Technical details of explicharr

Kuan Yu[1]

September 11, 2018

[1] kuanyu@uni-potsdam.de

explicharr[3]

- sentence simplification with
- character-level
- transformer[2]

"It is located in Potsdam ." $\longmapsto$ "It is in Potsdam ."

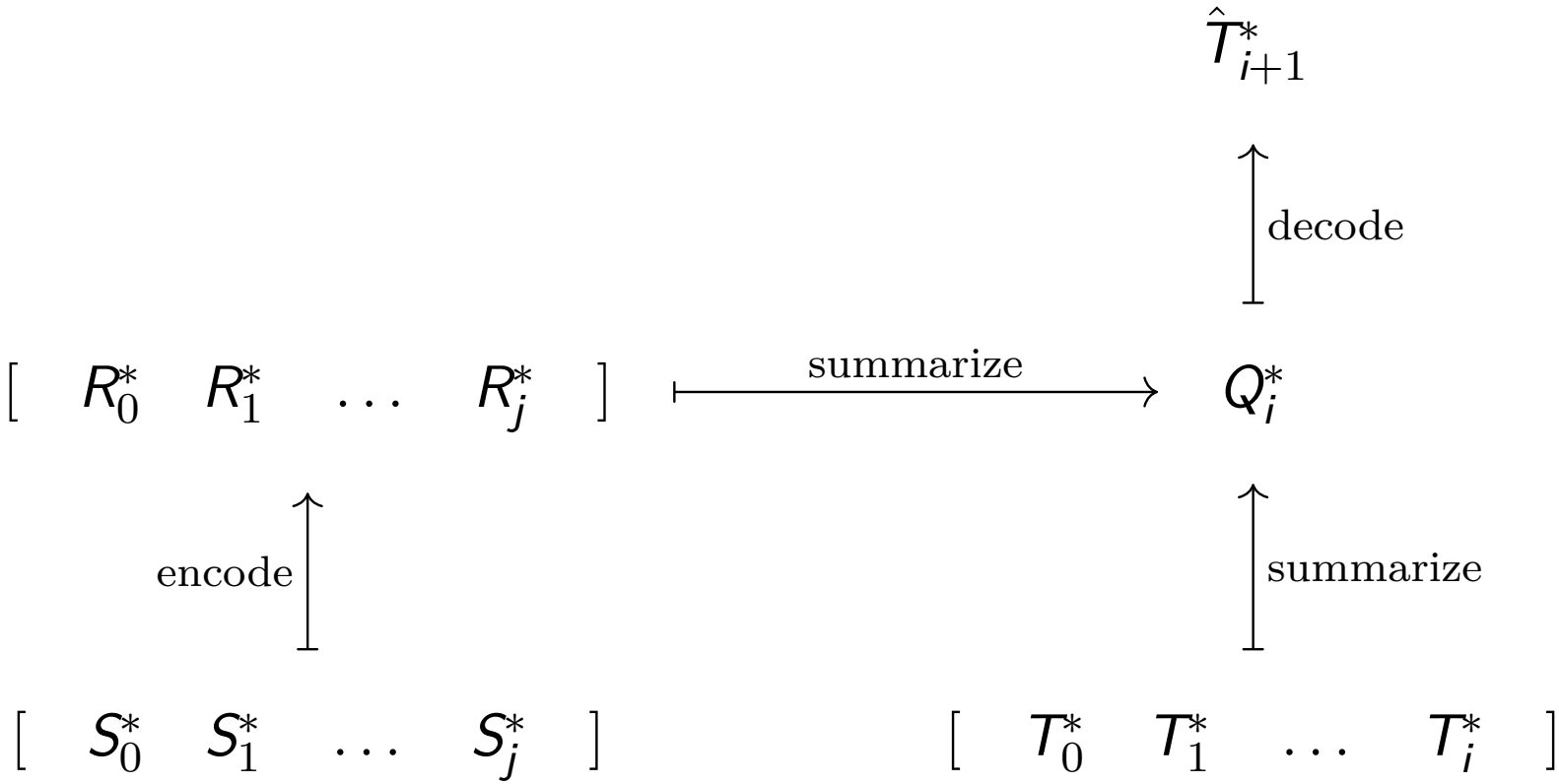
model: $S^* \to T^*$ where

- S = the source alphabet
- T = the target alphabet

[2] https://arxiv.org/abs/1706.03762

[3] https://github.com/srewai/explicharr

encoder-decoder, seq-to-seq, autoregressive



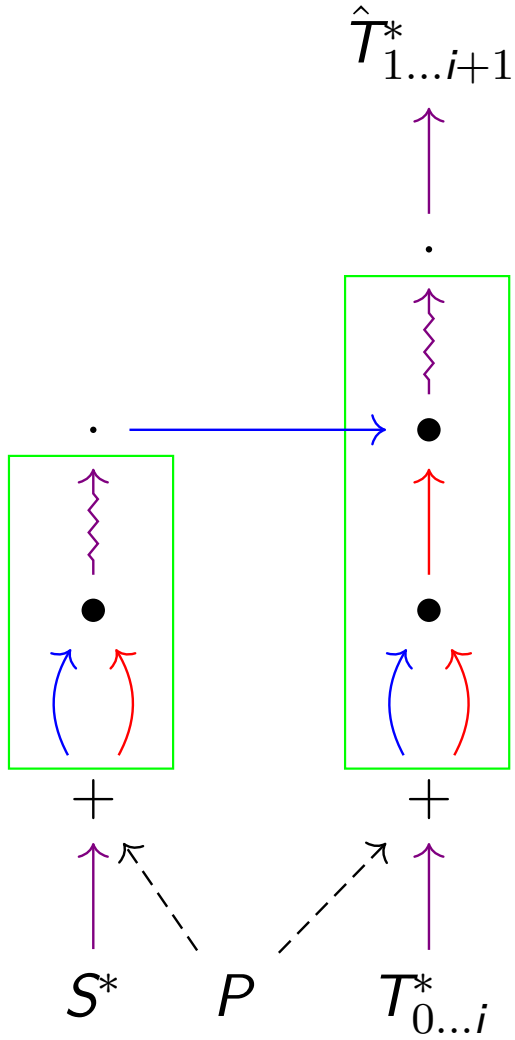
soft attention

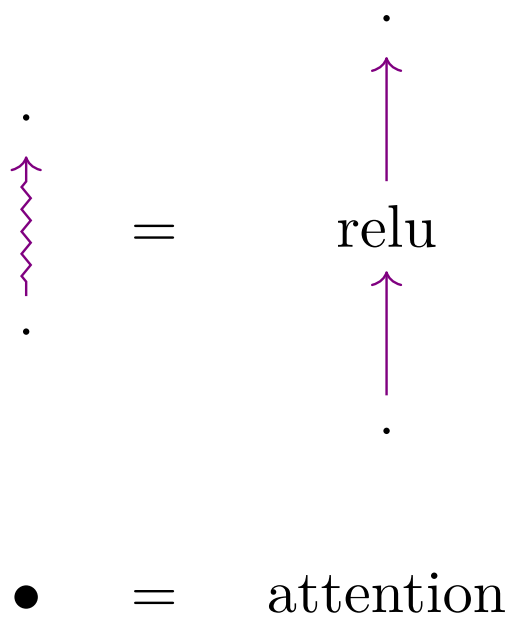
given a query vector and multiple value vectors

attention: 

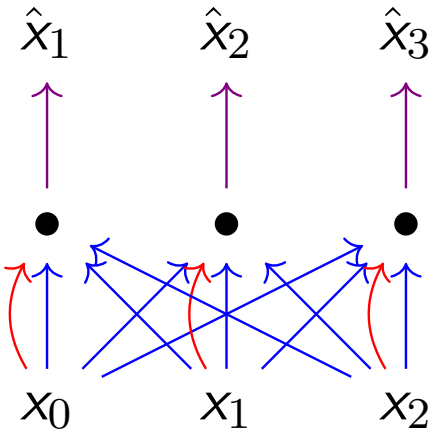
- compute a weight for each value, according to the query
- normalize the weights with softmax
- take the weighted sum of the values

transformer

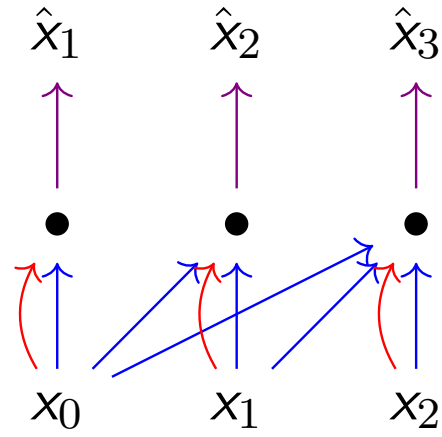




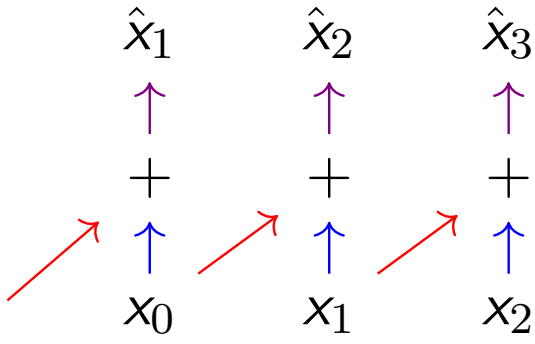
self-attention[4]



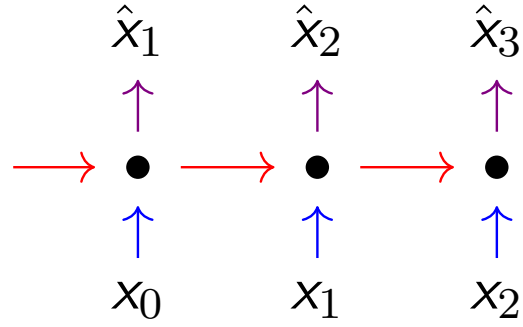
encoder self-attention



decoder self-attention



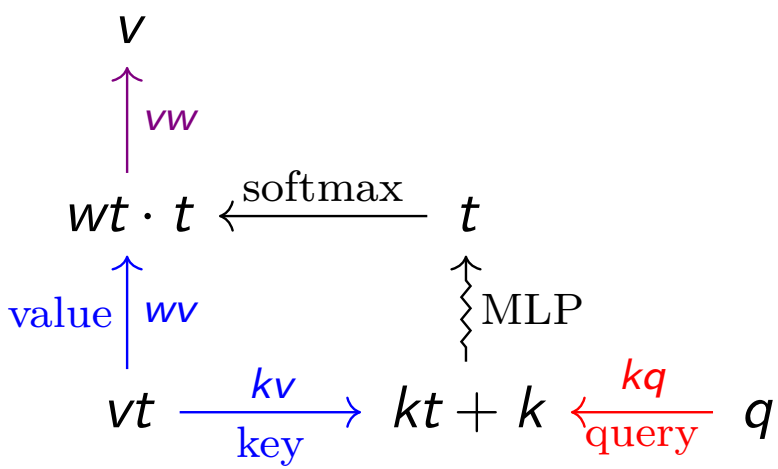
convolutional



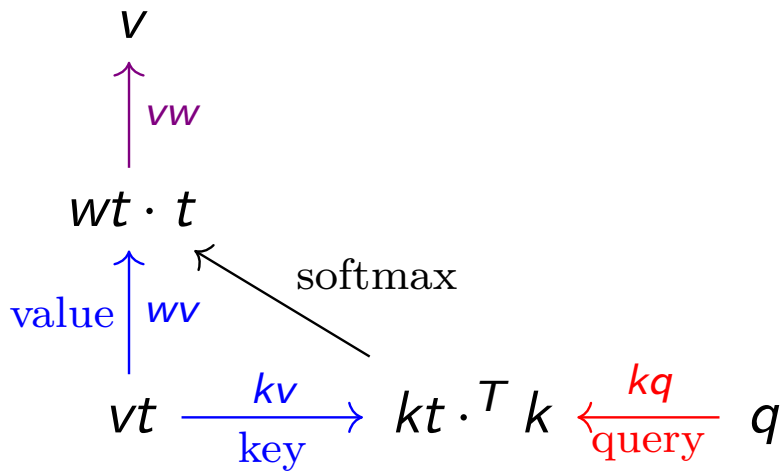
recurrent

[4] https://arxiv.org/abs/1606.01933

attention cells: additive[5] vs dot-product[6], key-value[7]



additive



dot-product

dimensions: **t**ime, **q**uery, **k**ey, **v**alue, **w** intermediate

$$A \cdot B := AB$$

$$A \cdot^T B := A^T B$$

[5] https://arxiv.org/abs/1409.0473
[6] https://arxiv.org/abs/1508.04025
[7] https://arxiv.org/abs/1702.04521

transformer attention

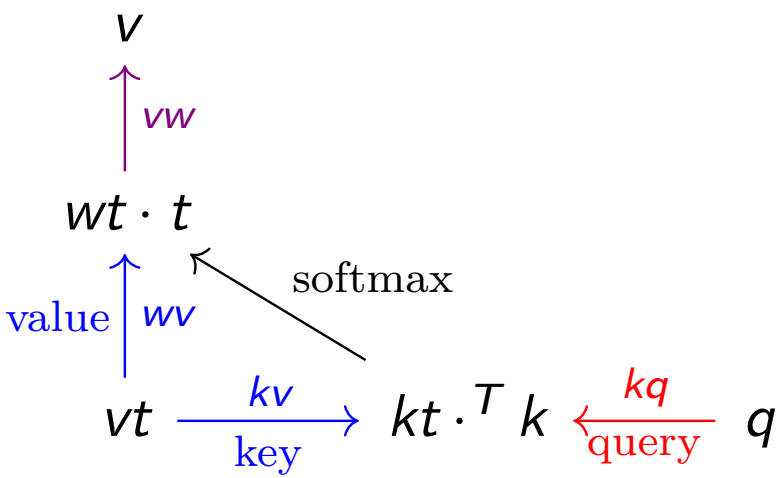
scaled dot-product

- divide weights by $\sqrt{k}$ before applying $\mathrm{softmax}$
- raise temperature
- lower variance

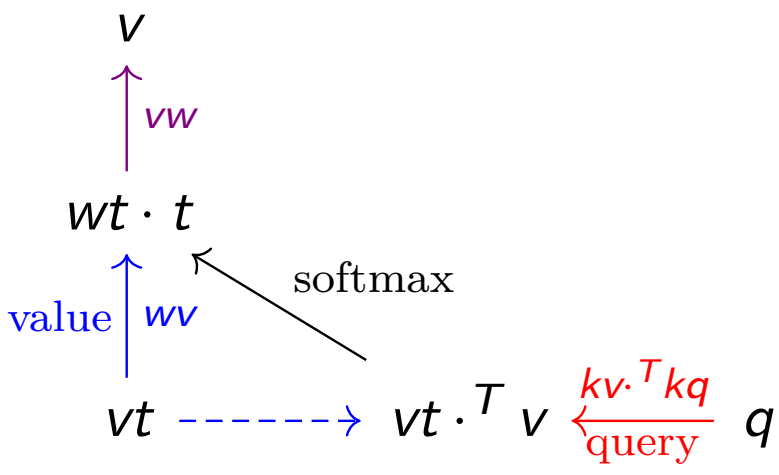
multi-head attention

- split spaces (query, value, key) into disjoint subspaces (subquery, subvalue, subkey)
- one attention head for each split
- concatenate the resulting subvectors

key transformation

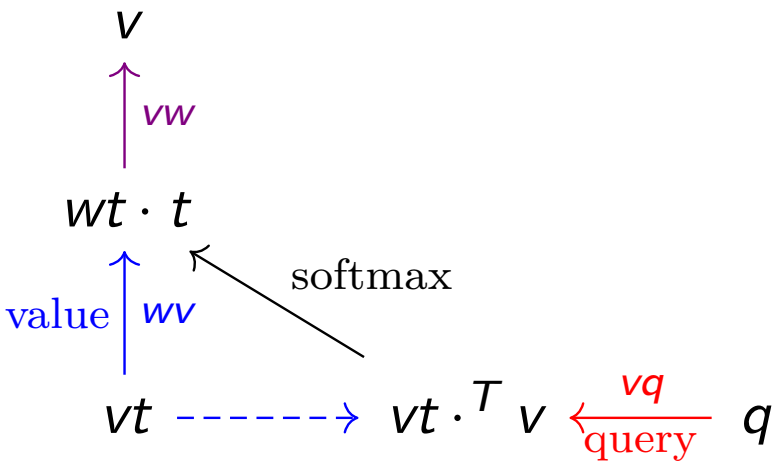


linear

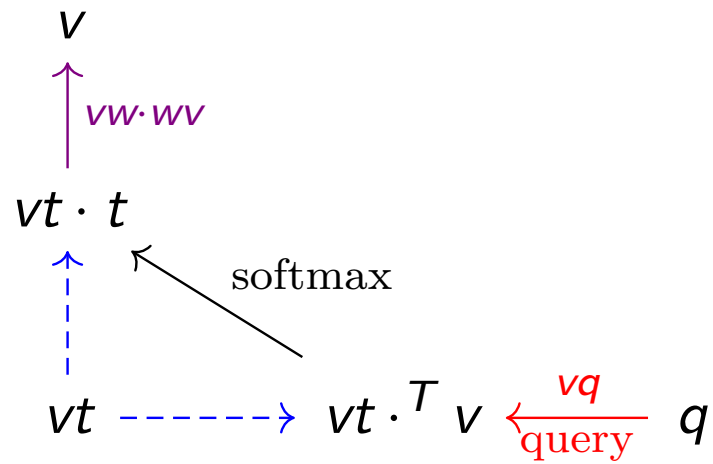


identity

value transformation

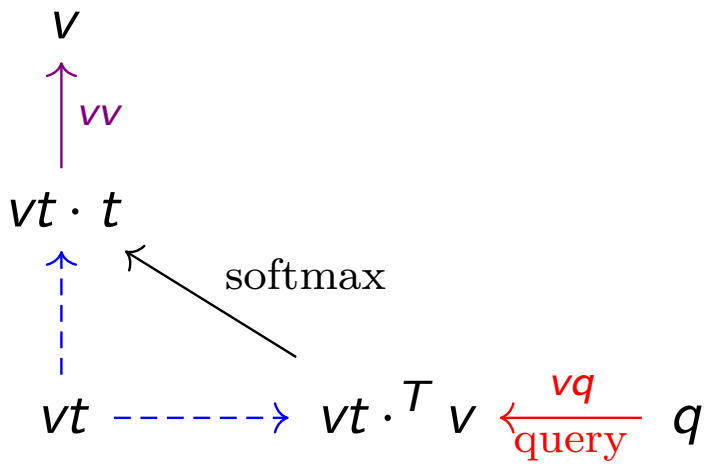


linear

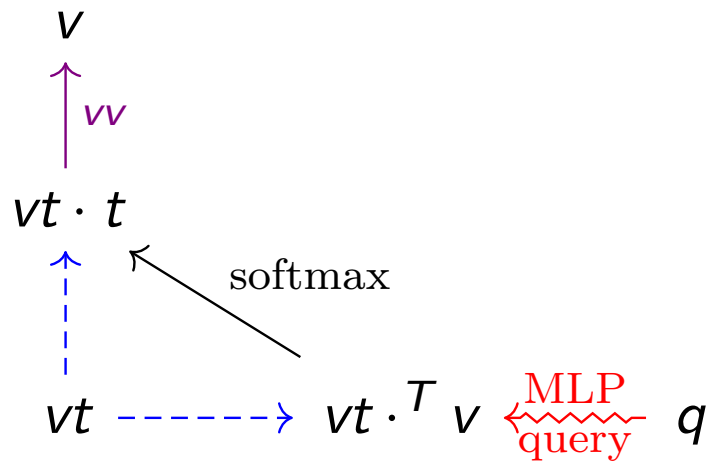


identity

query transformation

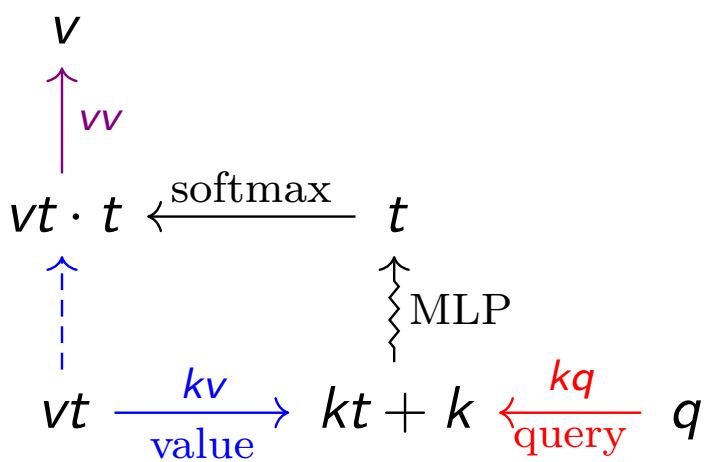


linear

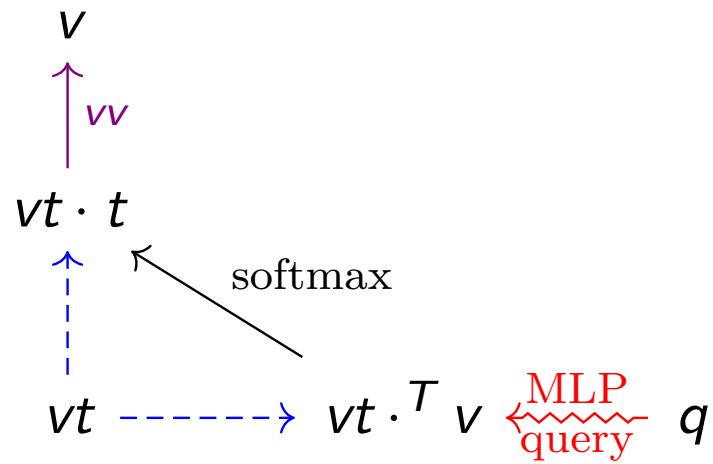


non-linear

additive vs dot-product with non-linear query

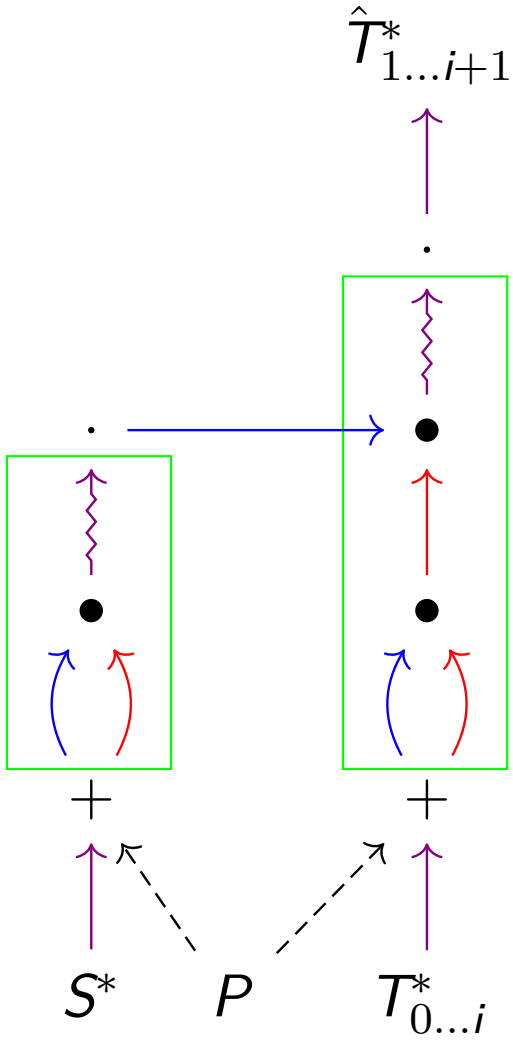


additive

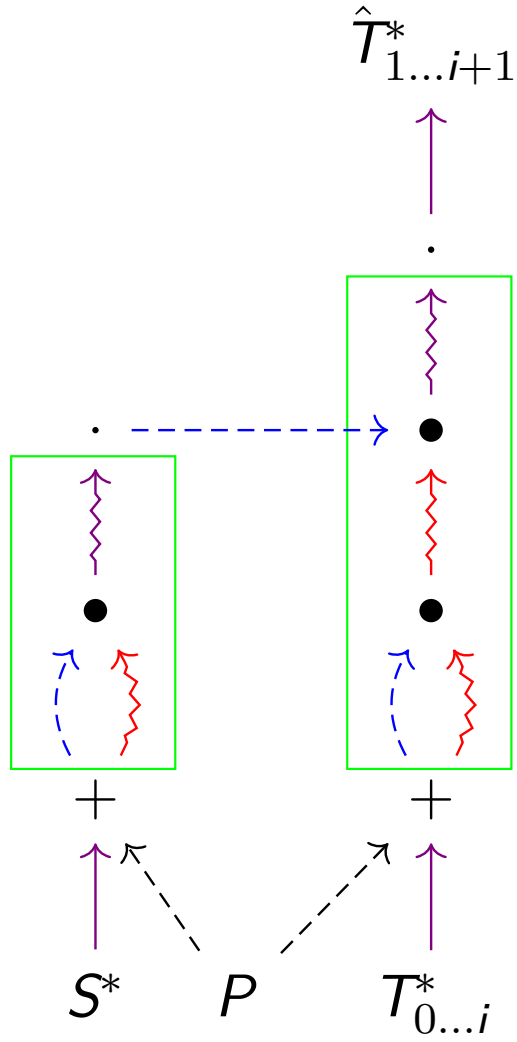


dot-product

transformer



old



new

architecture

- 2 encoder layers, 2 decoder layers
- 2 input embedding layers, 1 output softmax layer
- 256 representation dimension, 512 relu in MLPs
- single-head scaled dot-product attention
- dropout[8], residual connection[9], layer normalization[10] after each attention or MLP sublayer

[8] http://jmlr.org/papers/v15/srivastava14a.html
[9] https://arxiv.org/abs/1512.03385
[10] https://arxiv.org/abs/1607.06450

training

- cross entropy loss with label smoothing[11]
- teacher forcing
- batch size 64
- ~6 minutes per epoch (~223k instances)
- for 180 epochs

[11] https://arxiv.org/abs/1512.00567

introspection

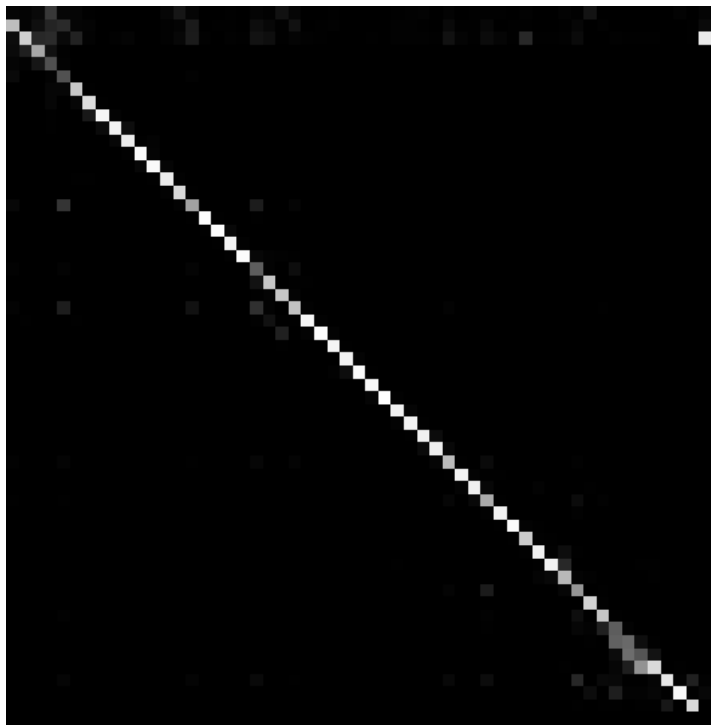
- greedy autoregressive decoding
- attention weight matrix

self-attention

- always a diagonal line
- encoder layer 1 and decoder layer 2 slightly fuzzy

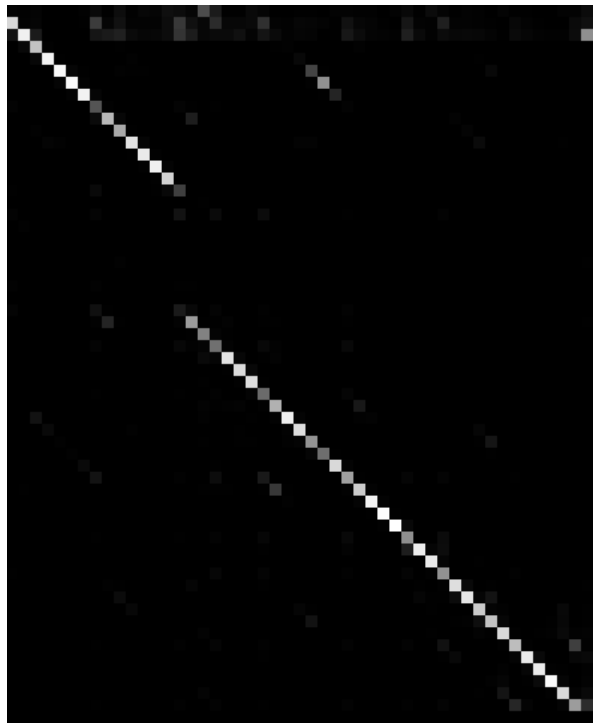
introspection: normal

```
The enshrined kami is Isonotakeru no mikoto ( 五十猛命? )

The enshrined kami is Isonotakeru no mikoto ( ____ ? )
```



introspection: skip

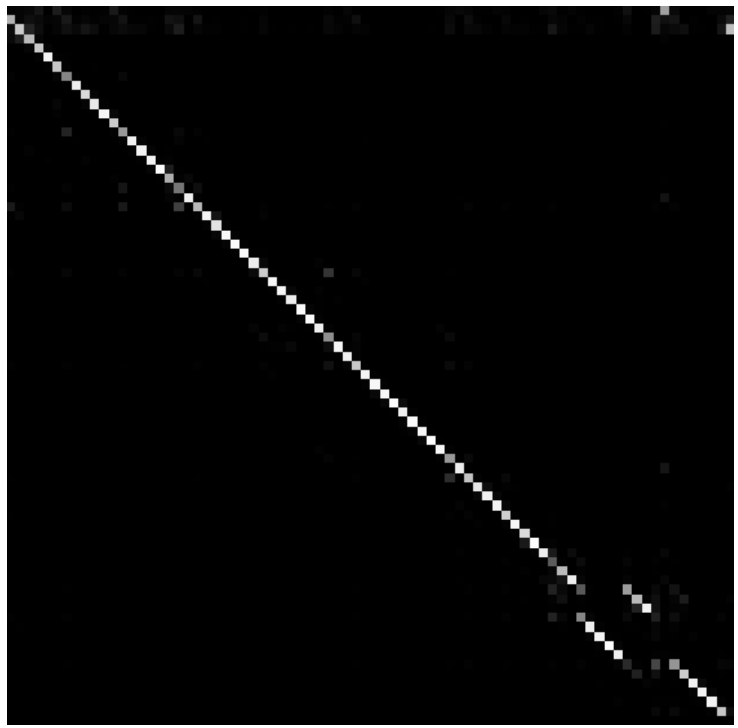
```
Sodium iodate ( NaIO3 ) is the sodium salt of iodic acid .

Sodium iodate is the sodium salt of iodic acid .
```



introspection: swap

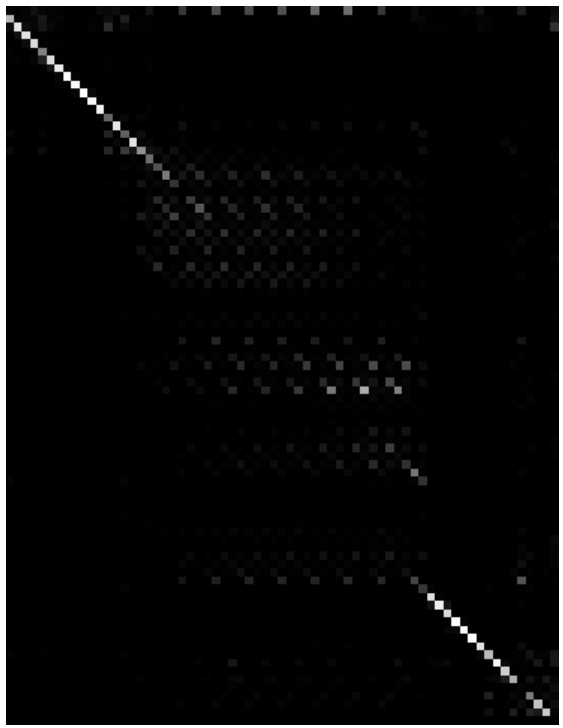
```
Their first child , Prince George of Cambridge , was born
on 22 July 2013 .
```

```
Their first child , Prince George of Cambridge , was born
on July 22 , 2013 .
```



introspection: confused

```
For example , 2 + 3 = 3 + 2 = 5 and 2 · 3 = 3 · 2 = 6 , but
23 = 8 , whereas 32 = 9 .

For example , 2 + 3 = 3 = 3 = 3 = 3 = 3 = 3 , whereas
32 = 9 .
```



introspection: really confused

```
1 and - 1 divide ( are divisors of ) every integer .

In the 19th century , the state of the state of the state
of the state of the state of the state of the state .
```