Neural Networks and Deep Learning

Transformer II

Shumin Wu

Department of Computer Science

shumin.wu@colorado.edu

March 2, 2020

Previous Lecture: Memory Network for Q/A

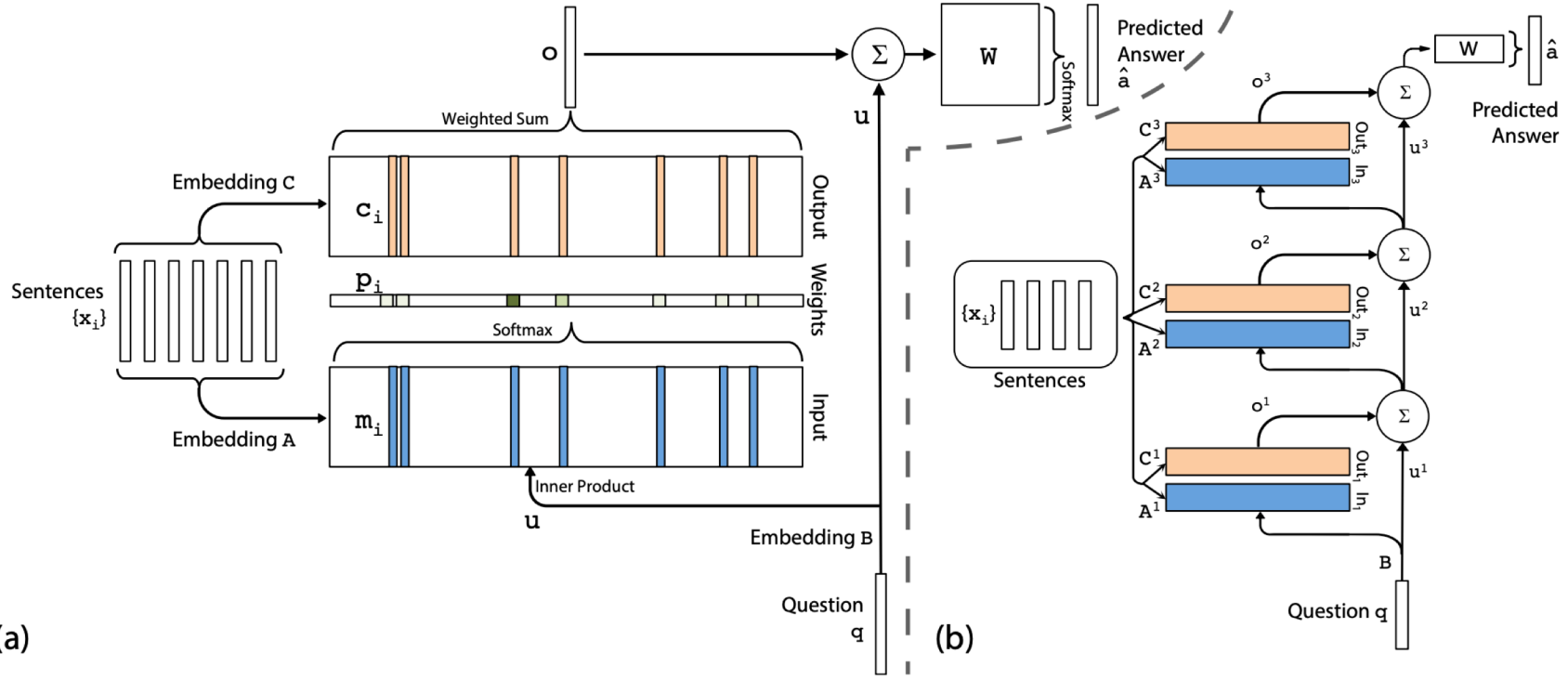
Sam walks into the kitchen.
Sam picks up an apple.
Sam walks into the bedroom.
Sam drops the apple.
Q: Where is the apple?
A. Bedroom

Brian is a lion.
Julius is a lion.
Julius is white.
Bernhard is green.
Q: What color is Brian?
A. White

Mary journeyed to the den.
Mary went back to the kitchen.
John journeyed to the bedroom.
Mary discarded the milk.
Q: Where was the milk before the den?
A. Hallway

Separate query/key (question/input) and value (output) space:

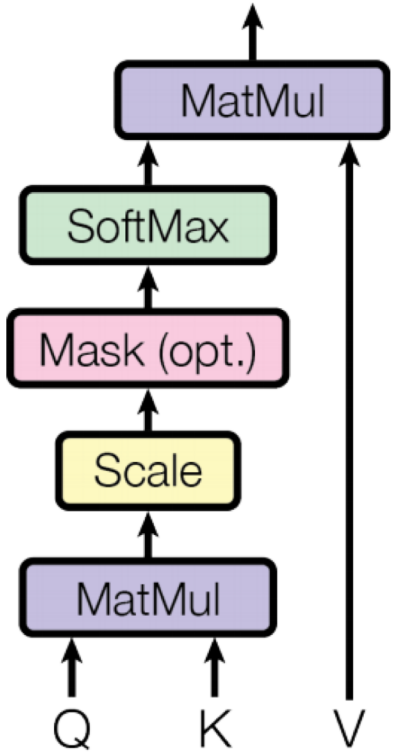
- query/key space may encode information types like location & color.
- value space may encode actual location & color values (like kitchen, bedroom, white, red, etc).



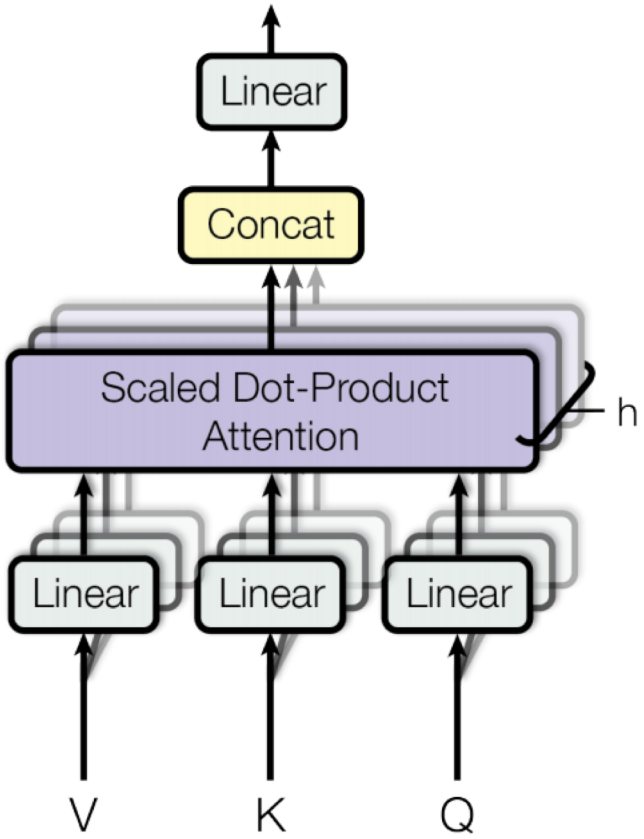
source: Sukhbaatar et al., 2015

Previous Lecture: Attention is All You Need

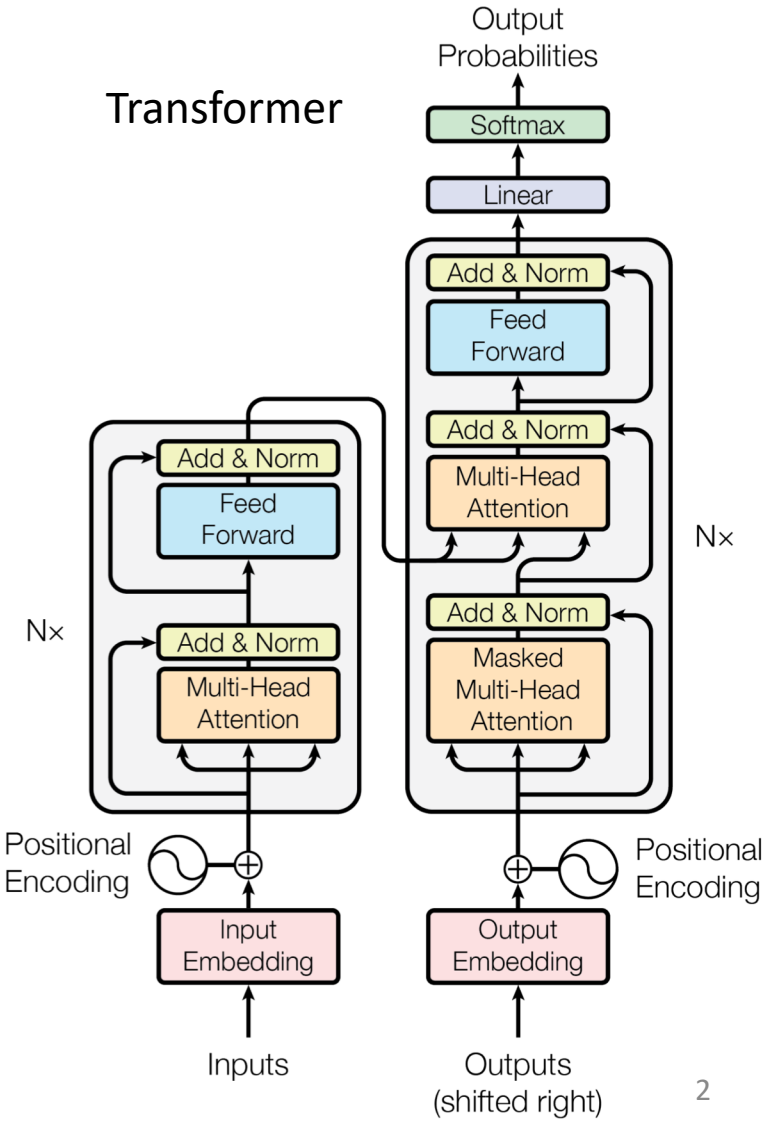
Scaled Dot-Product Attention



Multi-Head Attention



Transformer

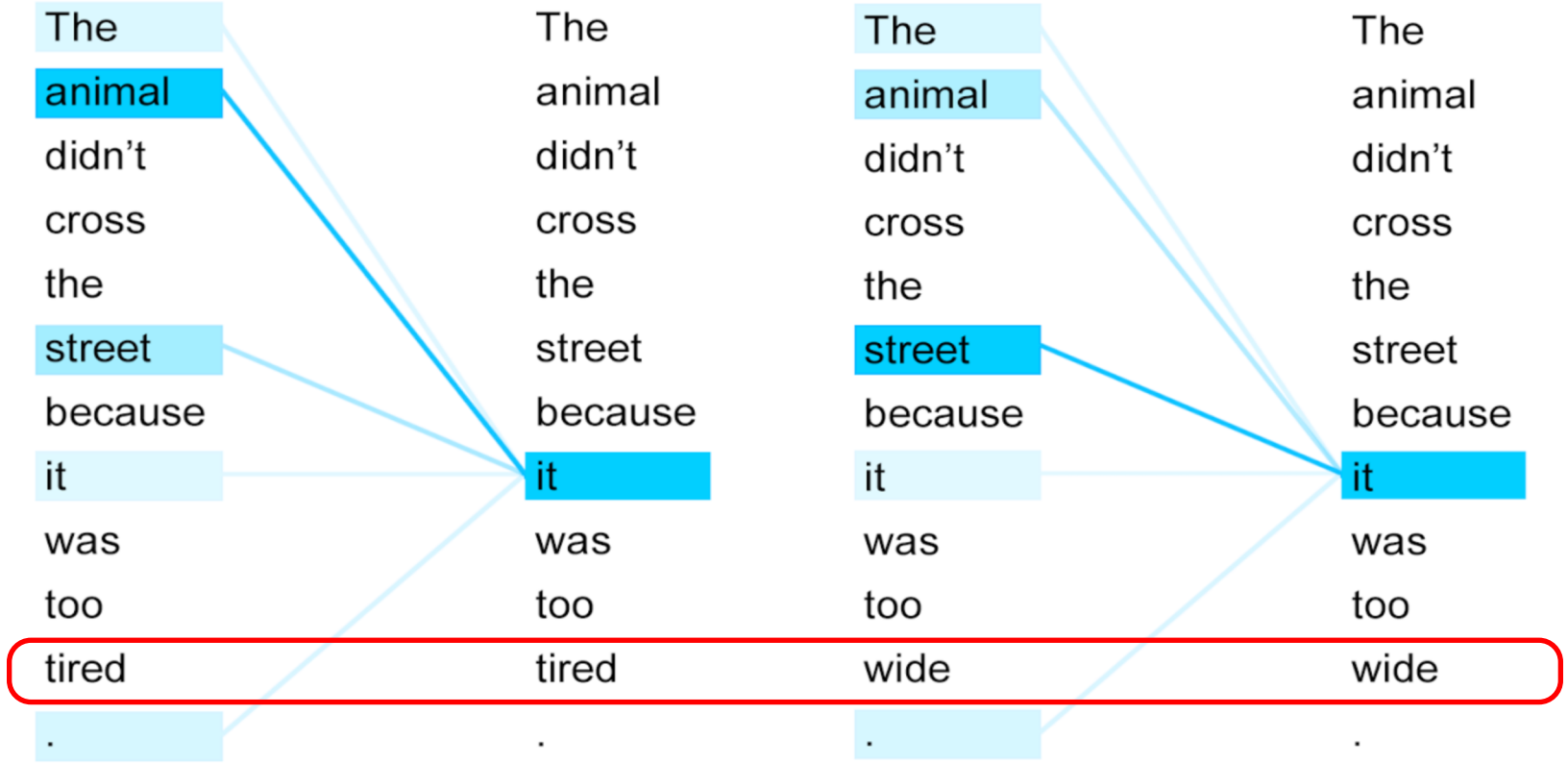


Source: [Vaswani et al, 2017]

Attention Visualization: Anaphora Resolution

Likely difficult for single head & single self-attention block:

- pronoun/noun type agreement
- subject/object compatibility



source: [Google Blog]

T2T Code Lab (Joe Bruce)

Is Attention Really “All” You Need?

What about tasks tracking states that can be expressed with simple iterative/recursive transformation:

- Parity bit computation: $h_i = \mathrm{XOR}(h_{i-1}, x_i)$
- Integer decimal digit sequence to value: $h_i = 10 \times h_{i-1} + x_i$

bAbi question answering (20 tasks):

Story:

John went to the hallway.
John went back to the bathroom.
John grabbed the milk there.
Sandra went back to the office.
Sandra journeyed to the kitchen.
Sandra got the apple there.
Sandra dropped the apple there.
John dropped the milk.

Question:

Where is the milk?

Answer:

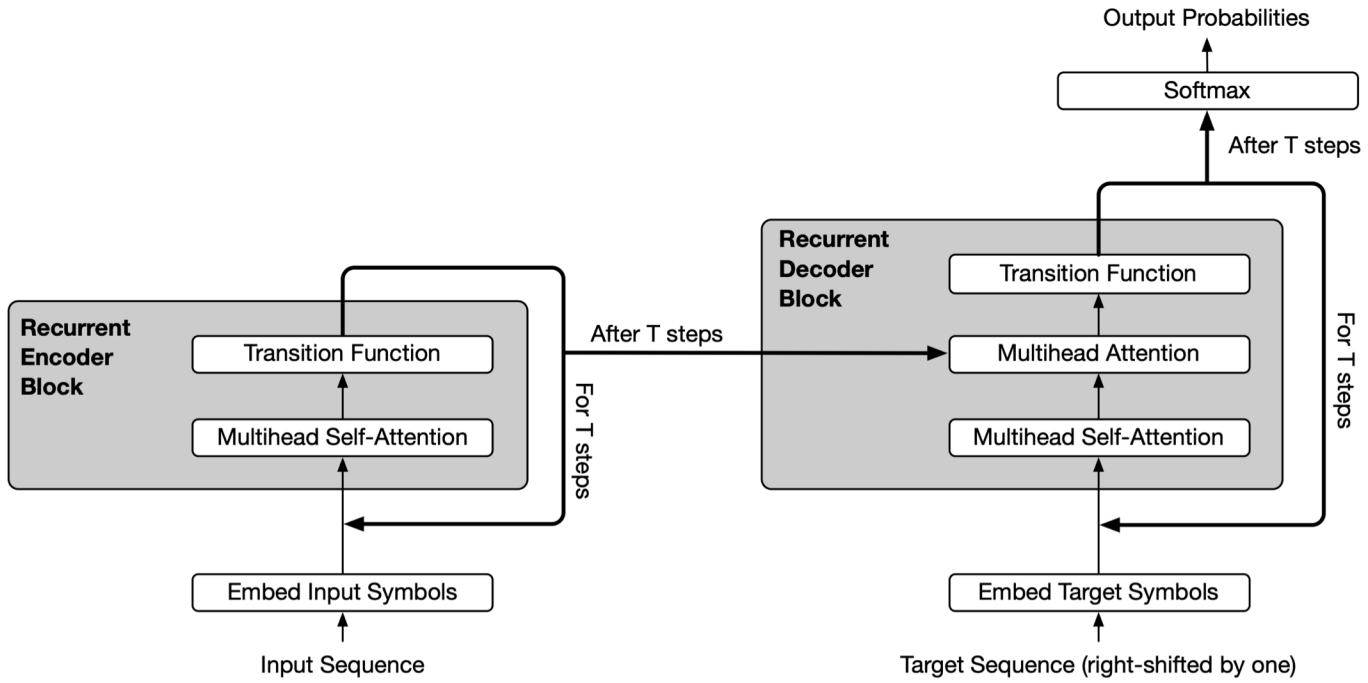
bathroom

Universal Transformer

- Add recurrence to transformer
  - Share parameters between all encoder blocks and between all decoder blocks
- Dynamic halting
  - Some symbols are more ambiguous than others and need more iterations
  - Learn marginal halting probability function for input states:

$$p_s = \sigma(W^s s)$$

$$p(\text{halting}_{s_i}) = p(\text{halting}_{s_{i-1}}) + p_{s_{i-1}}$$



source: [Dehghani et al., 2019]

Universal Transformer bABi Results

Model	10K examples		1K examples	
	train single	train joint	train single	train joint
Previous best results:				
QRNet (Seo et al., 2016)	0.3 (0/20)	-	-	-
Sparse DNC (Rae et al., 2016)	-	2.9 (1/20)	-	-
GA+MAGE Dhingra et al. (2017)	-	-	8.7 (5/20)	-
MemN2N Sukhbaatar et al. (2015)	-	-	-	12.4 (11/20)
Our Results:				
Transformer (Vaswani et al., 2017)	15.2 (10/20)	22.1 (12/20)	21.8 (5/20)	26.8 (14/20)
Universal Transformer (this work)	0.23 (0/20)	0.47 (0/20)	5.31 (5/20)	8.50 (8/20)
UT w/ dynamic halting (this work)	**0.21 (0/20)**	**0.29 (0/20)**	**4.55 (3/20)**	**7.78 (5/20)**

train single: 20 tasks trained individually
train joint: all tasks jointly trained

source: Dehghani et al., 2019

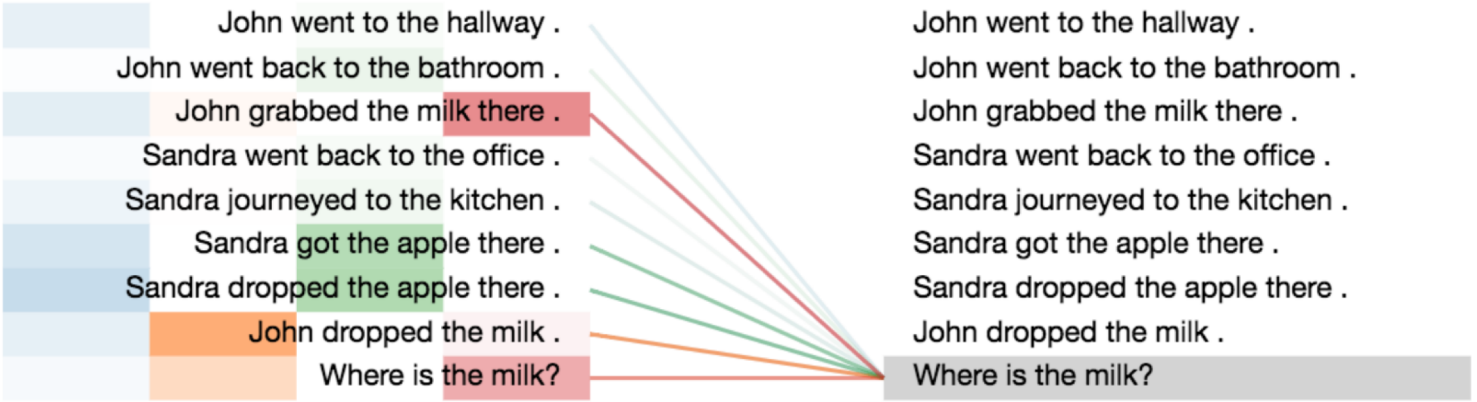
Universal Transformer Attention Visualization

Question:

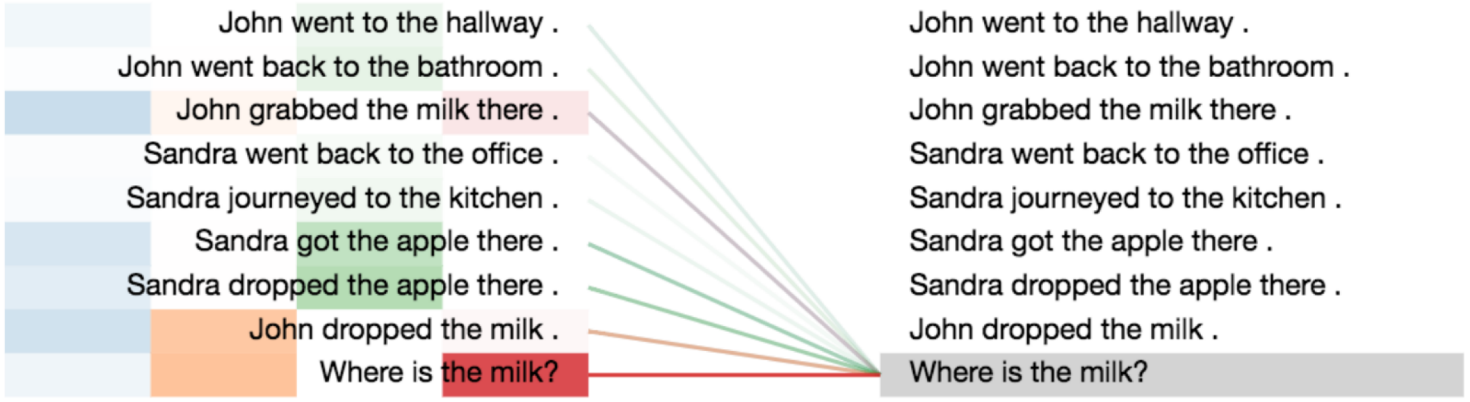
Where is the milk?

Answer:

bathroom



(a) Step 1



(b) Step 2

source: [Dehghani et al., 2019]

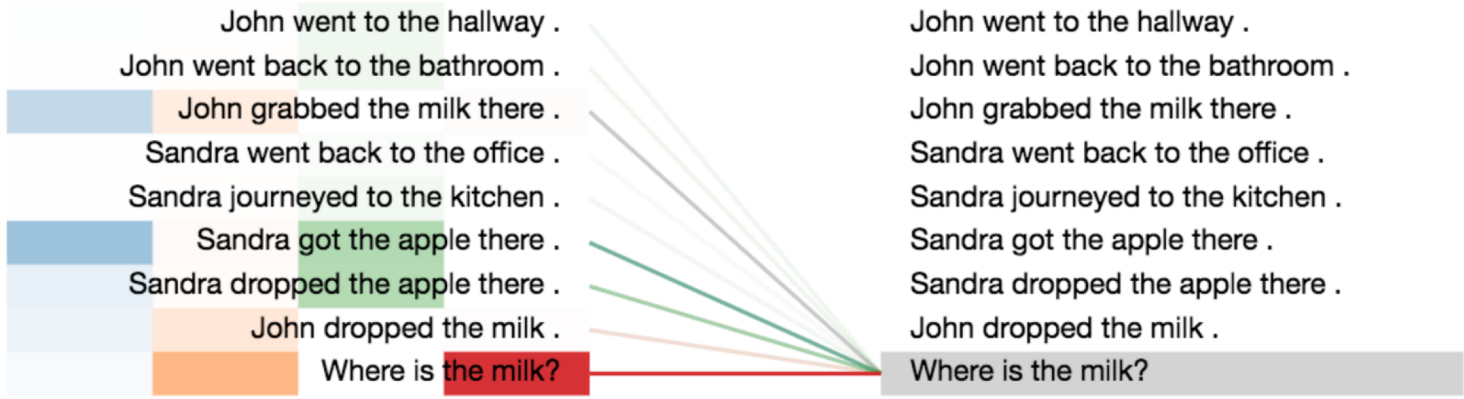
Universal Transformer Attention Visualization

Question:

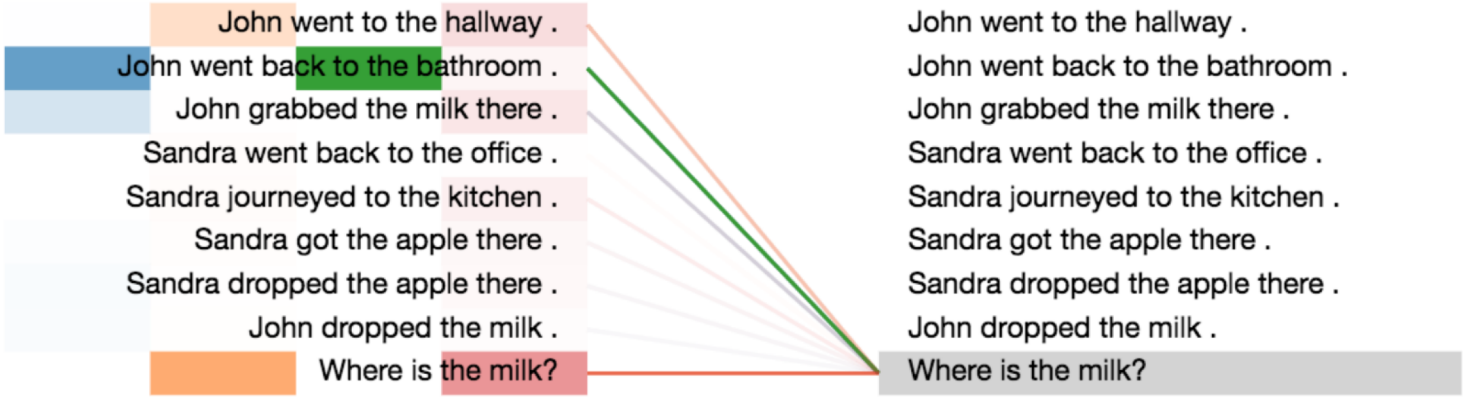
Where is the milk?

Answer:

bathroom



(c) Step 3



(d) Step 4

source: [Dehghani et al., 2019]

Transformer Complexity

Recall:

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

n: token length d: embedding dimension

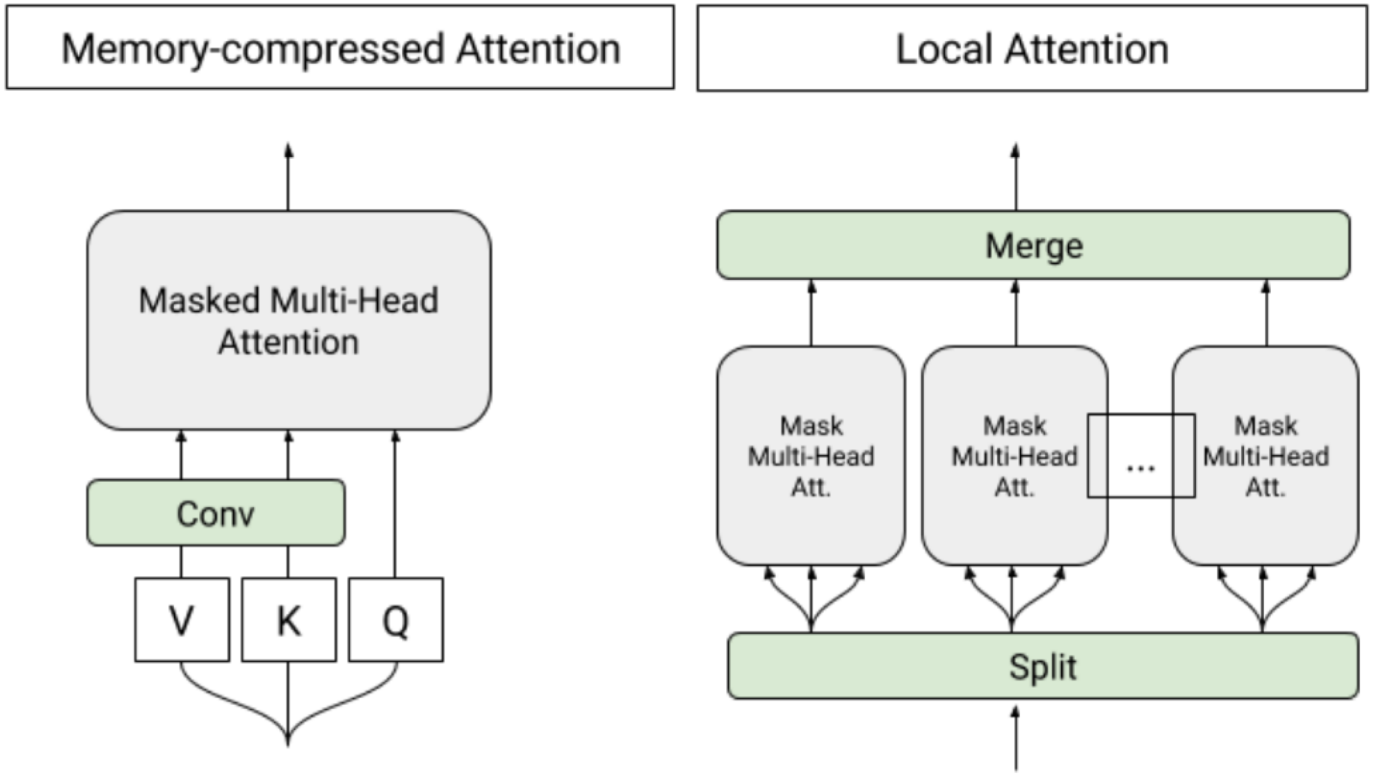
source: Vaswani et al, 2017

Transformer is more efficient than RNN when $n \ll d$

But what about when n is very large?

Long-Sequence Attention: Generating Wikipedia

- Extract paragraphs from articles
- Decoder-only seeded w/ extracted paragraphs.
- Memory-compressed Attention
  - CNN w/ 1x3 kernel, 3 stride on K, V
  - Q vector not compressed (allows token-specific query)
- Local Attention
  - Divide input into chunks of 256 tokens and restrict attention to within each chunk.
- Alternating local & memory-compressed attention blocks (5 total)



source: [Liu et al., 2018]

Generating Wikipedia Output

OUTPUT:
== Wings Over Kansas

==wings over kansas is the best aviation history website i have encountered on the world wide web . it is informative , entertaining , provides ever changing content , and is populated with the true voices of the mainstream aviation community . there is no better place to see where aviation has been and where it is going . centered in the midst of the greatest producer of aircraft in the world , wings over kansas reflects that aviation community to the rest of the world .

== Wings Over Kansas History

==wings over kansas was established with the mission of becoming the number one online kansas aviation resource showcasing the pioneers , educators , newsmakers , manufacturers , pilots and craftsmen , who have made kansas the world center for aviation production . wings over kansas was established with the mission of becoming the number one online kansas aviation resource showcasing the pioneers , educators , newsmakers , manufacturers , pilots and craftsmen , who have made kansas the world center for aviation production . wings over kansas has been recognized by mcgraw - hill as one of the top 500 best aviation web sites with visitors from over 225 countries . wings over

TARGET:
== Wings Over Kansas

==wings over kansas.com is an aviation website founded in 1998 by carl chance owned by chance communications , inc. to provide information and entertainment to aviation enthusiasts and professionals worldwide . the web site is based in wichita , kansas , known as the `` air capital of the world '' due to the many aircraft manufacturers located there . in 2003 , the site was upgraded to a data - based web site to better serve the needs of its members . `` wings over kansas '' has grown steadily and as of 2009 draws over a quarter of a million visitors yearly from over 125 countries .

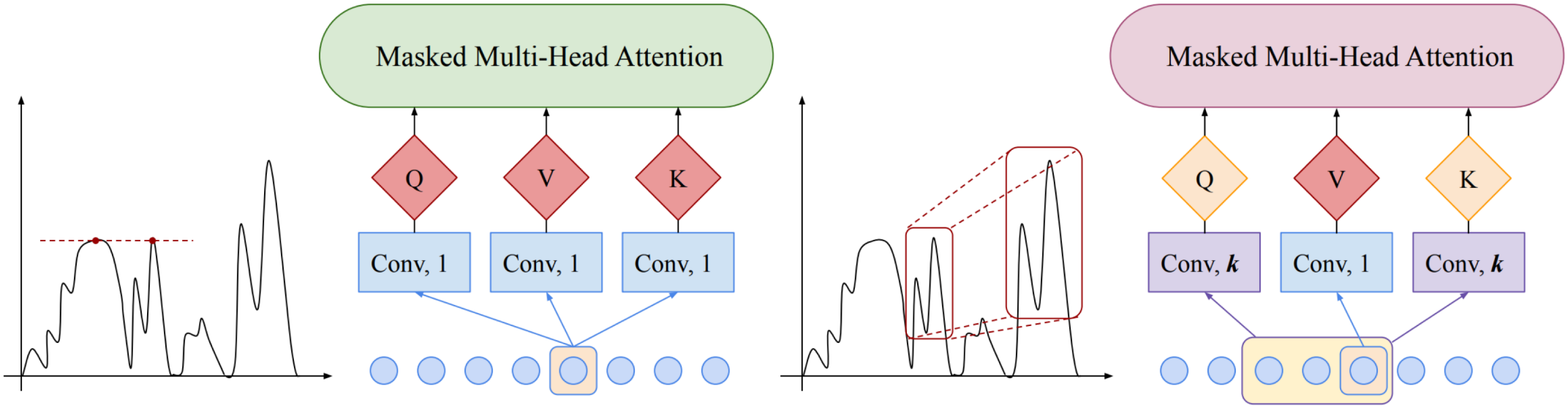
== Wings Over Kansas History

==wings over kansas.com was created in 1998 by wichita native carl chance , a broadcast professional and producer for the wingspan air & space channel . in his more than thirty years of experience , chance developed many relationships in the aviation community that have directly benefited the web site . he is a charter member and past trustee on the kansas aviation museum board of directors and a former member of the kansas aviation council . from 1998 to 2003 , the site underwent a number of modifications to improve its value and navigation .

source: Liu et al., 2018

Time Series: Attention Locality

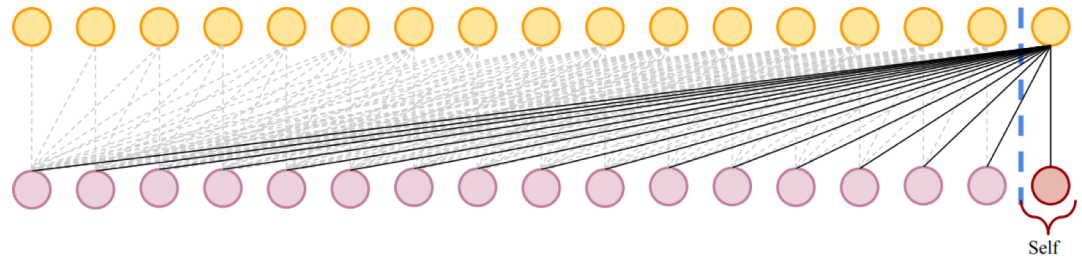
Convolution of Q, K for attention distribution that has stronger locality awareness



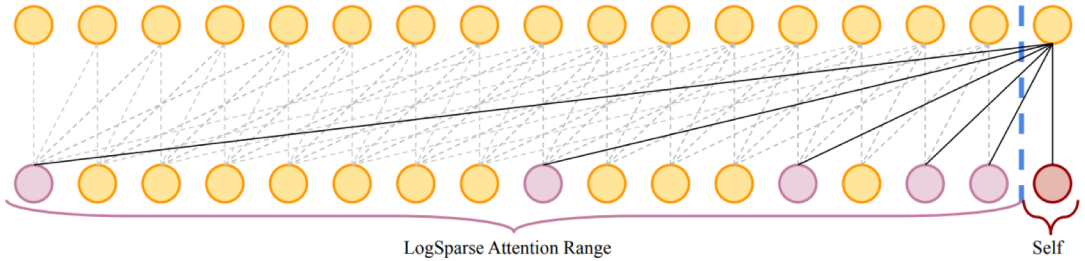
source: Li et al., 2019

Time Series: Long-Sequence Attention

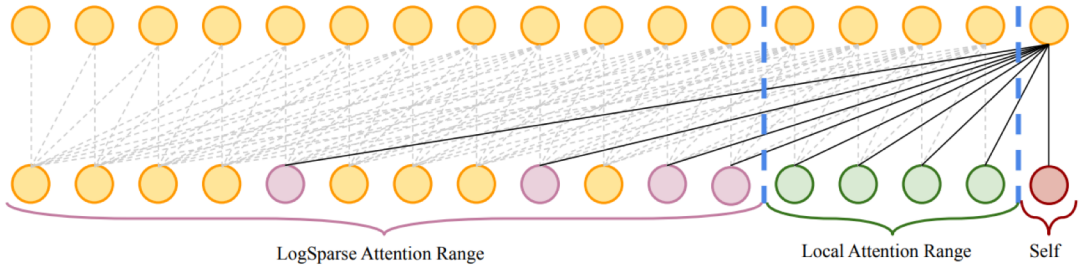
LogSparse attends to $O(\log L)$ tokens per block.
Needs $O(\log L)$ attention blocks to access all input tokens.



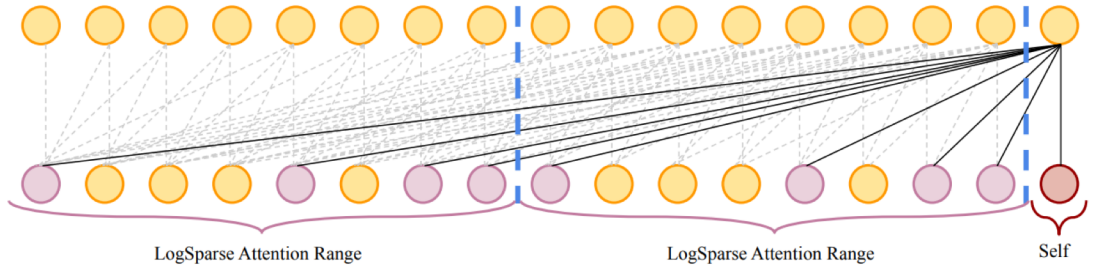
(a). Full Self Attention



(b). LogSparse Self Attention



(c). Local Attention + LogSparse Self Attention



(d). Restart Attention + LogSparse Self Attention

source: Li et al., 2019