Natural Language Processing

Self Attention and Transformers

Vidhisha Balachandran

vbalacha@cs.cmu.edu

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Carnegie Mellon University
Language Technologies Institute

Announcements

- Please fill our survey!
- 3 thing that you like about the course
- 3 things that could be improved

Readings

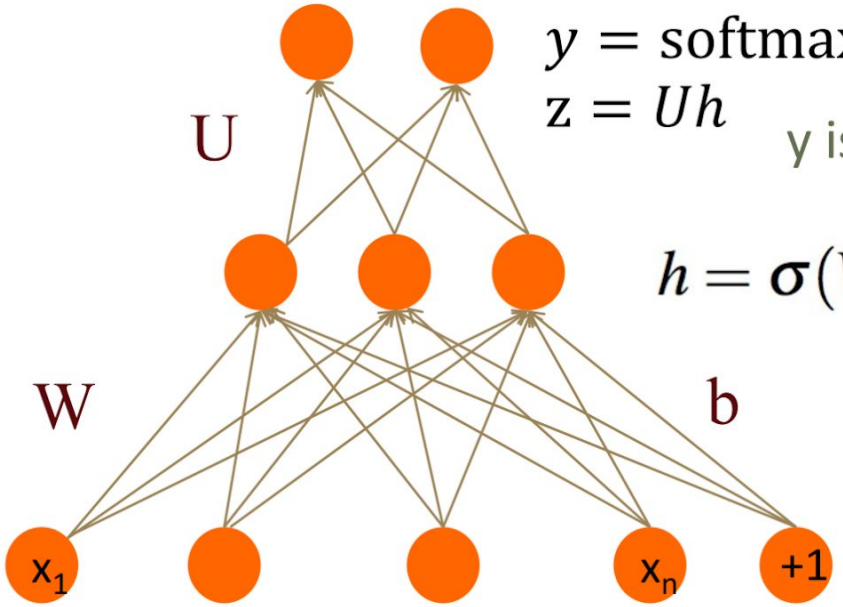
- Attention Is All You Need
- The Illustrated Transformer
- The Annotated Transformer
- Language Modeling with Transformers and PyTorch

Recap - 2 Layer MLP

Output layer
(σ node)

hidden units
(σ node)

Input layer
(vector)



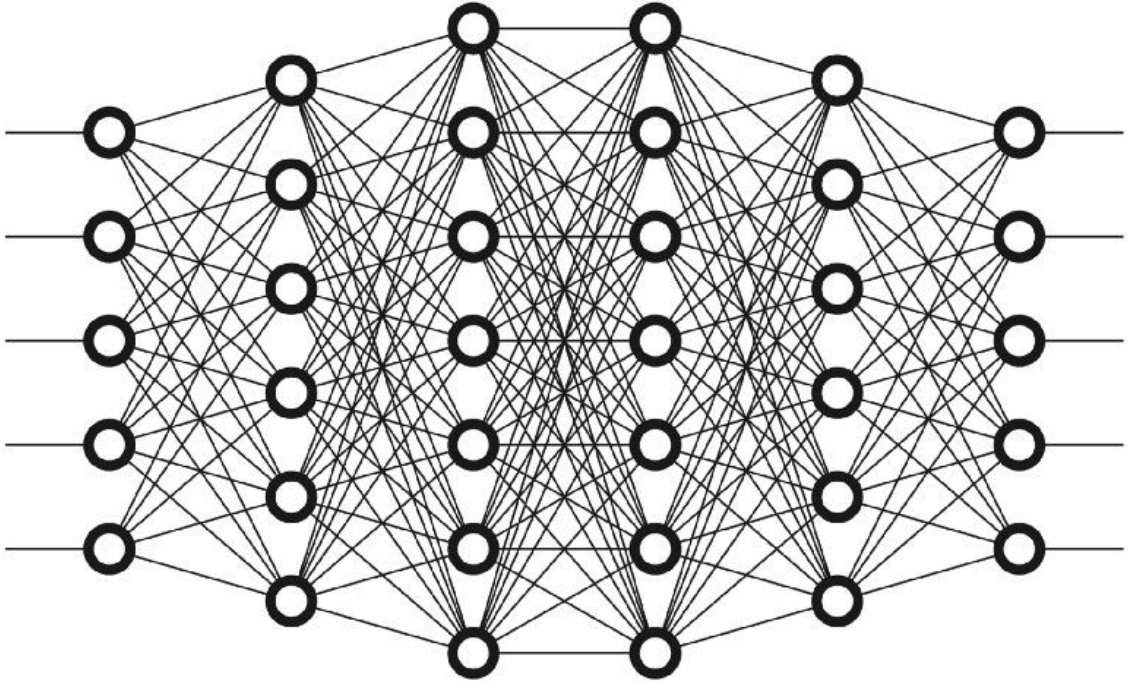
$$y = \text{softmax}(z)$$
$$z = Uh$$

y is a vector

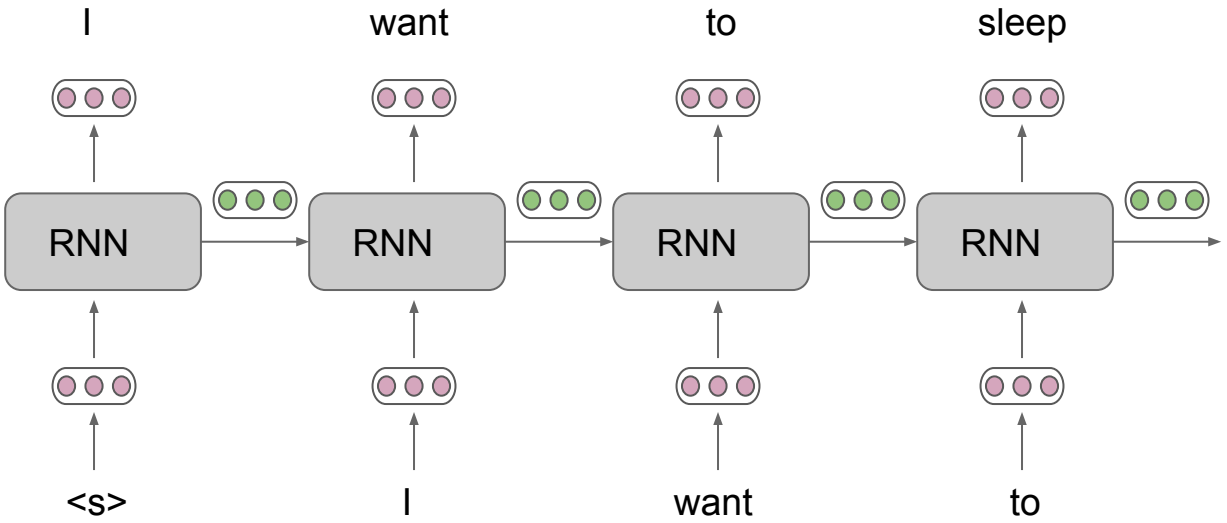
$$h = \sigma(Wx + b)$$

Could be ReLU
Or tanh

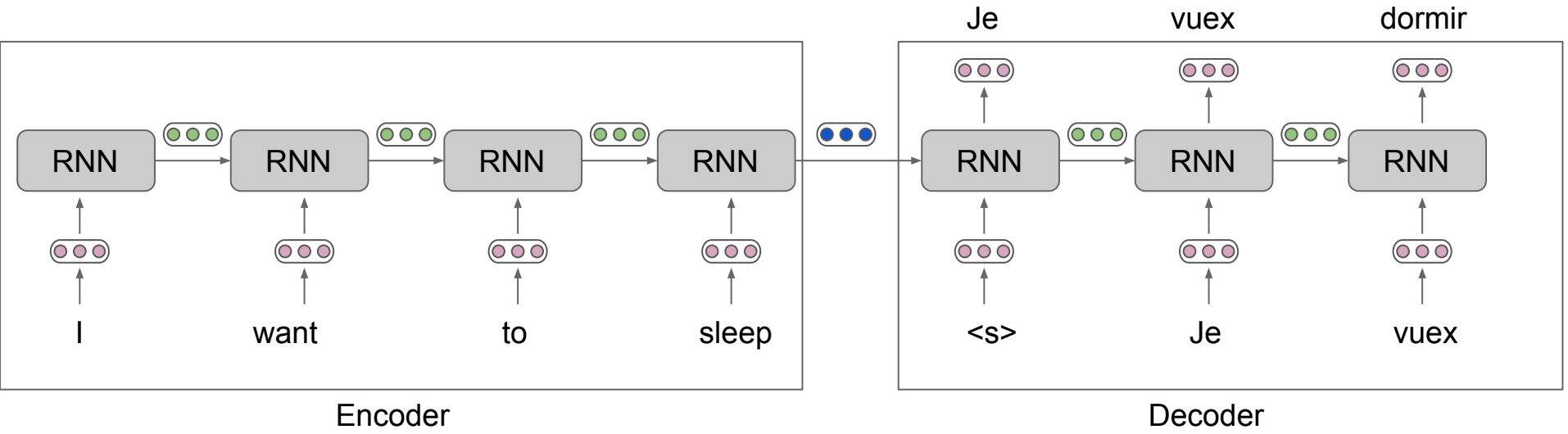
Deep MLP



Recurrent Neural Networks - RNNs



Encoder-Decoder Models



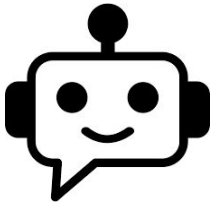
Limitations

- Long Range Dependencies
- Gradient vanishing / explosion
- Long time to converge
- Expensive computation

Long Range Dependencies

I'm want to watch Wicked! How does the weather in NYC look next week?

It looks sunny with some light rain during the weekend.



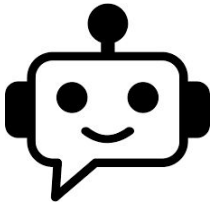
Oh! But I don't have a rain jacket :(Is there a store nearby?

There's a marshall's a mile away. They have the navy blue jacket you have been eyeing for a while!

Long Range Dependencies

I'm want to watch Wicked! How does the weather in NYC look next week?

It looks sunny with some light rain during the weekend.



Oh! But I don't have a rain jacket :(Is there a store nearby?

There's a marshall's a mile away. They have the navy blue jacket you have been eyeing for a while!

Ok! Looks like I can actually go! Book **the tickets** for next Wed!

Long Range Dependencies

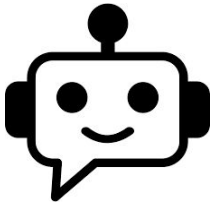
I'm want to watch **Wicked**! How does the weather in NYC look next week?

It looks sunny with some light rain during the weekend.

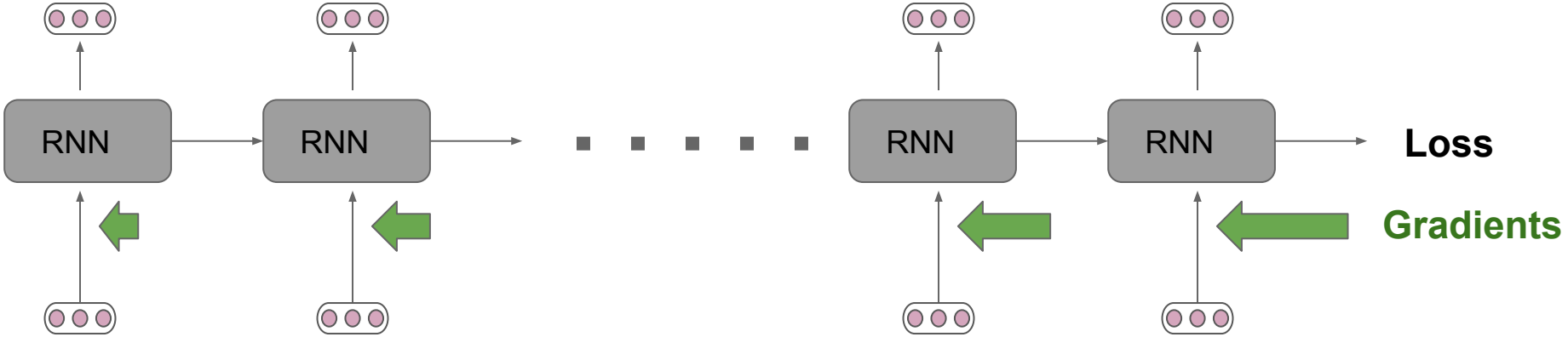
Oh! But I don't have a rain jacket :(Is there a store nearby?

There's a marshall's a mile away. They have the navy blue jacket you have been eyeing for a while!

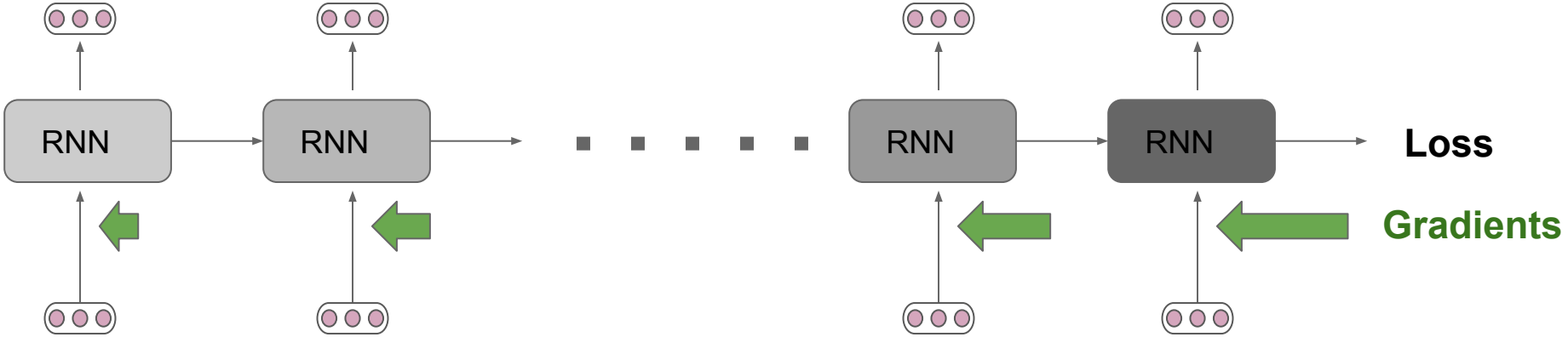
Ok! Looks like I can actually go! Book **the tickets** for next Wed!



Gradient vanishing / explosion



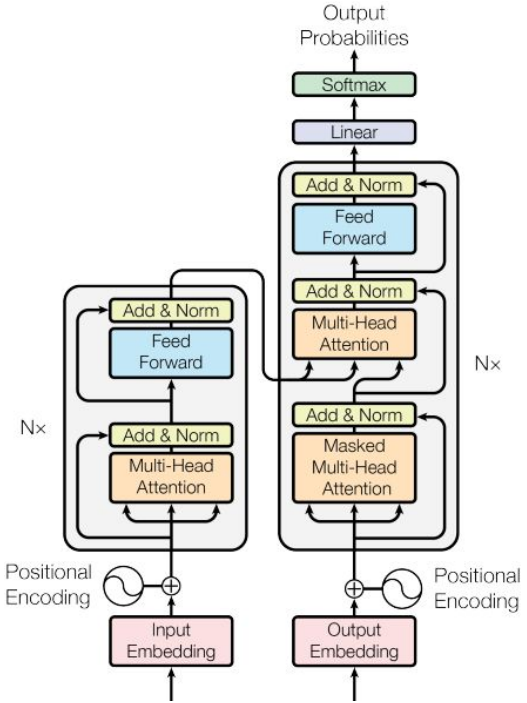
Gradient vanishing / explosion



Limitations

- Long Range Dependencies
- Gradient vanishing / explosion
- Long time to converge
- Expensive computation

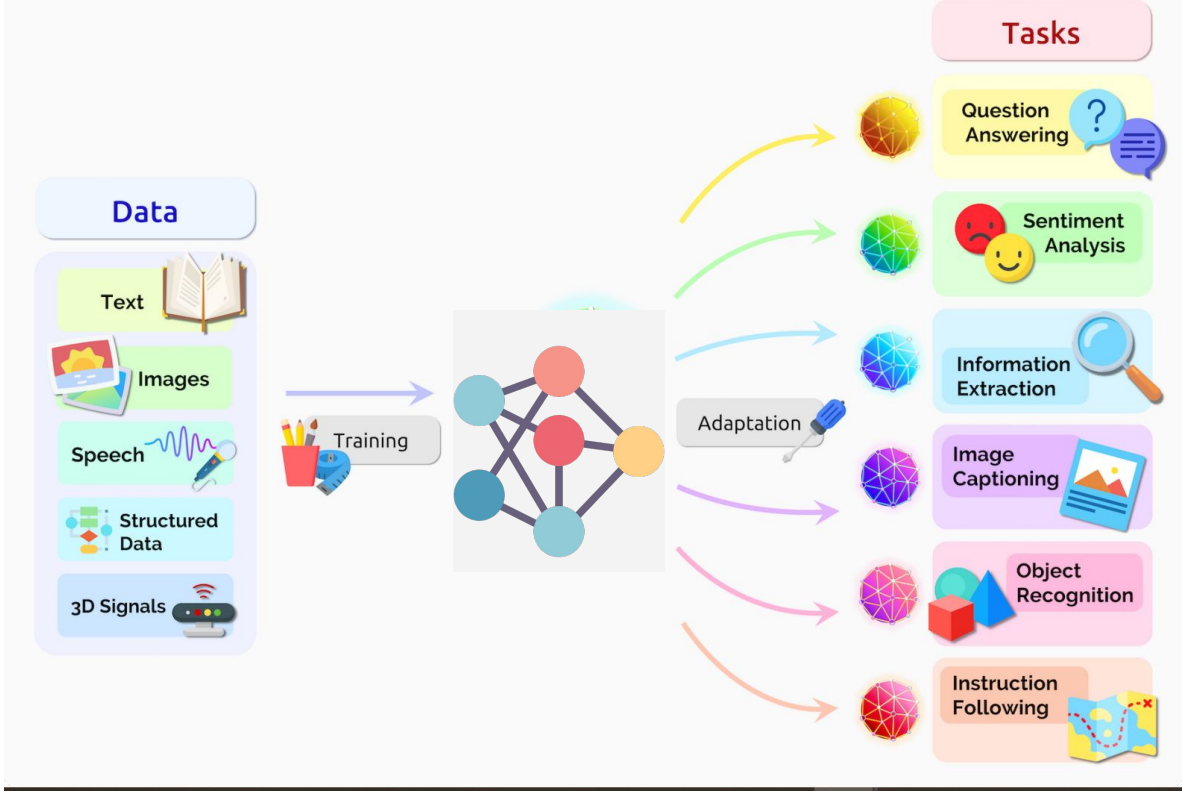
Transformer Model



Google Research

Attention is all you need (Vaswani et.al, 2017)

Wide Applications



https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/

Real World Impact

Machine Translation

Smart Assistants

Search Engines

Auto Transcription

google-research/
pegasus

Summarization Engines

Health Record Analysis

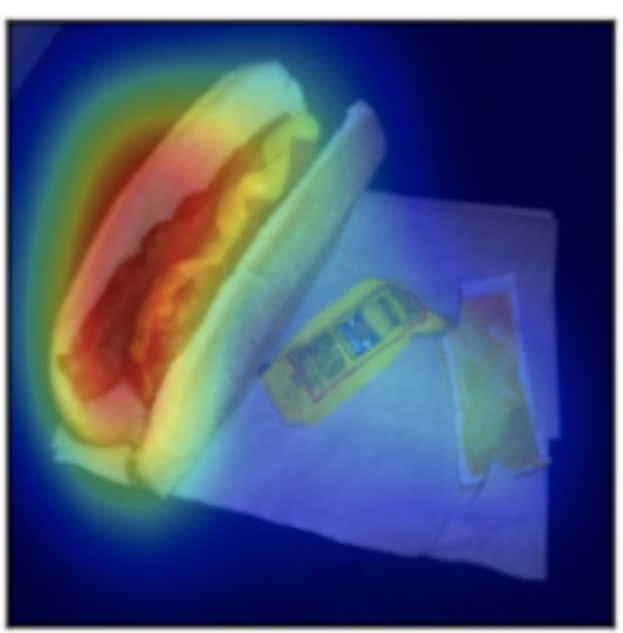
and many more ....

1

Questions?

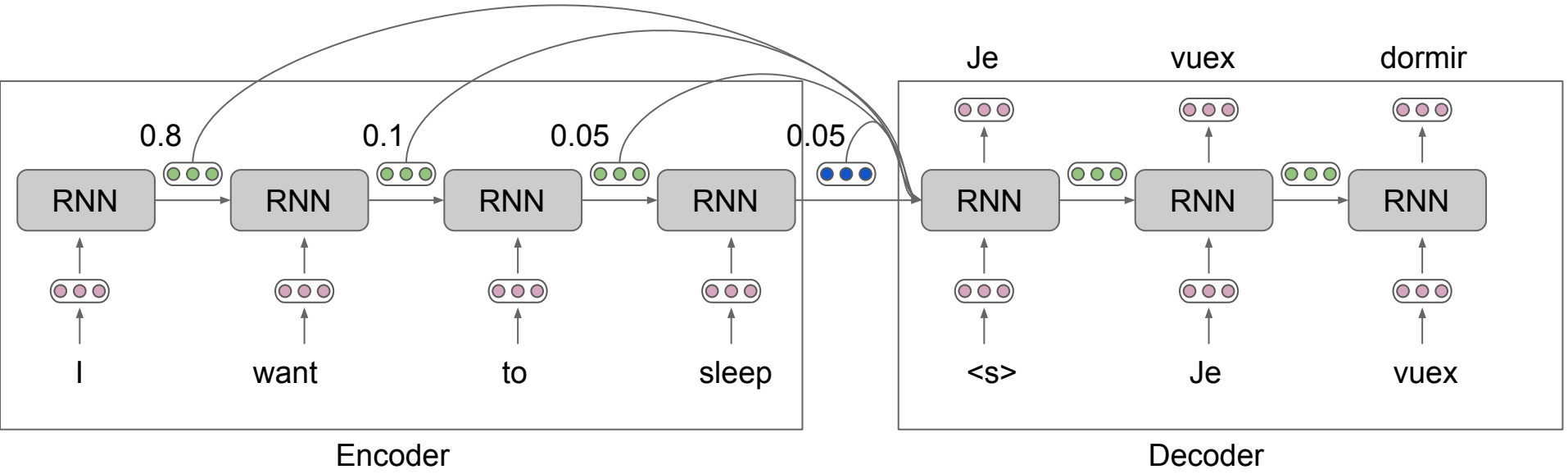
Visual Attention

What toppings are on the hot dog?



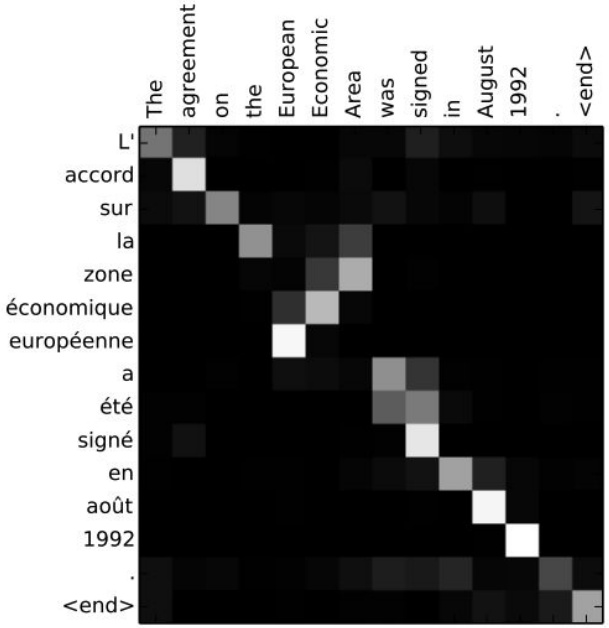
Differential Attention for Visual Question Answering (Patro et.al, 2018)

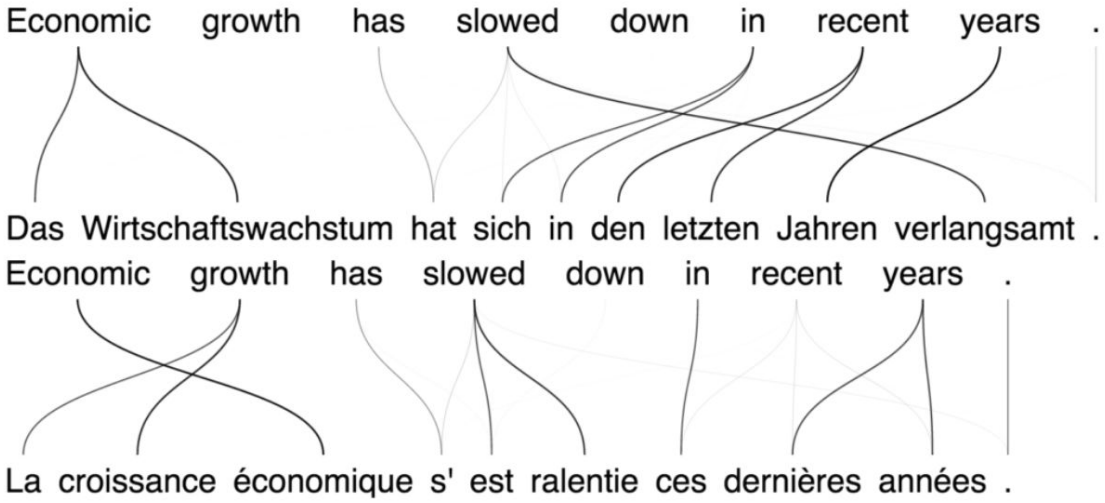
Cross Attention in NMT



Neural Machine Translation by Jointly Learning to Align and Translate. Bahdanau et al, 2015

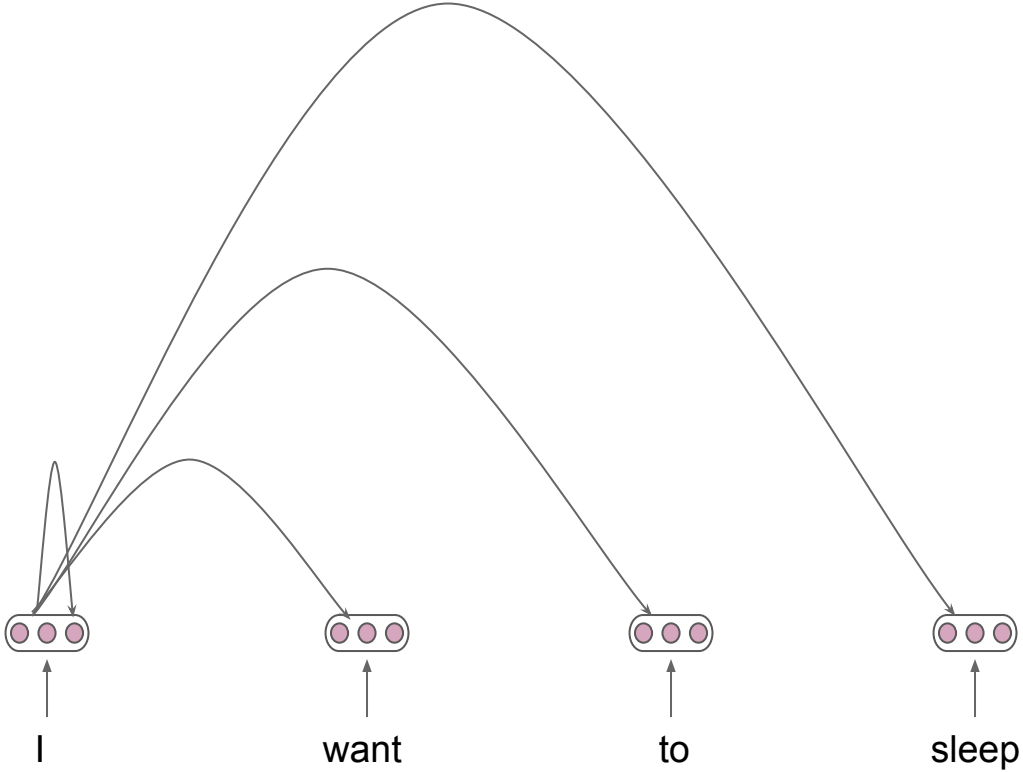
Attention in NMT



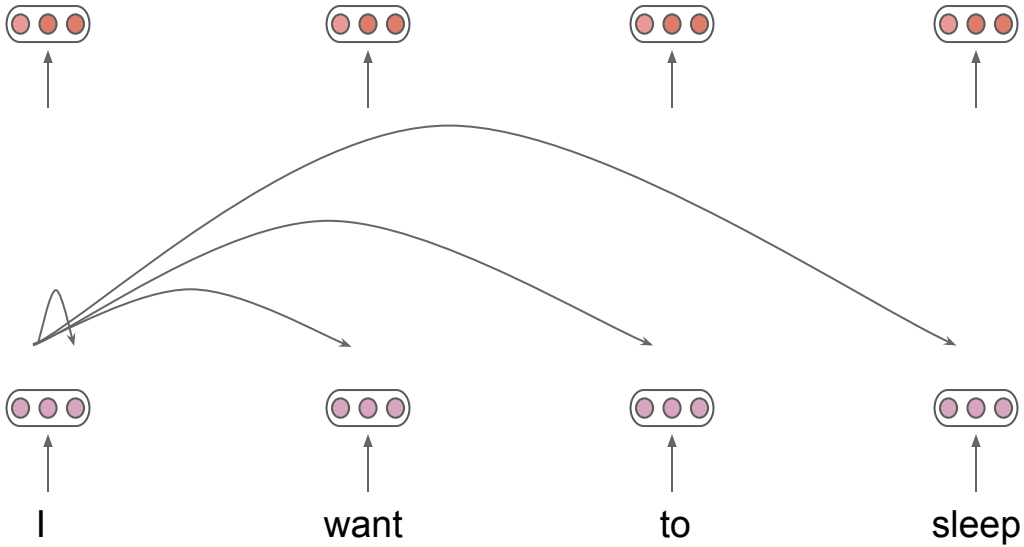


Neural Machine Translation by Jointly Learning to Align and Translate. Bahdanau et al, 2015
https://developer.nvidia.com/blog/introduction-neural-machine-translation-gpus-part-3/

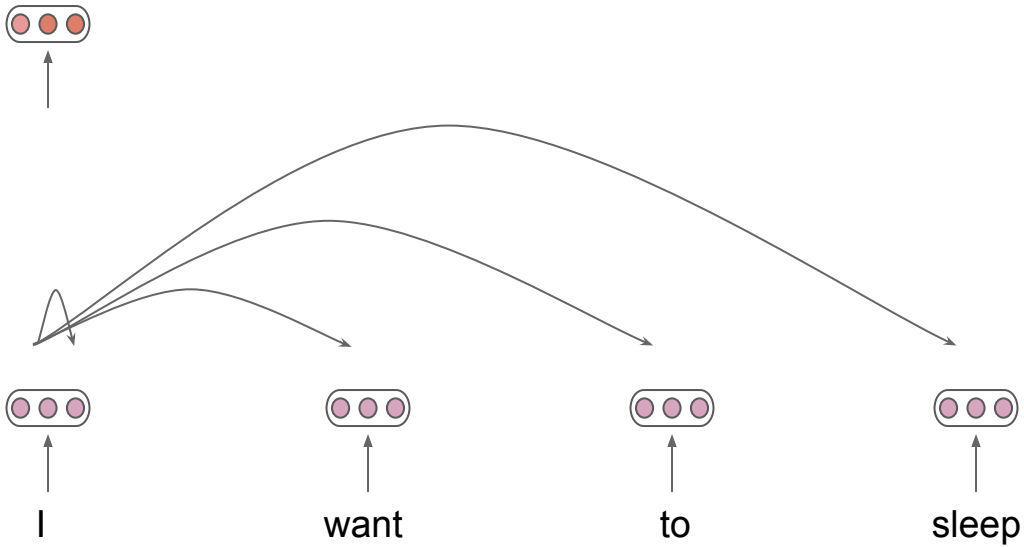
Self Attention



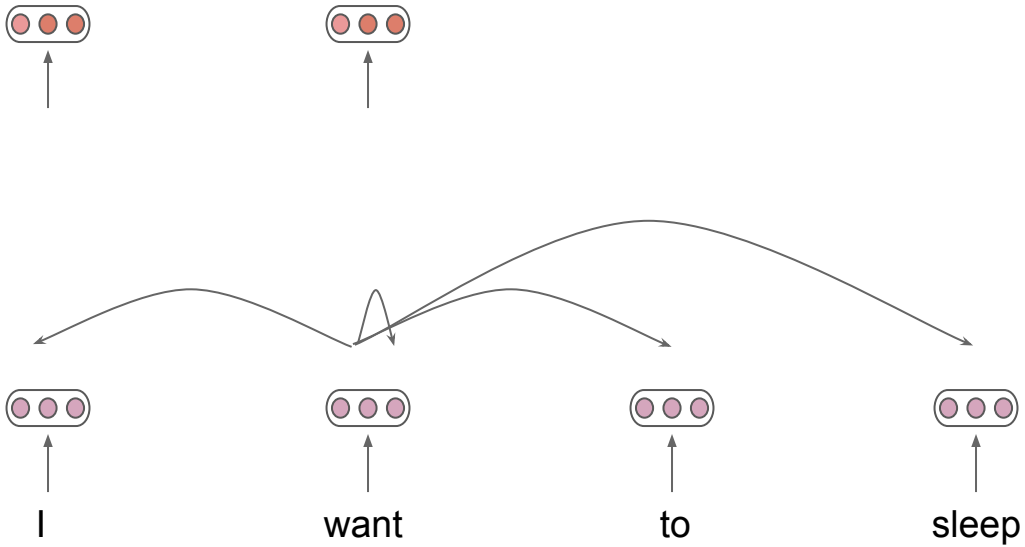
Self Attention - No more recurrence



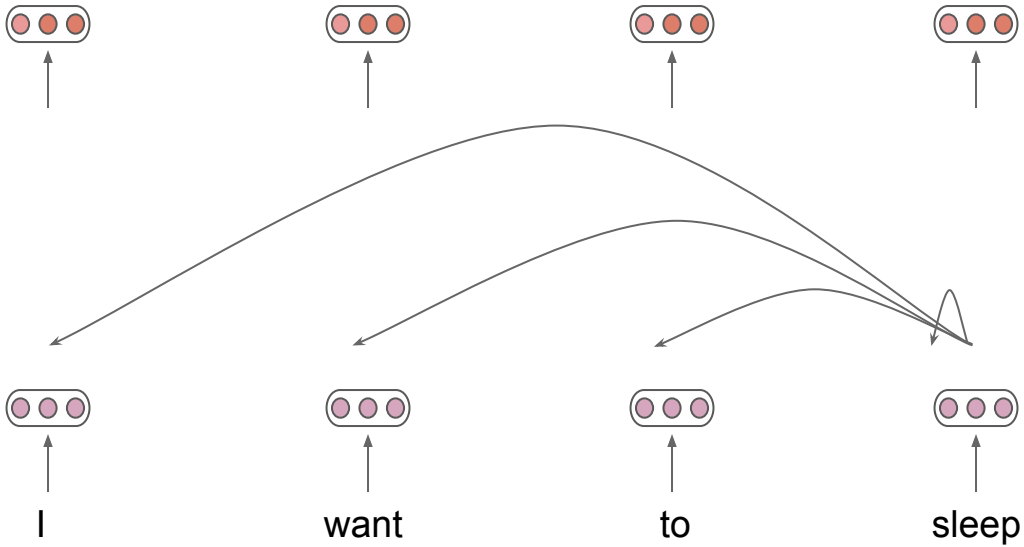
Self Attention - No more recurrence



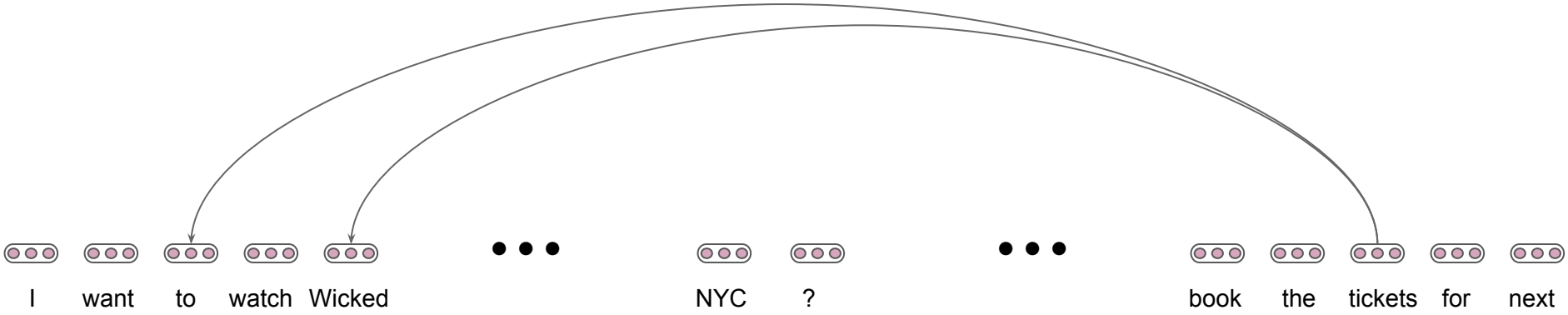
Self Attention - No more recurrence



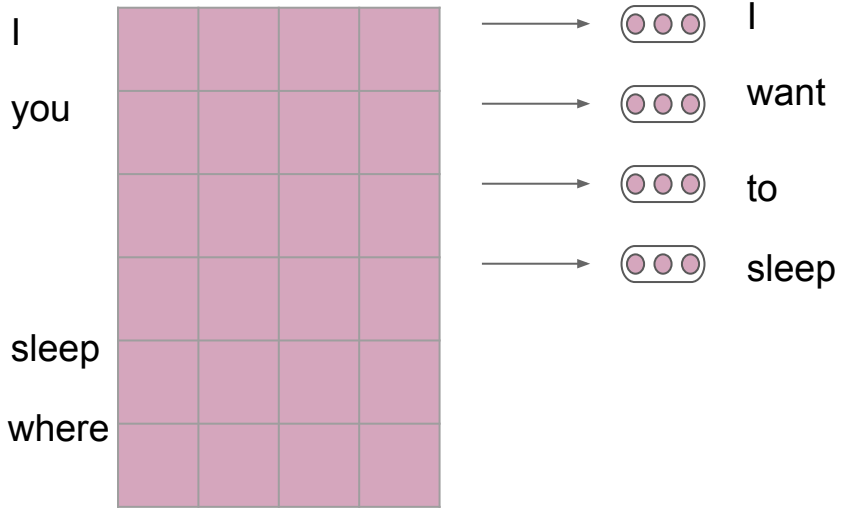
Self Attention - No more recurrence

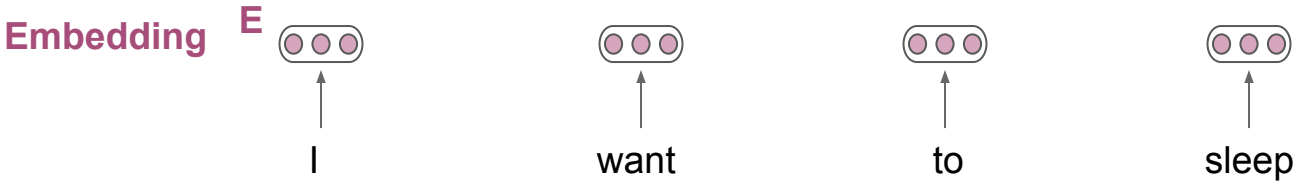


Self Attention for long sequences

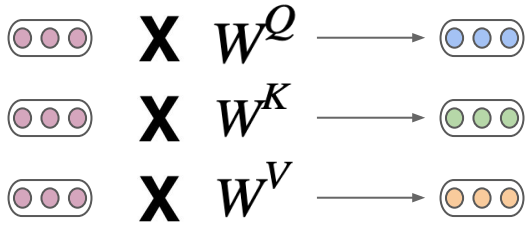


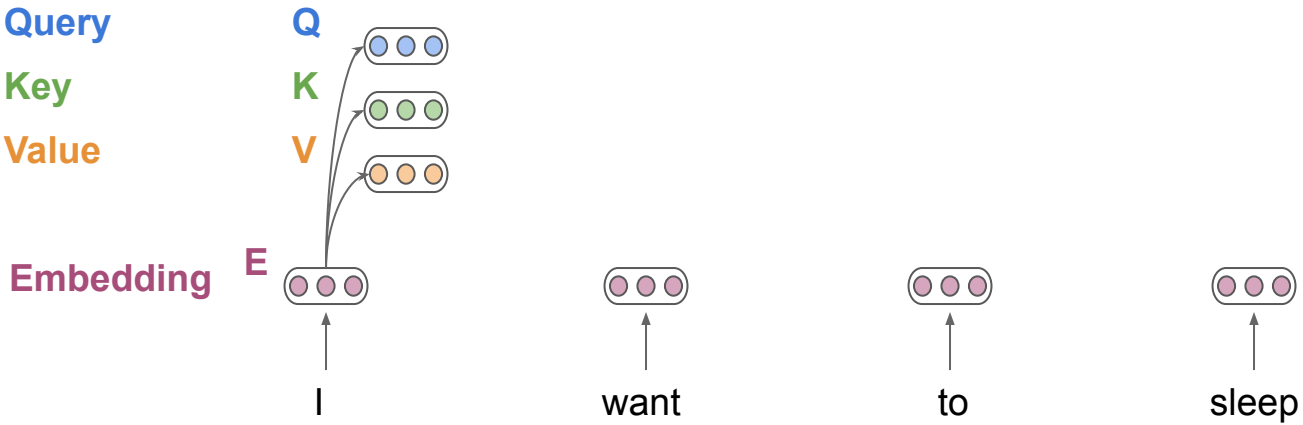
Self Attention - Word Embedding



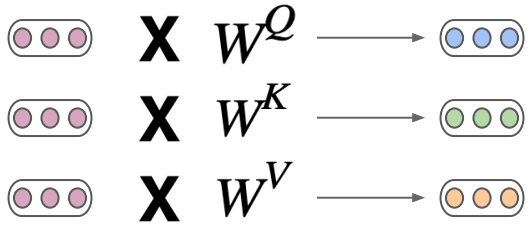


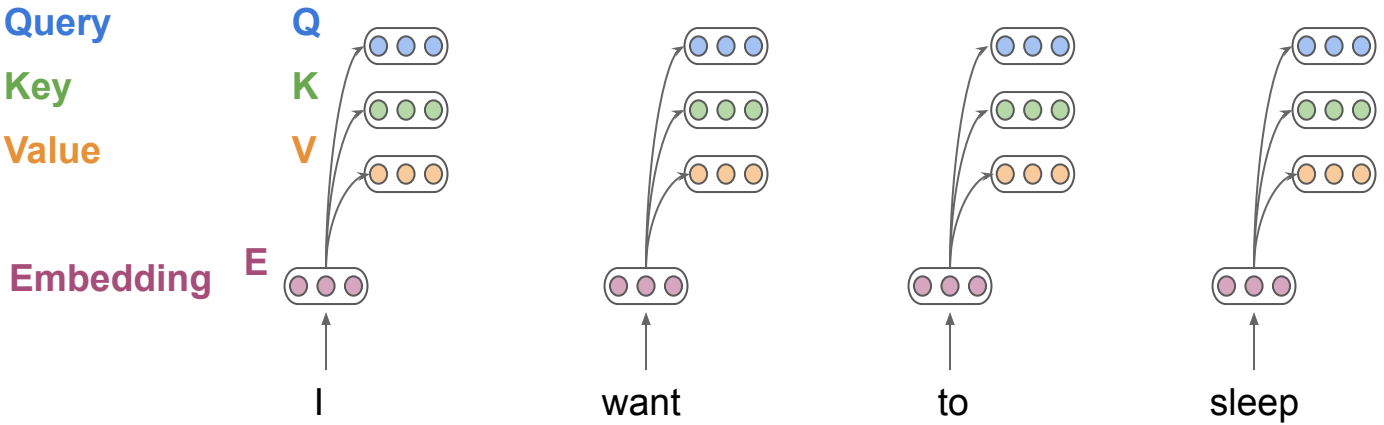
Self Attention - Projection Layer



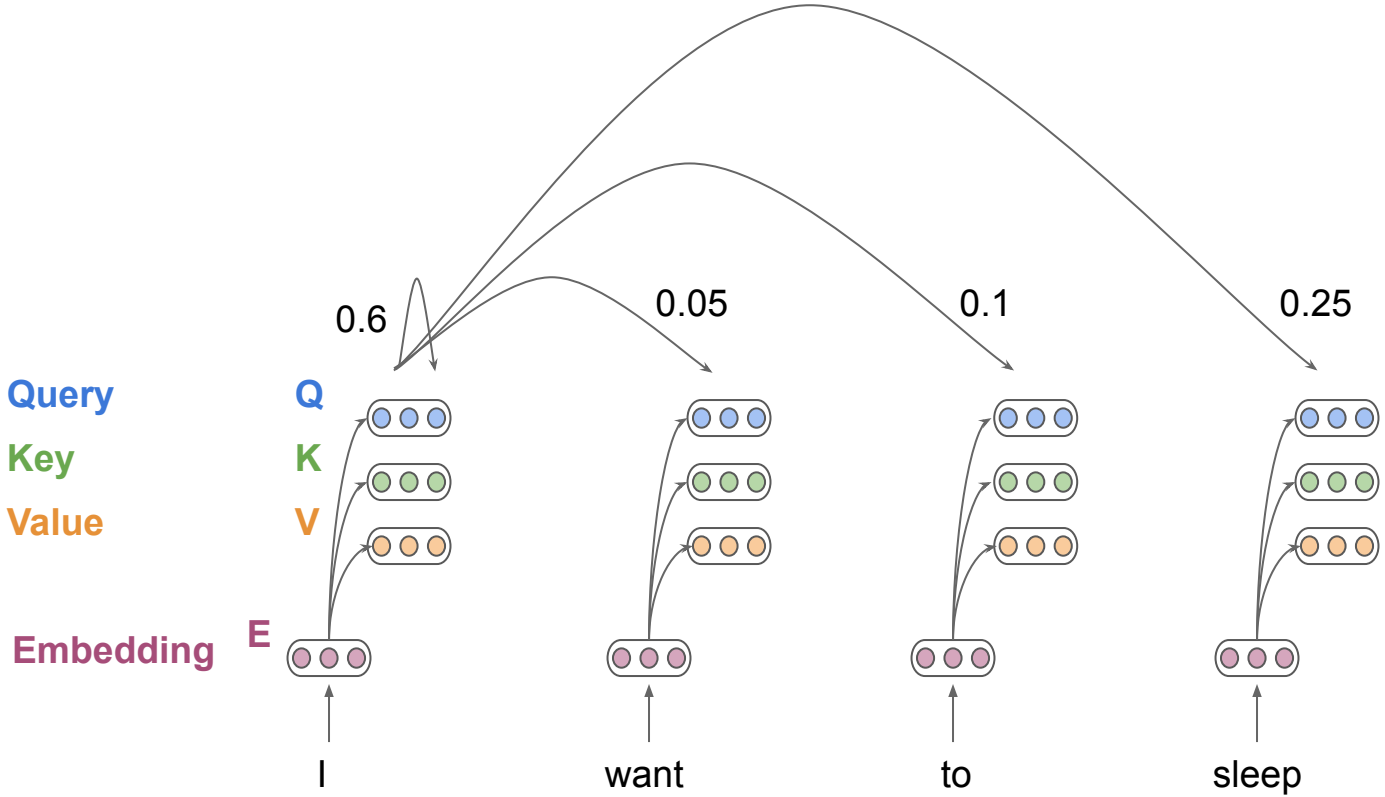


Self Attention - Projection Layer

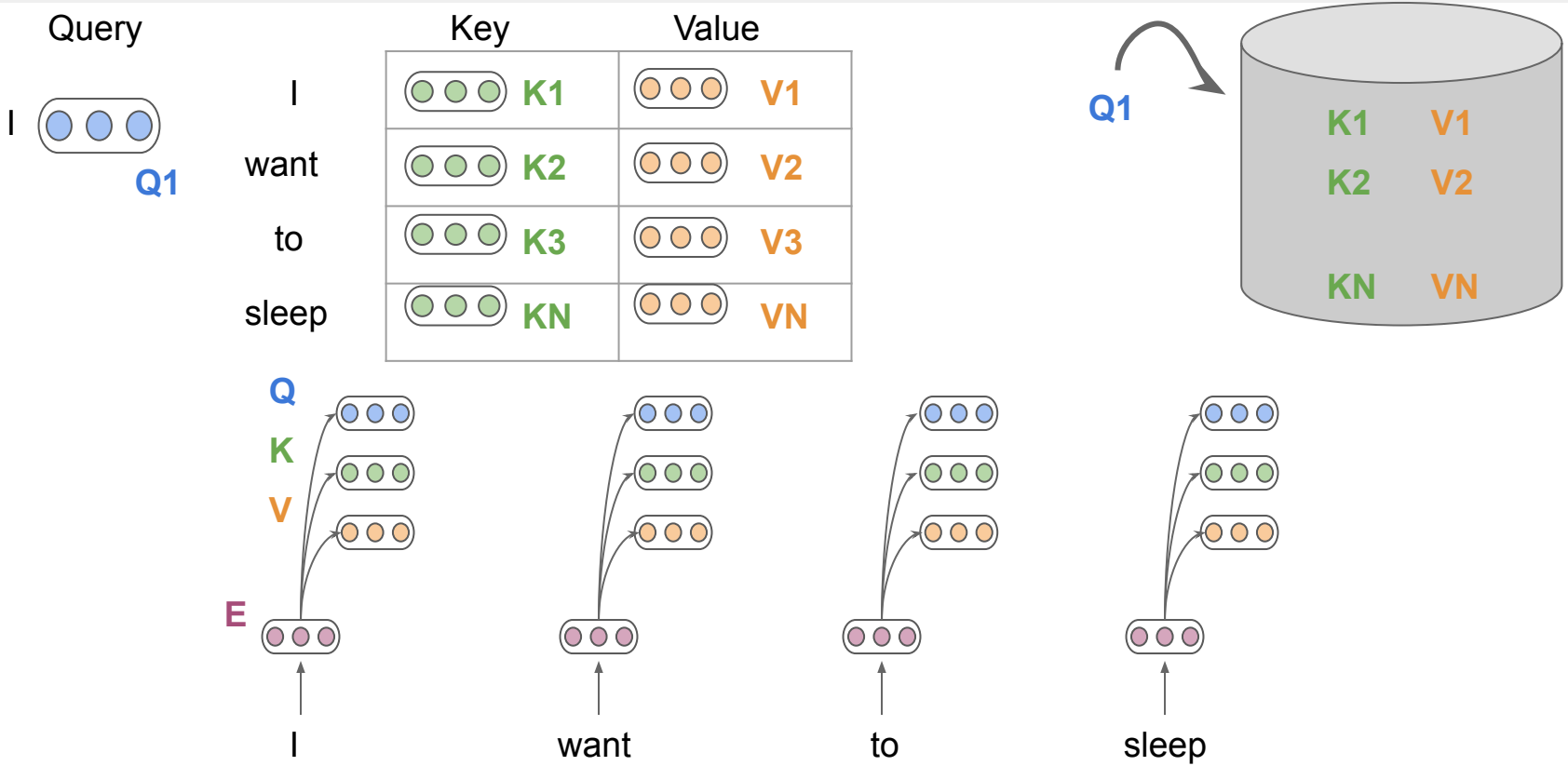




Self Attention - Attention Scores

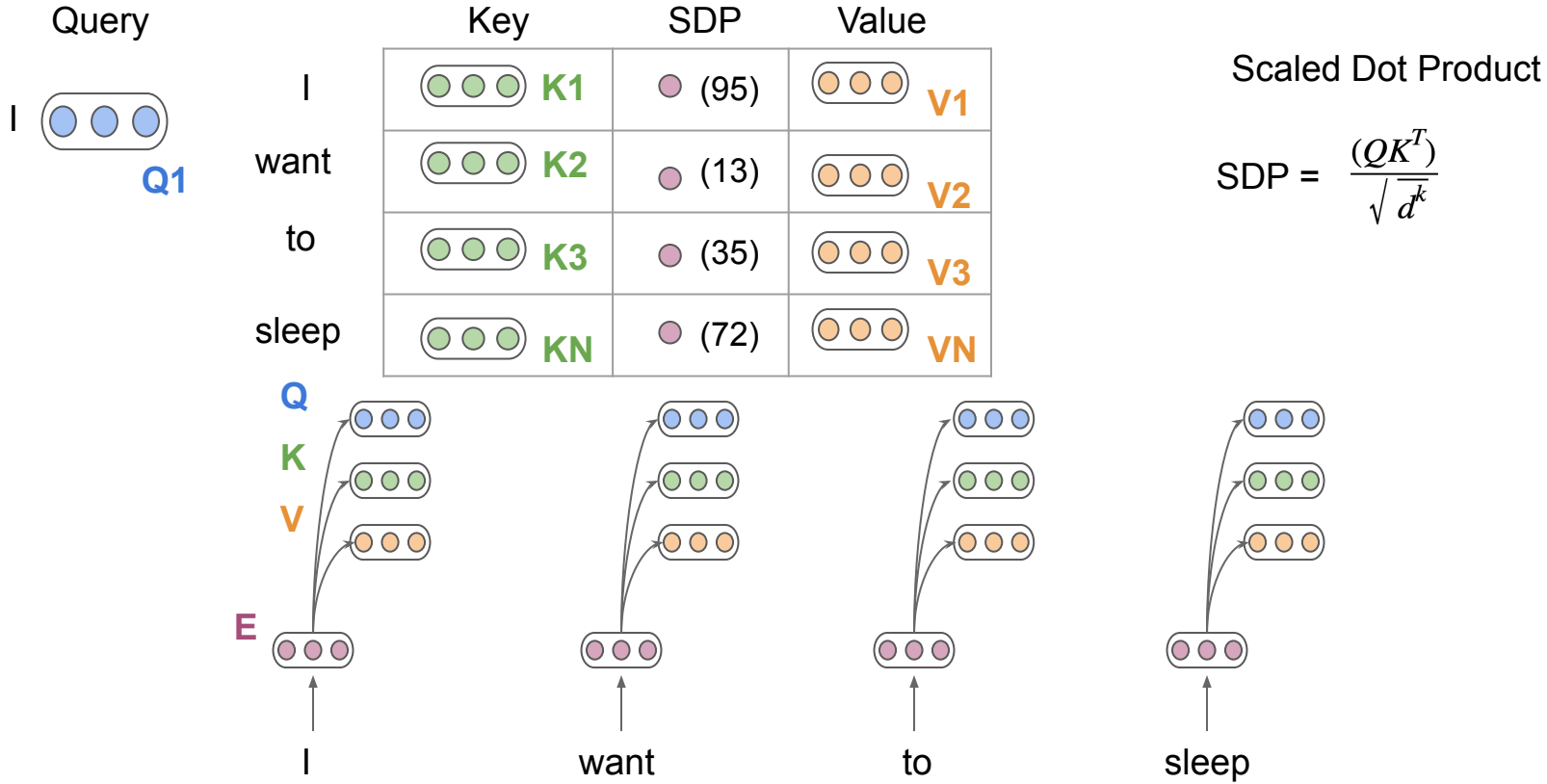


Self Attention



Questions?

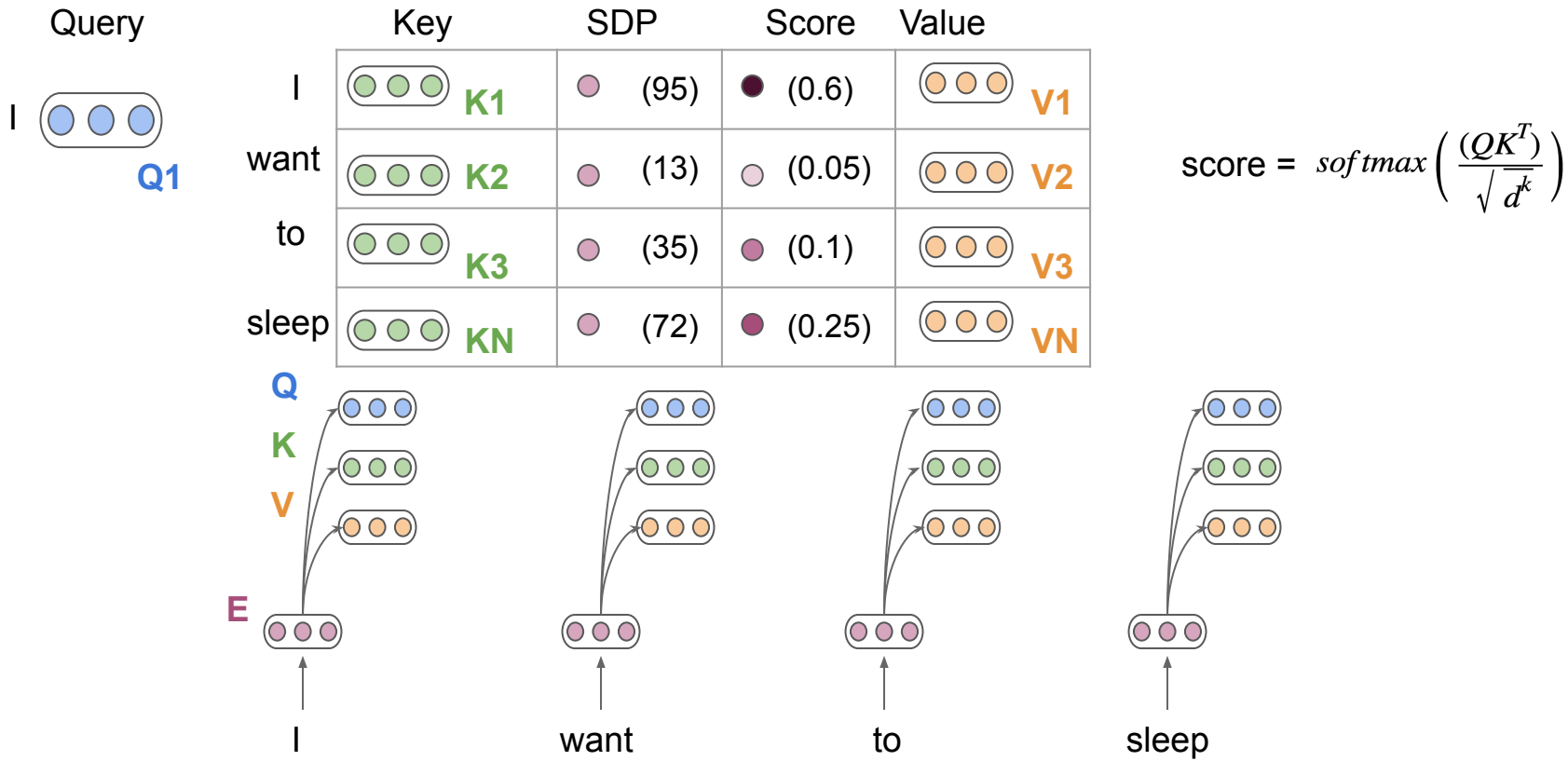
Self Attention - Scaled Dot Product



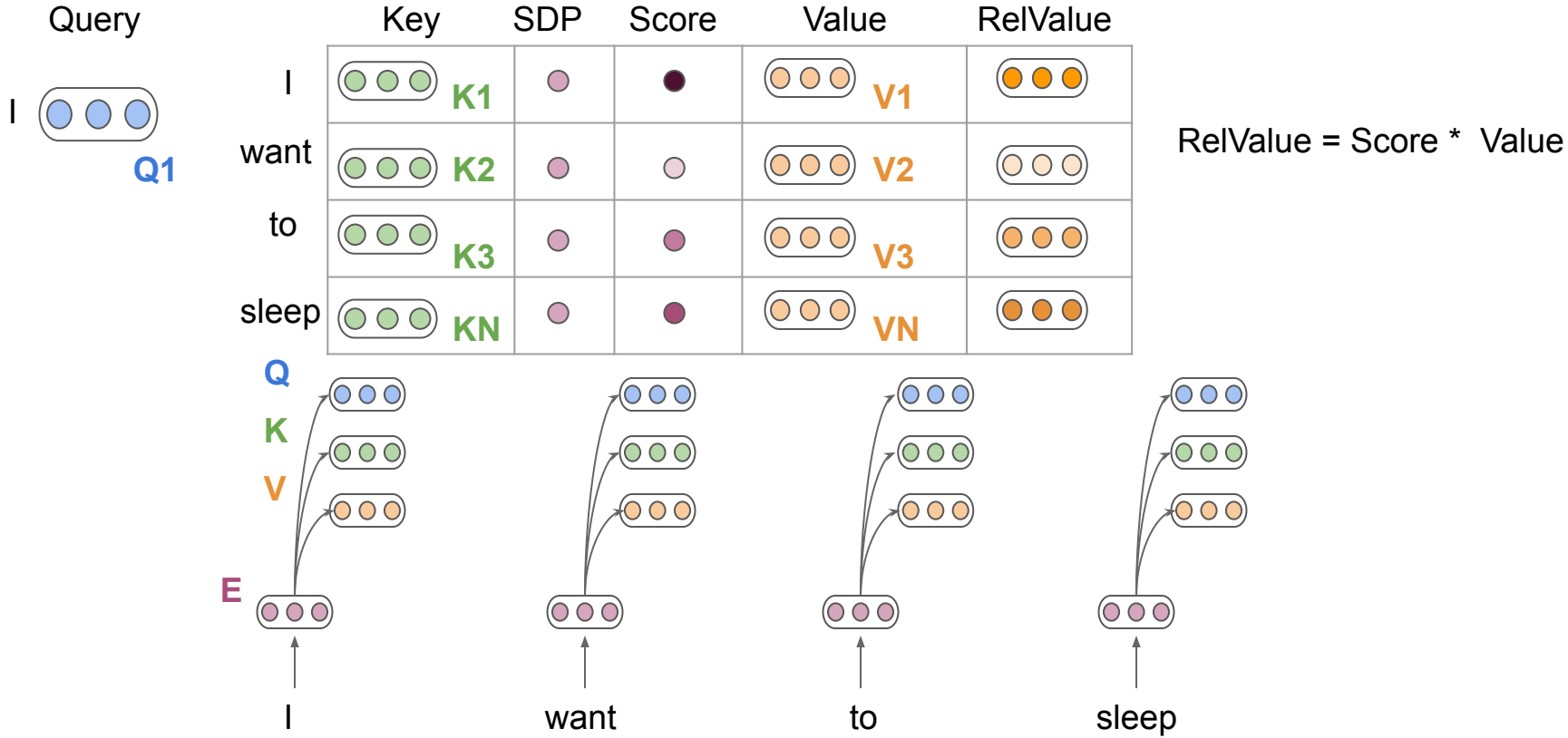
Scaled Dot Product

$$\text{SDP} = \frac{(QK^T)}{\sqrt{d^k}}$$

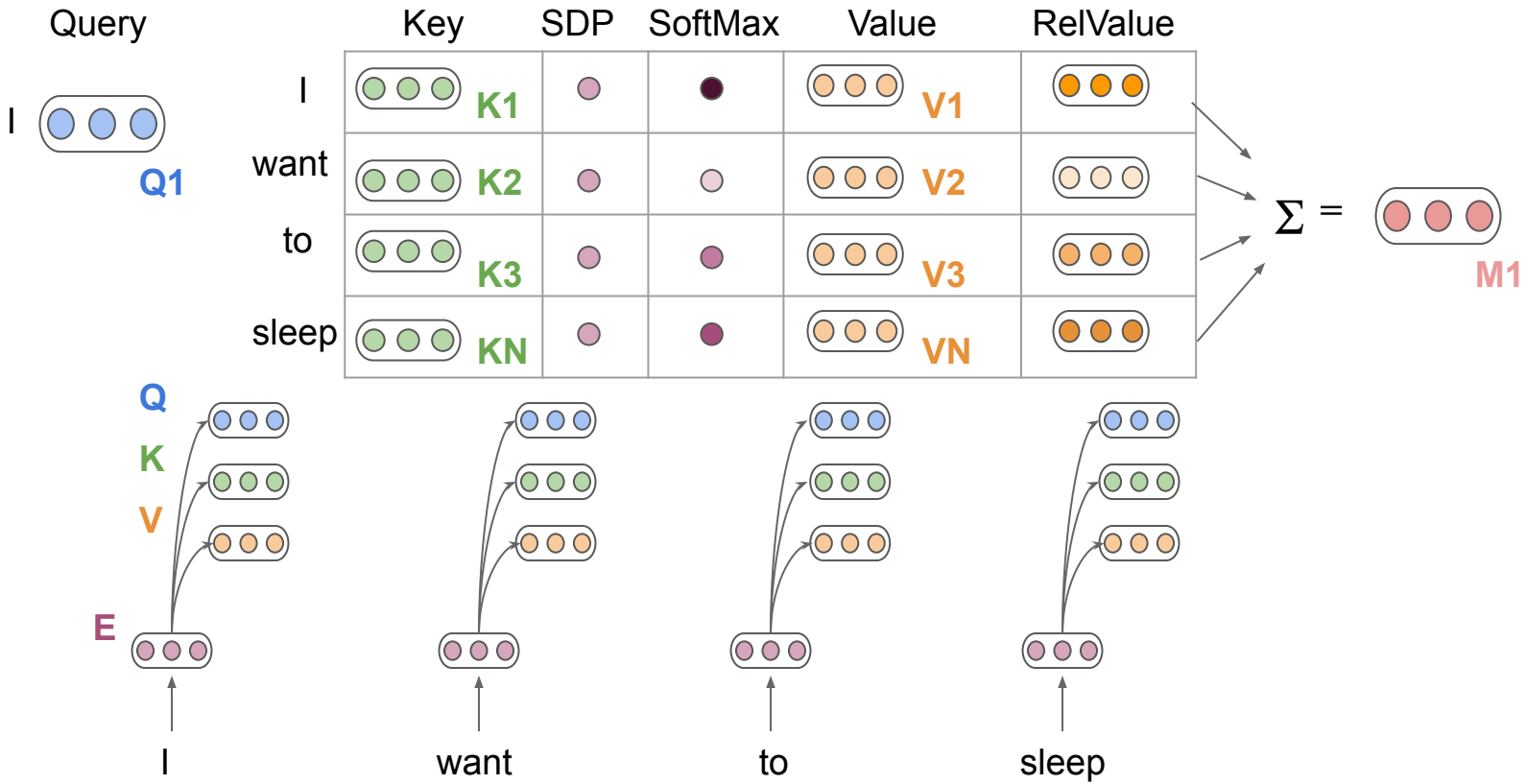
Self Attention - SoftMax



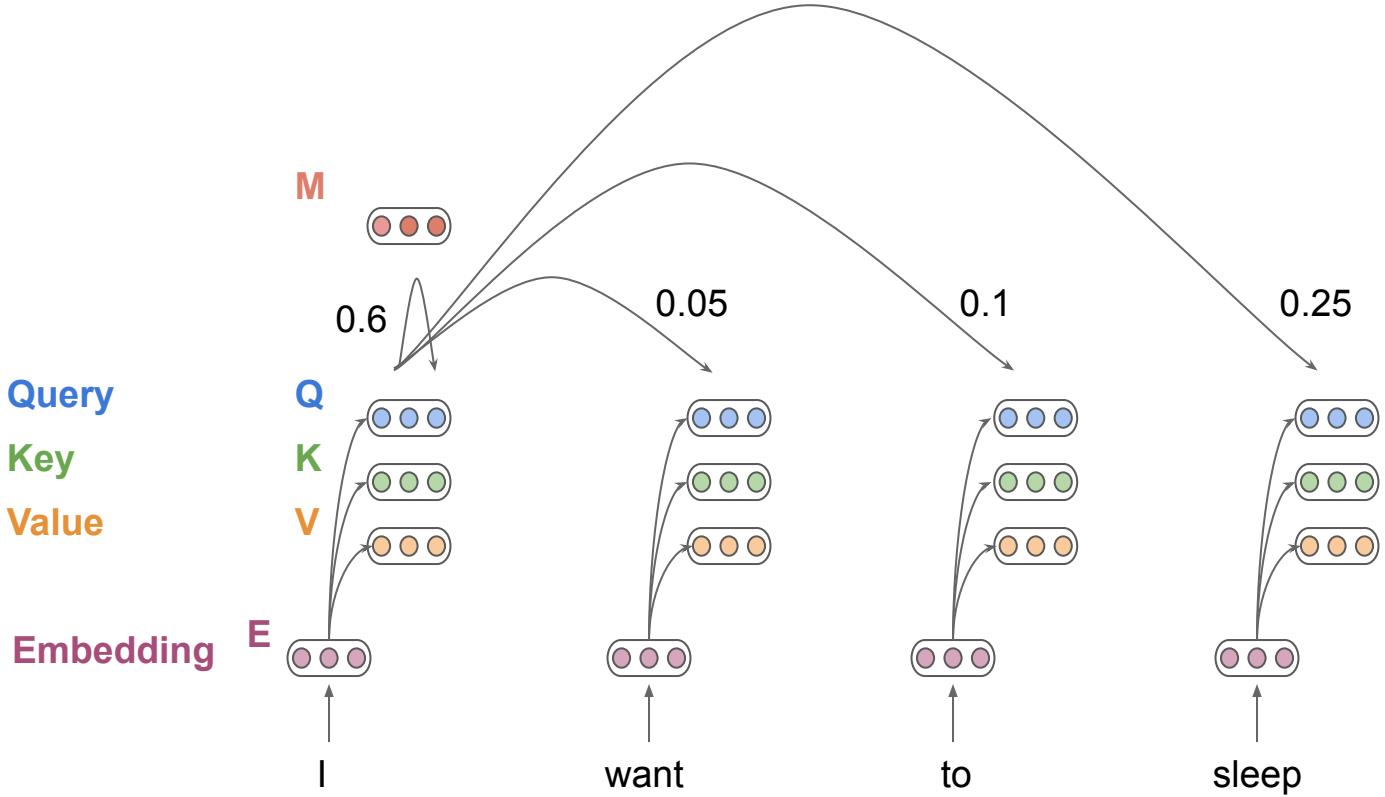
Self Attention - Soft (Relative) Values



Self Attention - Attended Repr

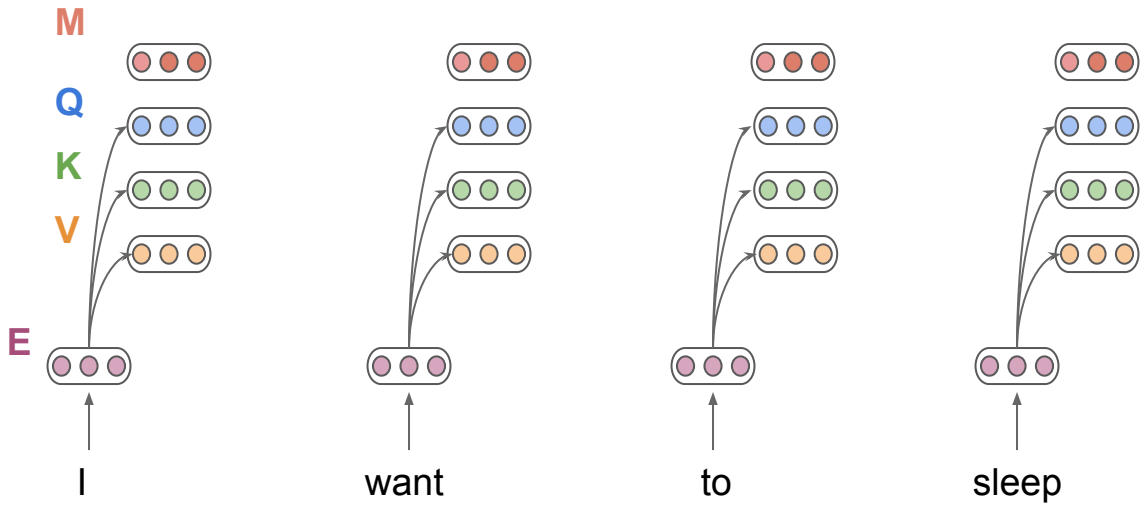


Self Attention - Attended Contextual Rep



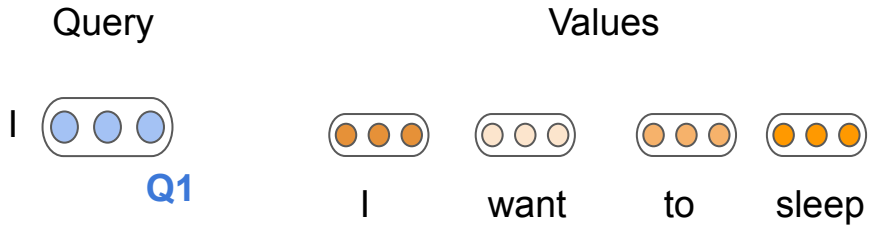
Questions?

Self Attention - Attended Contextual Rep



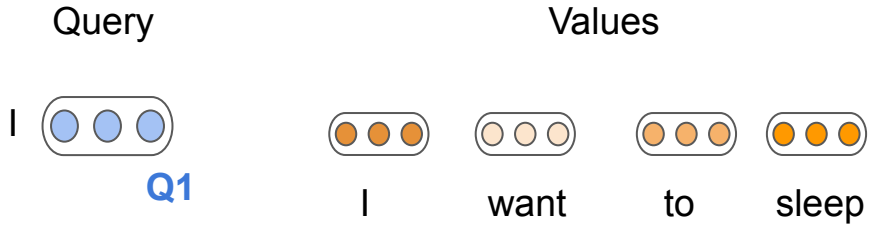
Problem with Self Attention

- Self Attention can focus heavily on the same word!



Problem with Self Attention

- Self Attention can focus heavily on the same word!



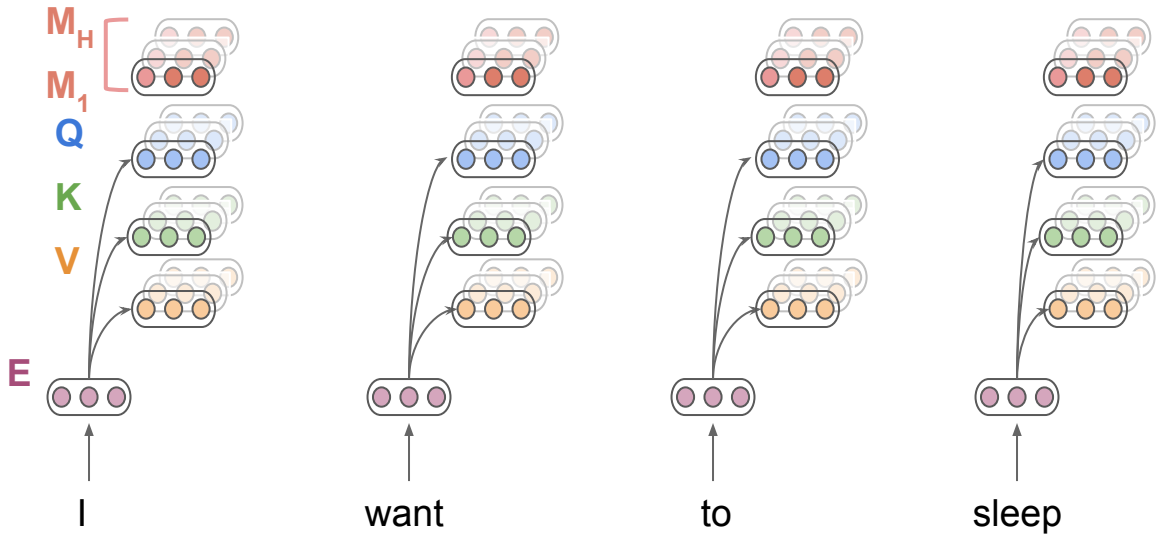
- Single representation

I like **Harry** **Potter** v/s I like **Harry** **Potter**

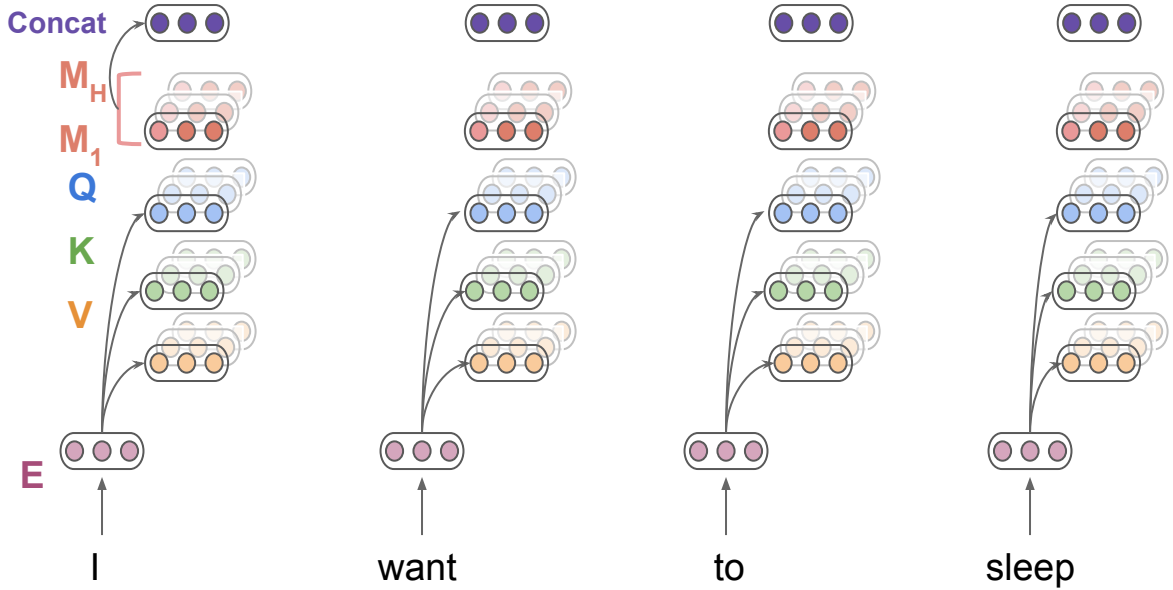
(Book) (Movie)

Multi-Headed Self Attention

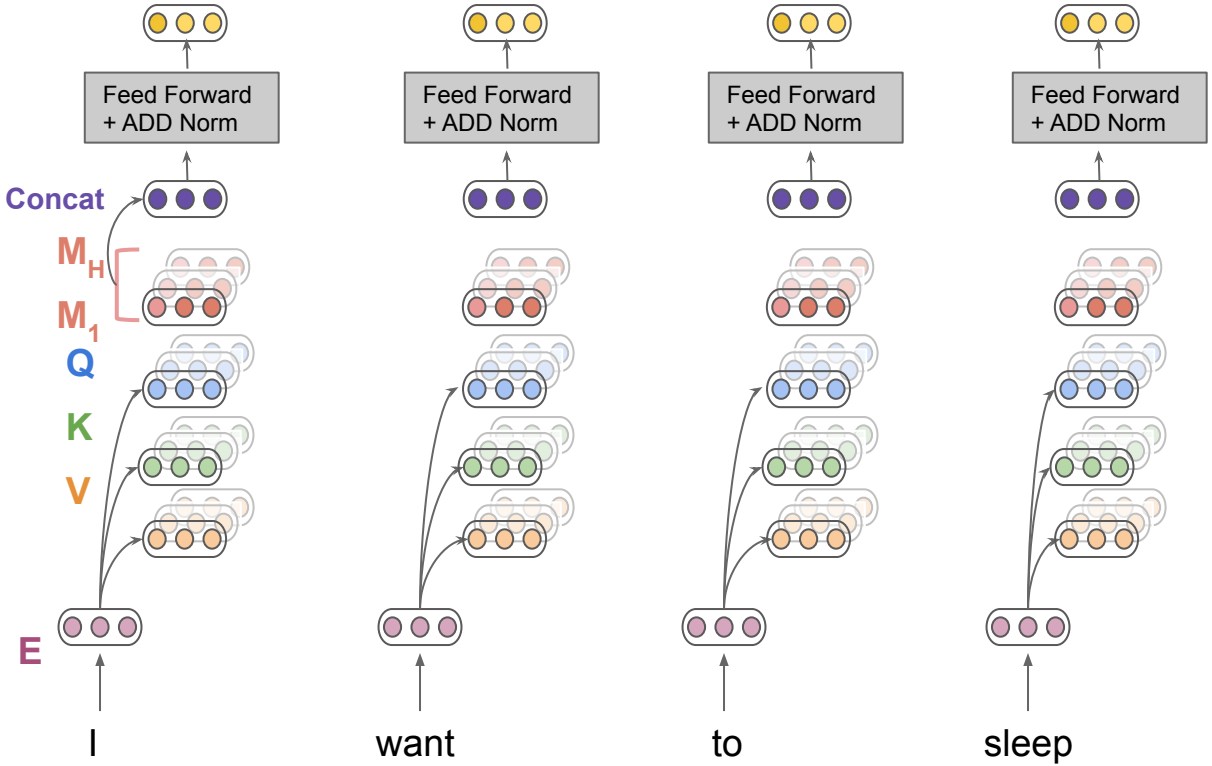
H (no: of heads) Different versions of Q,K,V
Each different repr -> Different attended repr



Multi-Headed Self Attention

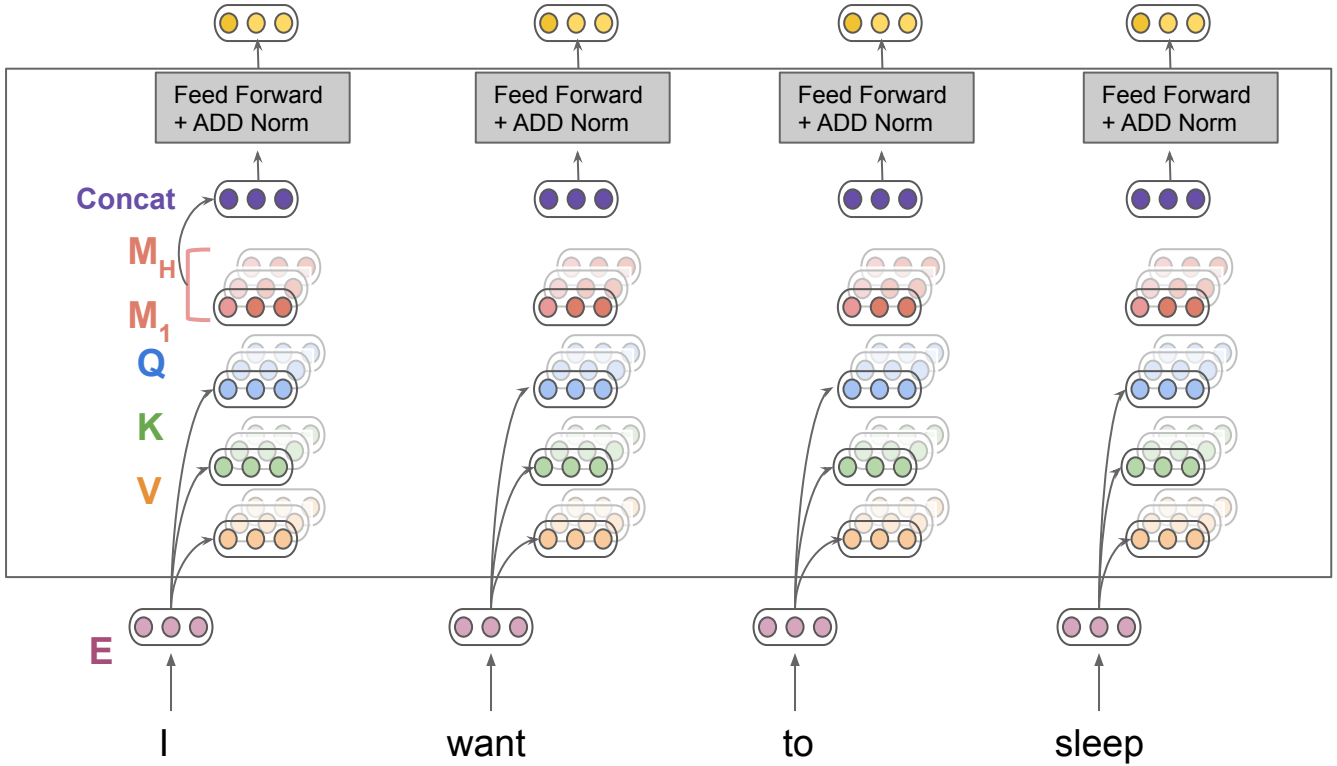


Multi-Headed Self Attention

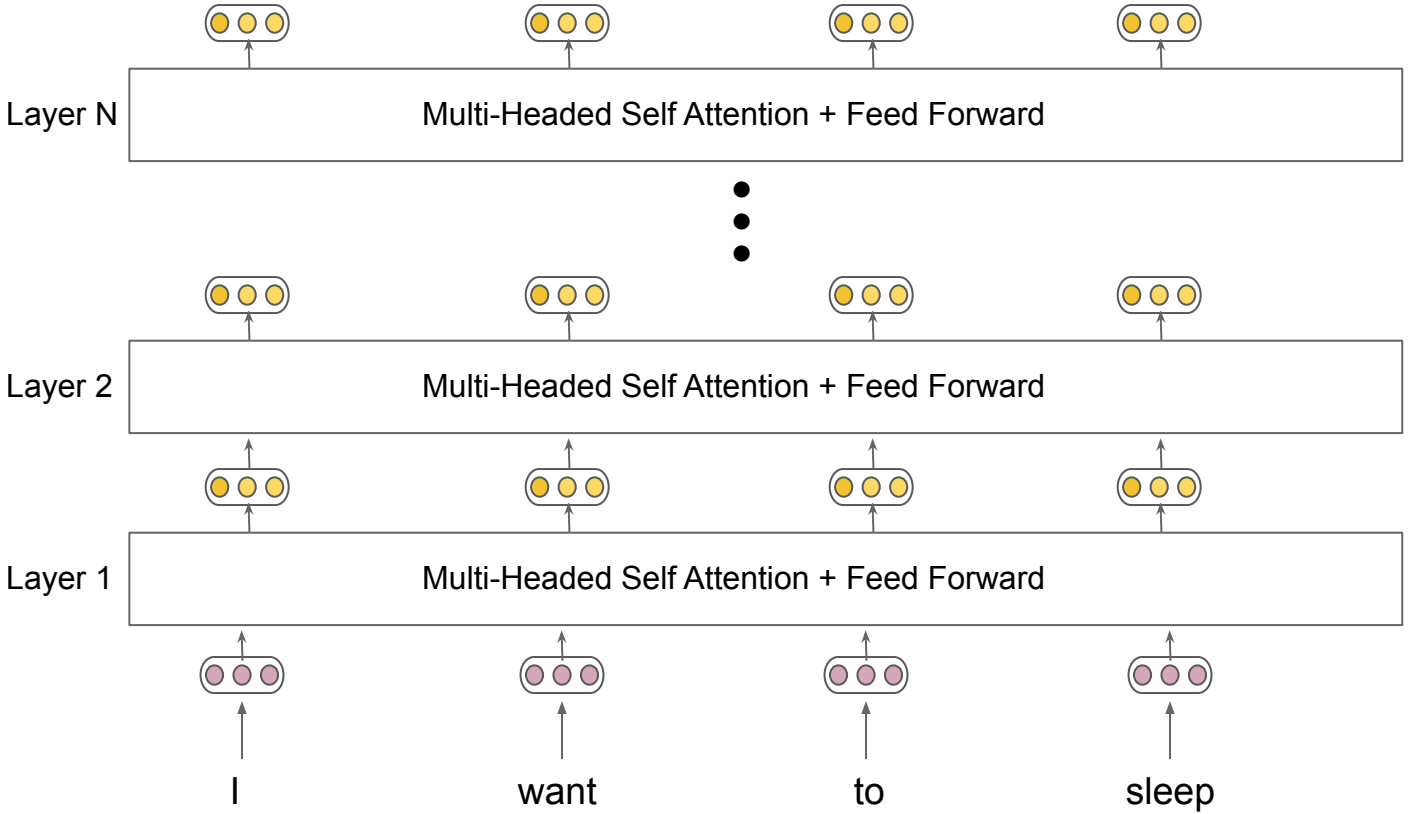


Multi-Headed Self Attention

Multi-Headed Self Attention + Feed Forward

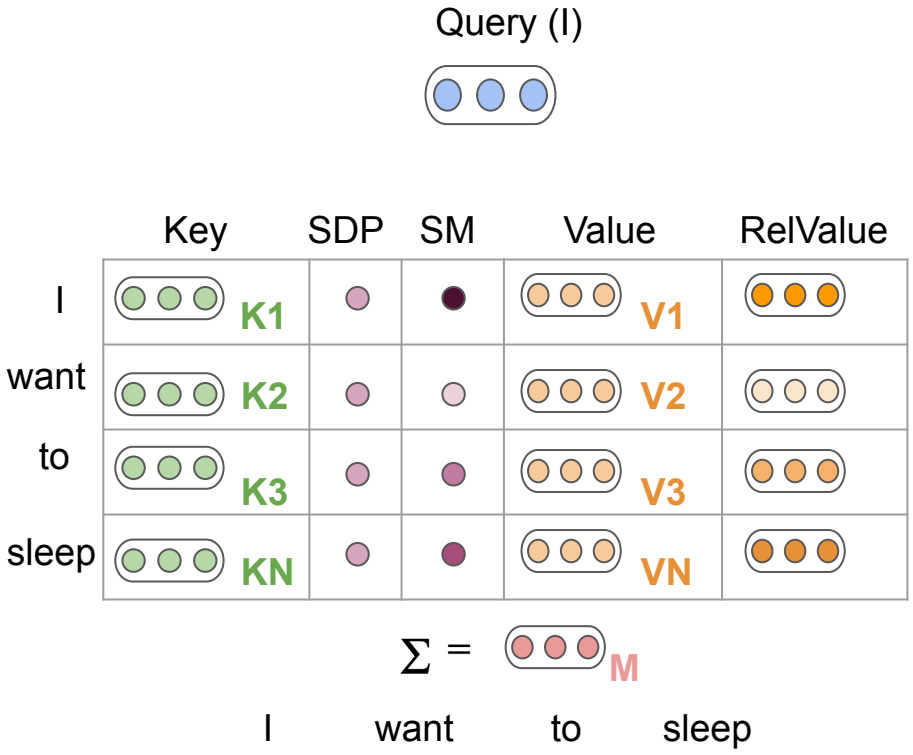


Multi-Headed Self Attention

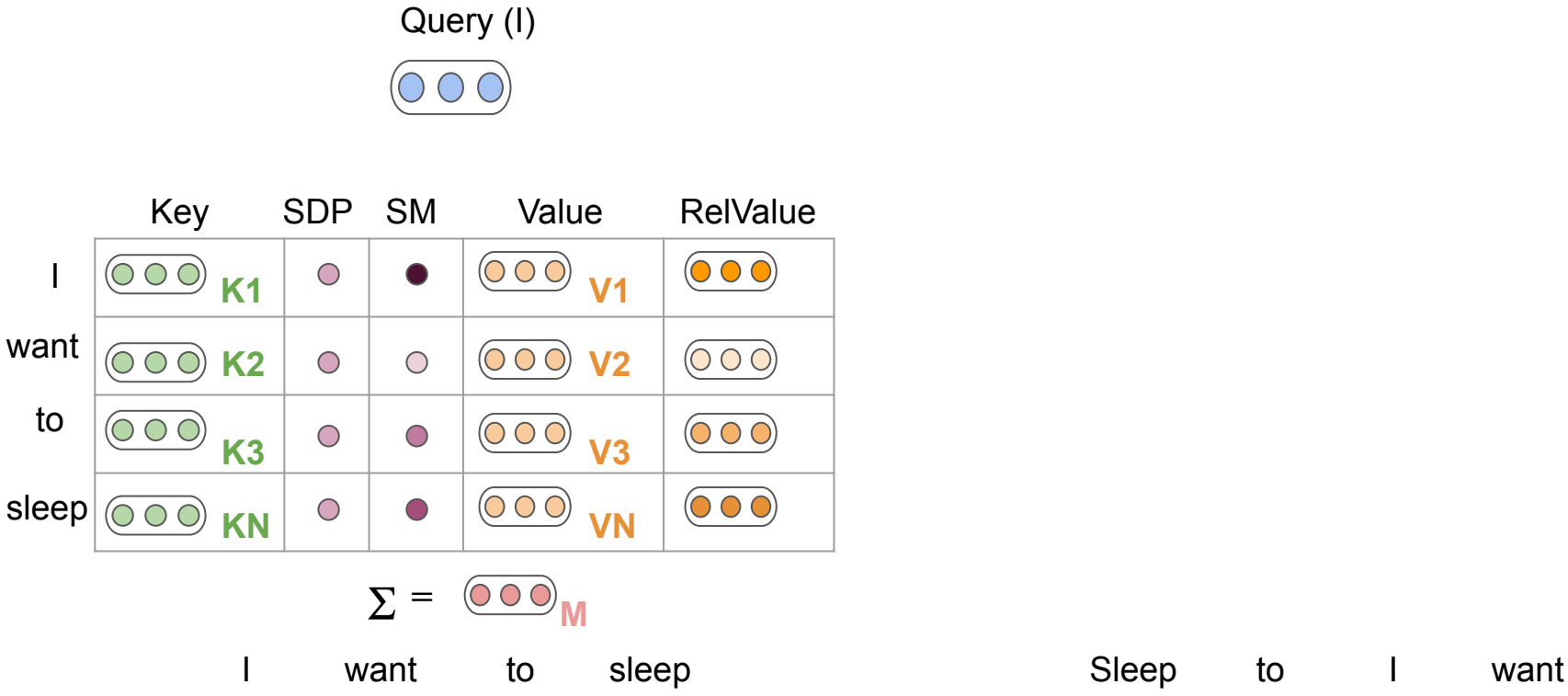


Questions?

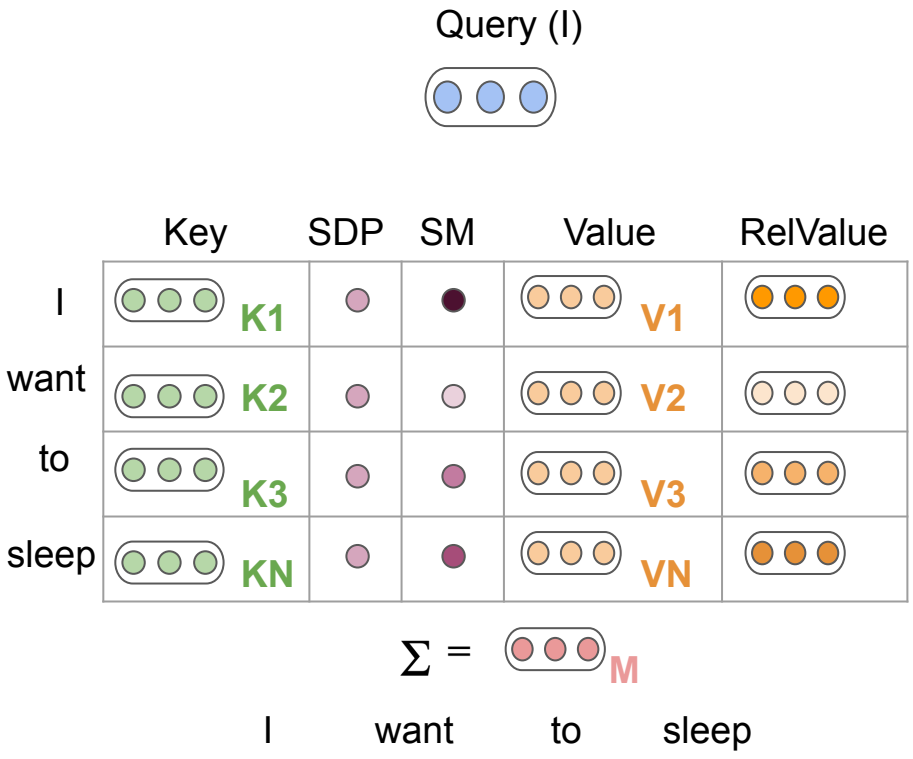
Revisiting Self Attention

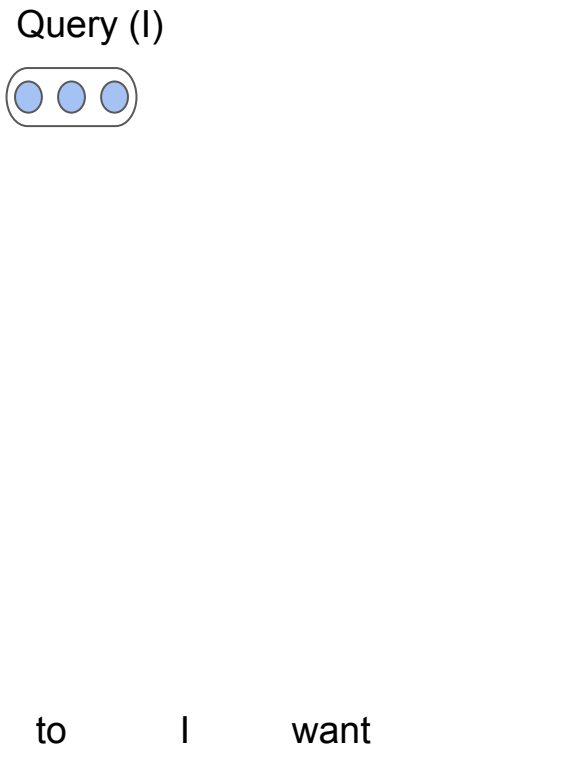


Revisiting Self Attention

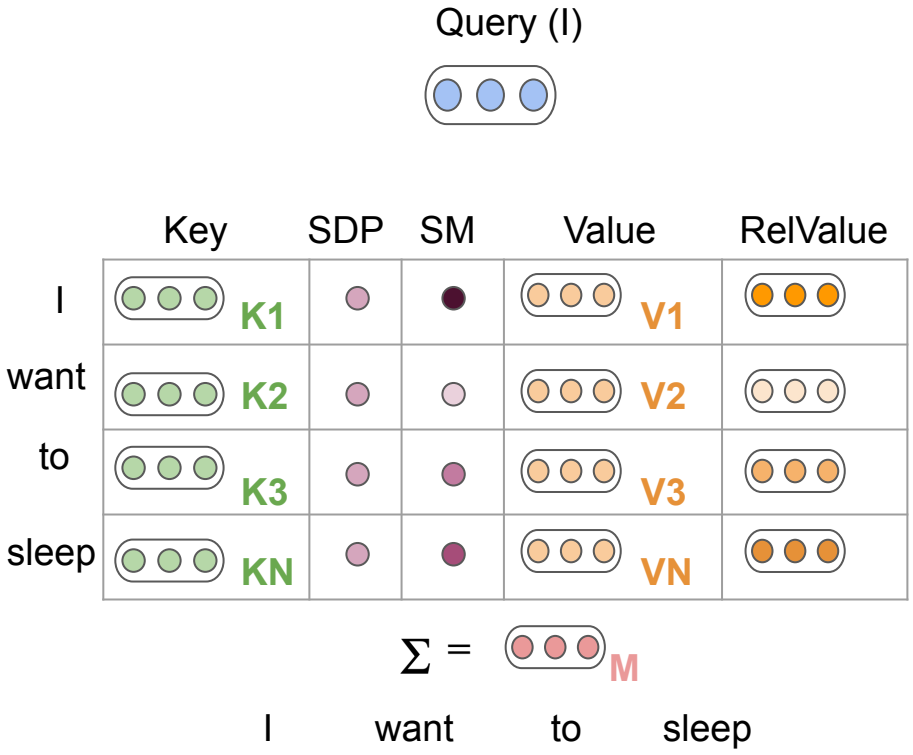


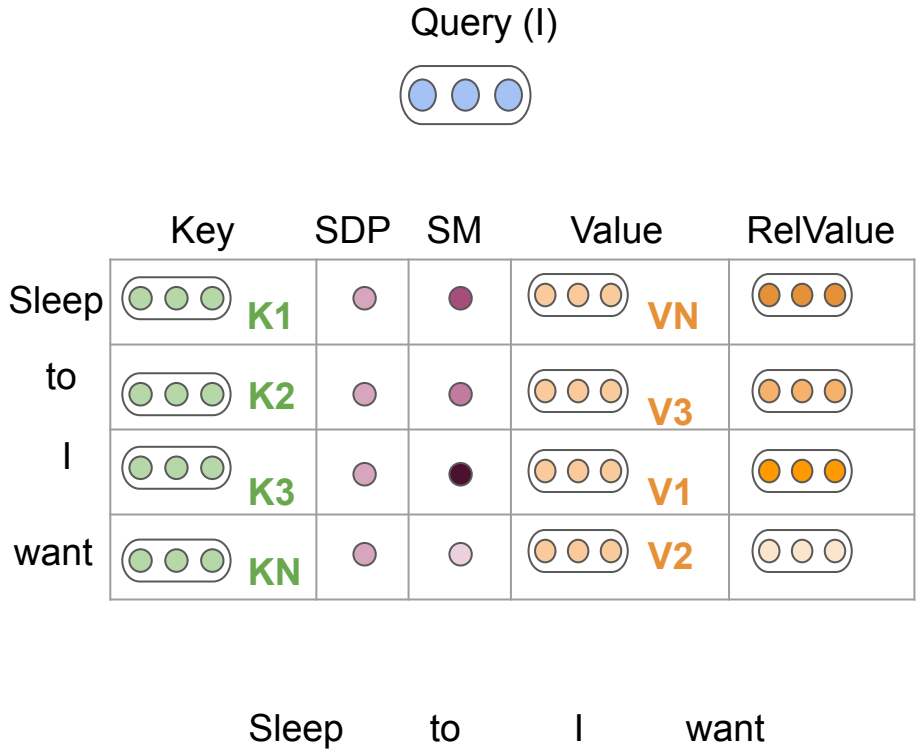
Revisiting Self Attention



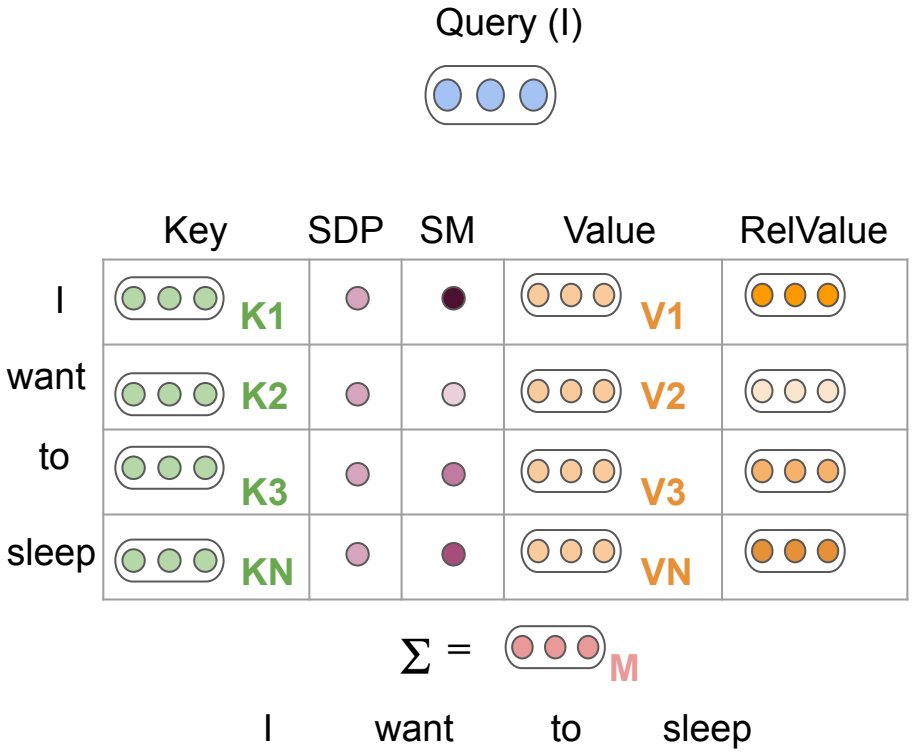


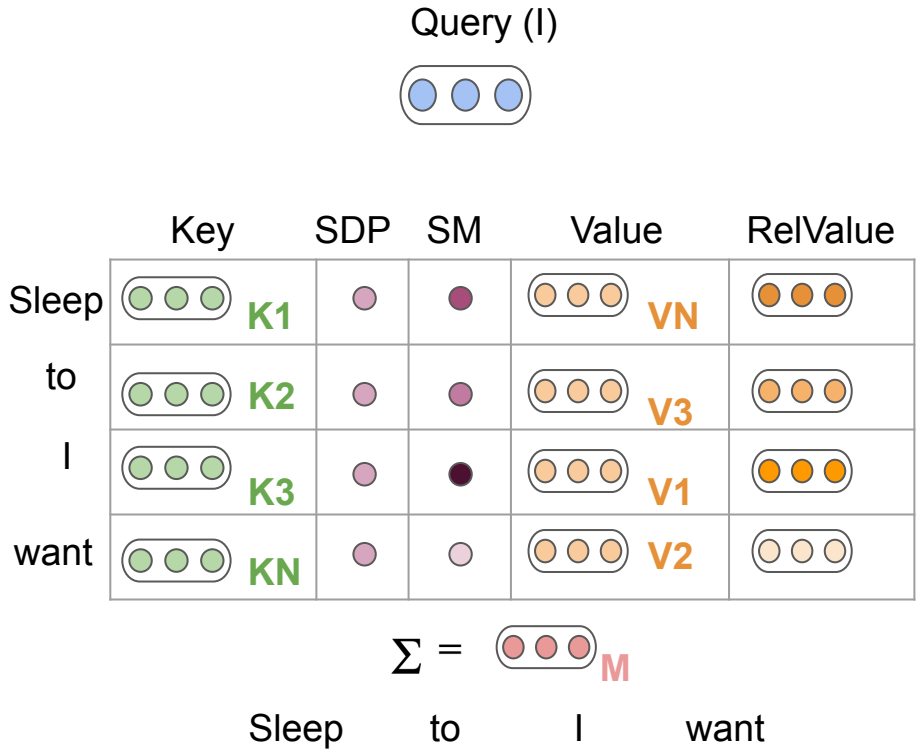
Revisiting Self Attention



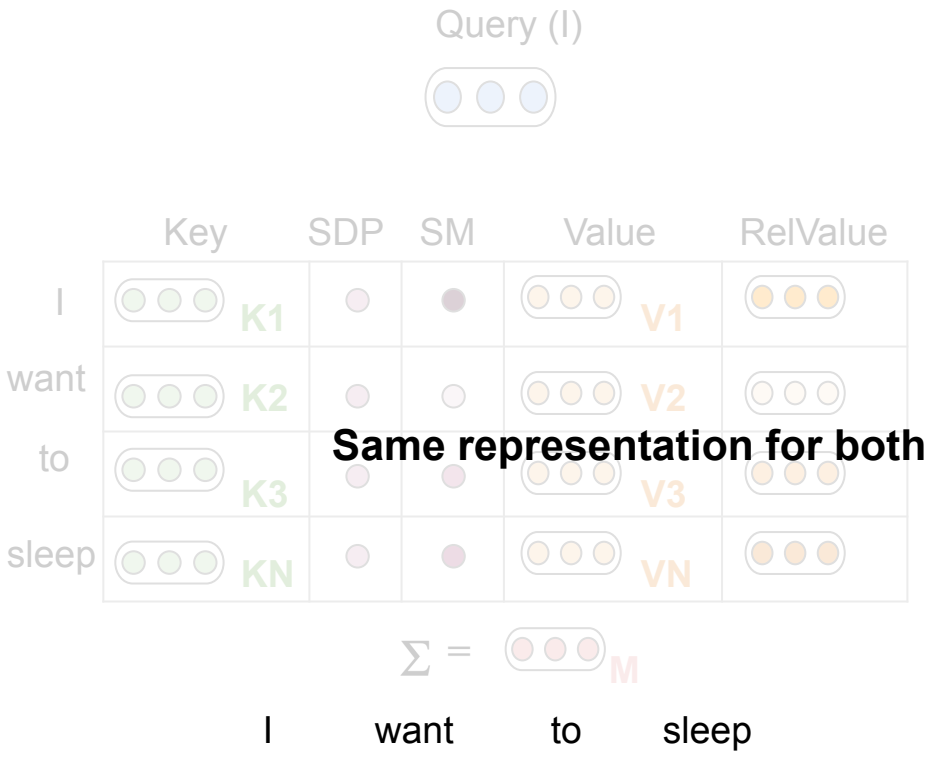


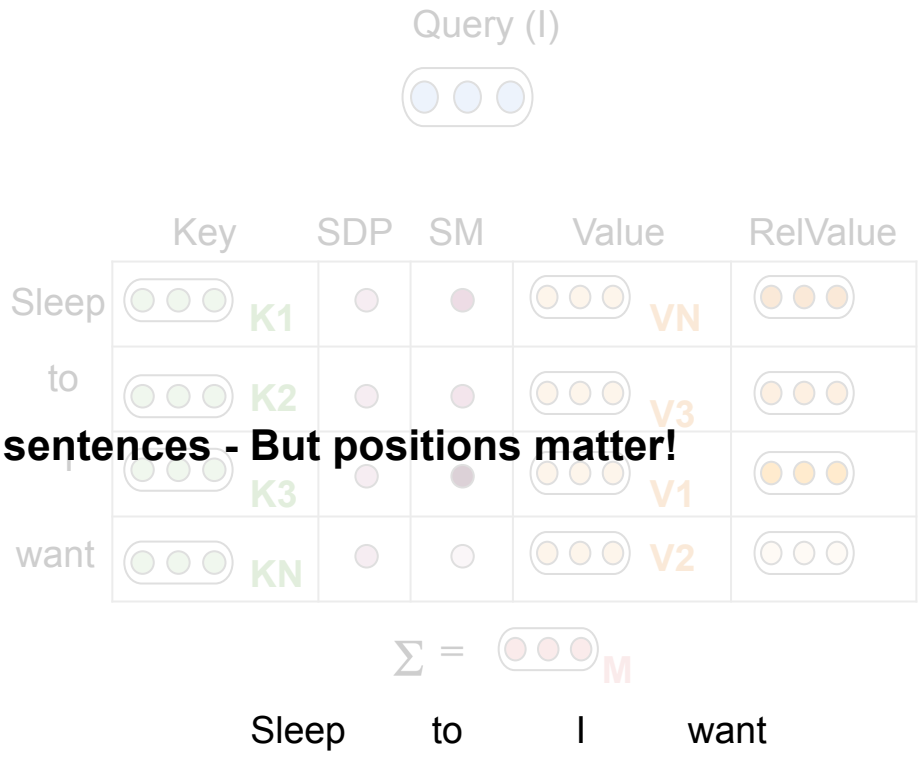
Revisiting Self Attention





Revisiting Self Attention

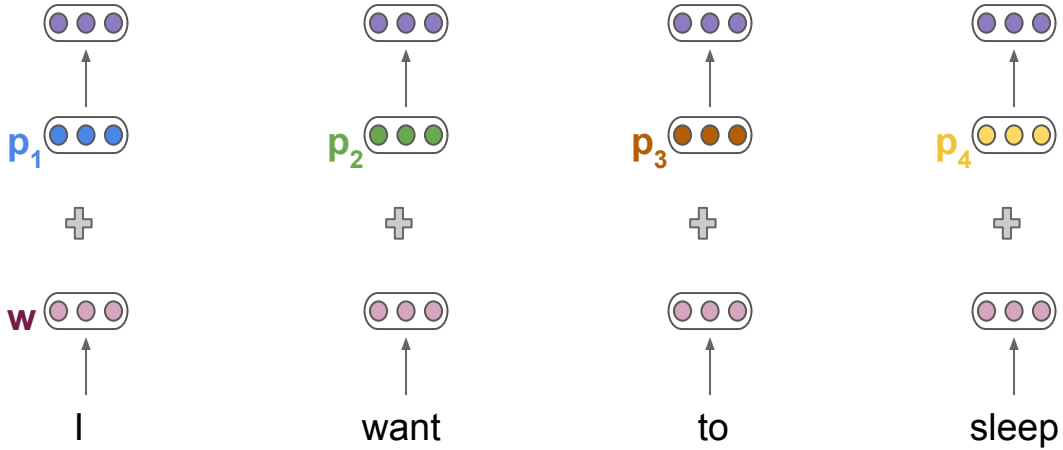




Same representation for both sentences - But positions matter!

Positional Encoding

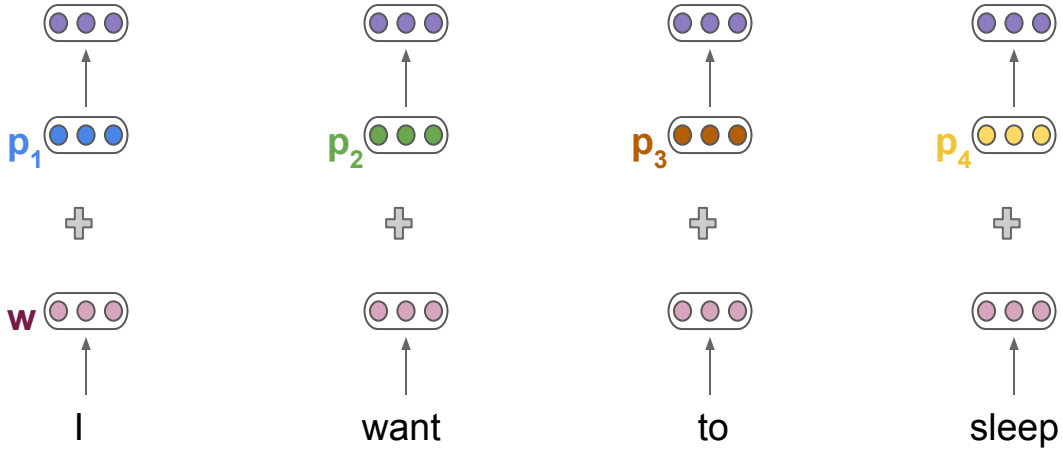
Position embeddings - each position number has an associated embedding



Positional Encoding

Sinusoidal Position embeddings - generalize to any sequence length

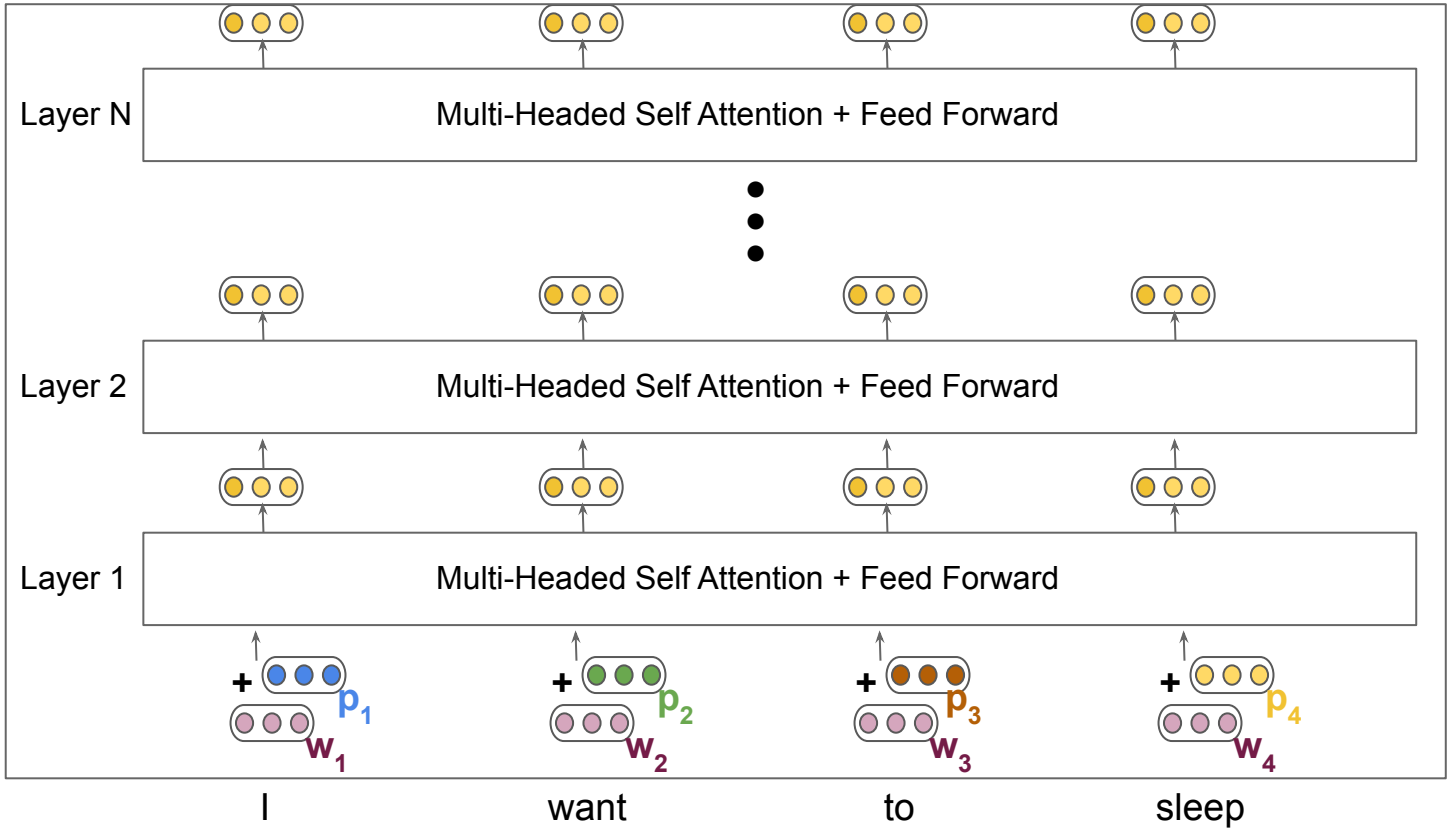
$$p = f(i, t)$$



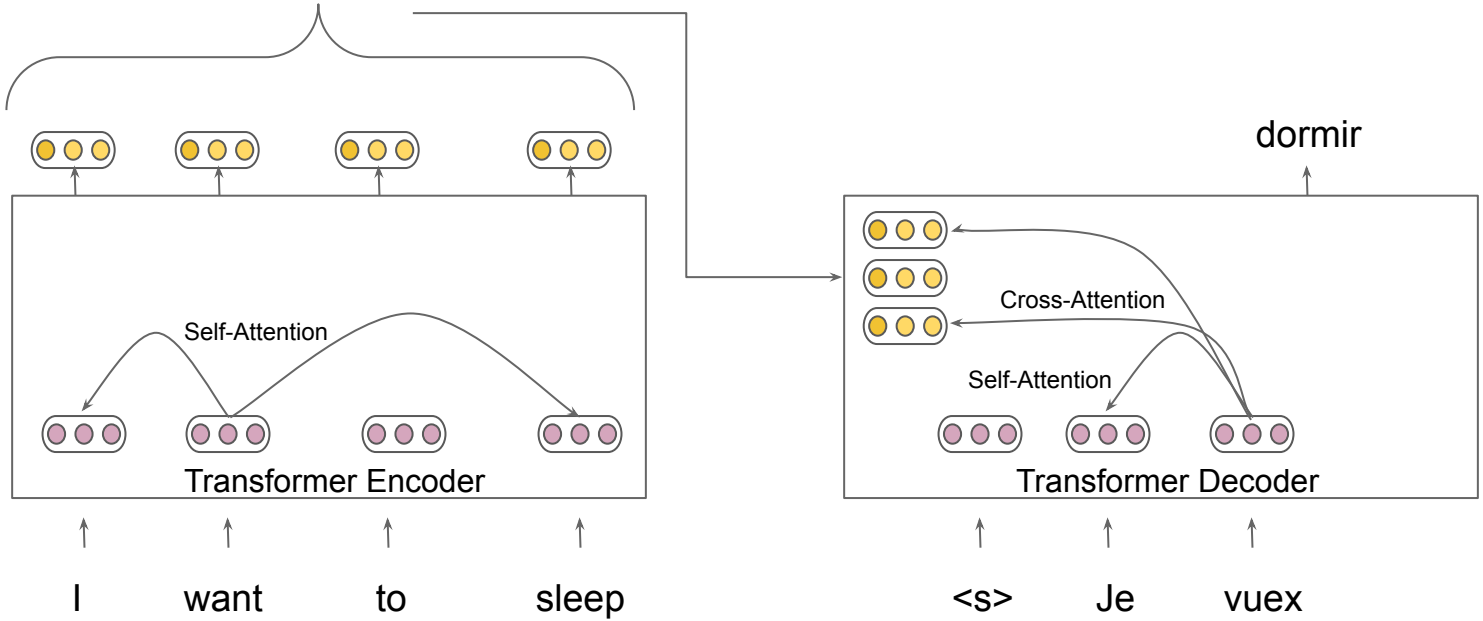
Questions?

Transformer Encoder

N-Layer Transformer Encoder



Transformer Encoder - Decoder

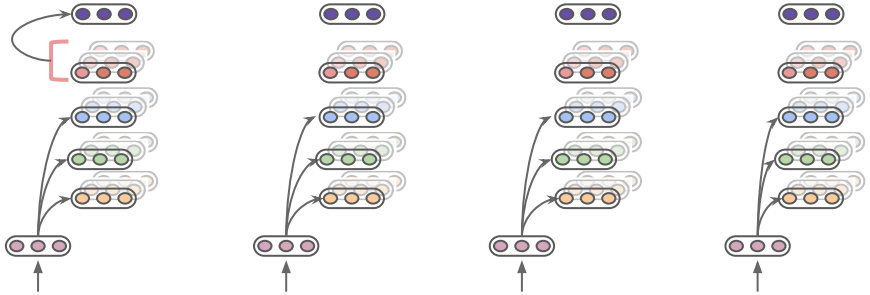


What's so great about Transformers?

- Parallelizable computation
  - Entire sequence, All queries, all attention heads computed in parallel
  - Benefits from fast matrix multiplication on GPUs
- Rich expressive power
  - Every token connected to every other token
  - Can form long range dependencies
- Depth not proportional to seq length
  - Reduces exploding/vanishing gradient problem
  - Converges faster

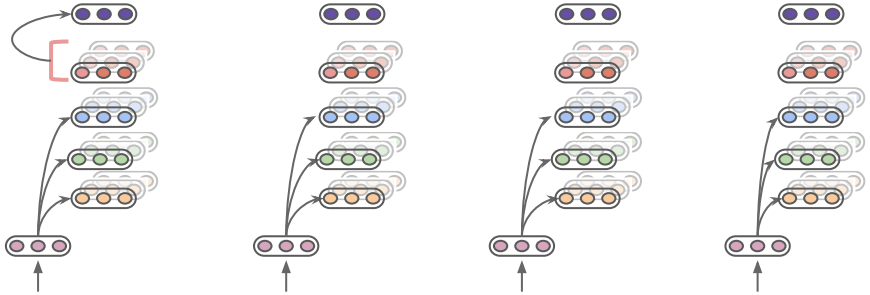
What's so great about Transformers?

- Parallelizable computation - Entire sequence can be processed in parallel

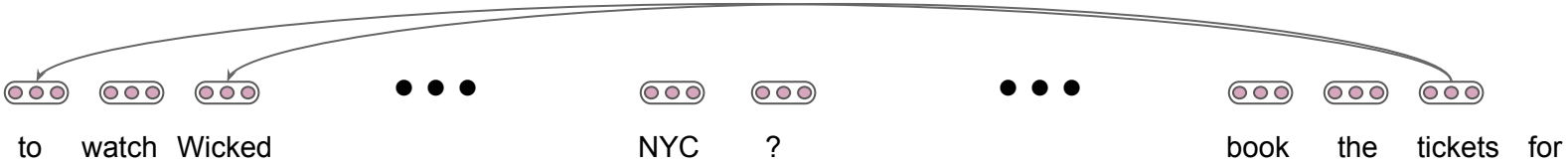


What's so great about Transformers?

- Parallelizable computation - Entire sequence can be processed in parallel

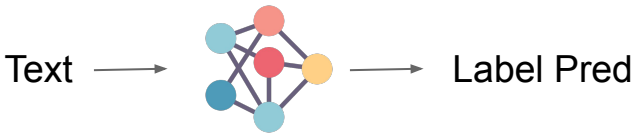


- Rich expressive power - long range dependencies

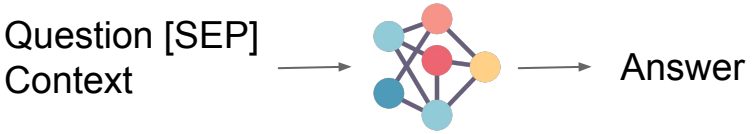


Impact - Wide Applications!

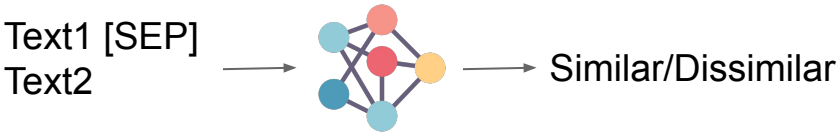
Classification



Question Answering



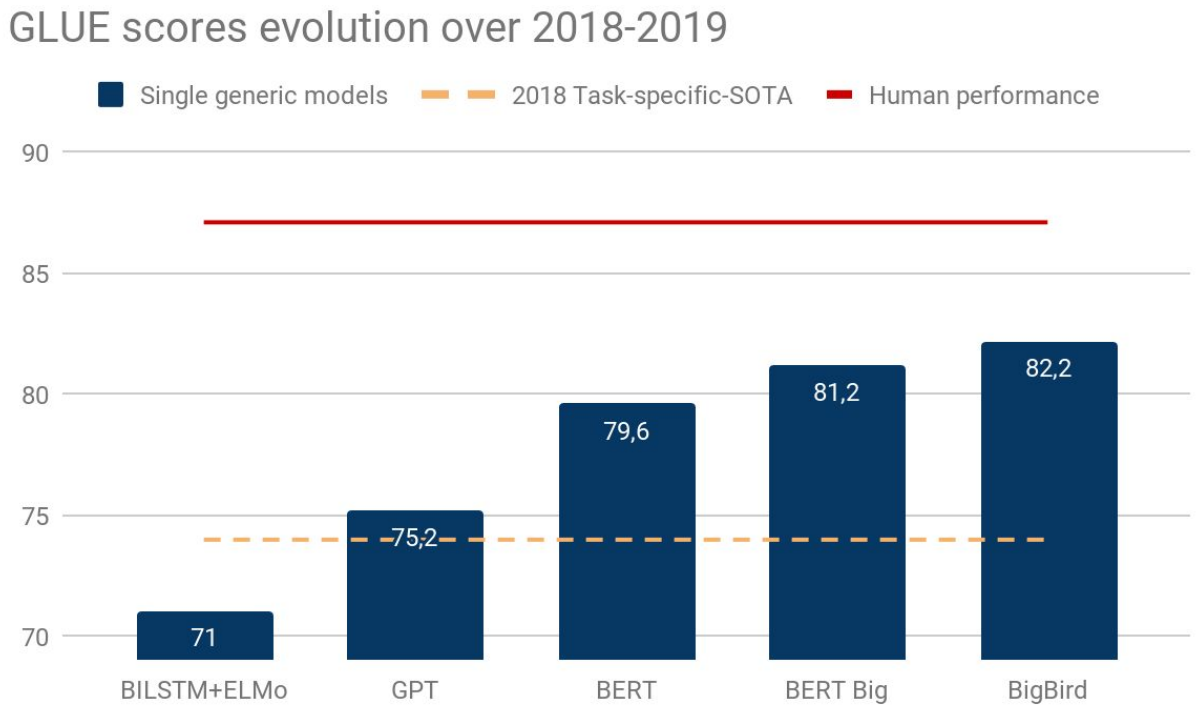
Sentence Similarity



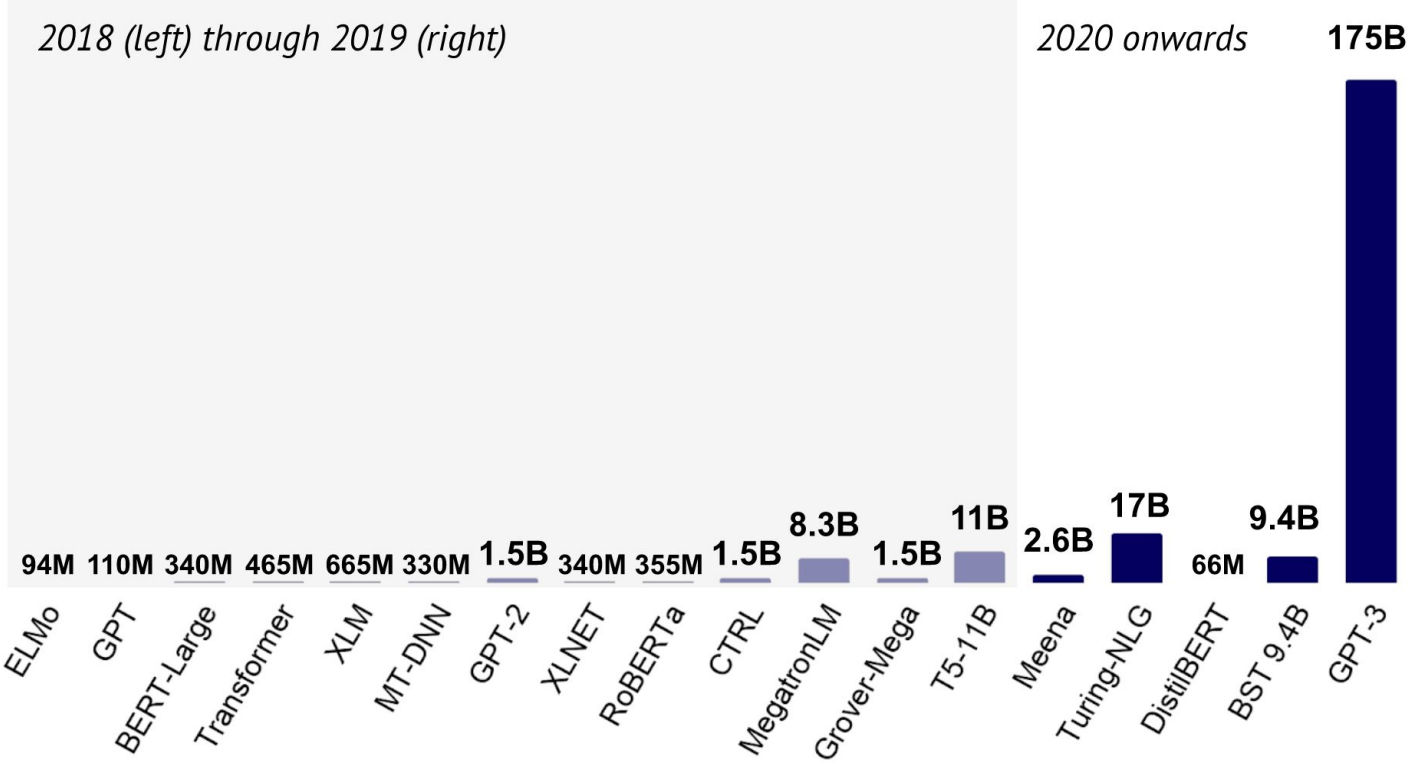
Translation

Larger Impact

Larger Impact



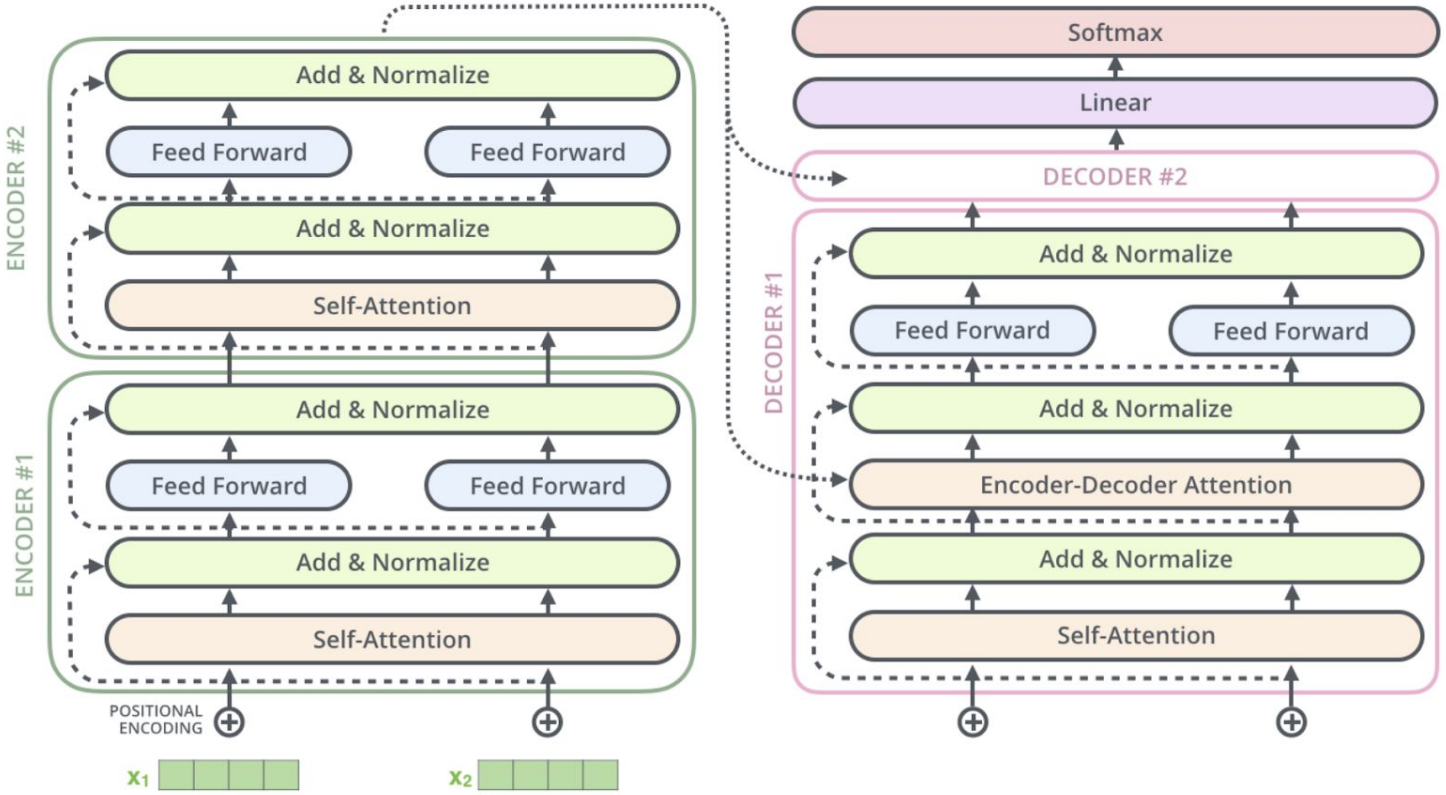
Larger Impact



Thank you!

vbalacha@cs.cmu.edu

Transformer Encoder-Decoder



Results/Impact

- Improves results, Establishes SOTA in various tasks!
  - Machine Translation
  - Constituency Parsing
  - Language Modeling
  - and more!

- Computationally faster!
  - No sequential computation - Entire sequence processed in parallel