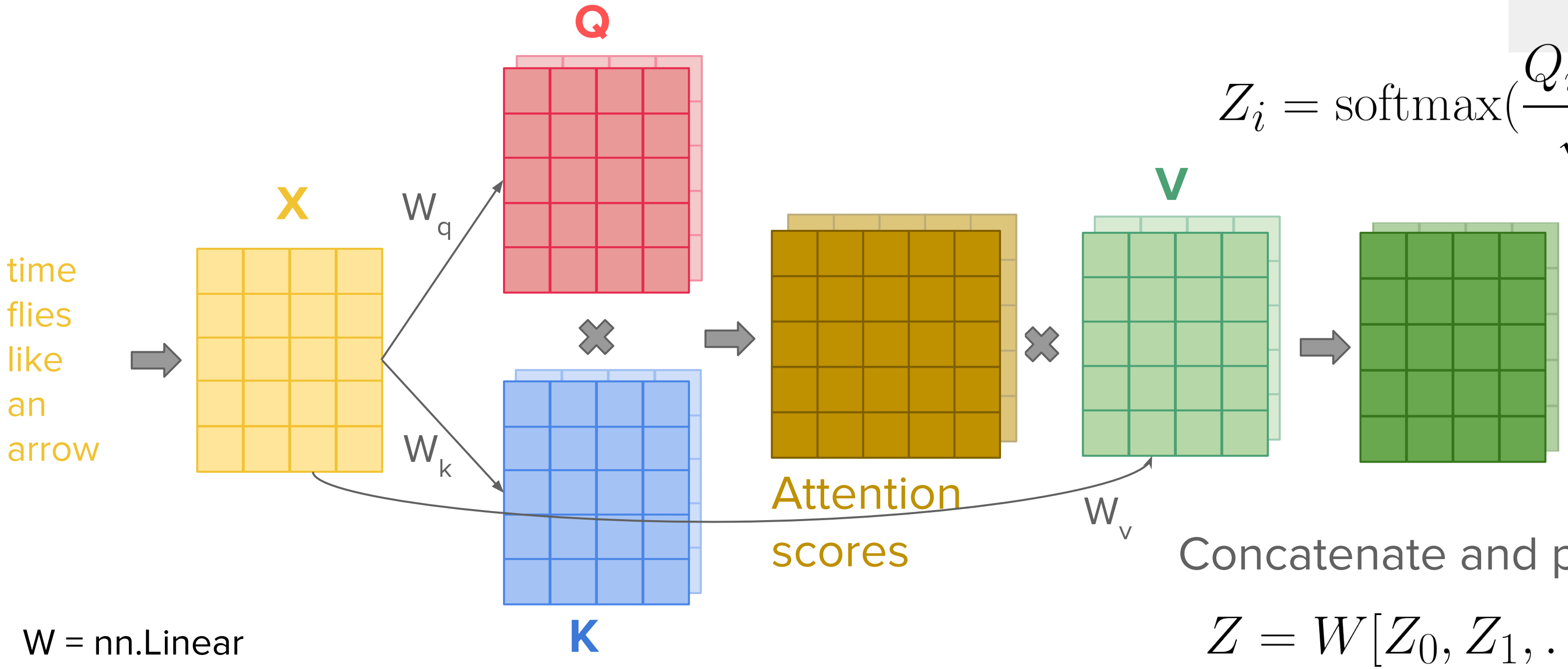


Single-head attention



$$\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

Multi-head attention



$$Z_i = \mathrm{softmax}(\frac{Q_i K_i^T}{\sqrt{d}})V_i$$

Concatenate and project:

$$Z = W[Z_0, Z_1, \ldots, Z_m]$$

W = nn.Linear

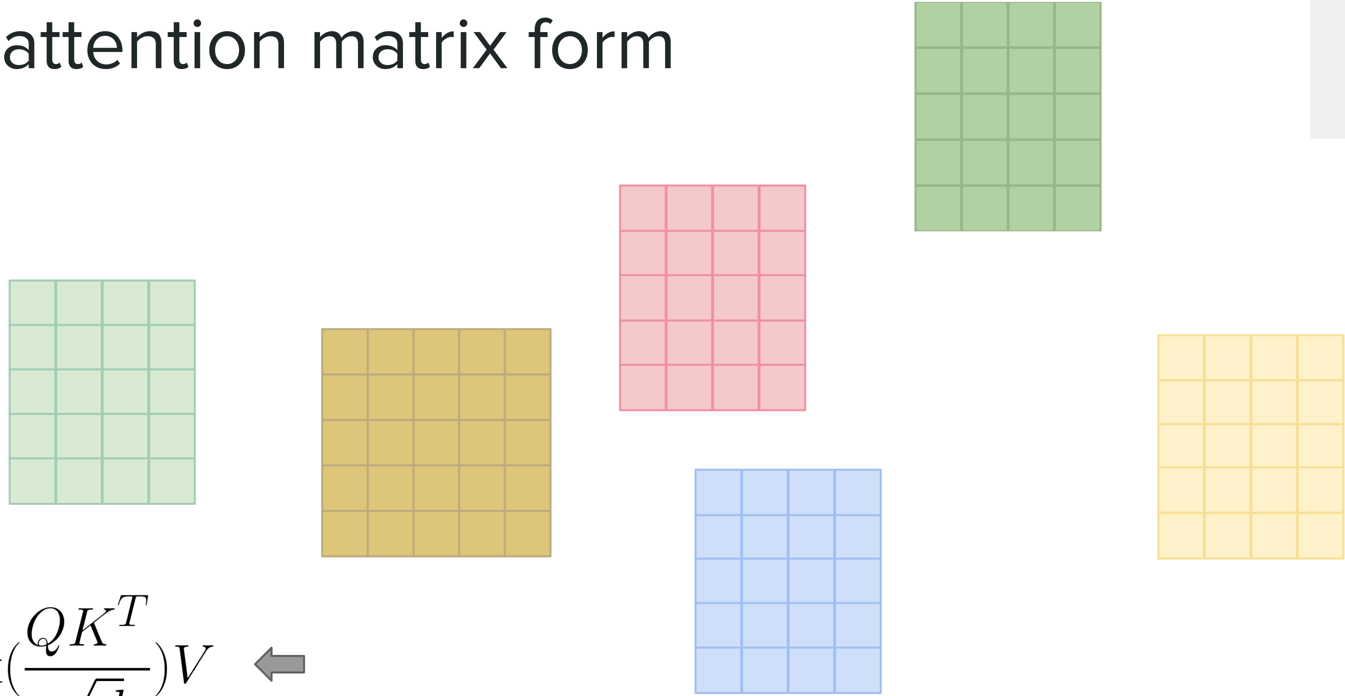
From the Authors

Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. With a single attention head, averaging inhibits this.

Instead of performing a single attention function with d_{model}-dimensional keys, values and queries, we found it beneficial to linearly project the queries, keys and values h times with different, learned linear projections to d_k, d_k and d_v dimensions, respectively. On each of these projected versions of queries, keys and values we then perform the attention function in parallel, yielding d_v-dimensional output values. These are concatenated and once again projected, resulting in the final values, as depicted in Figure 2.

https://doi.org/10.48550/arXiv.1706.03762

Self-attention matrix form



$$\text{softmax}(\frac{QK^T}{\sqrt{d}})V$$

Time complexity of multi-head attention

Concatenate and project:

$$Z = W[Z_0, Z_1, \ldots, Z_m]$$

$$Z_i = \mathrm{softmax}(\frac{Q_i K_i^T}{\sqrt{d}})V_i$$

- Problem size
  - Sequence length: n
  - Number of heads: m
  - Embedding size: d

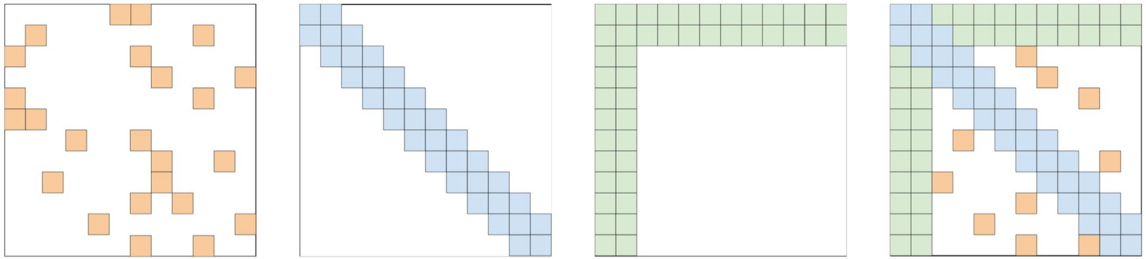
Expensive for long sequences!

- Time complexity
  - Attention score for a pair of word (dot product): O(d)
  - Self-attention (pairwise interaction): $O(n^2)$
  - Multi-head attention: O(m)
  - Overall: $O(mdn^2)$

Efficient transformer

Goal: reduce the $O(n^2)$ time/space complexity for long sequence problems

- Sparse attention



[Zaheer et al., 2021]

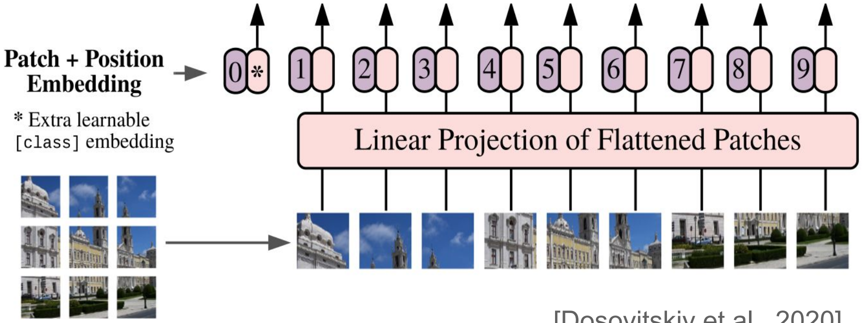
- Locality sensitive hashing
- Low-rank decomposition

$O(n^2) \rightarrow O(nk)$ where k is small

Transformer Overview I

Applications

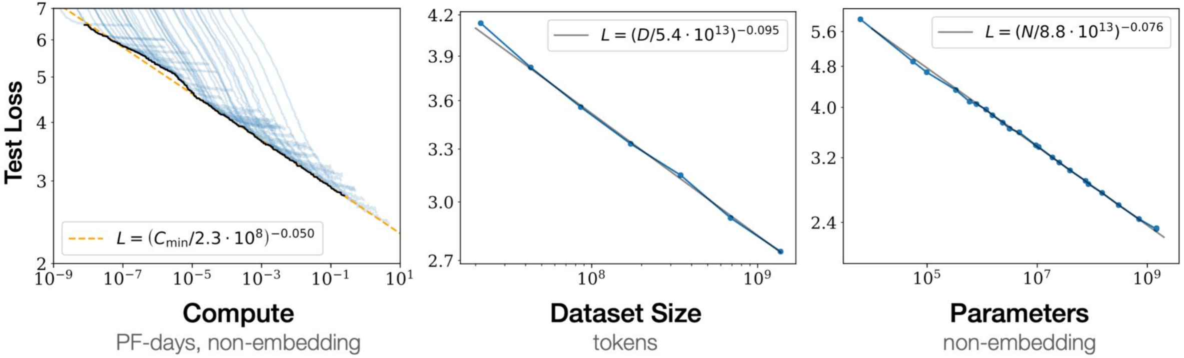
- Originally designed for machine translation (sequence-to-sequence)
- Now widely used in numerous **NLP** applications, e.g., text classification, generation, question answering
- Also used in **computer vision** and **speech**



[Dosovitskiy et al., 2020]

Scalability

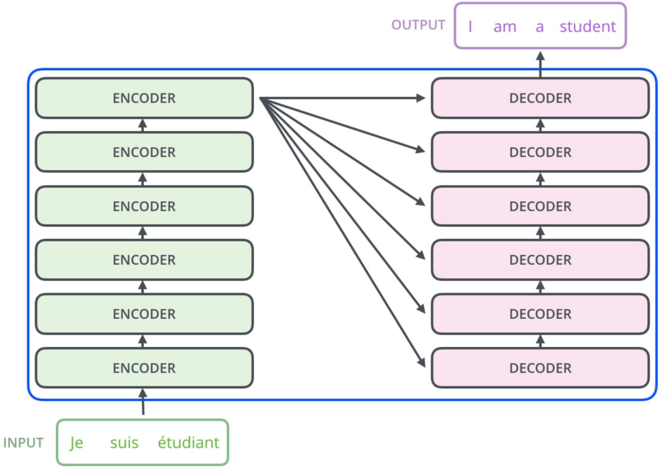
- Core component: **self-attention** (no recurrence)
- Scalable to large models and large data
- Backbone of current pre-trained models (BERT, GPT-3 etc.)



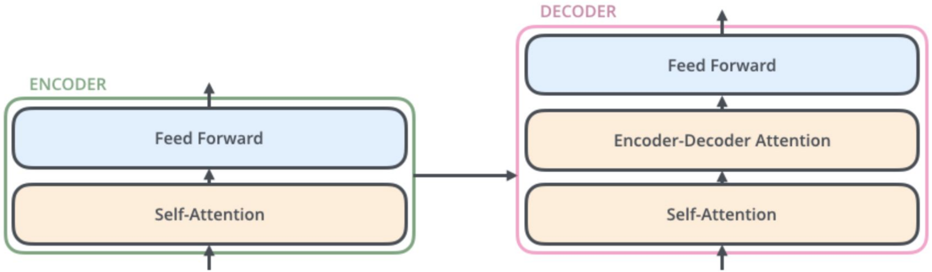
[Kaplan et al., 2020]

The high-level picture

- Multi-layer sequence-to-sequence model



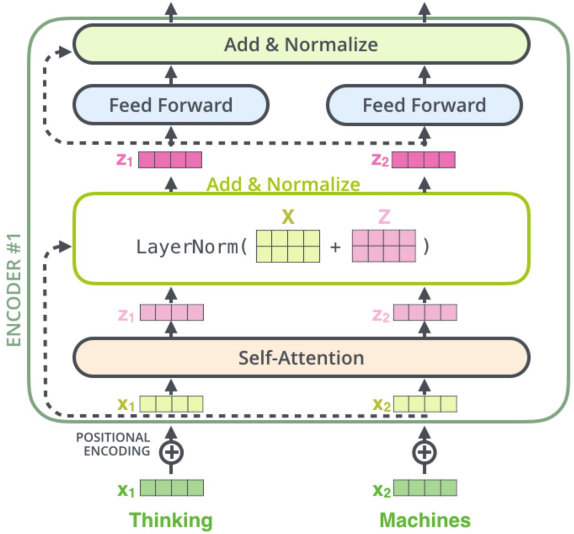
- Self-attention based sequence representation



[https://jalammar.github.io/illustrated-transformer/]

The Transformer block

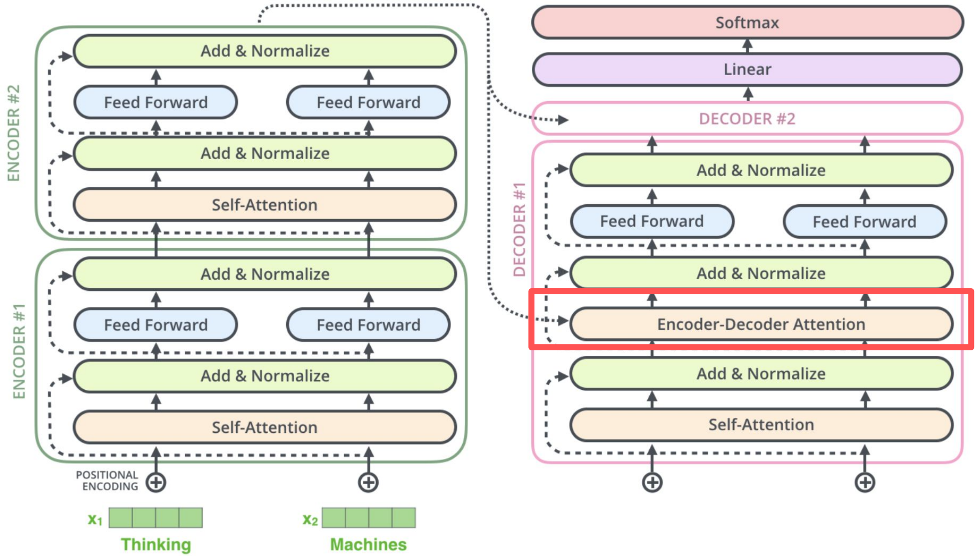
- Multi-head self-attention
  - Capture dependence among inputs
- Positional encoding
  - Capture order information
- Residual connection and layer normalization
  - Efficient optimization



[https://jalammar.github.io/illustrated-transformer/]

Transformer Overview II

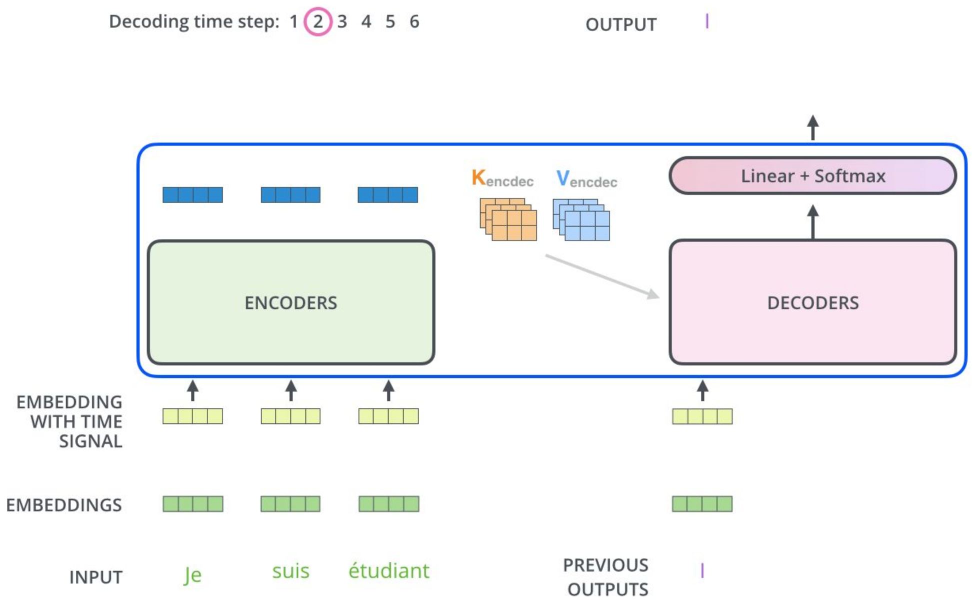
Connect the decoder



- Same as the encoder with an additional attention module
- **Encoder-decoder attention**
  - Query: decoder state
  - Value/key: encoder embeddings

[https://jalammar.github.io/illustrated-transformer/]

Encoder-decoder model



[https://jalammar.github.io/illustrated-transformer/]

- **Auto-regressive** model (word-by-word generation from left to right)
- Output: p(next word | prefix, input)
- Decoder self-attention only in the previous outputs
- Trained by maximum likelihood estimation

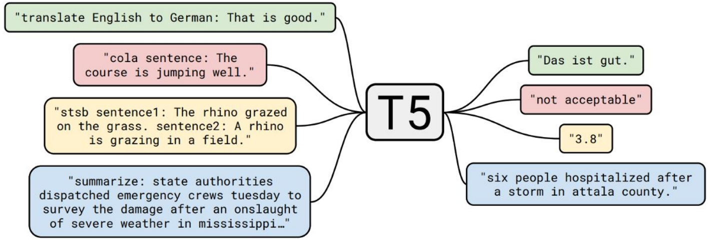
Classification

We can already do quite many cool things with the encoder!

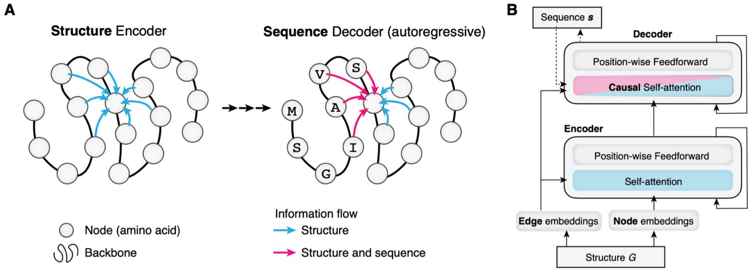
- Task
  - Text classification, e.g., sentiment, topic
  - Pair classification, e.g. textual entailment, paraphrase identification
  - Image classification
- Architecture
  - Only using the encoder (output an embedding for each word)
  - Aggregate over all words, e.g., mean pooling
  - Classifier, e.g., linear + cross-entropy loss

Sequence prediction

- Sequence labeling
  - Extractive question answering
  - Named entity recognition
- Sequence generation
  - Text generation: machine translation, summarization
  - Music generation
  - Protein/Molecule generation



[Raffel et al., 2020]



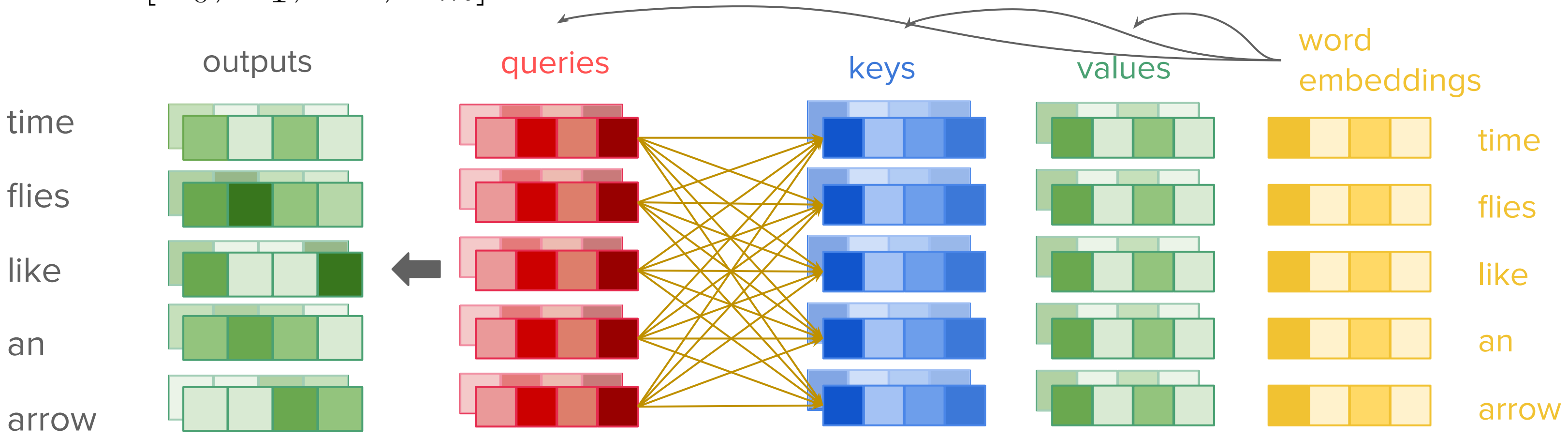
[Ingraham et al., 2019]

Multi-head attention matrix form

Concatenate and project:

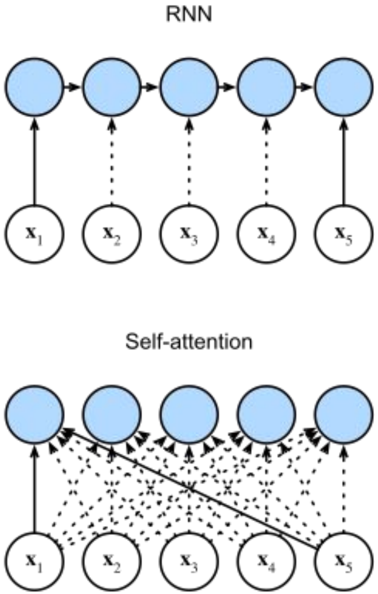
$$Z = W[Z_0, Z_1, \ldots, Z_m]$$

$$Z_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right)V_i$$



Positional Encoding

Positional encoding



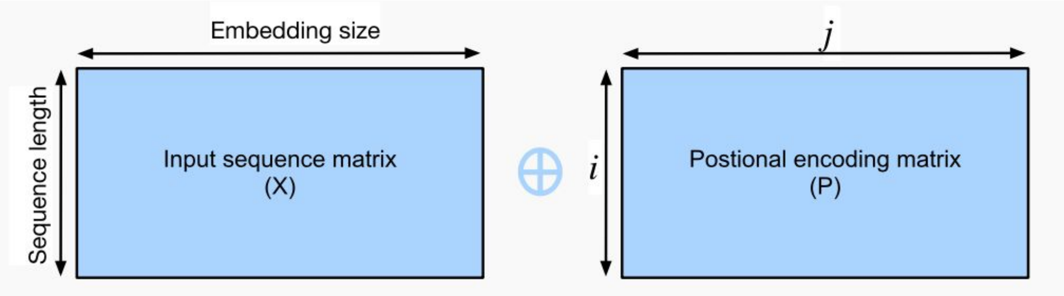
Self-attention is not sensitive to word ordering!

Does it matter?

All dogs are smart and some are dumb.
All dogs are dumb and some are smart.

Represent position

Solution: add a positional embedding to the input word embedding



[https://www.d2l.ai]

- Encode absolute and relative positions of a word
- Same dimension as the word embedding (for addition)
- Learned or deterministic

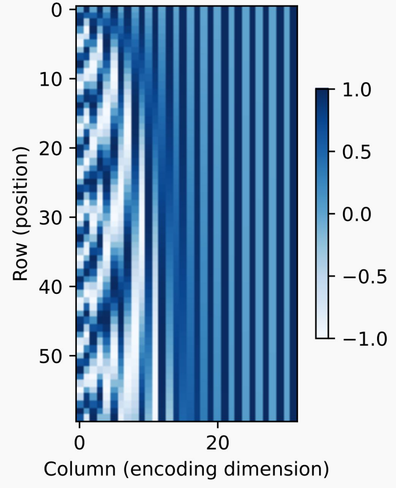
Sinusoidal position embedding

Intuition: binary encoding

- The frequency of bit flips increases from left to right

0 → 000
1 → 001
2 → 010
3 → 011
4 → 100
5 → 101
6 → 110
7 → 111

Continuous version



[https://www.d2l.ai]

Col 1: $\sin(w_1 t)$
Col 2: $\cos(w_1 t)$
Col 3: $\sin(w_2 t)$
Col 4: $\cos(w_2 t)$
...

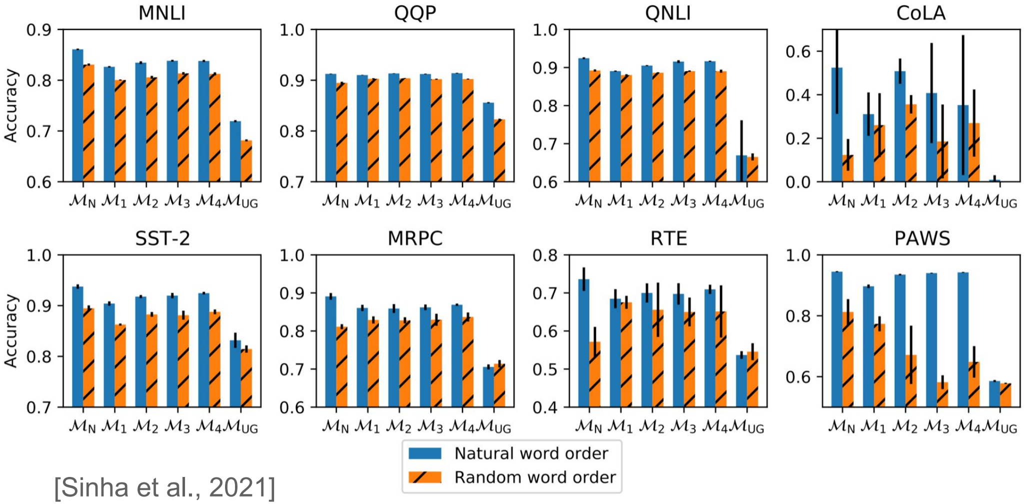
w_i: frequency
t: position

How important is word ordering?

Reasonable performance when trained on permuted n-grams!

Are word ordering unimportant?

- May need better evaluation of “understanding”
- Results are only on English



[Sinha et al., 2021]

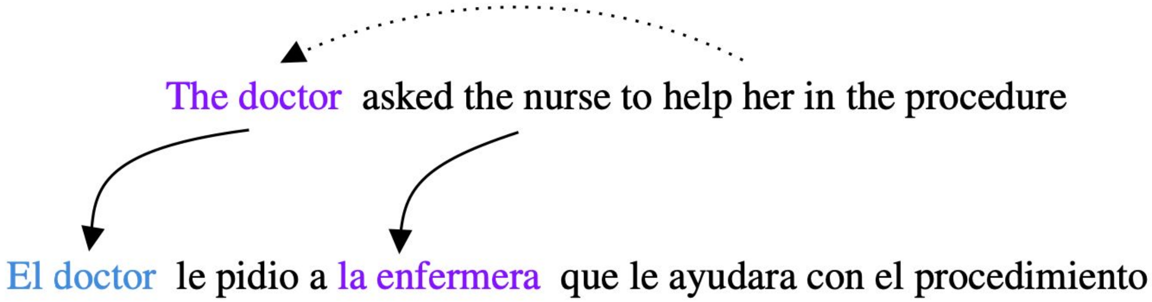
Ethics

Sources of social bias in model

- Various **prejudices and stereotypes in text**, derived from biases in human society
  - *It was a very important discovery, one you wouldn't expect from a female astrophysicist*

- The training data has **selection bias**
  - Most benchmarks are in English but "natural language" ≠ English!

Gender bias in machine translation



Translation error due to gender stereotypes when translating into languages with grammatical gender

[Stanovsky et al., 2019]

Social bias in pre-trained language models

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

Text generation inherits biases in pre-trained language models

[Sheng et al., 2019]

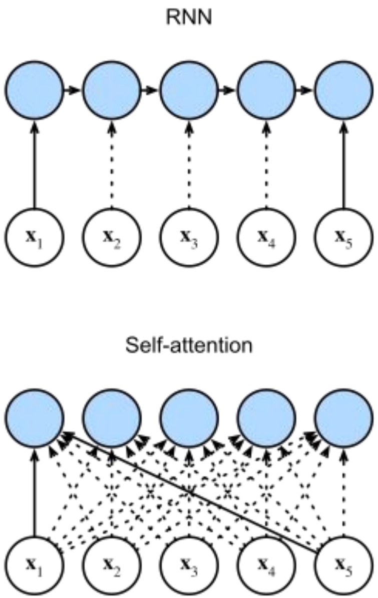
What can we do?

- Bias mitigation is an active research field but don’t solely rely on it

- Be aware of potential harms of your system!

- Provide explicit statements of
    - The training data (e.g. language, source, potential bias)
    - What type of system behavior is harmful, in what ways and to whom

Parting remarks



[https://www.d2l.ai]

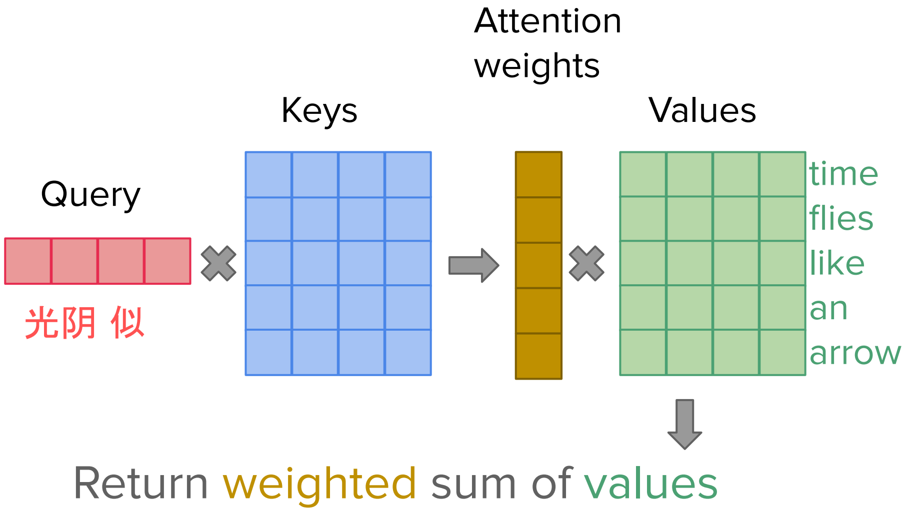
- Attention models dependence between two parts
- Self-attention is an efficient way to model (long-range) dependencies in sequence data
- Attention is a centerpiece of today's large-scale NLP models

Attention works on embeddings

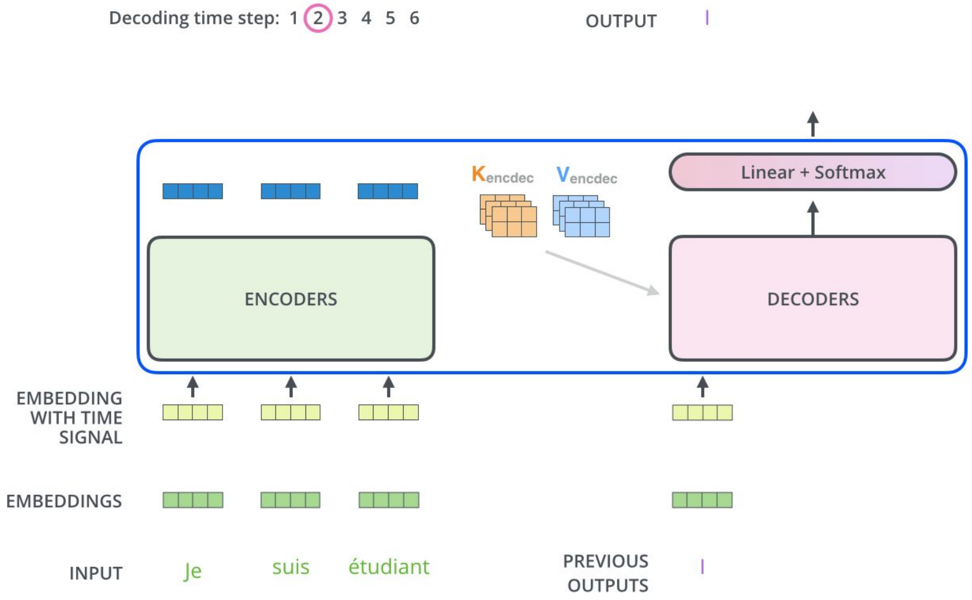
Query: Embeddings without context (input)

Keys: The surrounding embeddings (context)

Values : Embeddings with context (Output)



Transformers are everywhere!



[https://jalammar.github.io/illustrated-transformer/]

A transformer is:

1. Encoder: a way to find a low dimensional representation (embedding) of input space
2. Decoder: an embedding of output space

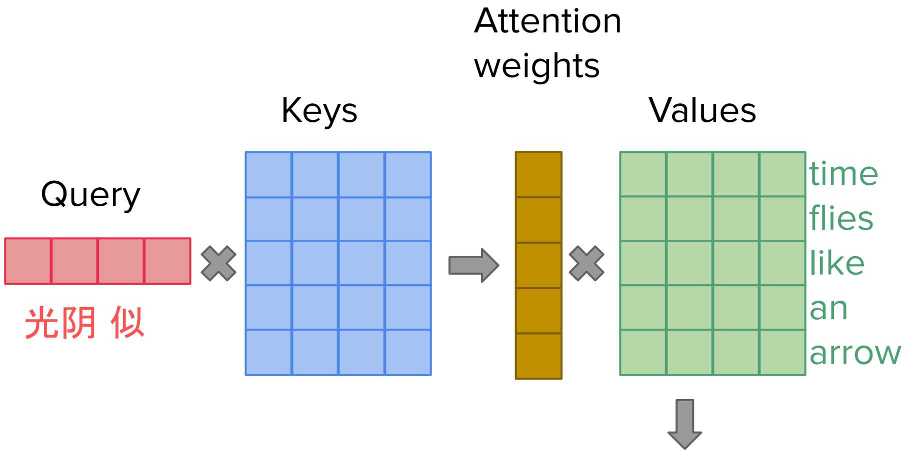
Transformers aren’t just for Language

Attention works on embeddings

Query: Embeddings without context (input)

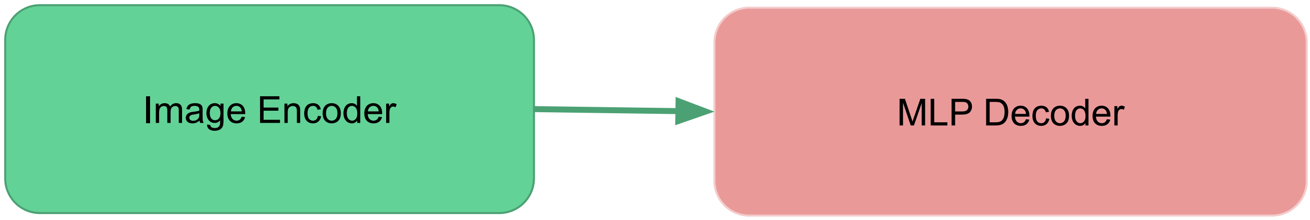
Keys: The surrounding embeddings (context)

Values : Embeddings with context (Output)

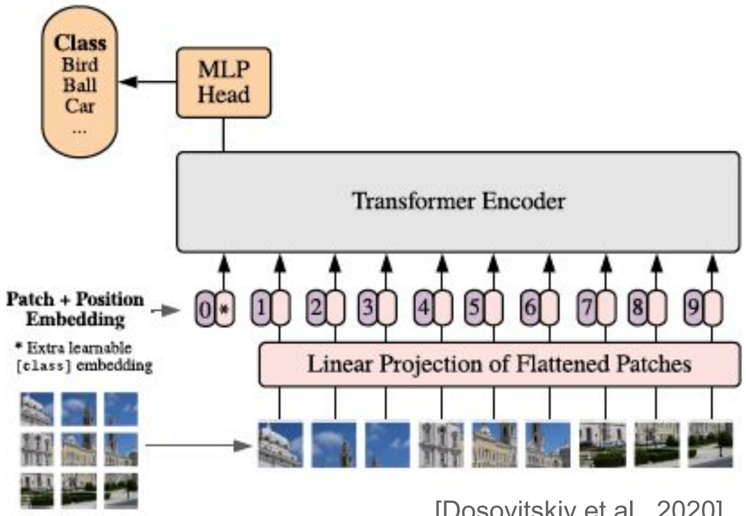


Return weighted sum of values

An image is worth 16 X 16 words: Transformers for image classification

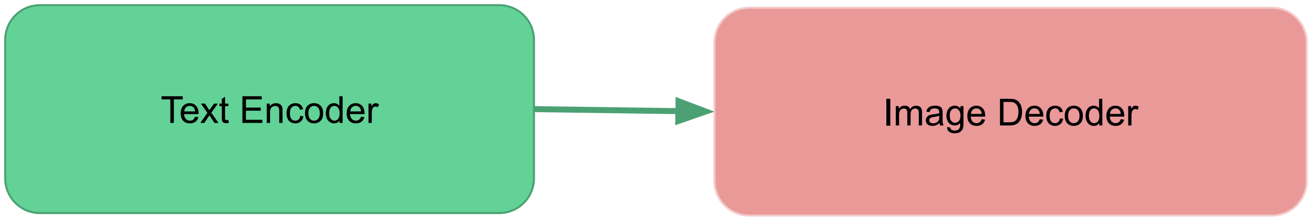


An image is worth 16 X 16 words: Transformers for image classification

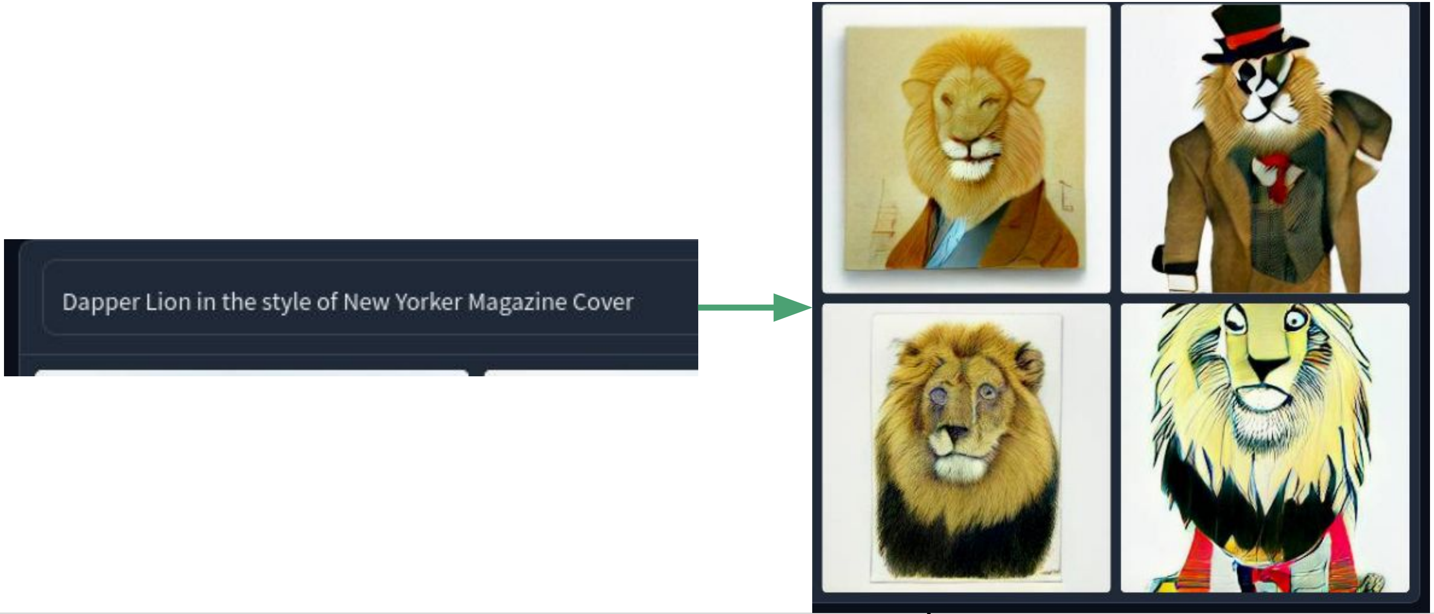


[Dosovitskiy et al., 2020]

DALL-E transforms text into images



DALL-E transforms text into images



Parting remarks

- Self-supervised representation learning enables non-task-specific models
- The learning paradigm today: pre-train then fine-tune
- Scaling doesn't solve all problems