Memory

Tim Rocktäschel & Sebastian Riedel
COMP0087 Natural Language Processing

Question Answering

Who is Sebastian Riedel?

All Images News Shopping Videos More Settings Tools

See results about

Sebastian Riedel (Musical artist)
Born: 2 March 1978 (age 40 years), Tychy, Poland
Music group: Cree (Since 1993)

Question Answering

How far is Gliese 1214 b away?

All Images Maps News Videos More Settings Tools

About 26,600 results (0.57 seconds)

Gliese 1214 b / Distance to Earth

40 light years

People also search for

Gliese 1214
47.62 light years

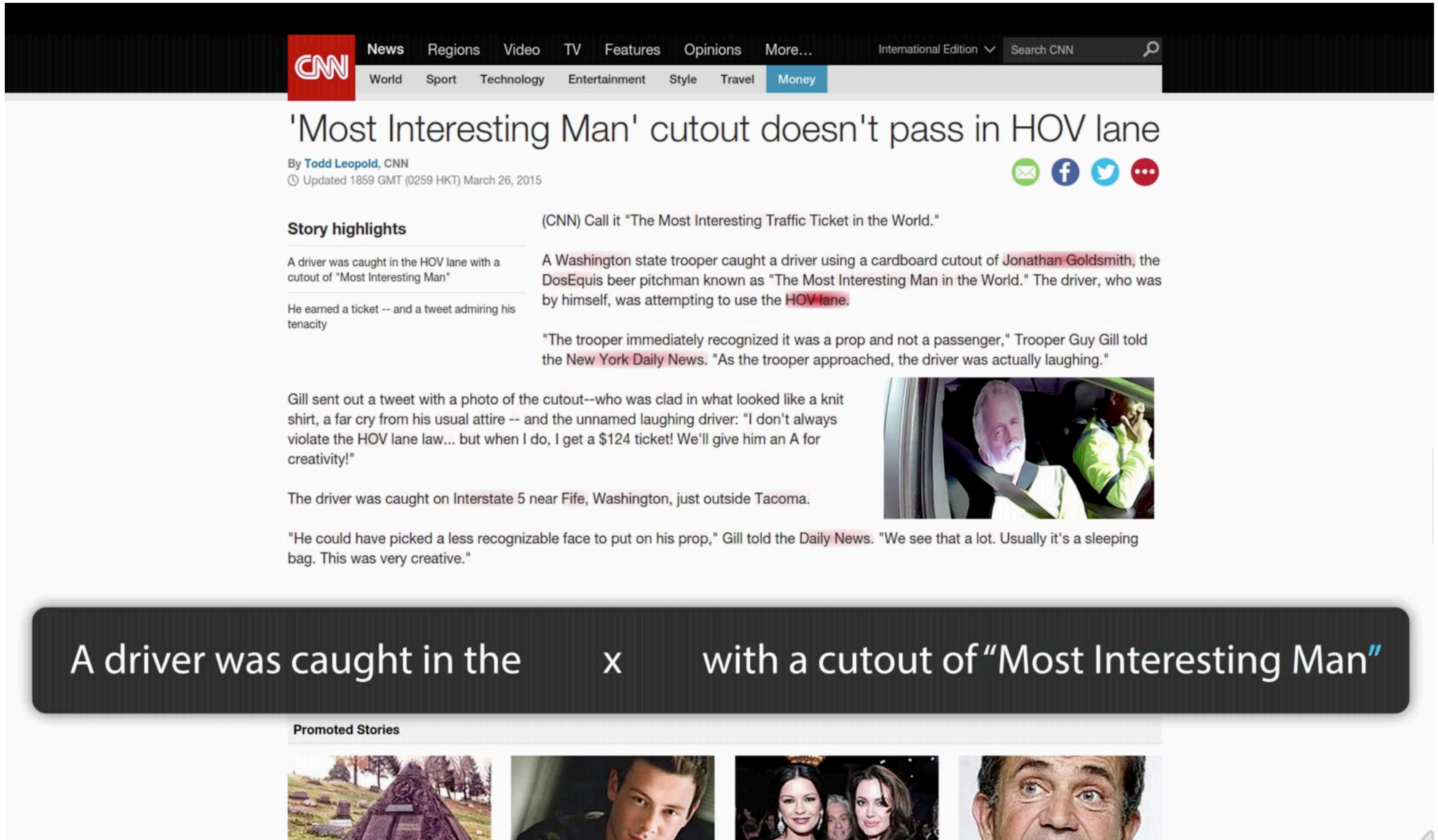
Gliese 436 b
33.01 light years

COROT-7b
489 light years

Question Answering

- Given a question (+ context) produce an answer
- Many Flavors
  - Extractive vs Generative Answers
  - Colze-style vs Natural Questions
  - Open vs Closed Domain
- Many applications
  - Search
  - Information Extraction
  - Dialog agents
  - ...

"Dataset Engineering"

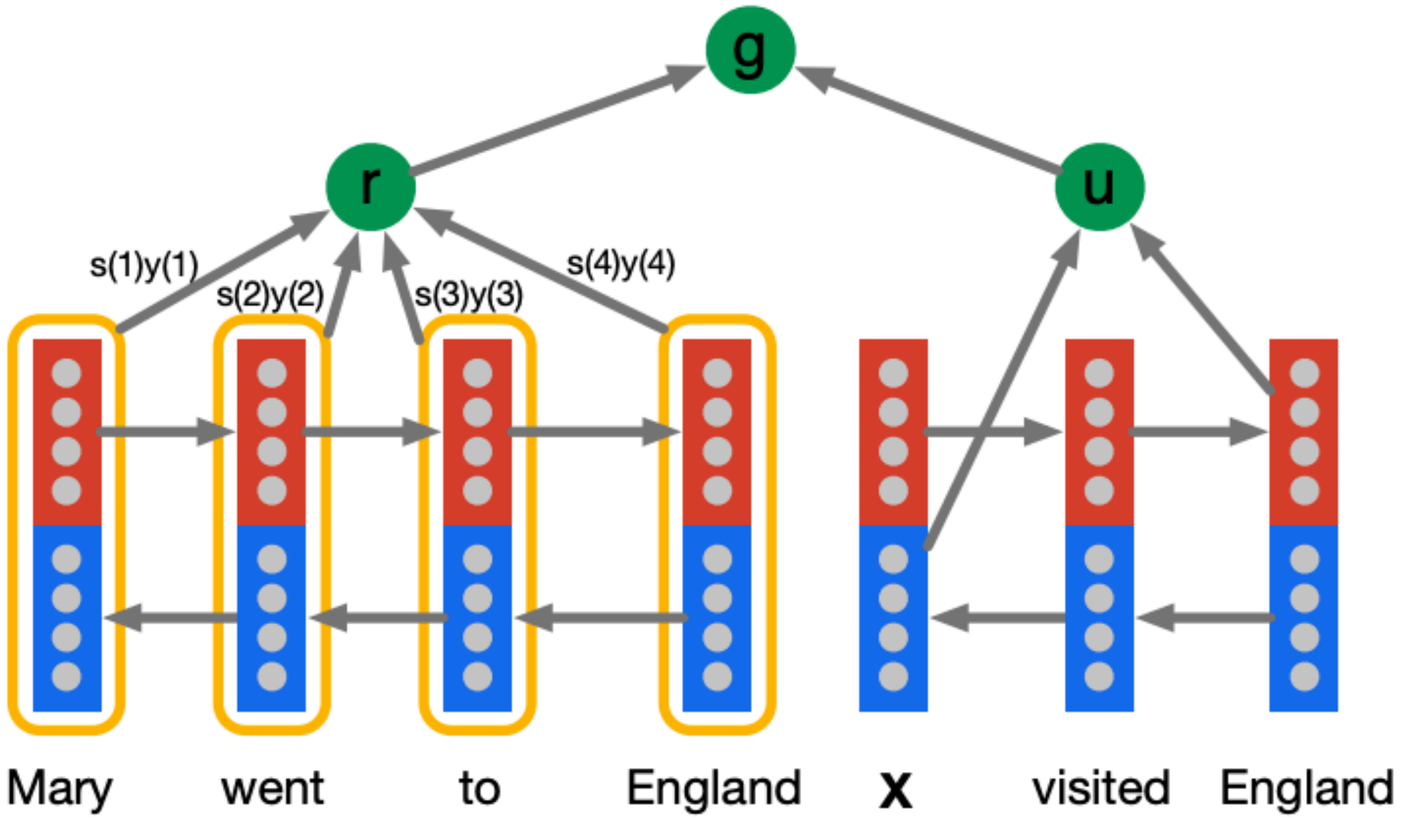


Machine Reading — CNN/Daily Mail

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." . . .	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " . . .
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	
Oisin Tymon	*ent193*

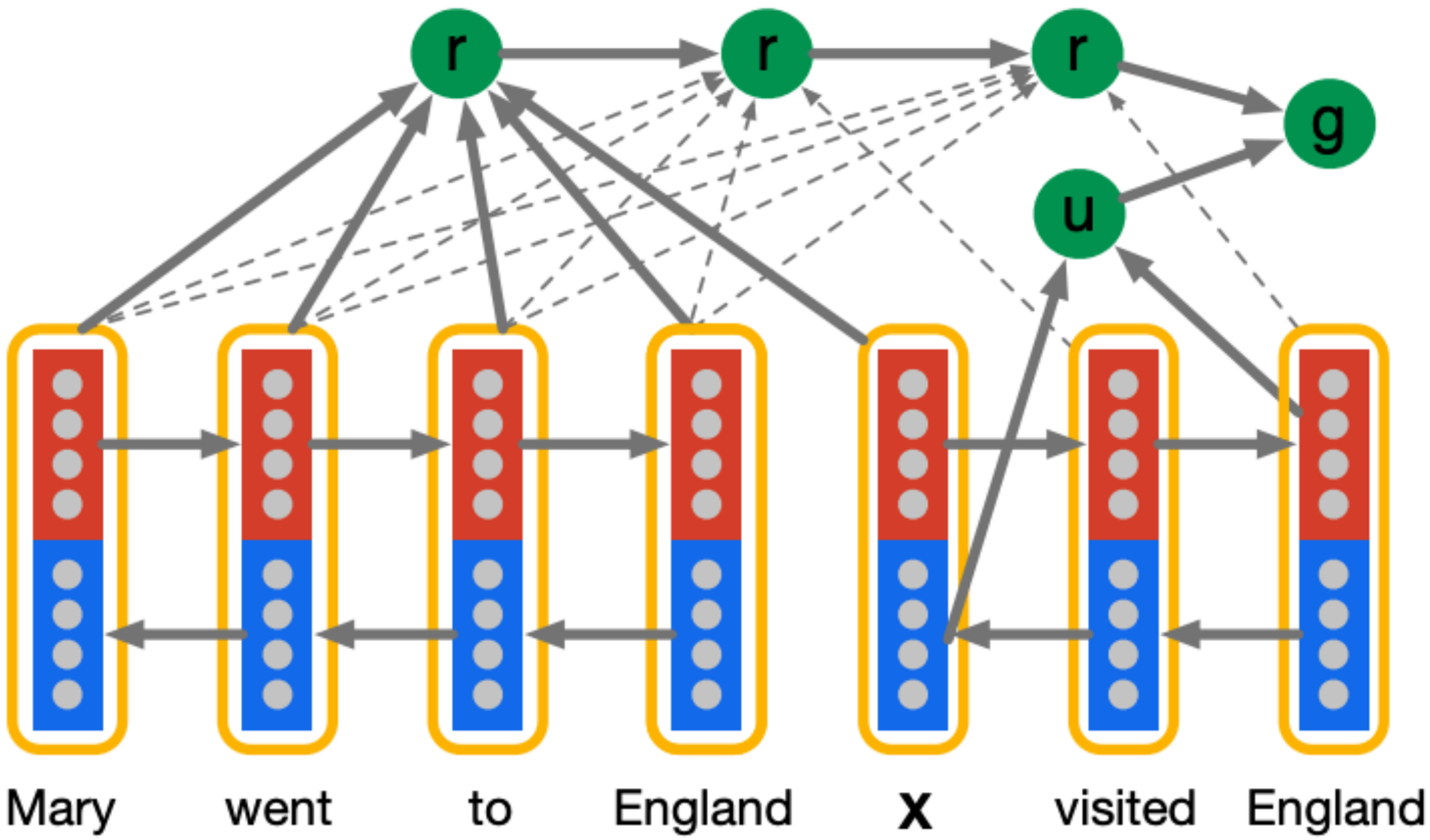
Hermann et al. Teaching Machines to Read and Comprehend. NIPS 2015.

Attentive Reader



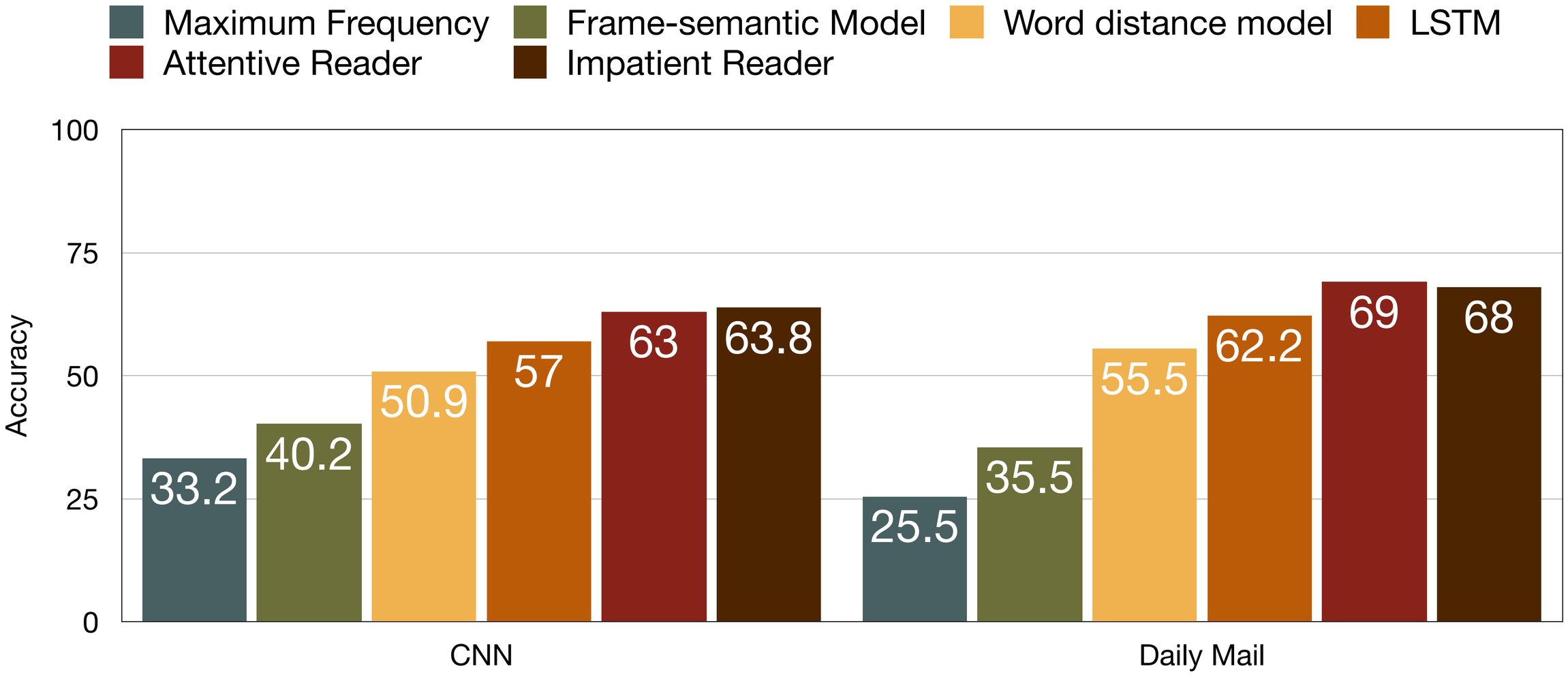
Hermann et al. Teaching Machines to Read and Comprehend. NIPS 2015.

Impatient Reader



Hermann et al. Teaching Machines to Read and Comprehend. NIPS 2015.

Results



Hermann et al. Teaching Machines to Read and Comprehend. NIPS 2015.

Attention Maps

by *ent423* , *ent261* correspondent updated 9:49 pm et , thu march 19 , 2015 (*ent261*) a *ent114* was killed in a parachute accident in *ent45* , *ent85* , near *ent312* , a *ent119* official told *ent261* on wednesday . he was identified thursday as special warfare operator 3rd class *ent23* , 29 , of *ent187* , *ent265* . `` *ent23* distinguished himself consistently throughout his career . he was the epitome of the quiet professional in all facets of his life , and he leaves an inspiring legacy of natural tenacity and focused

. . .

ent119 identifies deceased sailor as **X** , who leaves behind a wife

Attention Maps

by *ent270* , *ent223* updated 9:35 am et , mon march 2 , 2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday , dedicating its collection to `` mamma '' with nary a pair of `` mom jeans '' in sight . *ent164* and *ent21* , who are behind the *ent196* brand , sent models down the runway in decidedly feminine dresses and skirts adorned with roses , lace and even embroidered doodles by the designers ' own nieces and nephews . many of the looks featured saccharine needlework phrases like `` i love you ,

. . .

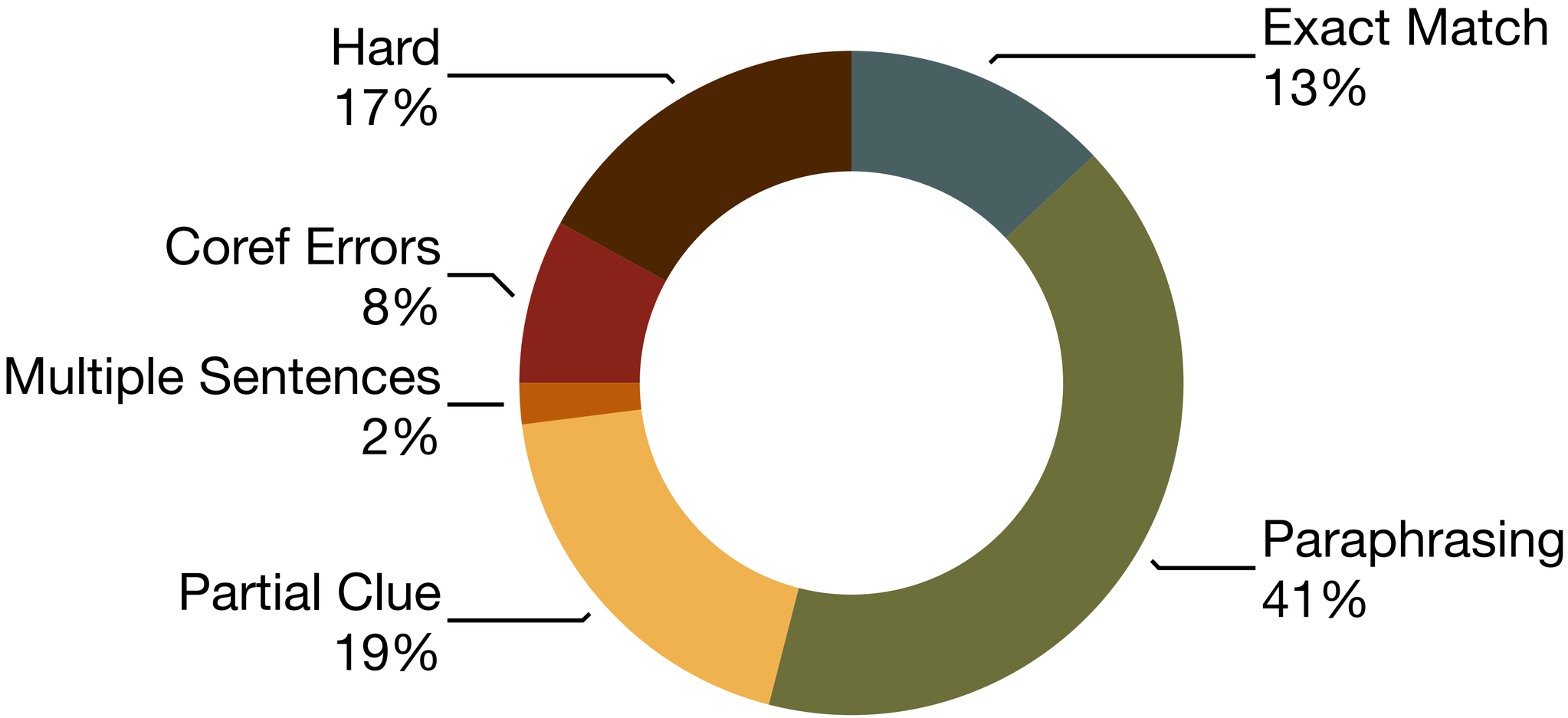
X dedicated their fall fashion show to moms

Look at the data!

Category	Question	Passage
Exact Match	*it 's clear @entity0 is leaning toward* **@placeholder** , says an expert who monitors @entity0	. . . @entity116 , who follows @entity0 's operations and propaganda closely , recently told @entity3 , *it 's clear @entity0 is leaning toward* **@entity60** in terms of doctrine , ideology and an emphasis on holding territory after operations . . . .
Paraphrase	**@placeholder** says he understands why @entity0 wo n't play at his tournament	. . . @entity0 called me personally to let me know that he would n't be playing here at @entity23 , " **@entity3** said on his @entity21 event 's website . . . .
Partial clue	a tv movie based on @entity2 's book **@placeholder** casts a @entity76 actor as @entity5	. . . to @entity12 @entity2 professed that his **@entity11** is not a religious book . . . .
Multiple sent.	he 's doing a his - and - her duet all by himself , @entity6 said of **@placeholder**	. . . we got some groundbreaking performances , here too , tonight , @entity6 said . we got **@entity17** , who will be doing some musical performances . he 's doing a his - and - her duet all by himself . . . .
Coref. Error	rapper **@placeholder** " disgusted , " cancels upcoming show for @entity280	. . . with hip - hop star **@entity246** saying on @entity247 that he was canceling an upcoming show for the @entity249 . . . . (but @entity249 = @entity280 = SAEs)
Hard	pilot error and snow were reasons stated for **@placeholder** plane crash	. . . a small aircraft carrying **@entity5** , @entity6 and @entity7 the @entity12 @entity3 crashed a few miles from @entity9 , near @entity10 , @entity11 . . . .

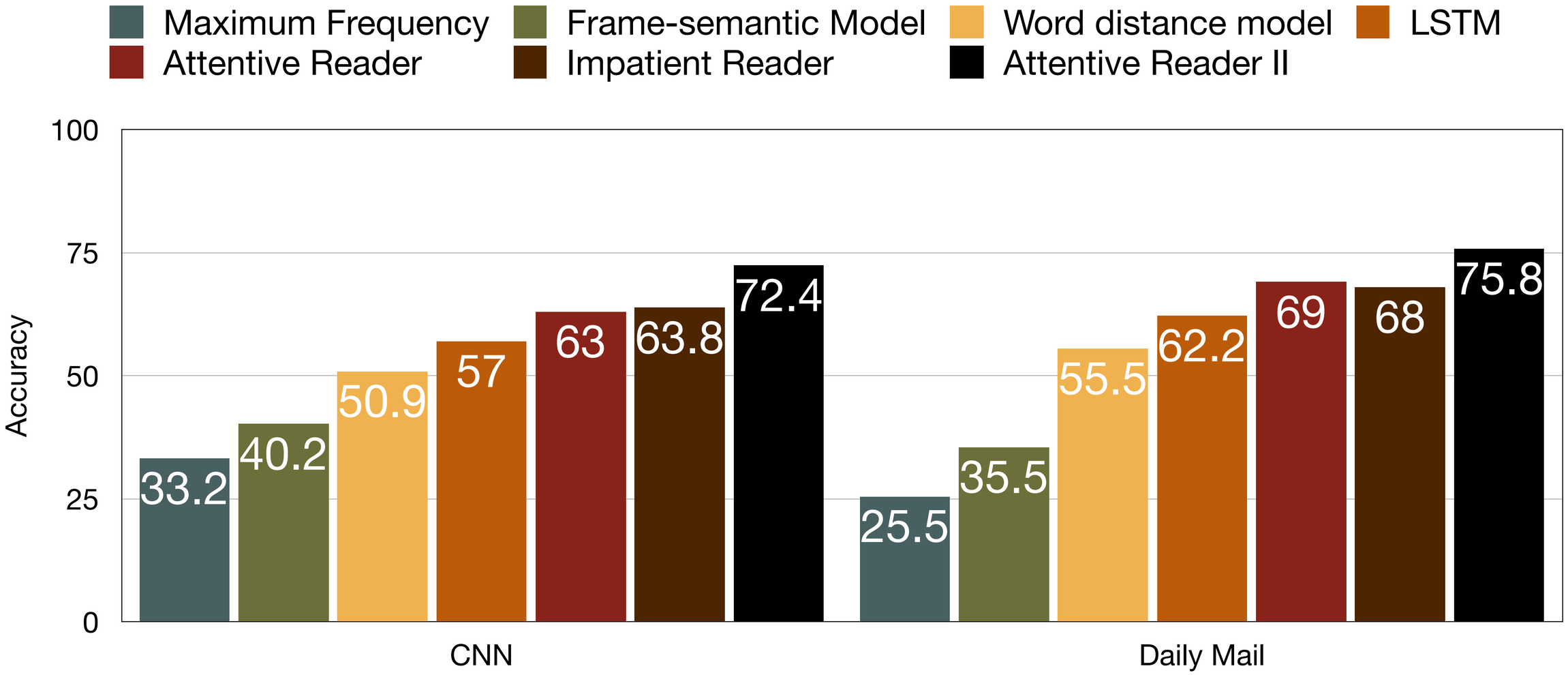
Chen et al. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. ACL 2016.

Look at the data!



Chen et al. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. ACL 2016.

Results



Chen et al. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. ACL 2016.

Differences

- **Bilinear attention**, instead of a tanh layer

$$\alpha_i = \textbf{softmax}(\mathbf{w}^\top \tanh(\mathbf{W}[\mathbf{q}; \mathbf{h}_i] + \mathbf{b})) \qquad \textbf{vs} \qquad \alpha_i = \textbf{softmax}(\mathbf{q}^\top \mathbf{W} \mathbf{h}_i)$$

- Use **weighted contextual embeddings for direct prediction** (Hermann et al. 2015 combined it with the question embedding via another non-linear layer before making final predictions)
- The original model considers all the words from the vocabulary in making predictions—instead **only predict among entities which appear in the passage**

Stanford Question Answering Dataset (SQuAD)

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... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035
5 Jan 14, 2019	BERT + MMFT + ADA (single model) *Microsoft Research Asia*	83.040	85.892
6 Jan 10, 2019	BERT + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	82.972	85.810

Amazon Mechanical Turk

Spend around 4 minutes on the following paragraph to ask 5 questions! If you can't ask 5 questions, ask 4 or 3 (worse), but do your best to ask 5. Select the answer from the paragraph by clicking on 'Select Answer', and then highlight the smallest segment of the paragraph that answers the question.

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O
2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

When asking questions, **avoid using** the same words/phrases as in the paragraph. Also, you are encouraged to pose **hard questions**.

Ask a question here. Try using your own words

Select Answer

Ask a question here. Try using your own words

Select Answer

Synthetic Corpora

Task 1: Single Supporting Fact

Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts

John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts

John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations

The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Task 5: Three Argument Relations

Mary gave the cake to Fred.
Fred gave the cake to Bill.
Jeff was given the milk by Bill.
Who gave the cake to Fred? A: Mary
Who did Fred give the cake to? A: Bill

Task 6: Yes/No Questions

John moved to the playground.
Daniel went to the bathroom.
John went back to the hallway.
Is John in the playground? A:no
Is Daniel in the bathroom? A:yes

Task 7: Counting

Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A: two

Task 8: Lists/Sets

Daniel picks up the football.
Daniel drops the newspaper.
Daniel picks up the milk.
John took the apple.
What is Daniel holding? milk, football

Task 9: Simple Negation

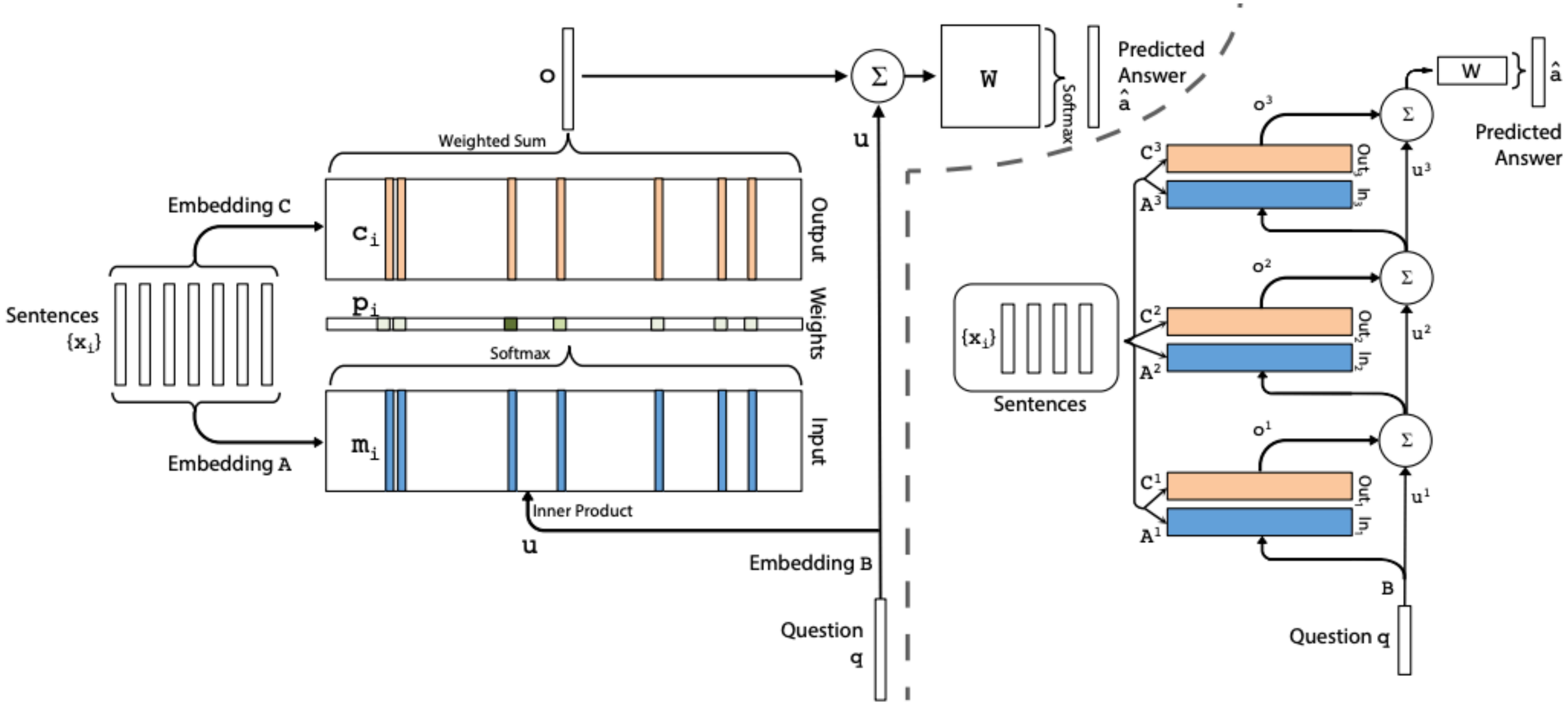
Sandra travelled to the office.
Fred is no longer in the office.
Is Fred in the office? A:no
Is Sandra in the office? A:yes

Task 10: Indefinite Knowledge

John is either in the classroom or the playground.
Sandra is in the garden.
Is John in the classroom? A:maybe
Is John in the office? A:no

Weston et al. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. arXiv 2015.

Memory Networks—Text as Knowledge Base



Sukhbaatar et al. End-to-End Memory Networks. NIPS 2015.

Attention Maps

Story (1: 1 supporting fact)	Support	Hop 1	Hop 2	Hop 3
Daniel went to the bathroom.		0.00	0.00	0.03
Mary travelled to the hallway.		0.00	0.00	0.00
John went to the bedroom.		0.37	0.02	0.00
John travelled to the bathroom.	yes	0.60	0.98	0.96
Mary went to the office.		0.01	0.00	0.00
Where is John? Answer: bathroom Prediction: bathroom				

Story (2: 2 supporting facts)	Support	Hop 1	Hop 2	Hop 3
John dropped the milk.		0.06	0.00	0.00
John took the milk there.	yes	0.88	1.00	0.00
Sandra went back to the bathroom.		0.00	0.00	0.00
John moved to the hallway.	yes	0.00	0.00	1.00
Mary went back to the bedroom.		0.00	0.00	0.00
Where is the milk? Answer: hallway Prediction: hallway				

Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow Prediction: yellow				

Story (18: size reasoning)	Support	Hop 1	Hop 2	Hop 3
The suitcase is bigger than the chest.	yes	0.00	0.88	0.00
The box is bigger than the chocolate.		0.04	0.05	0.10
The chest is bigger than the chocolate.	yes	0.17	0.07	0.90
The chest fits inside the container.		0.00	0.00	0.00
The chest fits inside the box.		0.00	0.00	0.00
Does the suitcase fit in the chocolate? Answer: no Prediction: no				

Sukhbaatar et al. End-to-End Memory Networks. NIPS 2015.

Key-Value Memory Networks

Doc: Wikipedia Article for Blade Runner (partially shown)

Blade Runner is a 1982 American neo-noir dystopian science fiction film directed by Ridley Scott and starring Harrison Ford, Rutger Hauer, Sean Young, and Edward James Olmos. The screenplay, written by Hampton Fancher and David Peoples, is a modified film adaptation of the 1968 novel "Do Androids Dream of Electric Sheep?" by Philip K. Dick. The film depicts a dystopian Los Angeles in November 2019 in which genetically engineered replicants, which are visually indistinguishable from adult humans, are manufactured by the powerful Tyrell Corporation as well as by other "mega-corporations" around the world. Their use on Earth is banned and replicants are exclusively used for dangerous, menial, or leisure work on off-world colonies. Replicants who defy the ban and return to Earth are hunted down and "retired" by special police operatives known as "Blade Runners". . . .

KB entries for Blade Runner (subset)

Blade Runner *directed_by* Ridley Scott
Blade Runner *written_by* Philip K. Dick, Hampton Fancher
Blade Runner *starred_actors* Harrison Ford, Sean Young, . . .
Blade Runner *release_year* 1982
Blade Runner *has_tags* dystopian, noir, police, androids, . . .

IE entries for Blade Runner (subset)

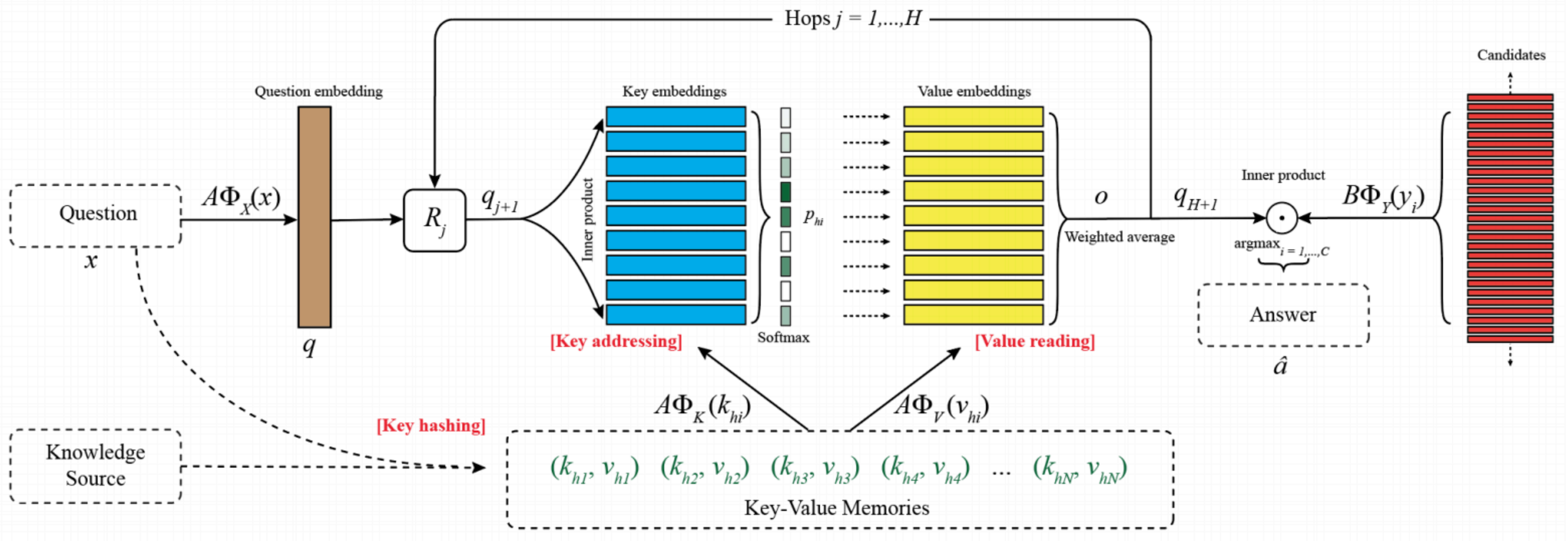
Blade Runner, Ridley Scott *directed* dystopian, science fiction, film
Hampton Fancher *written* Blade Runner
Blade Runner *starred* Harrison Ford, Rutger Hauer, Sean Young. . .
Blade Runner *labelled* 1982 neo noir
special police, Blade *retired* Blade Runner
Blade Runner, special police *known* Blade

Questions for Blade Runner (subset)

Ridley Scott directed which films?
What year was the movie Blade Runner released?
Who is the writer of the film Blade Runner?
Which films can be described by dystopian?
Which movies was Philip K. Dick the writer of?
Can you describe movie Blade Runner in a few words?

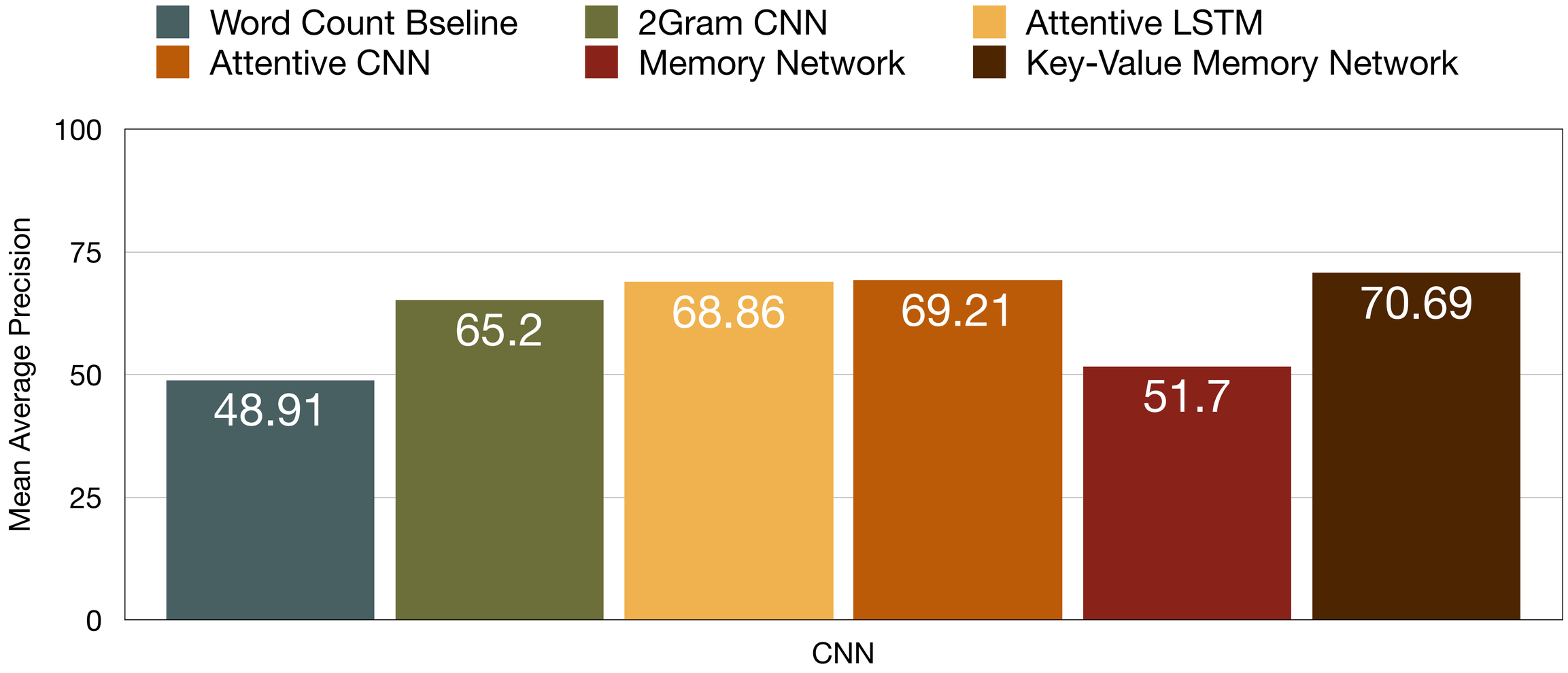
Miller et al. Key-Value Memory Networks for Directly Reading Documents. EMNLP 2016.

Key-Value Memory Networks



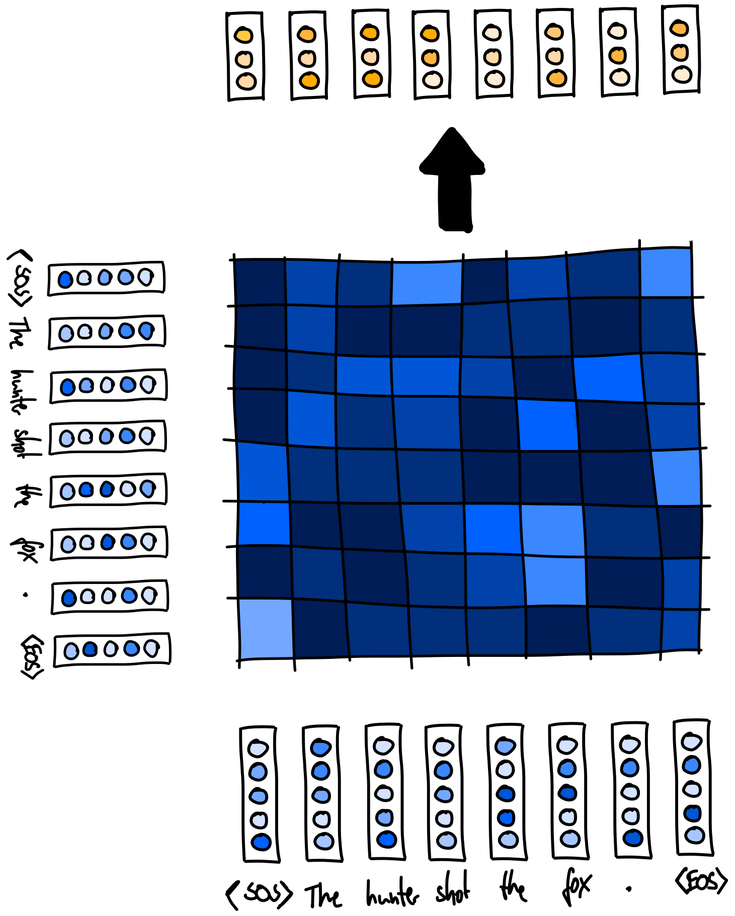
Miller et al. Key-Value Memory Networks for Directly Reading Documents. EMNLP 2016.

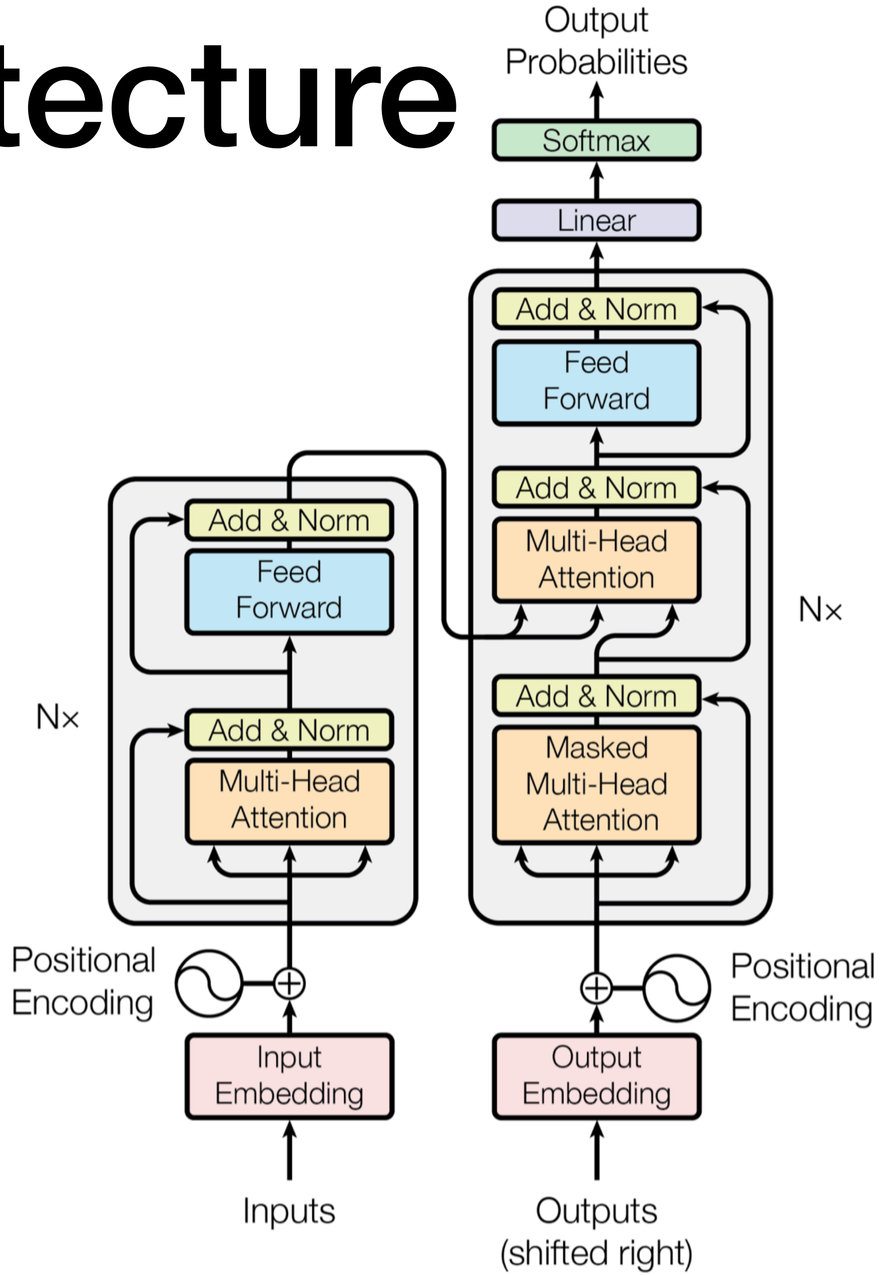
Results on WikiQA



Transformer Architecture

- Can we replace RNNs entirely by self attention?



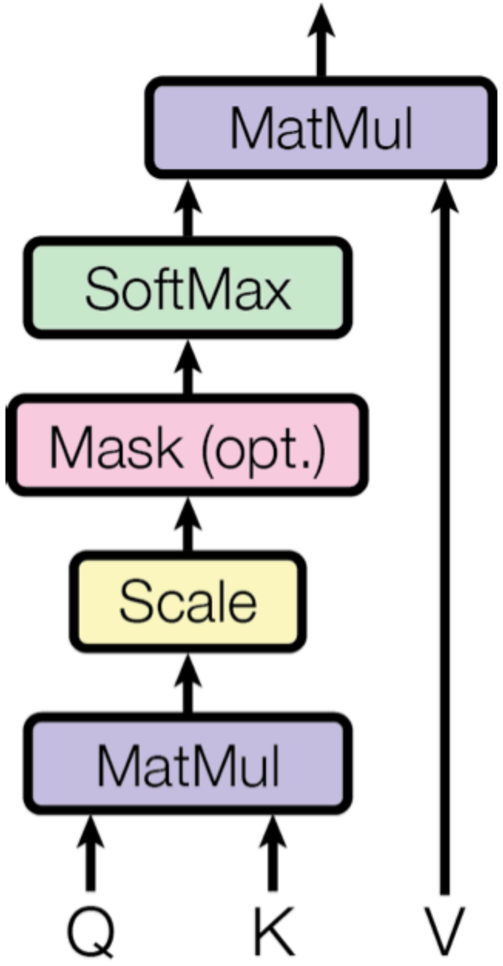


Vaswani et al. Attention is all you need. NIPS 2017.

Scaled Attention

- Observation: If your vectors have large dimension d_k, then the dot-product becomes very large, pushing the softmax function into regions of very small gradients
- Idea: Rescale by one over square-root of number of dimensions

$$\alpha = \textbf{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$



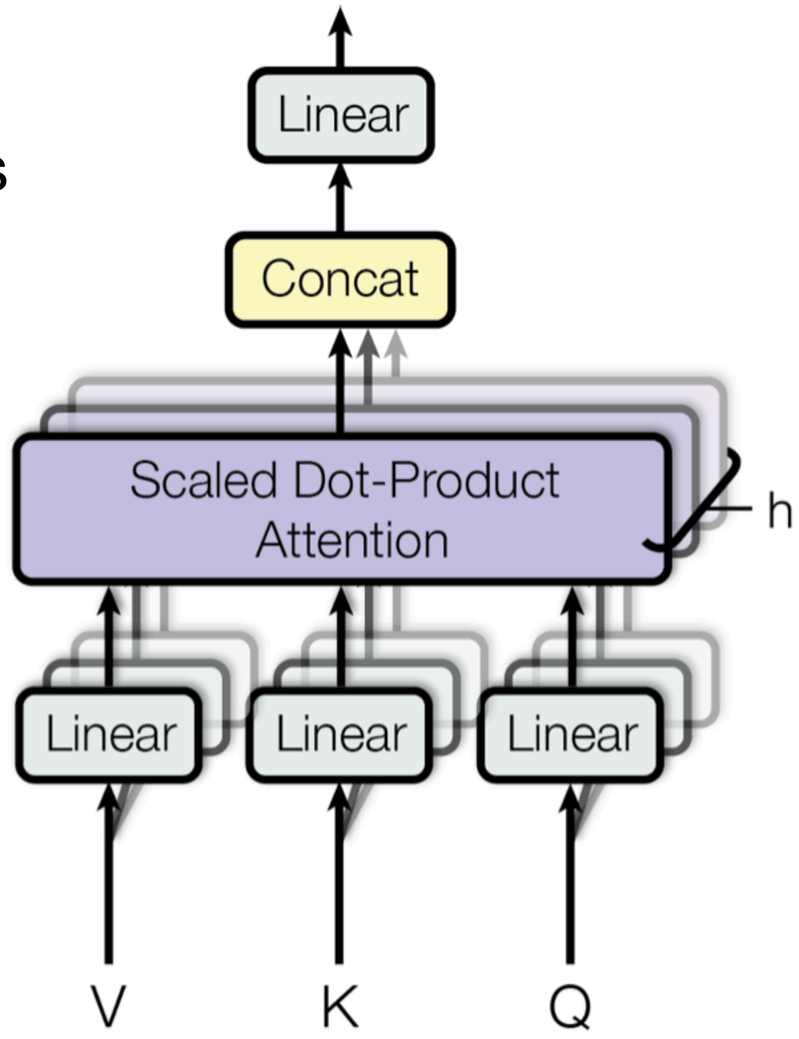
Vaswani et al. Attention is all you need. NIPS 2017.

Multi-head Attention

- Idea: Use h-many different attention functions that can specialize in matching different key/values pairs

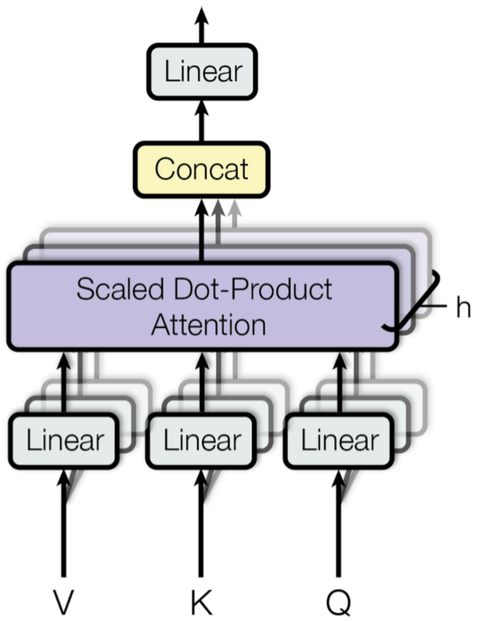
$$\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_h]\mathbf{W}^O$$

where $\alpha_i =$ **softmax** $\left(\frac{(\mathbf{W}^Q\mathbf{Q})(\mathbf{W}^K\mathbf{K}^\top)}{\sqrt{d_k}} \right)(\mathbf{W}^V\mathbf{V})$



Vaswani et al. Attention is all you need. NIPS 2017.

Implementation Sketch




```python
def forward(self, q, k, v, mask=None):
    residual = q
    q = rearrange(self.w_qs(q), 'b l (head k) -> head b l k', head=self.n_head)
    k = rearrange(self.w_ks(k), 'b t (head k) -> head b t k', head=self.n_head)
    v = rearrange(self.w_vs(v), 'b t (head v) -> head b t v', head=self.n_head)
    attn = torch.einsum('hblk,hbtk->hblt', [q, k]) / np.sqrt(q.shape[-1])
    if mask is not None:
        attn = attn.masked_fill(mask[None], -np.inf)
    attn = torch.softmax(attn, dim=3)
    output = torch.einsum('hblt,hbtv->hblv', [attn, v])
    output = rearrange(output, 'head b l v -> b l (head v)')
    output = self.dropout(self.fc(output))
    output = self.layer_norm(output + residual)
    return output, attn
```

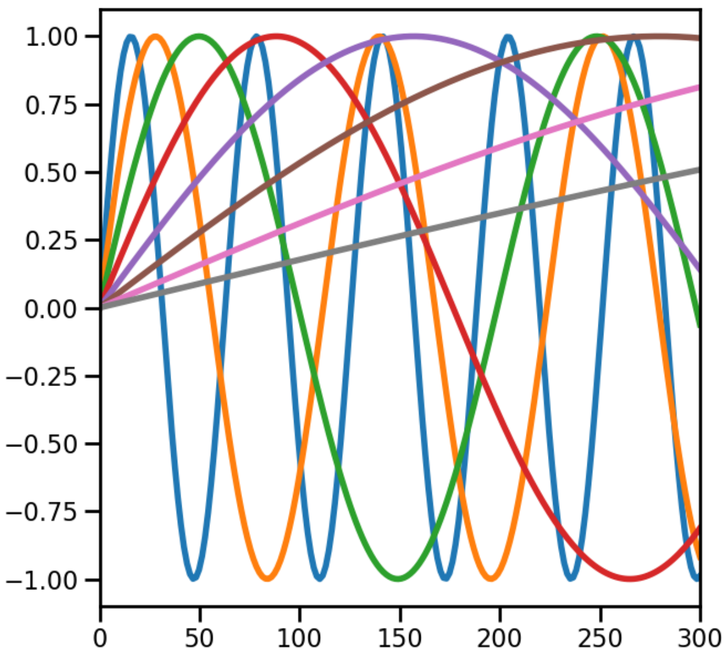

https://arogozhnikov.github.io/einops/pytorch-examples.html

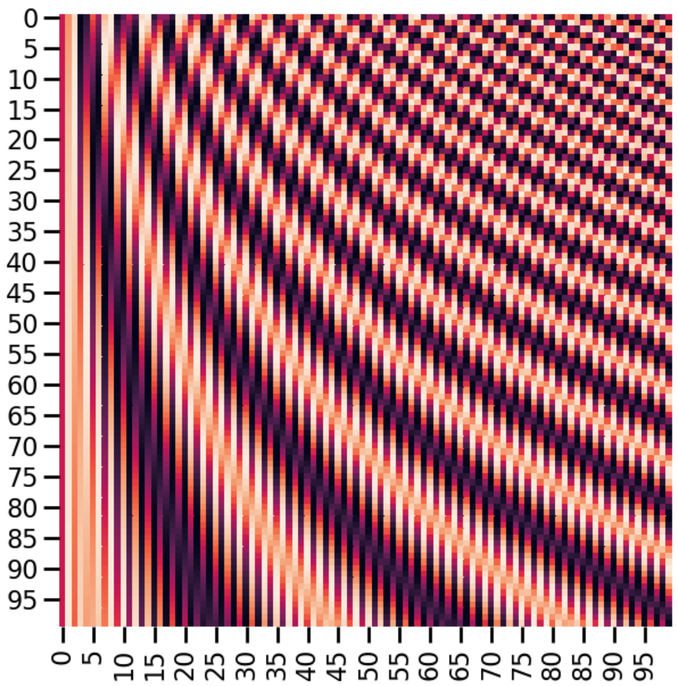
Positional Encoding

- Observation: By not having a recurrence, we loose information about the position of words in the sequence
- Idea: Remedy this by providing a positional encoding using sin and cos

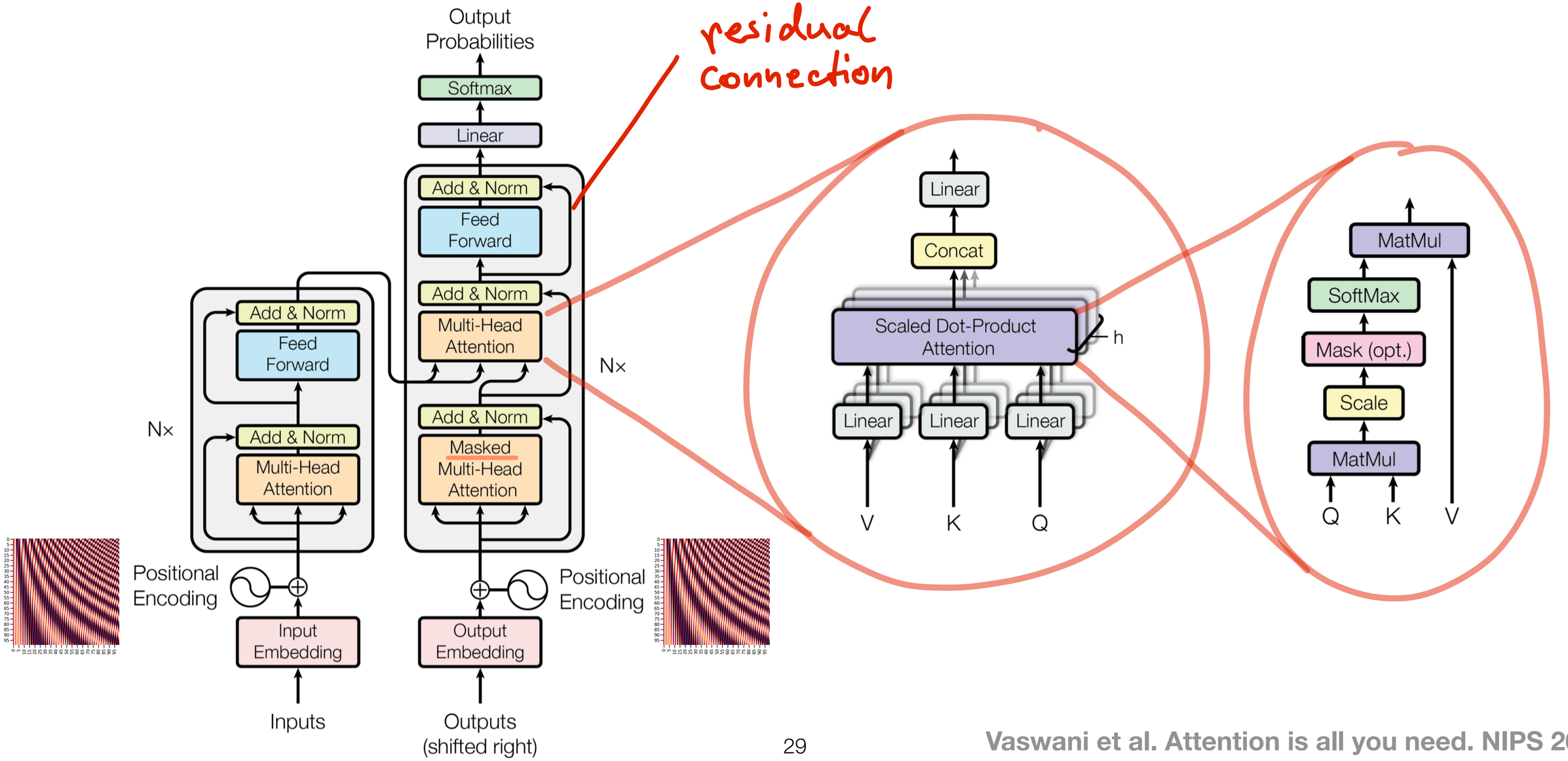
$$\mathbf{p}_{td} = \sin\left(\frac{t}{10000^{\frac{2d}{d_k}}}\right)$$

$$\mathbf{p}_{td+1} = \cos\left(\frac{t}{10000^{\frac{2d+1}{d_k}}}\right)$$



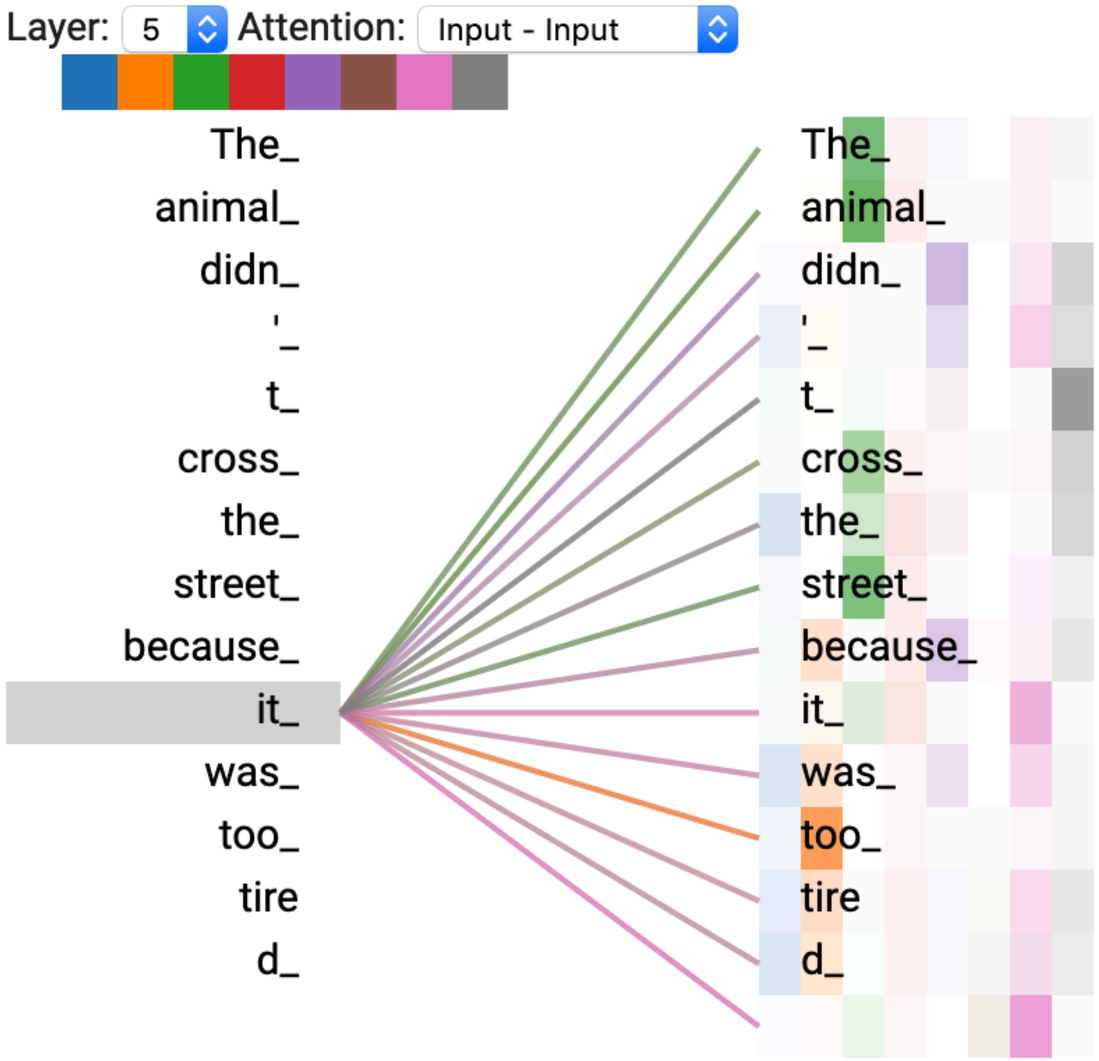


Putting Everything Together

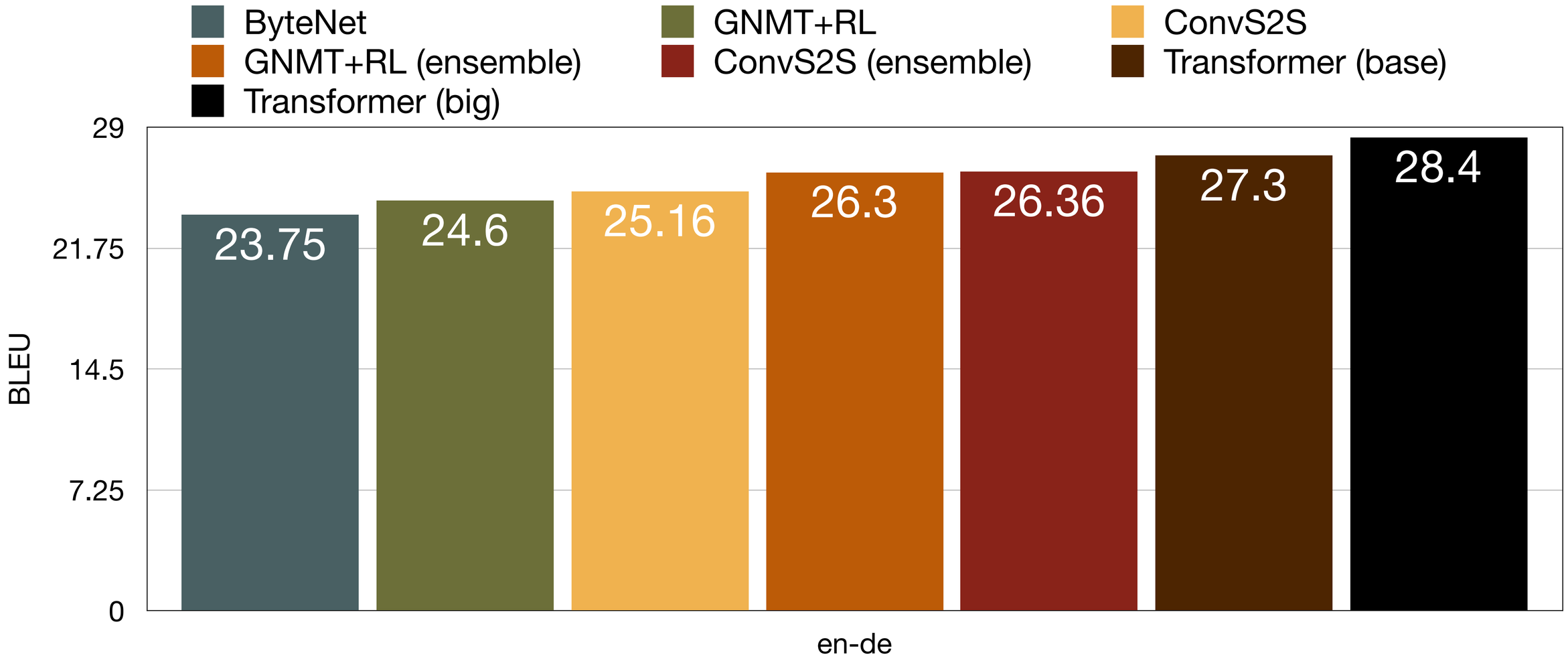


Vaswani et al. Attention is all you need. NIPS 2017.

Attention Maps



Results


ByteNet
GNMT+RL
ConvS2S
GNMT+RL (ensemble)
ConvS2S (ensemble)
Transformer (base)
Transformer (big)
29
21.75
14.5
7.25
0
BLEU
23.75
24.6
25.16
26.3
26.36
27.3
28.4
en-de

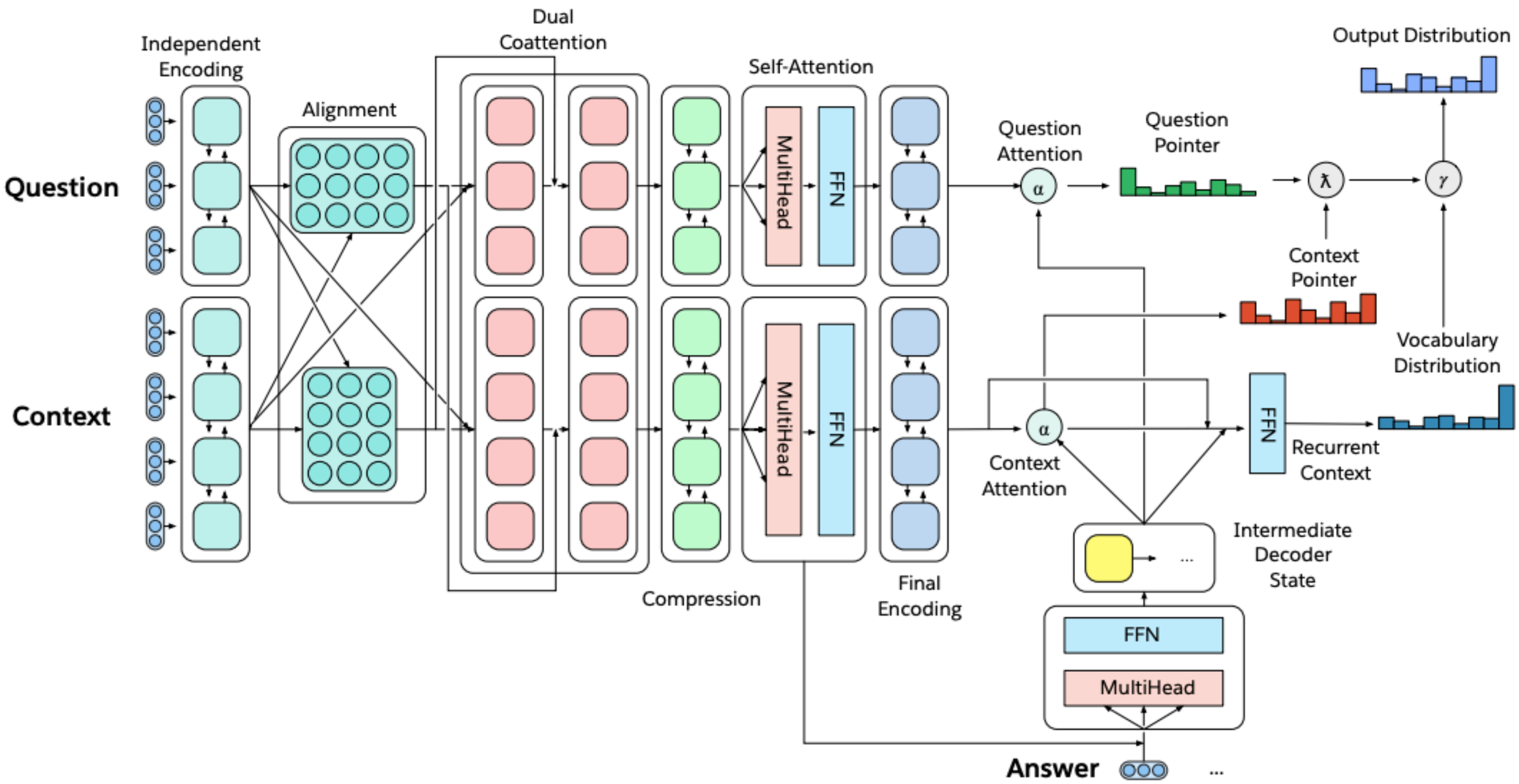
Everything is Q&A

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

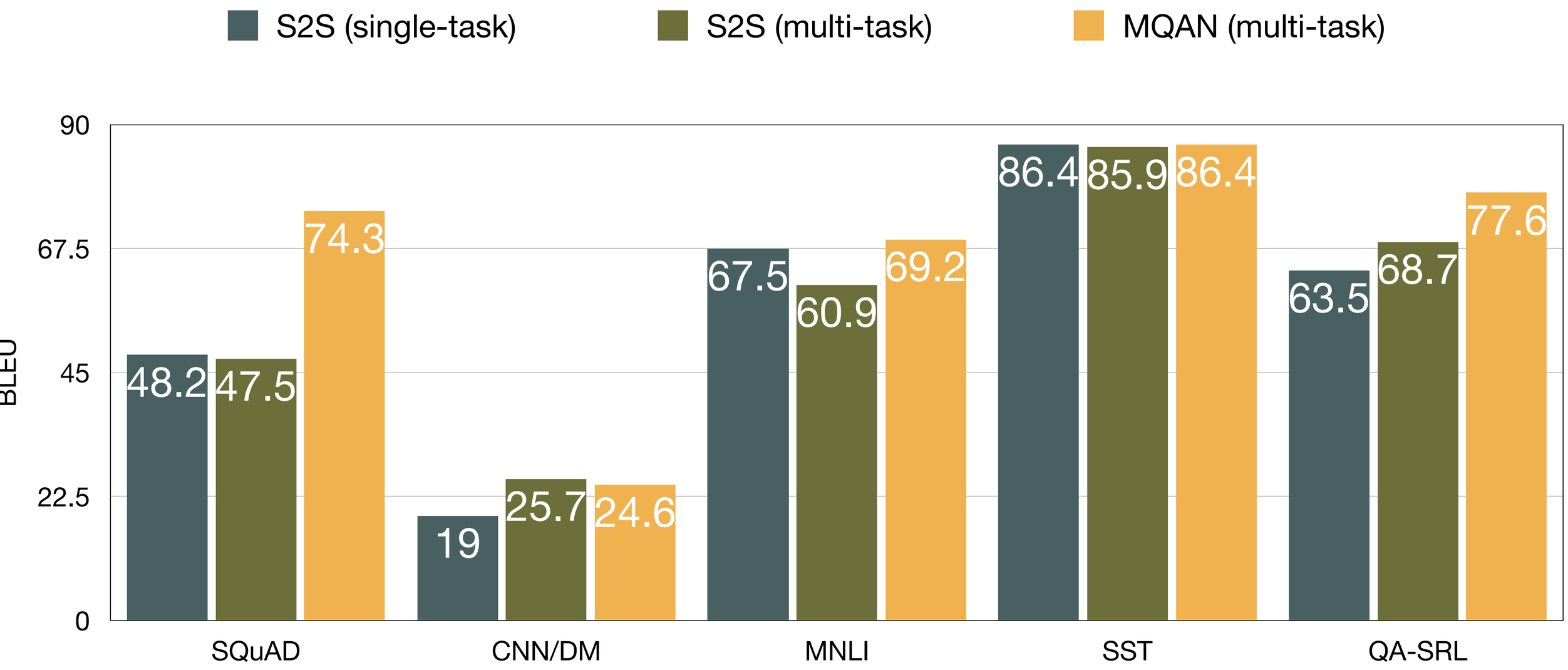
McCann et al. The Natural Language Decathlon: Multitask Learning as Question Answering. arXiv 2018.

Multitask Question Answering Network — Putting Everything Together



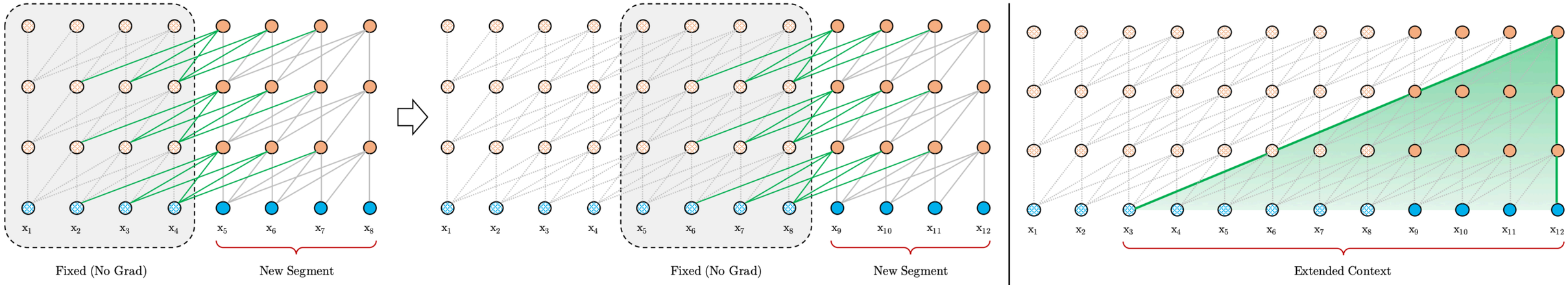
McCann et al. The Natural Language Decathlon: Multitask Learning as Question Answering. arXiv 2018.

Results


S2S (single-task)
S2S (multi-task)
MQAN (multi-task)
BLEU
90
67.5
45
22.5
0
48.2
47.5
74.3
19
25.7
24.6
67.5
60.9
69.2
86.4
85.9
86.4
63.5
68.7
77.6
SQuAD
CNN/DM
MNLI
SST
QA-SRL

Transformer XL

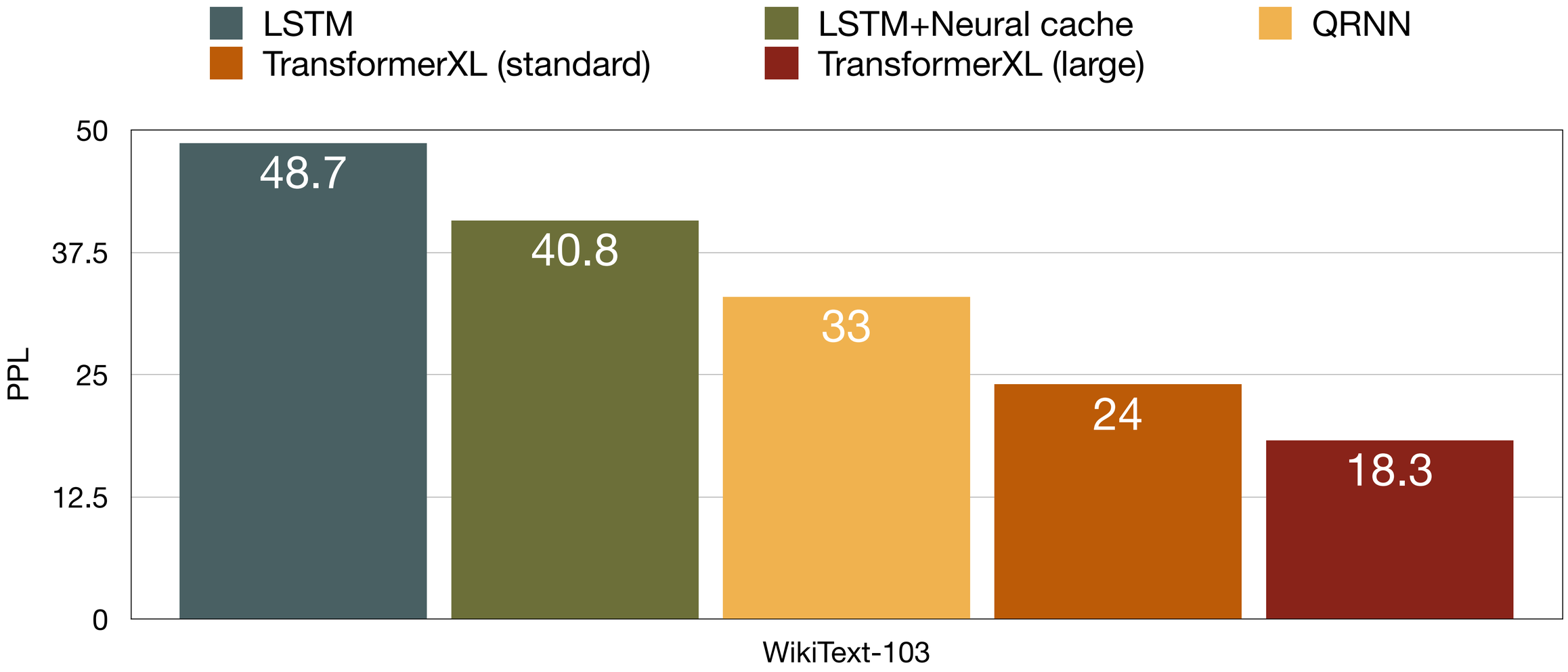
- Large-scale self-attention model for language modeling
- Idea: **Cache** outputs of Transformer for previous **segments**
- Allow TransformerXL to **attend over cached outputs** without backpropagating beyond current segment



(a) Training phase. (b) Evaluation phase.

Dai et al. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. arXiv 2019.

Results



Dai et al. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. arXiv 2019.

OpenAI GPT

- Use decoder in Transformer architecture for Language Modeling
- v2
  - Trained on 8 million documents (40GB of text)
  - Scraped from outgoing links of Reddit articles with at least 3 karma
  - 50k vocab
  - 1024 tokens context
  - >1.5 billion parameters

"Dataset Engineering"

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”

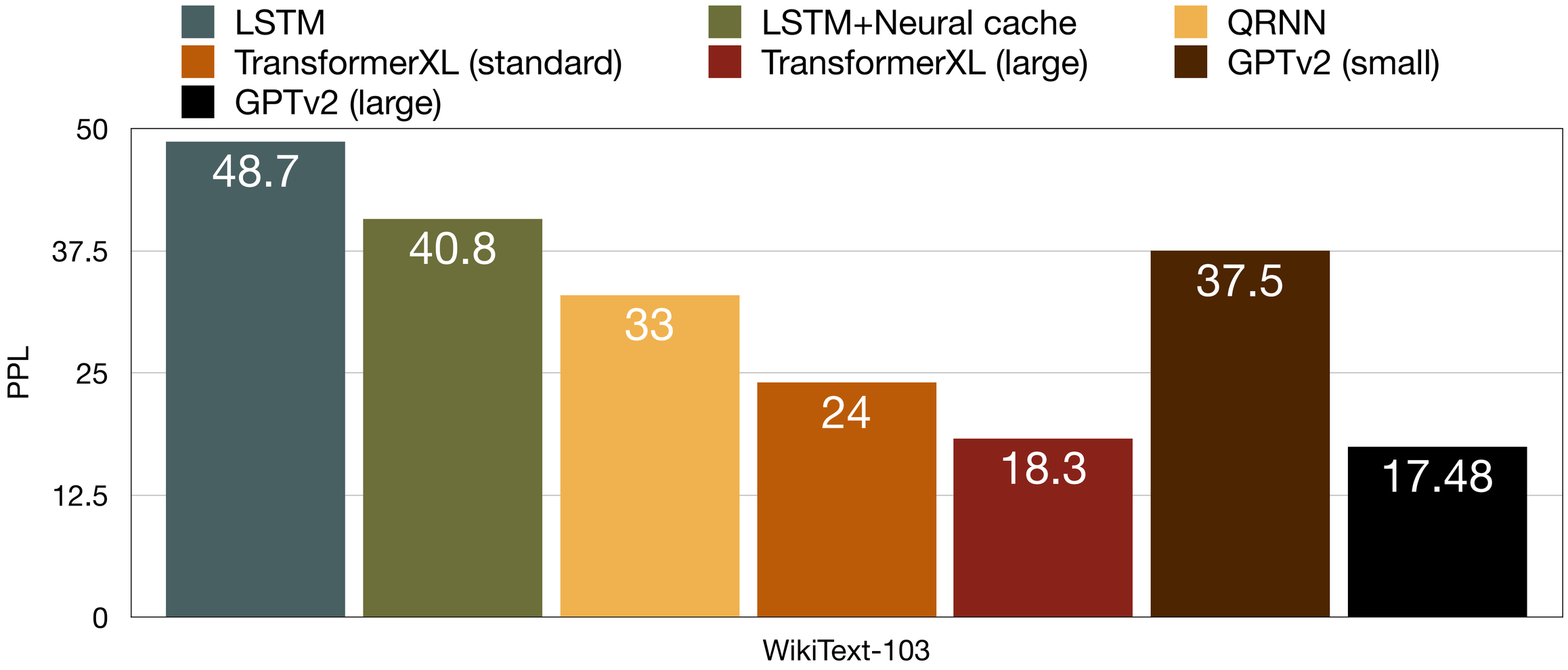
“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

Results



Dai et al. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. arXiv 2019.

Generating “Wikipedia”

Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by **John Clair**, **An Imperial Japanese Revolt**, associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the **Protestant Immineners**, which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to **Antioch, Perth, October 25|21** to note, the Kingdom of Costa Rica, unsuccessful fashioned the **Thrales**, **Cynth's Dajoard**, known in western **Scotland**, near Italy to the conquest of India with the conflict. Copyright was the succession of independence in the slop of Syrian influence that was a famous German movement based on a more popular servicious, non-doctrinal and sexual power post.

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Redford et al. Language Models are Unsupervised Multitask Learners. ??? 2019.

Talk to Transformer

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. Learn more below.

Follow @AdamDanielKing for more neat neural networks.

Custom prompt

You might not now this, but Sebastian was in fact a gangsta rapper in Hamburg before becoming an academic. His biggest hits include

GENERATE ANOTHER

Completion

You might not now this, but Sebastian was in fact a gangsta rapper in Hamburg before becoming an academic. His biggest hits include Krak Superstar and Bitch. Now, he's returned to the Czech Republic where he teaches the language, "It's difficult, but we work with it, take it seriously and don't look at it as our job." That last point is particularly useful, since many old school students tell him the language is actually in dire need of an overhaul.

The University of Hamburg has an existing native speaker, the world-renowned playwright Richard Cairns, but Sebastian wants more. He believes his academic language courses have inspired other linguists to take up the study. In October, Sebastian was given a Grammy award, for

Concerns

Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter: "we expect that safety and security concerns will reduce our traditional publishing in the future, while increasing the importance of sharing safety, policy, and standards research," and we see this current work as potentially representing the early beginnings of such concerns, which we expect may grow over time.