QANet: Towards Efficient Human-Level Reading Comprehension on SQuAD

Adams Wei Yu

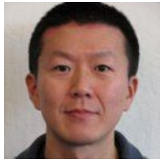
Deview 2018, Seoul

Collaborators

David Dohan

Thang Luong

Rui Zhao

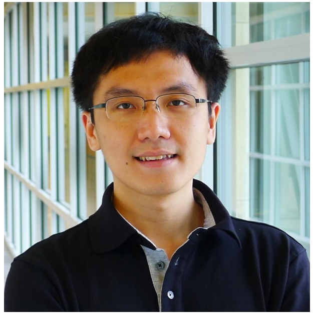


Kai Chen

Mohammad Norouzi

Quoc Le

Bio

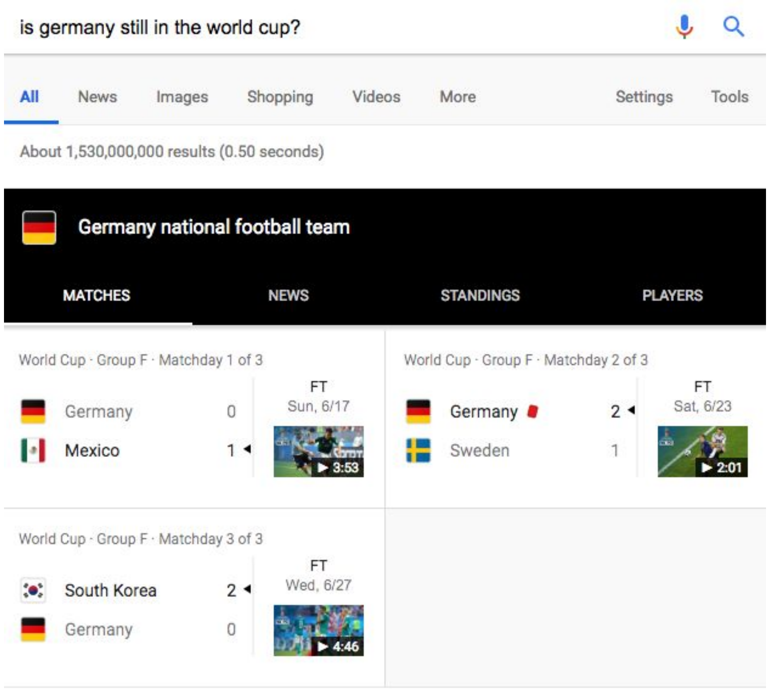


[Adams Wei Yu]

- Ph.D Candidate @ MLD, CMU
  - Advisor: Jaime Carbonell, Alex Smola
  - Large scale optimization
  - Machine reading comprehension

Question Answering

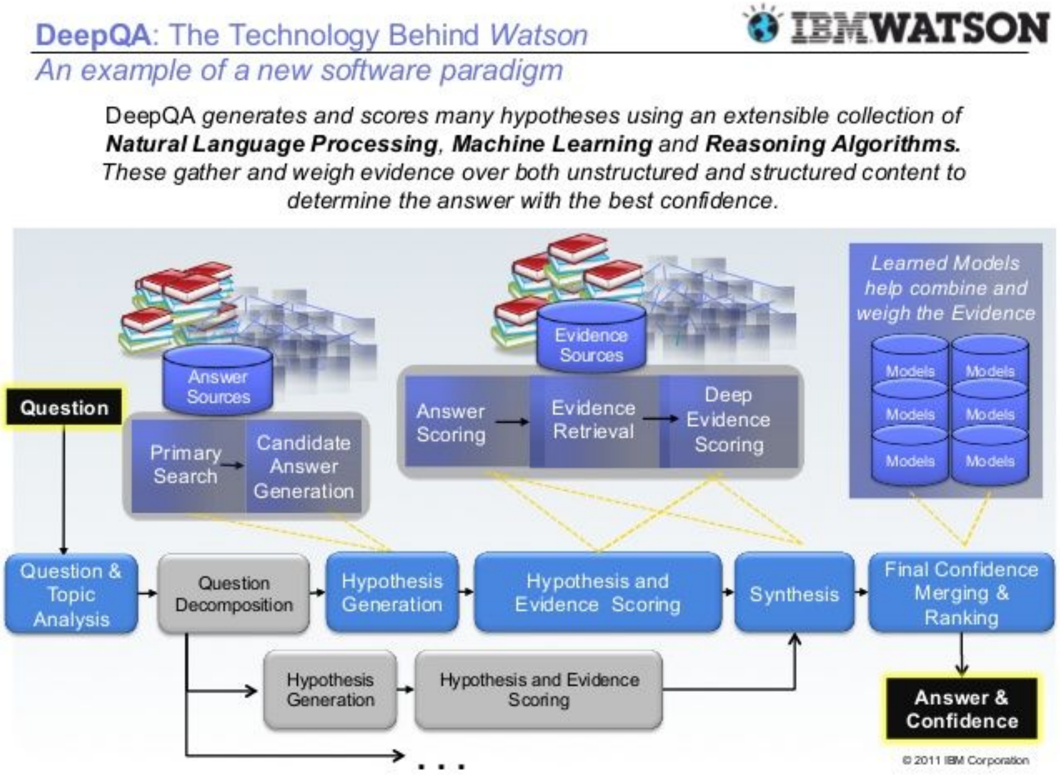
Concrete Answer



No clear answer

Early Success

Watson: complex multi-stage system



http://www.aaai.org/Magazine/Watson/watson.php

Moving towards end-to-end systems

- Translation
- Question Answering

Lots of Datasets Available

SQuAD2.0
The Stanford Question Answering Dataset

Narrative QA

MS Marco

QuAC
Question Answering in Context

CoQA
A Conversational Question Answering Challenge

TriviaQA

Stanford Question Answer Dataset (SQuAD)

Data: Crowdsourced 100k question-answer pairs on 500 Wikipedia articles.

Passage: In education, teachers facilitate student learning, often in a school or academy or perhaps in another environment such as outdoors. A teacher who teaches on an individual basis may be described as a tutor.

Question: **What is the role of teachers in education?**

Groundtruth: **facilitate student learning**

Prediction 1: **facilitate student learning** — **EM = 1, F1 = 1**

Prediction 2: **student learning** — **EM = 0, F1 = 0.8**

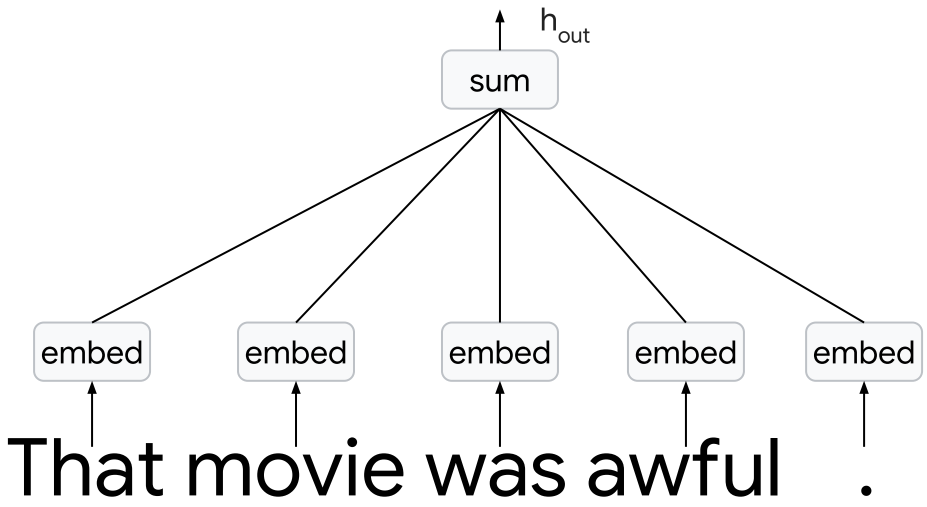
Prediction 3: **teachers facilitate student learning** — **EM = 0, F1 = 0.86**

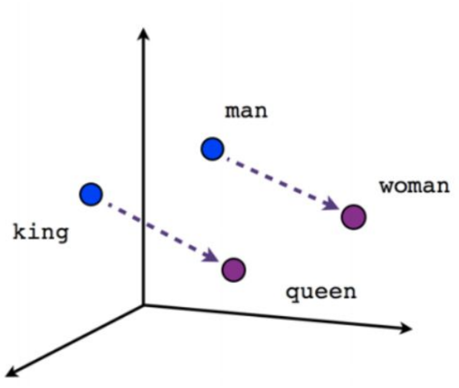
Roadmap

- Models for text
- General neural structures for QA
- Building blocks for QANet
  - Fully parallel (CNN + Self-attention)
  - data augmentation via back-translation
  - transfer learning from unsupervised tasks

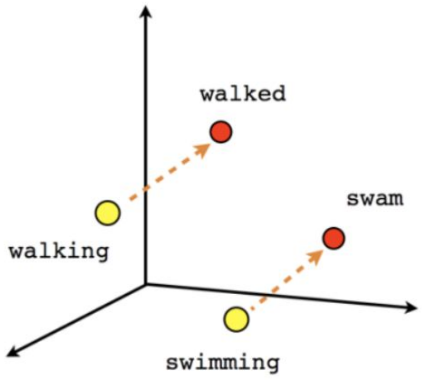
That movie was awful.

Bag of words

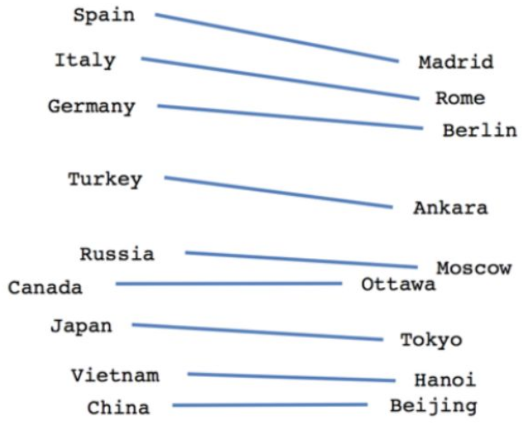




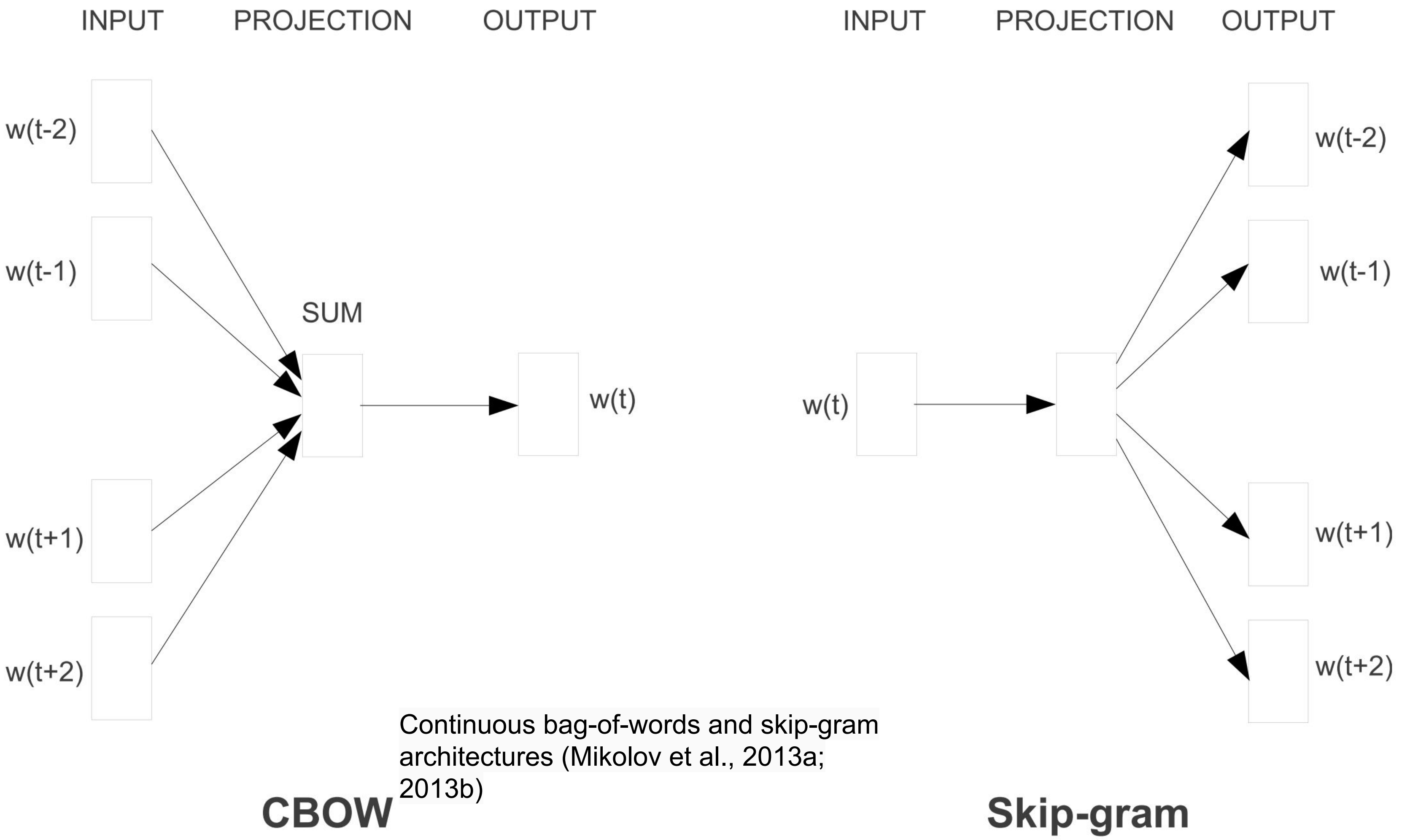
Male-Female



Verb tense

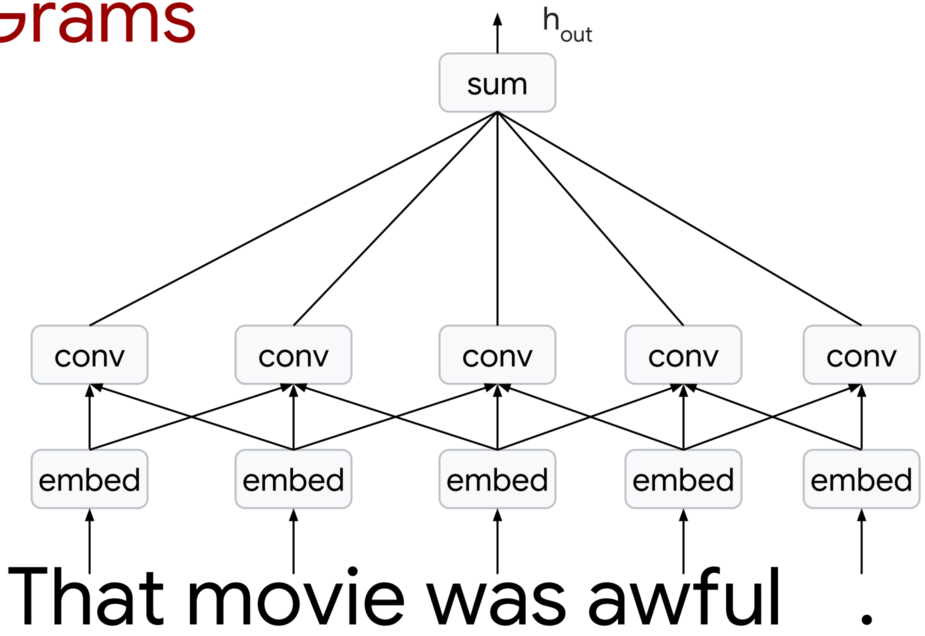


Country-Capital

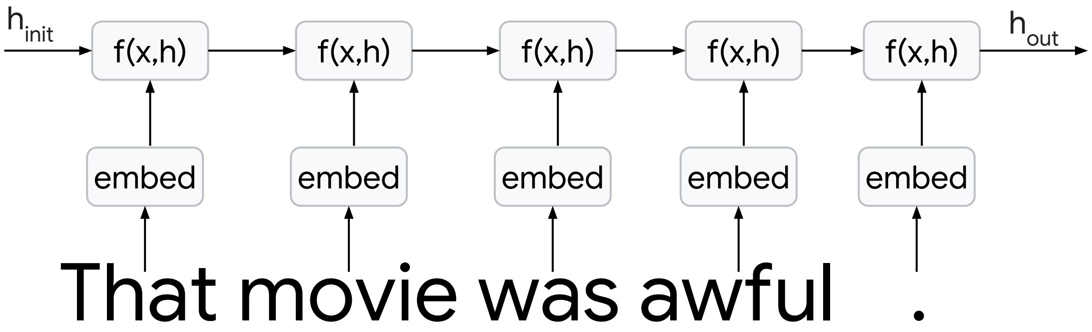


Continuous bag-of-words and skip-gram architectures (Mikolov et al., 2013a; 2013b)

Bag of N-Grams

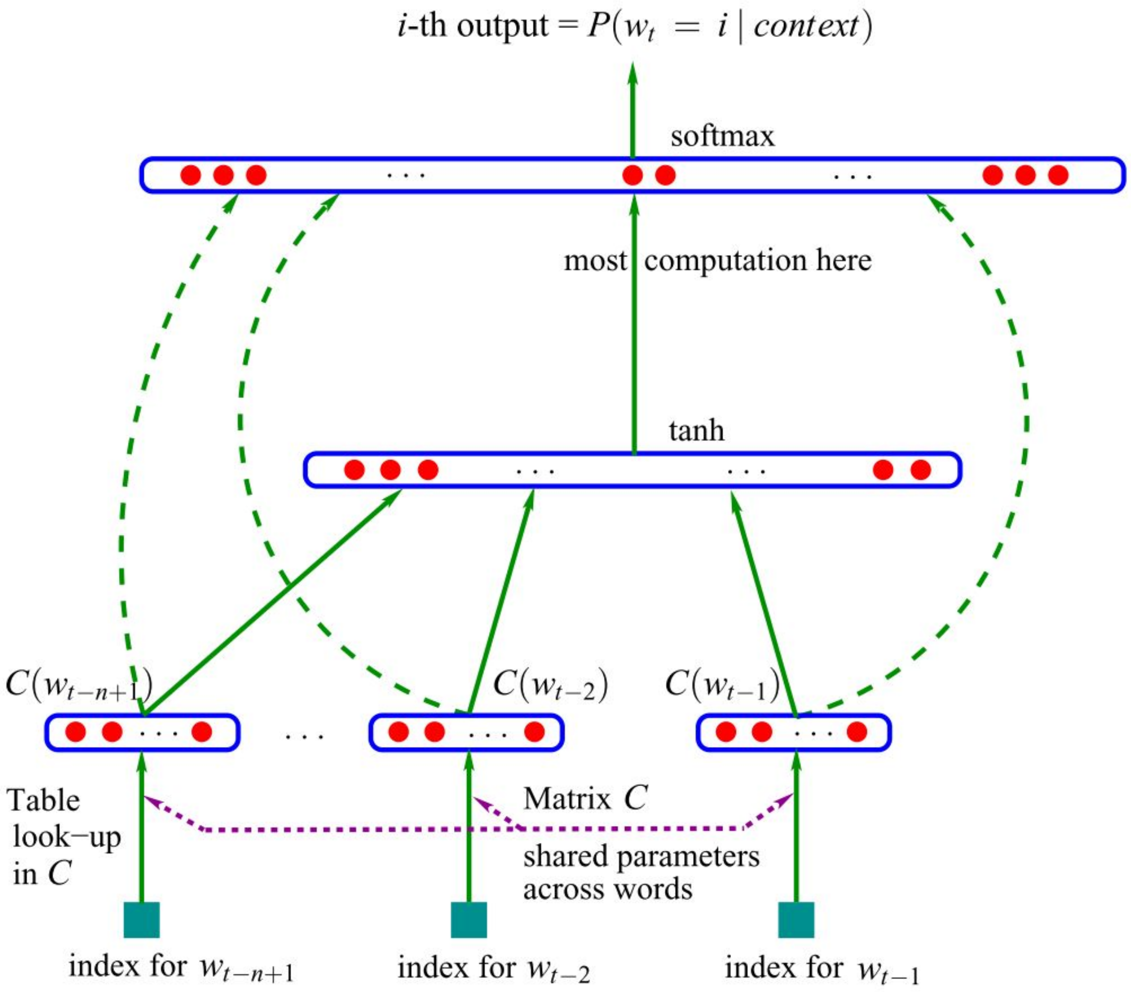


Recurrent Neural Networks



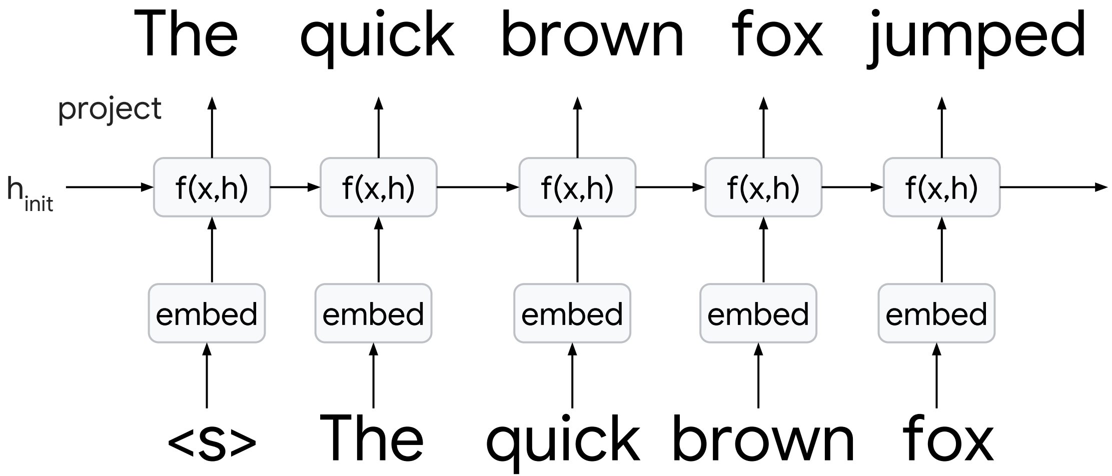
The quick brown fox jumped over the lazy ____

The quick brown fox jumped over the lazy dog

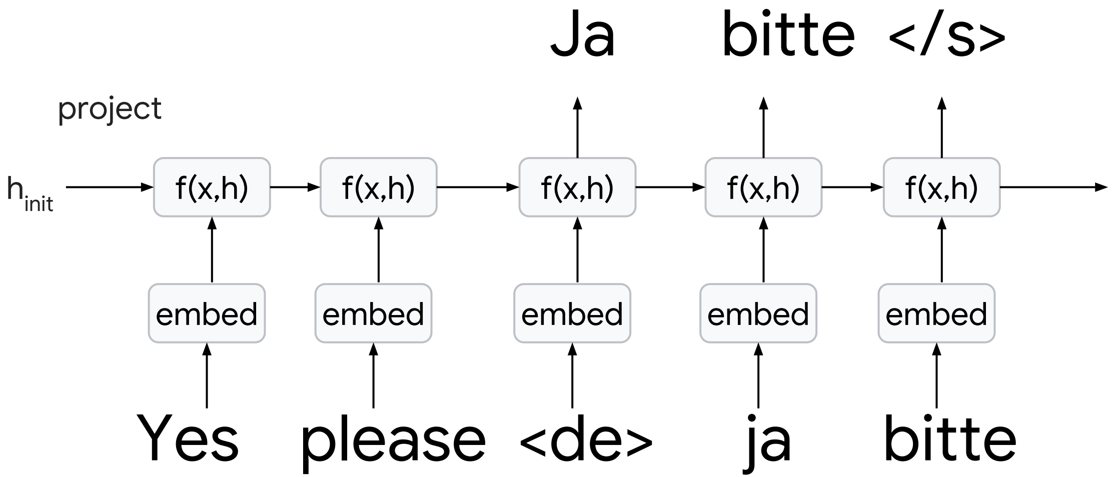


A feed-forward neural network language model (Bengio et al., 2001; 2003)

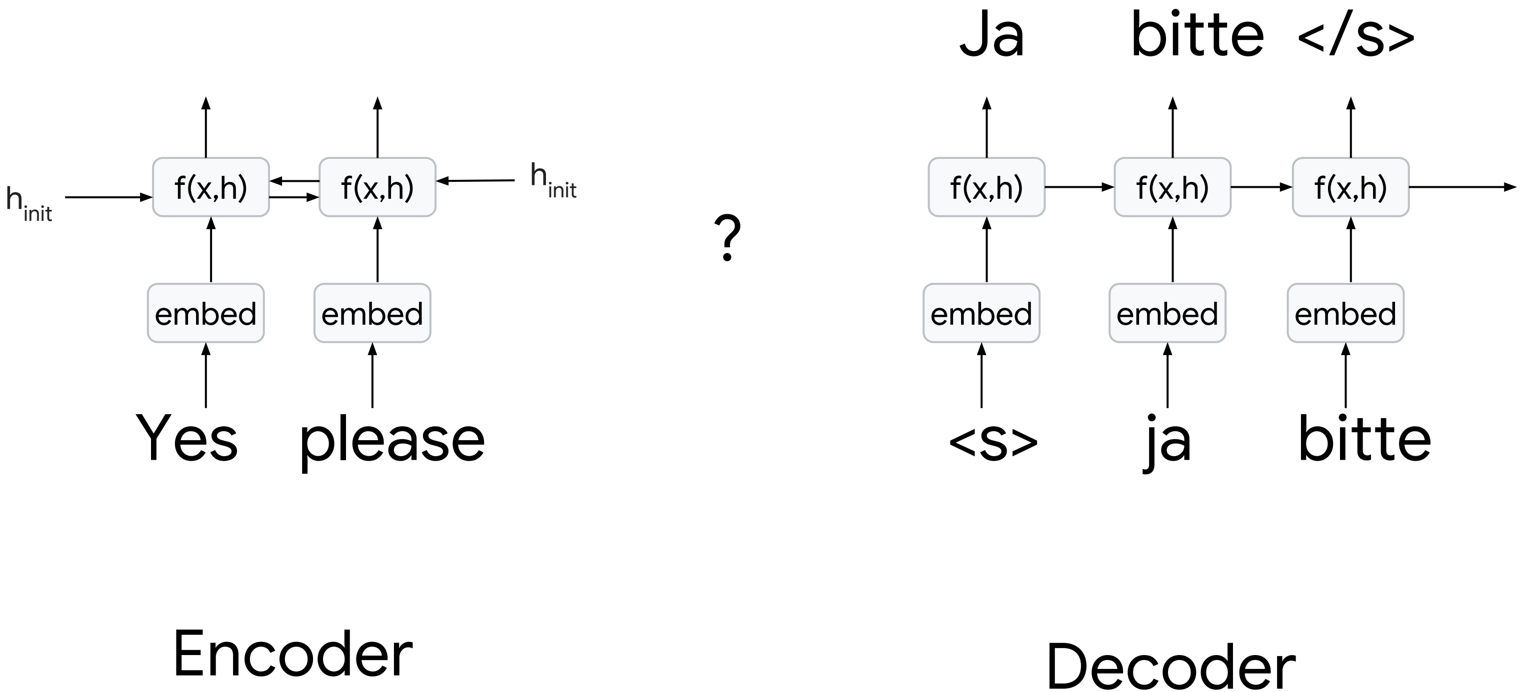
Language Models

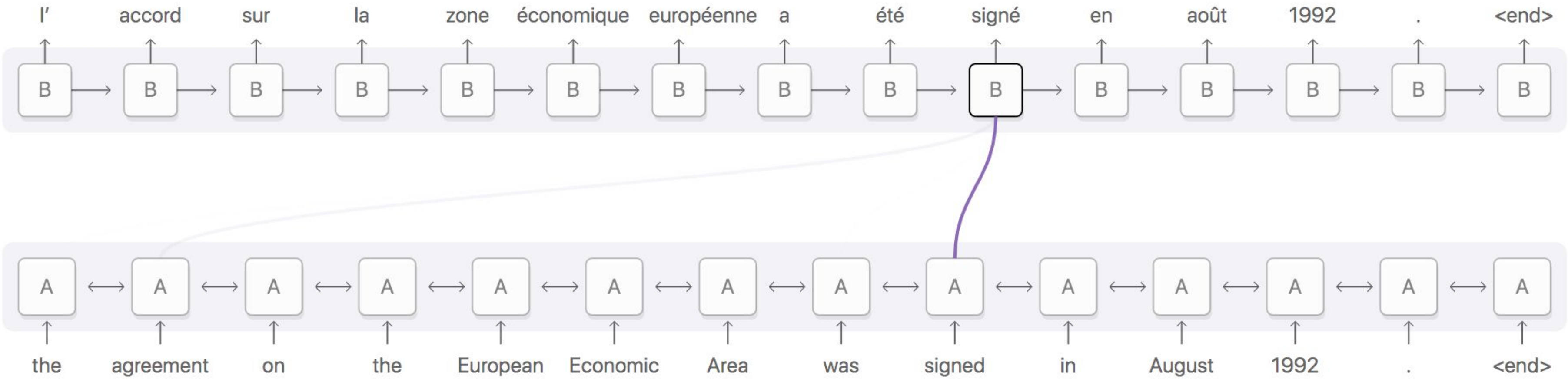


~~Language Models~~ Seq2Seq



Seq2Seq + Attention



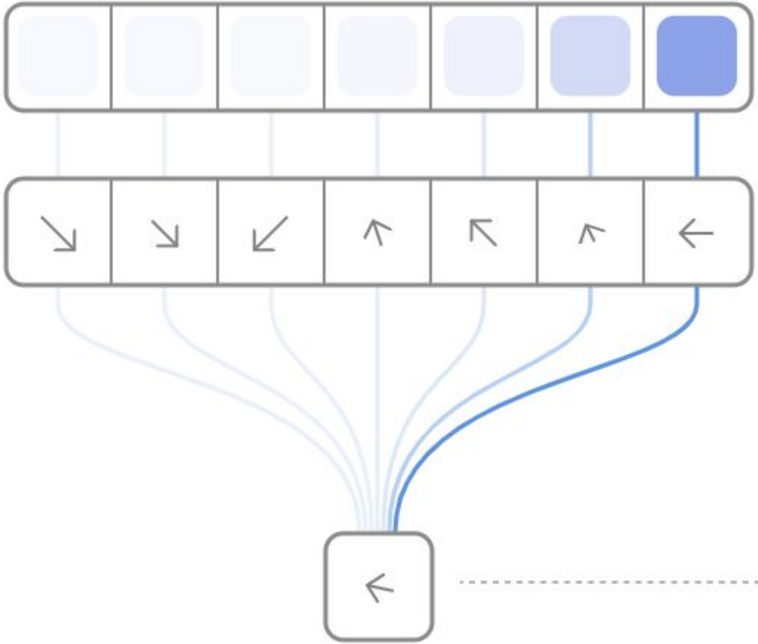


A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

https://distill.pub/2016/augmented-rnns/#



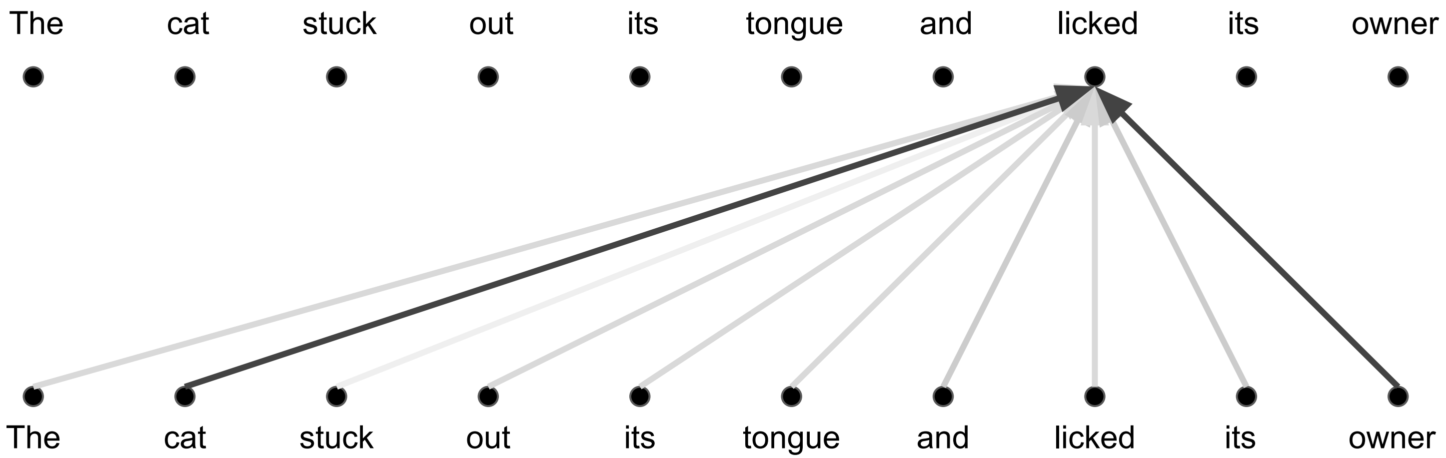
The RNN gives an attention distribution which describe how we spread out the amount we care about different memory positions.

The read result is a weighted sum.

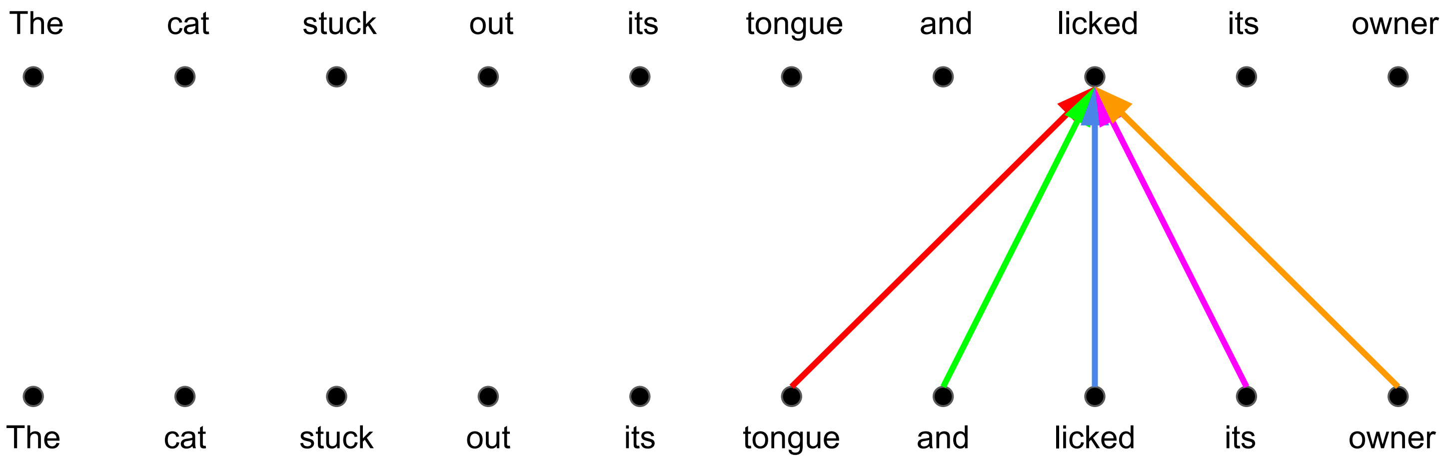
$$r \leftarrow \sum_i a_i M_i$$

https://distill.pub/2016/augmented-rnns/#

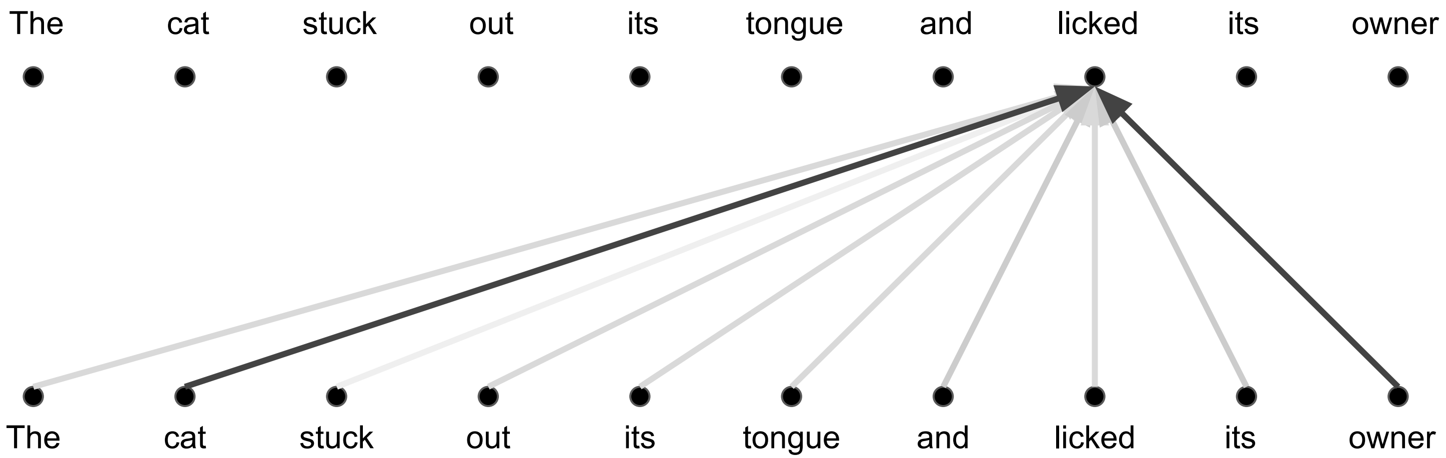
Attention: a weighted average



Convolution:
Different linear transformations by relative position.

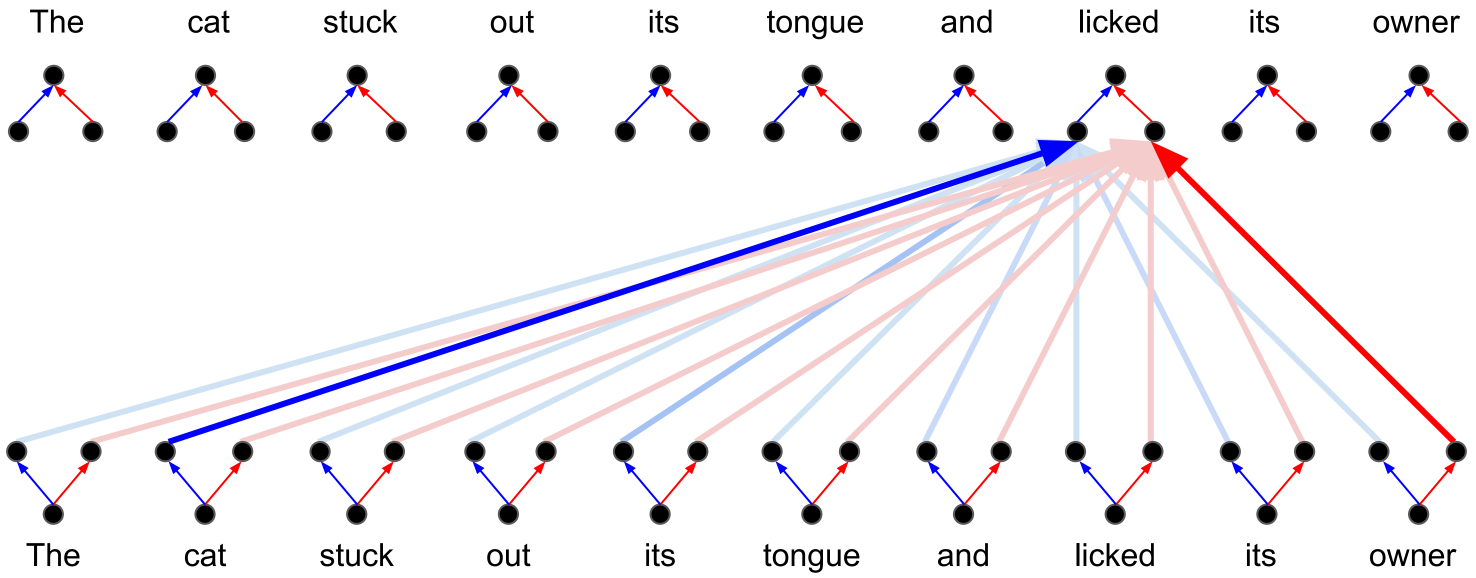


Attention: a weighted average

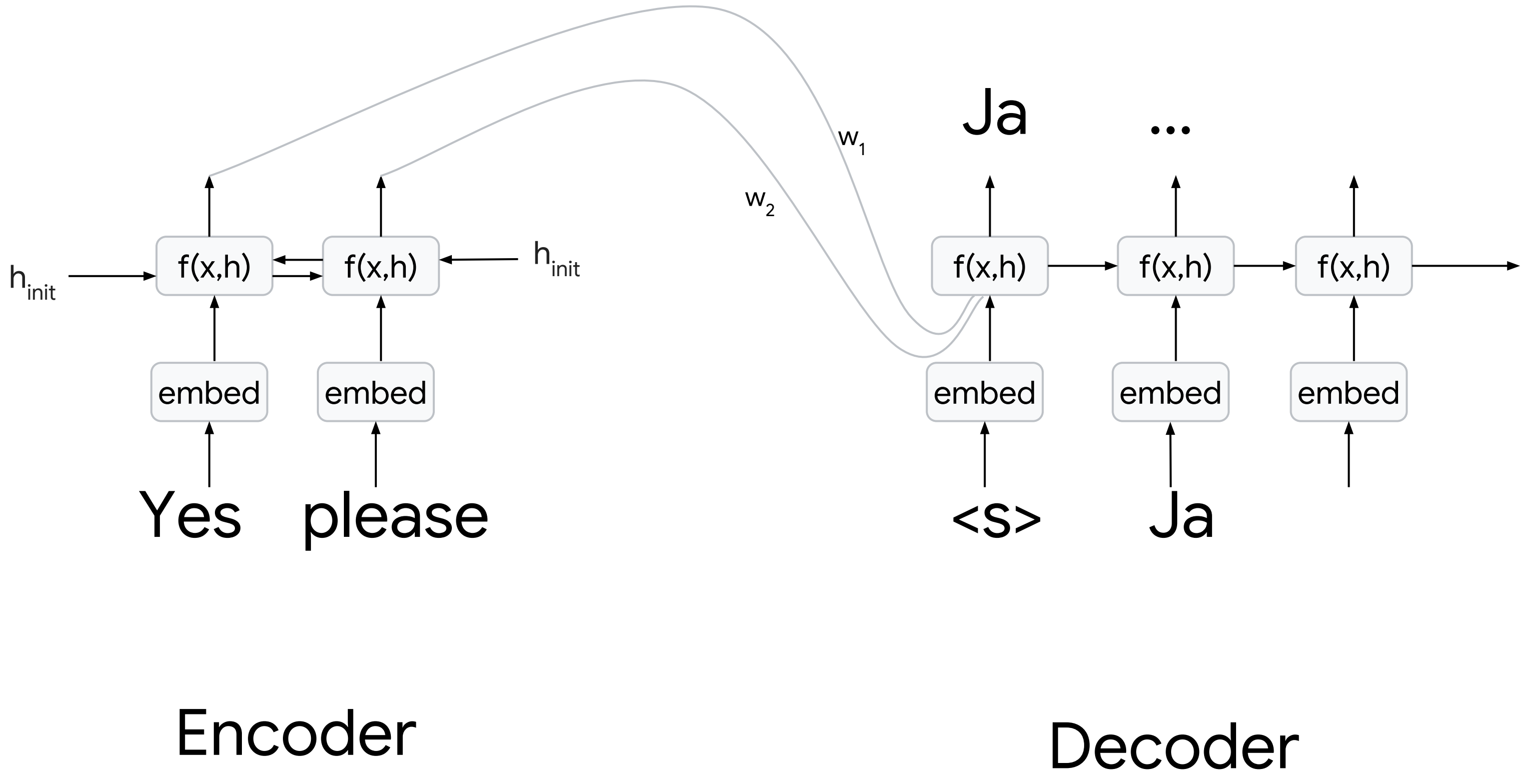


Multi-head Attention

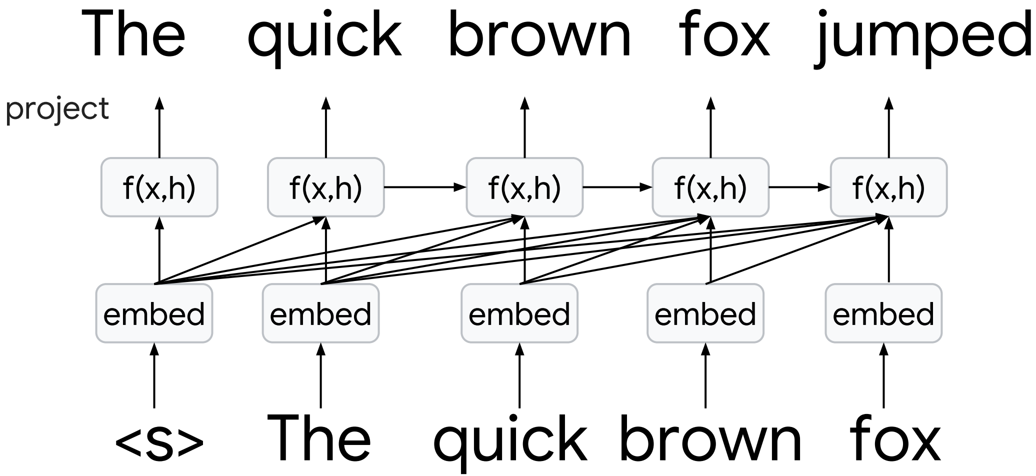
Parallel attention layers with different linear transformations on input and output.



Seq2Seq + Attention



Language Models with attention



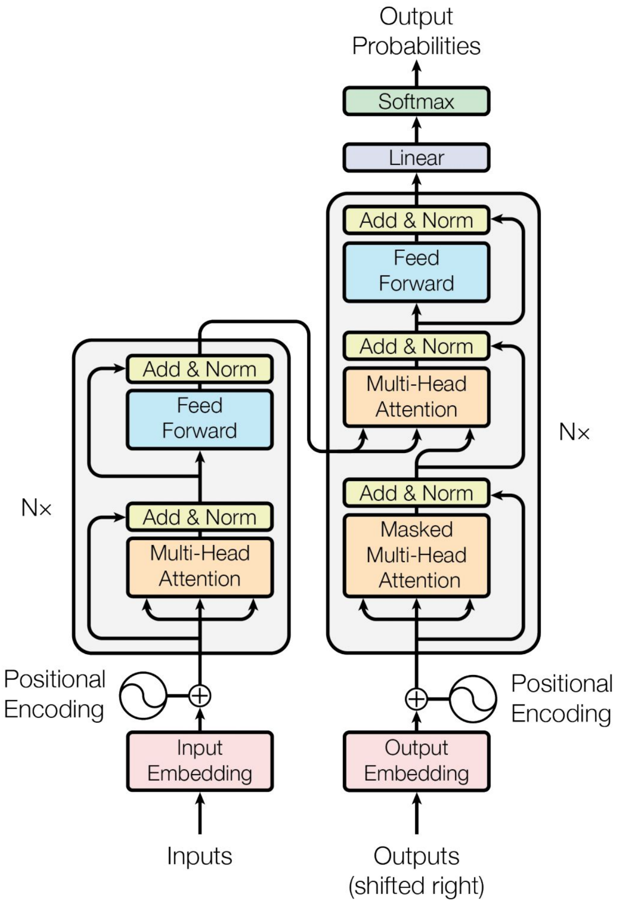
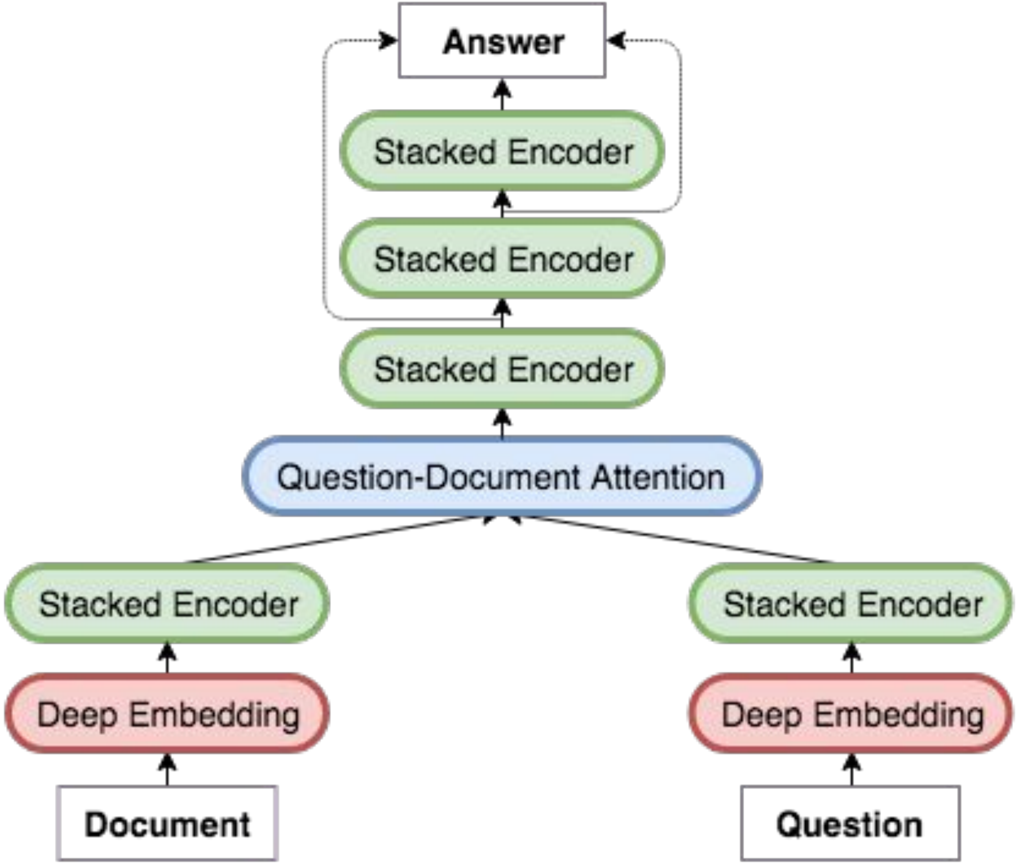


Figure 1: The Transformer - model architecture.

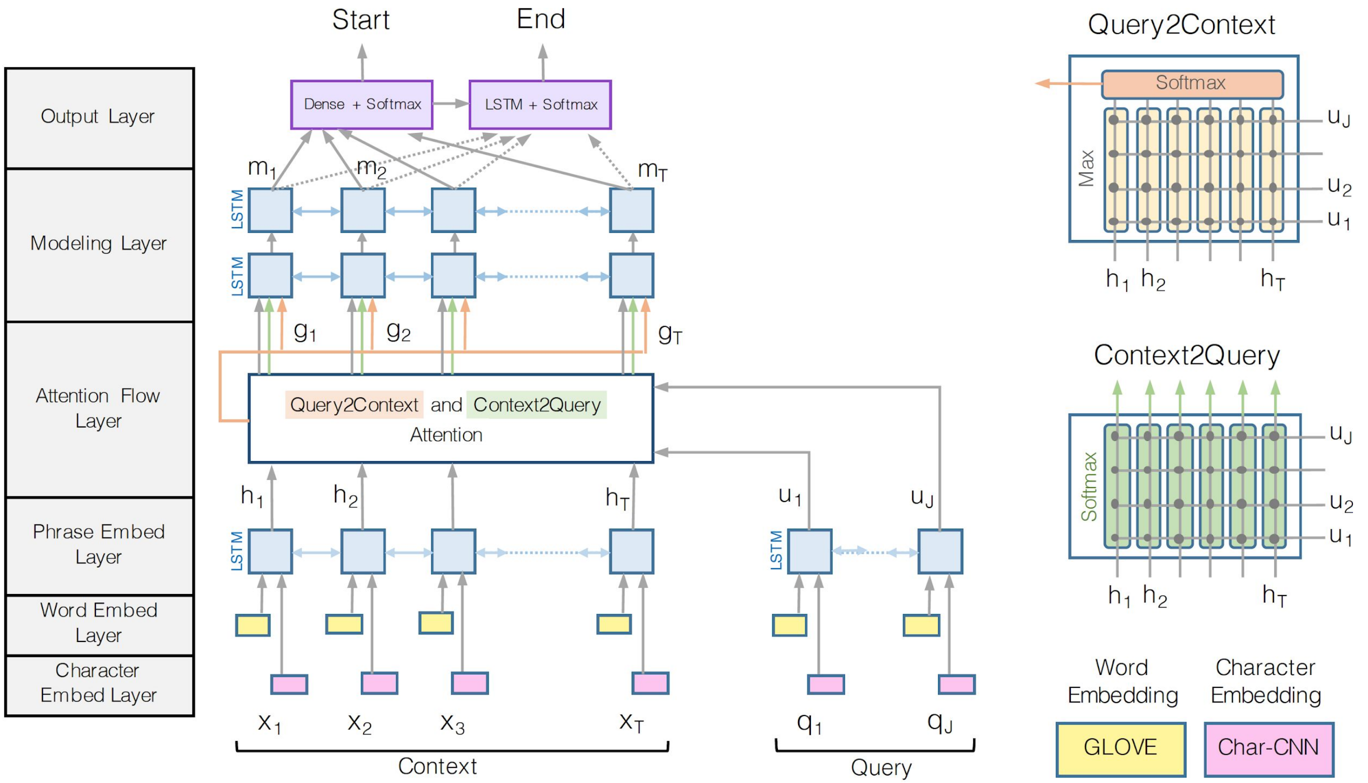
Roadmap

- Models for text
- General neural structures for QA
- Building blocks for QANet
  - Fully parallel (CNN + Self-attention)
  - data augmentation via back-translation
  - transfer learning from unsupervised tasks

General (Doc, Question) → Answer Model



General framework neural QA Systems



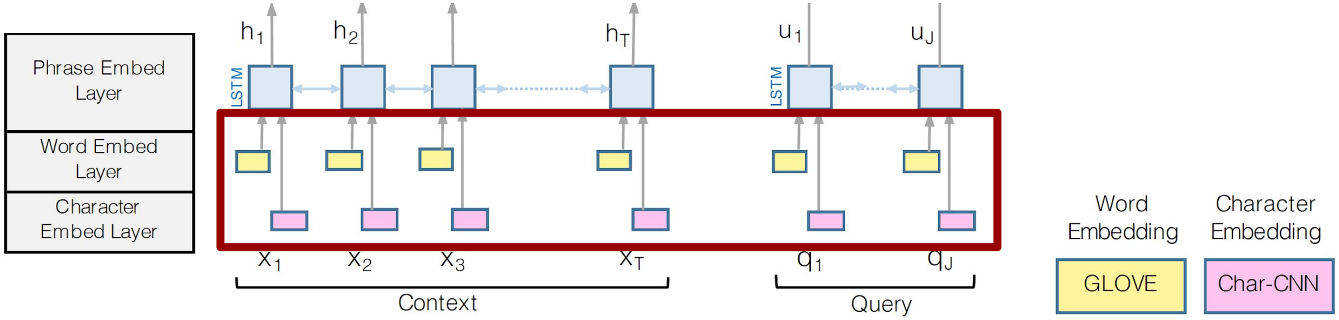
Bi-directional Attention Flow (BiDAF)

[Seo et al., ICLR'17]

Base Model (*BiDAF*)

Similar general architectures:

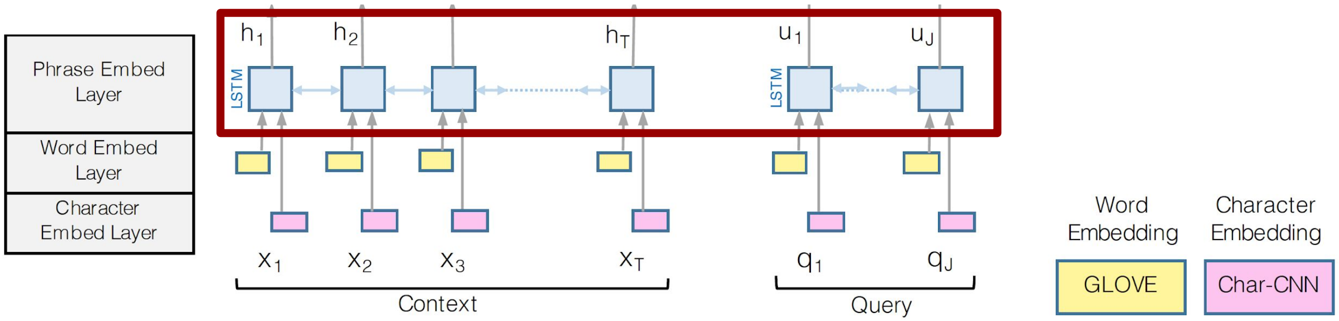
- R-Net [Wang et al, ACL'17]
- DCN [Xiong et al., ICLR'17]



Base Model (*BiDAF*)

Similar general architectures:

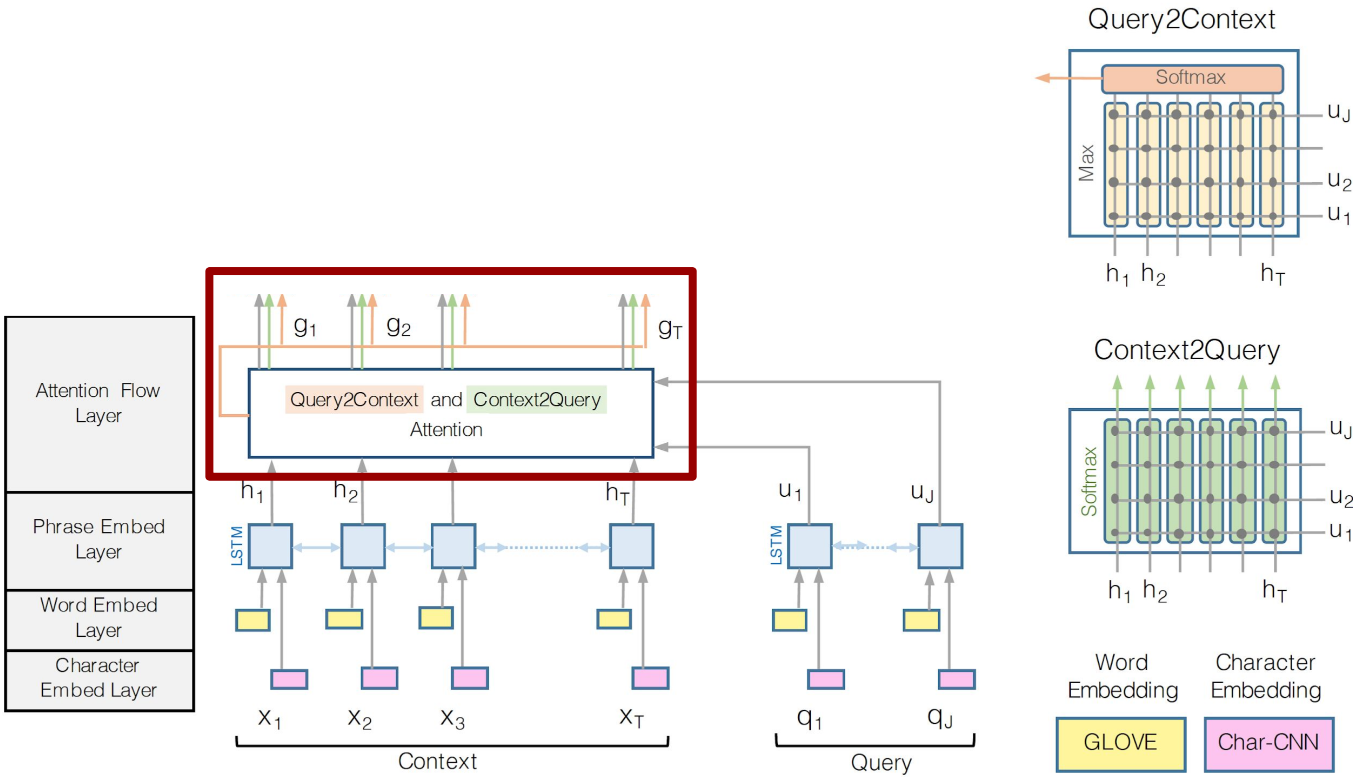
- R-Net [Wang et al, ACL'17]
- DCN [Xiong et al., ICLR'17]



Base Model (*BiDAF*)

Similar general architectures:

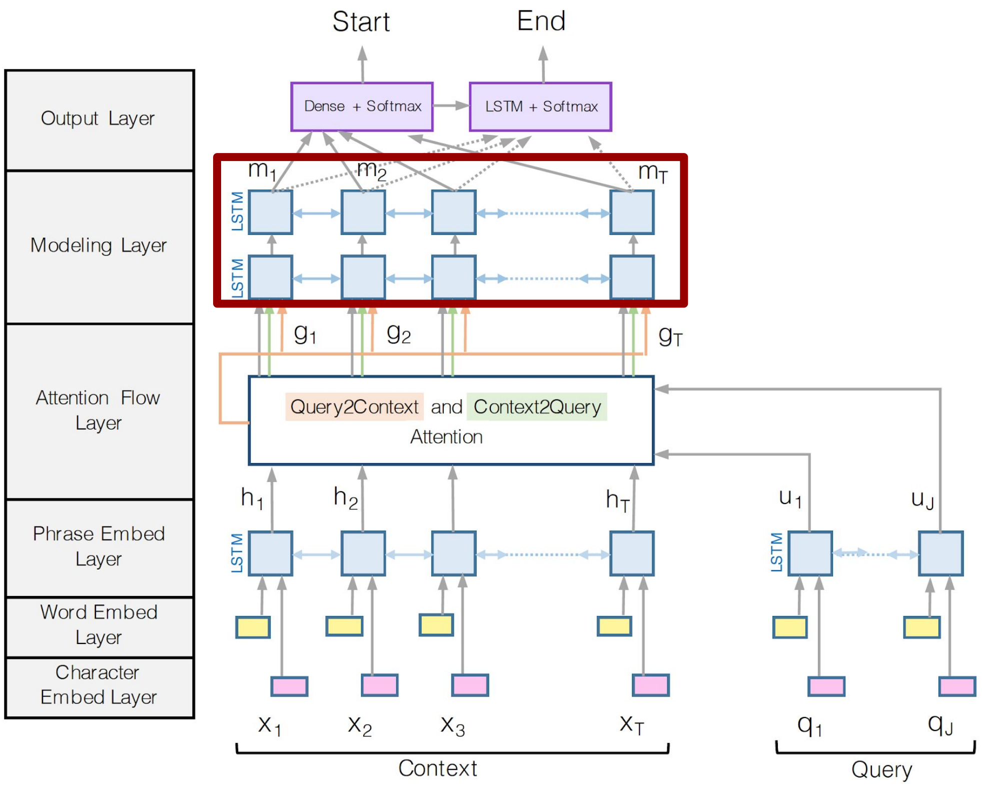
- R-Net [Wang et al, ACL'17]
- DCN [Xiong et al., ICLR'17]



Base Model (*BiDAF*)

Similar general architectures:

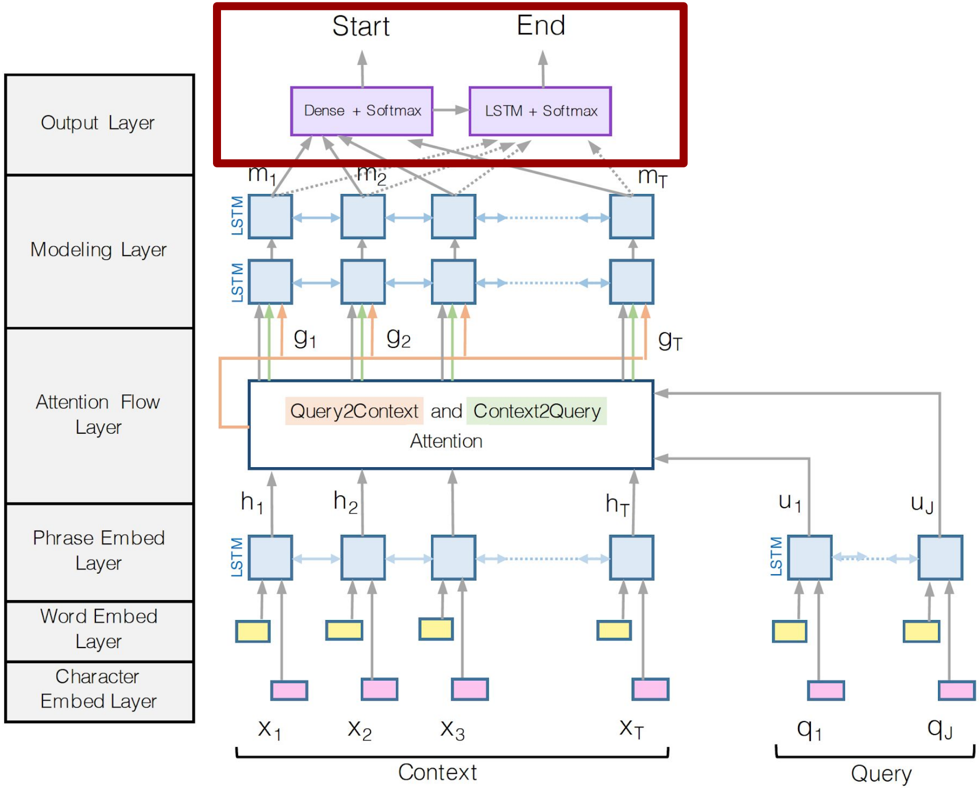
- R-Net [Wang et al, ACL'17]
- DCN [Xiong et al., ICLR'17]



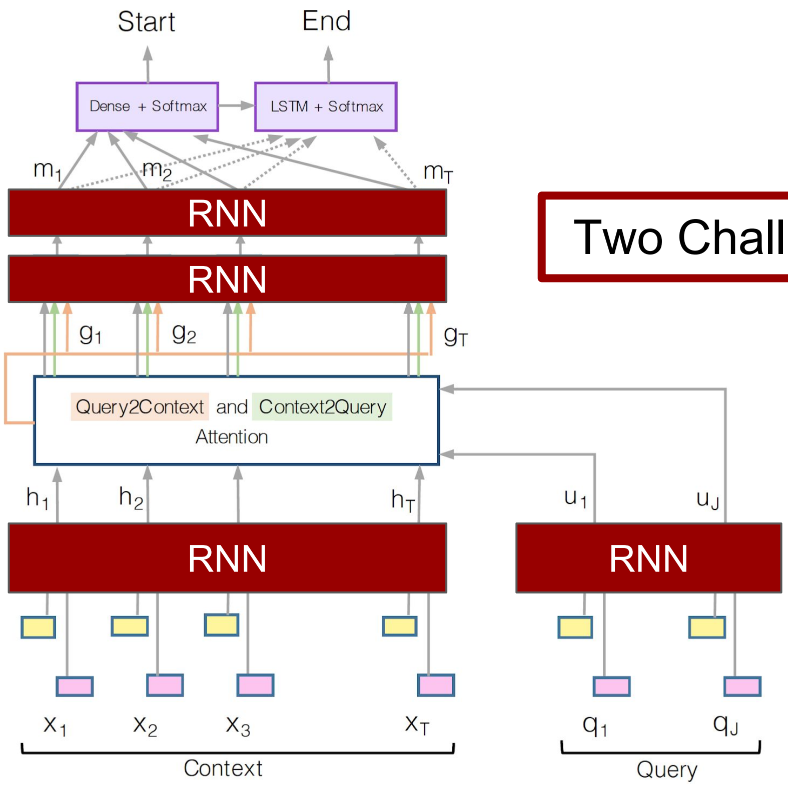
Base Model (*BiDAF*)

Similar general architectures:

- R-Net [Wang et al, ACL'17]
- DCN [Xiong et al., ICLR'17]



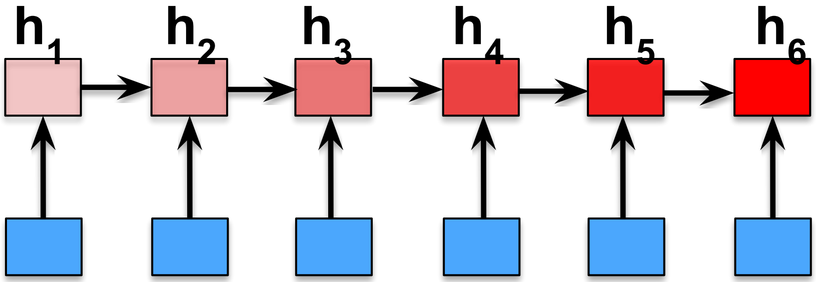
Base Model (*BiDAF*)



Two Challenges with RNNs Remain...

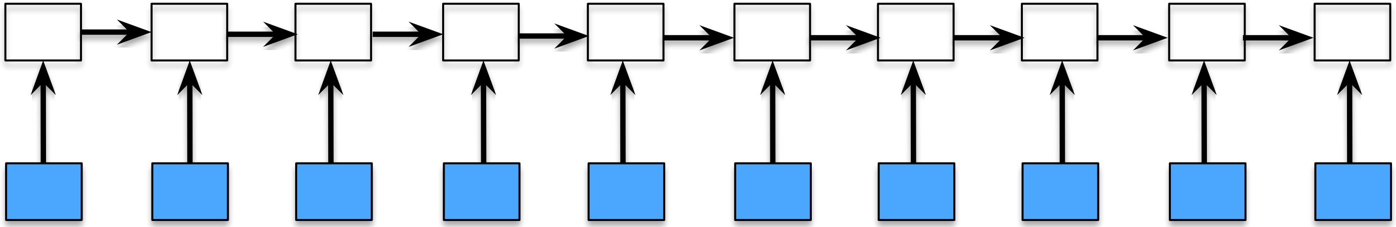
First challenge: hard to capture long dependency

Being a long-time fan of Japanese film, I expected more than this. I can't really be bothered to write too much, as this movie is just so poor. The story might be the cutest romantic little something ever, pity I couldn't stand the awful acting, the mess they called pacing, and the standard "quirky" Japanese story. If you've noticed how many Japanese movies use characters, plots and twists that seem too "different", forcedly so, then steer clear of this movie. Seriously, a 12-year old could have told you how this movie was going to move along, and that's not a good thing in my book. Fans of "Beat" Takeshi: his part in this movie is not really more than a cameo, and unless you're a rabid fan, you don't need to suffer through this waste of film.

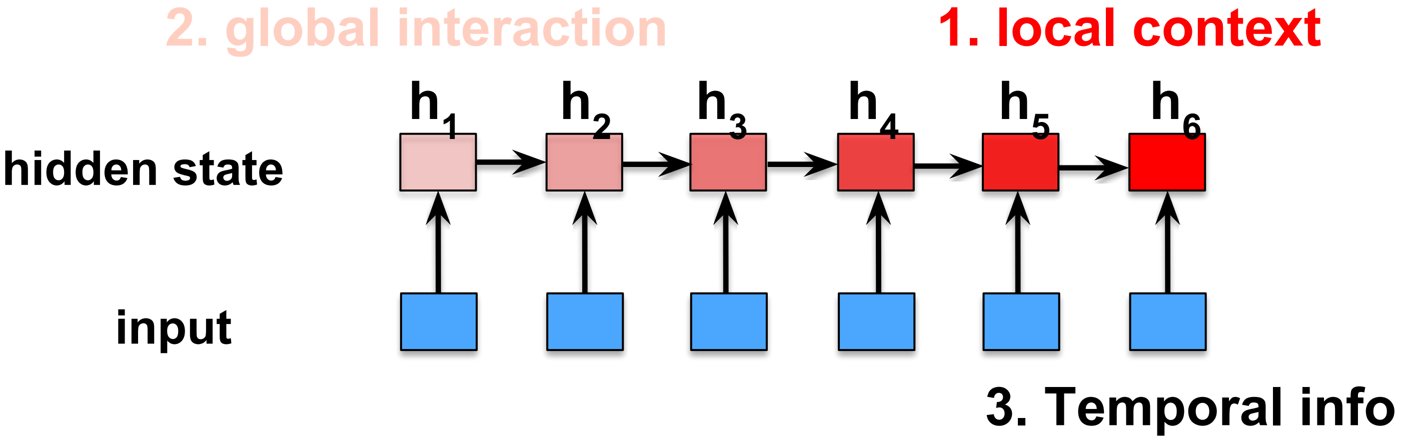


Second challenge: hard to compute in parallel

Strictly Sequential!



What do RNNs Capture?



Substitution?

Roadmap

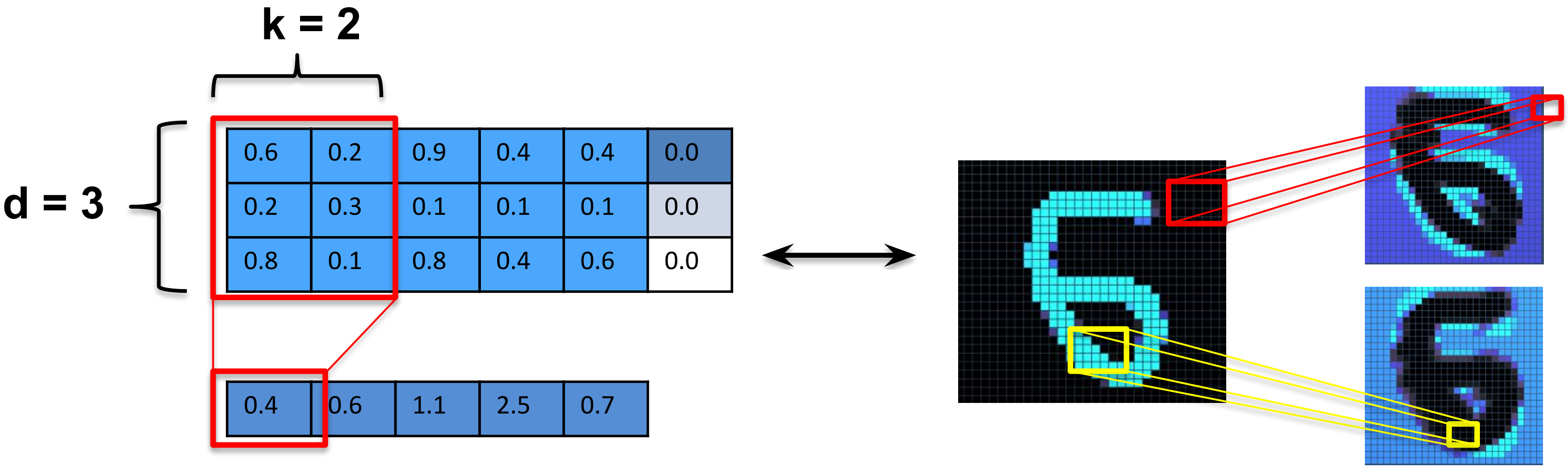
- Models for text
- General neural structures for QA
- Building blocks for QANet
  - Fully parallel (CNN + Self-attention)
  - data augmentation via back-translation
  - transfer learning from unsupervised tasks

Convolution: Capturing Local Context

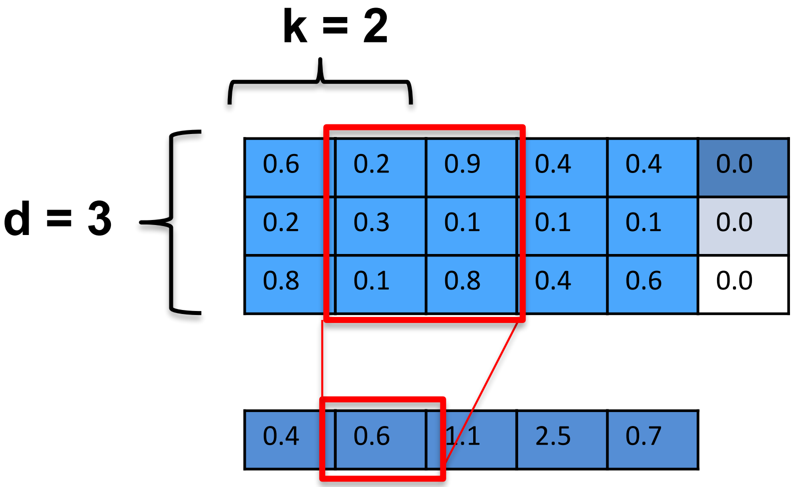
0.6	0.2	0.9	0.4	0.4
0.2	0.3	0.1	0.1	0.1
0.8	0.1	0.8	0.4	0.6
The	**weather**	**is**	**nice**	**today**

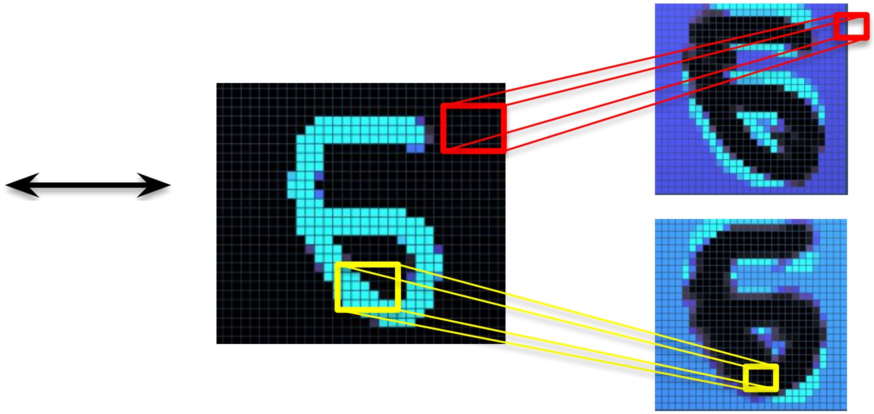
0.6	0.2	0.9	0.4	0.4
0.2	0.3	0.1	0.1	0.1
0.8	0.1	0.8	0.4	0.6

Convolution: Capturing Local Context

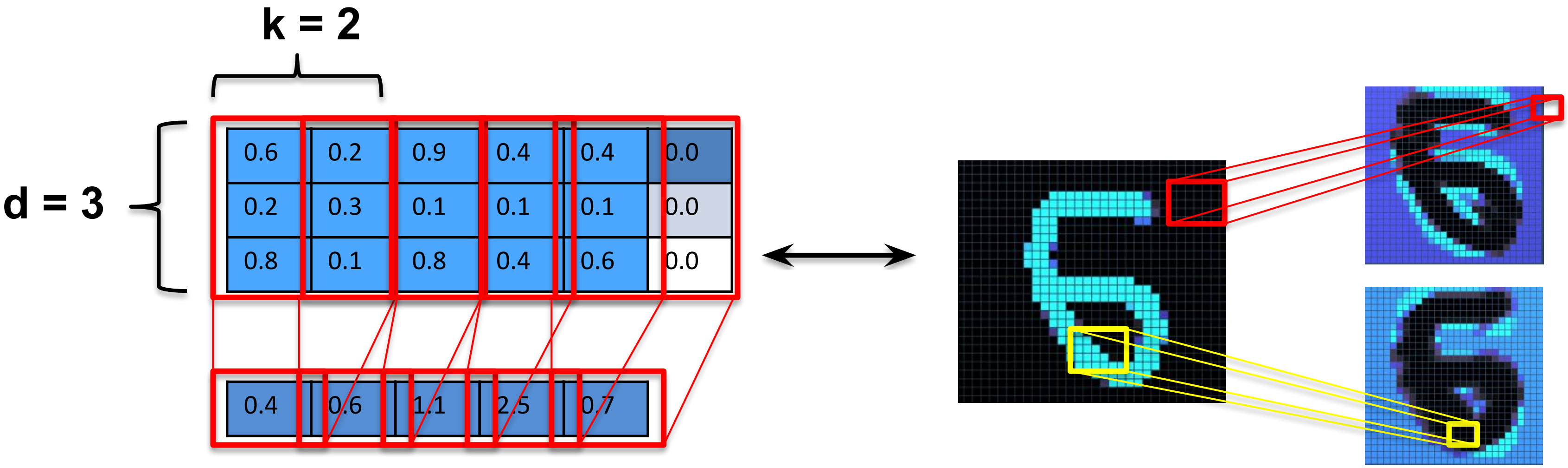


Convolution: Capturing Local Context

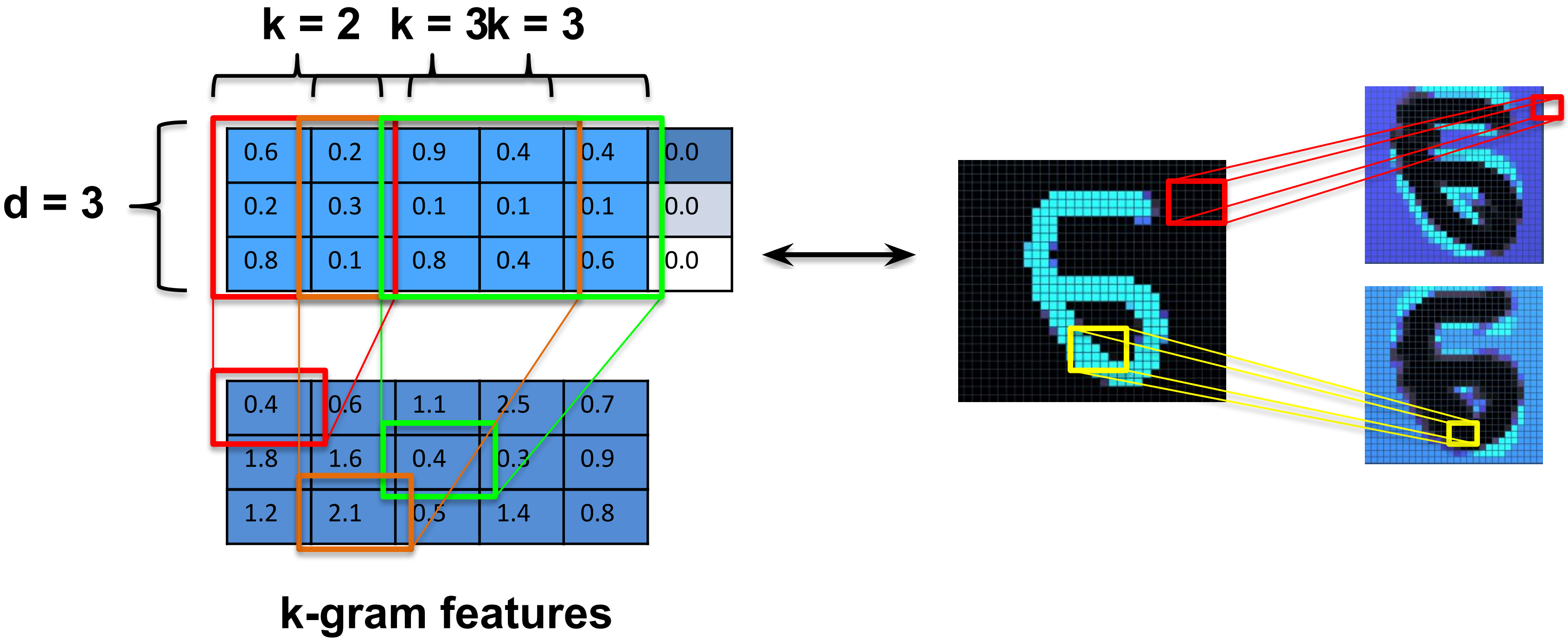




Convolution: Capturing Local Context

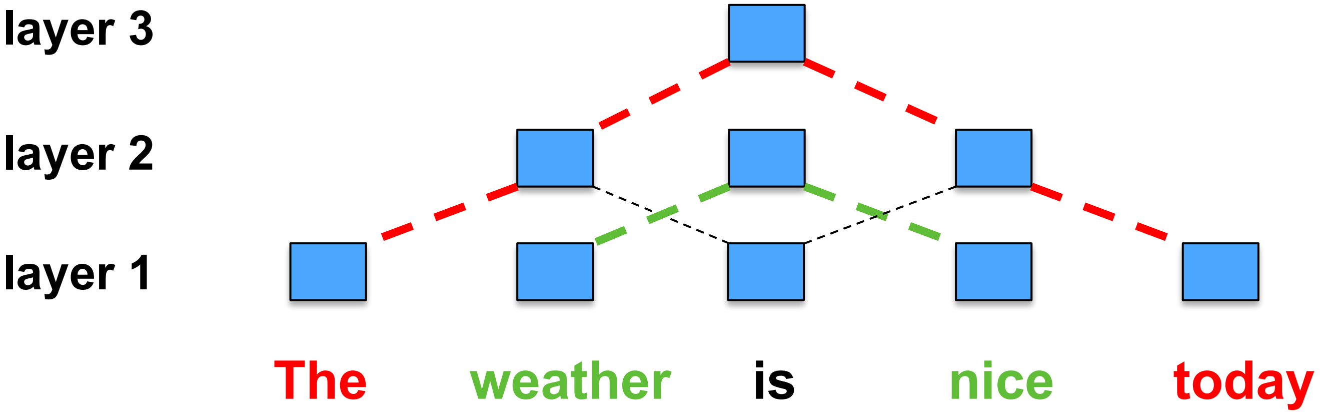


Convolution: Capturing Local Context



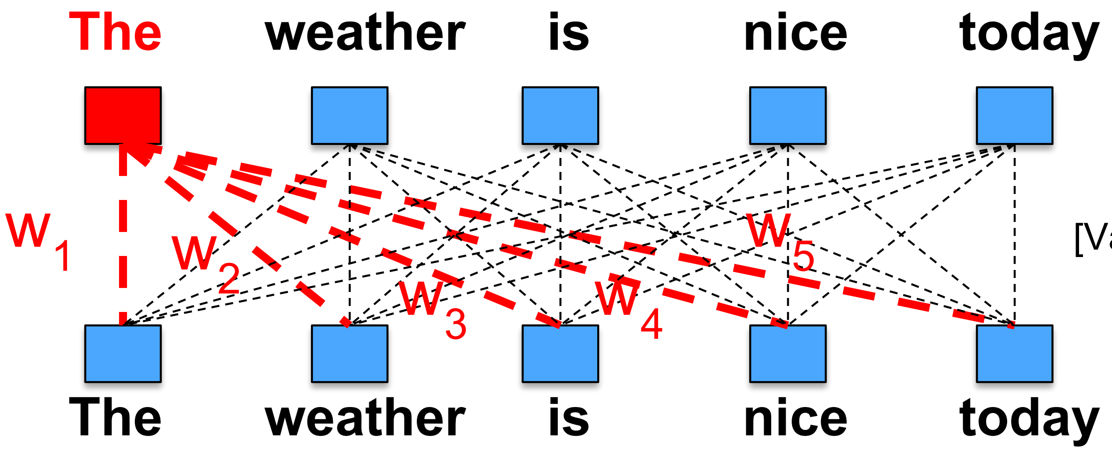
Fully parallel!

How about Global Interaction?



N: Seq length.
k: Filter size.

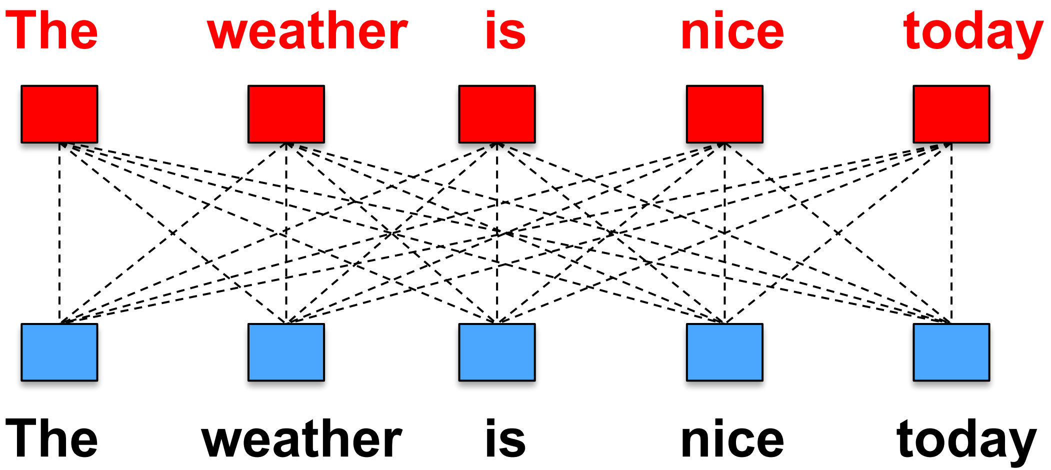
1. May need $O(\log_k N)$ layers
2. Interaction may become weaker



[Vaswani et al., NIPS'17]

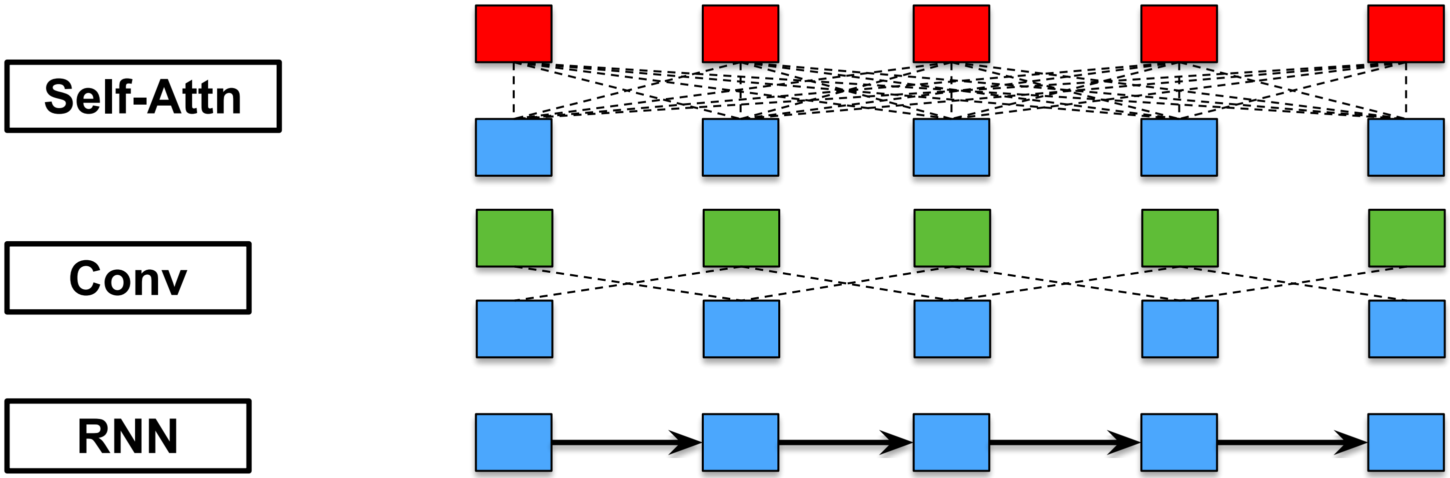
$$\underbrace{\begin{bmatrix}1.8\\2.3\\0.4\end{bmatrix}}_{\text{The}} = w_1 \times \underbrace{\begin{bmatrix}0.6\\0.2\\0.8\end{bmatrix}}_{\text{The}} + w_2 \times \underbrace{\begin{bmatrix}0.2\\0.3\\0.1\end{bmatrix}}_{\text{weather}} + w_3 \times \underbrace{\begin{bmatrix}0.9\\0.1\\0.8\end{bmatrix}}_{\text{is}} + w_4 \times \underbrace{\begin{bmatrix}0.4\\0.1\\0.4\end{bmatrix}}_{\text{nice}} + w_5 \times \underbrace{\begin{bmatrix}0.4\\0.1\\0.6\end{bmatrix}}_{\text{today}}$$

$$w_1, w_2, w_3, w_4, w_5 = \mathrm{softmax}\left(\underbrace{\begin{bmatrix}0.6 & 0.2 & 0.8\end{bmatrix}}_{\text{The}} \times \begin{array}{c}\begin{bmatrix}0.6 & 0.2 & 0.9 & 0.4 & 0.4\\ 0.2 & 0.3 & 0.1 & 0.1 & 0.1\\ 0.8 & 0.1 & 0.8 & 0.4 & 0.6\end{bmatrix}\\ \begin{matrix}\text{The} & \text{weather} & \text{is} & \text{nice} & \text{today}\end{matrix}\end{array} \right)$$



Self-attention is fully parallel & all-to-all!

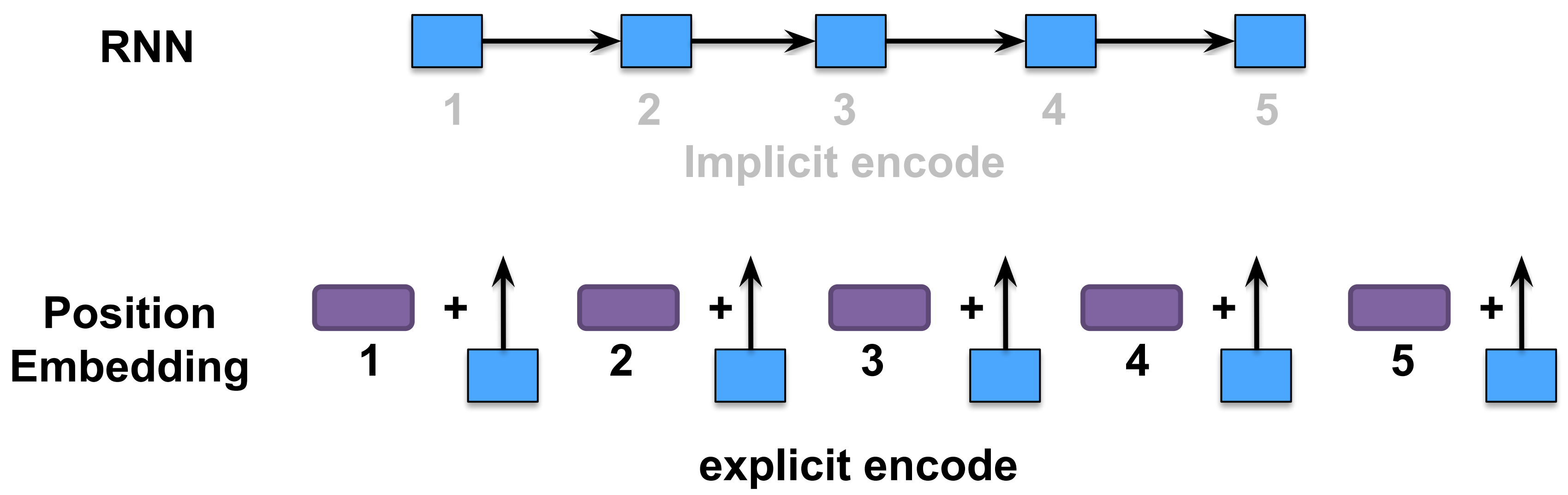
Complexity



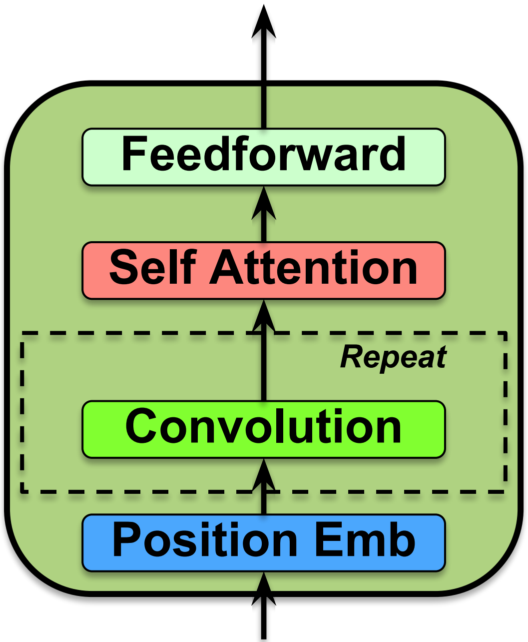
N: Seq length.
d: Dim. (N > d)
k: Filter size.

	Per Unit	Total Per Layer	Sequential Op (Path Memory)
Self-Attn	$O(Nd)$	$O(N^2d)$	$O(1)$
Conv	$O(kd^2)$	$O(kNd^2)$	$O(1)$
RNN	$O(d^2)$	$O(Nd^2)$	$O(N)$

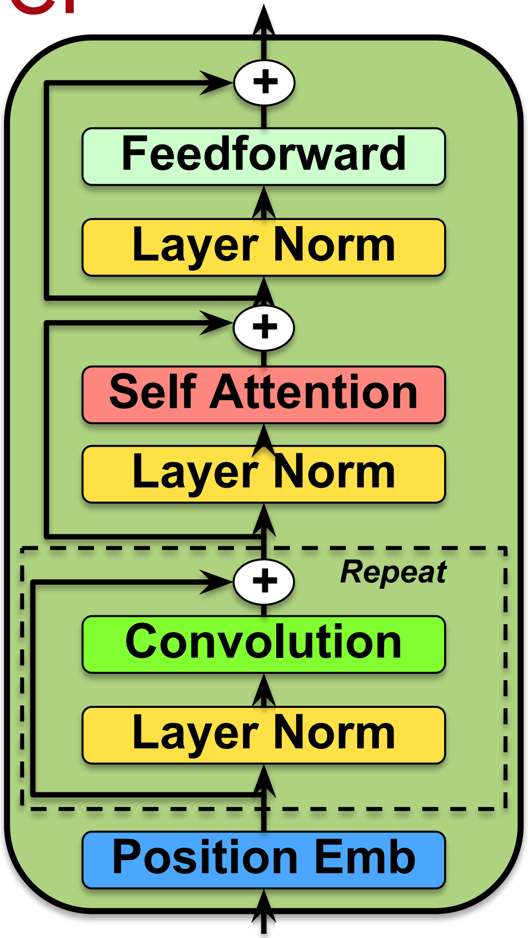
Explicitly Encode Temporal Info



QANet Encoder

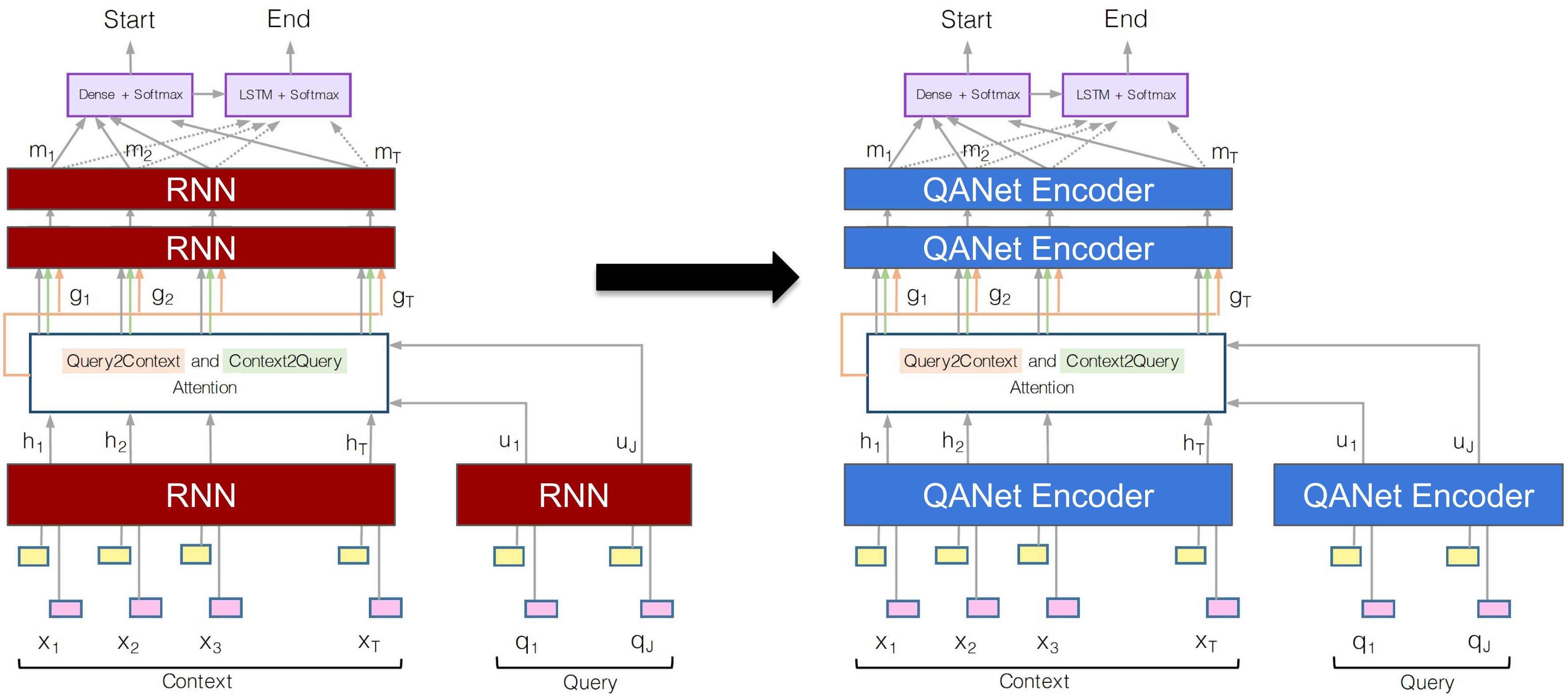


if you want to go deeper



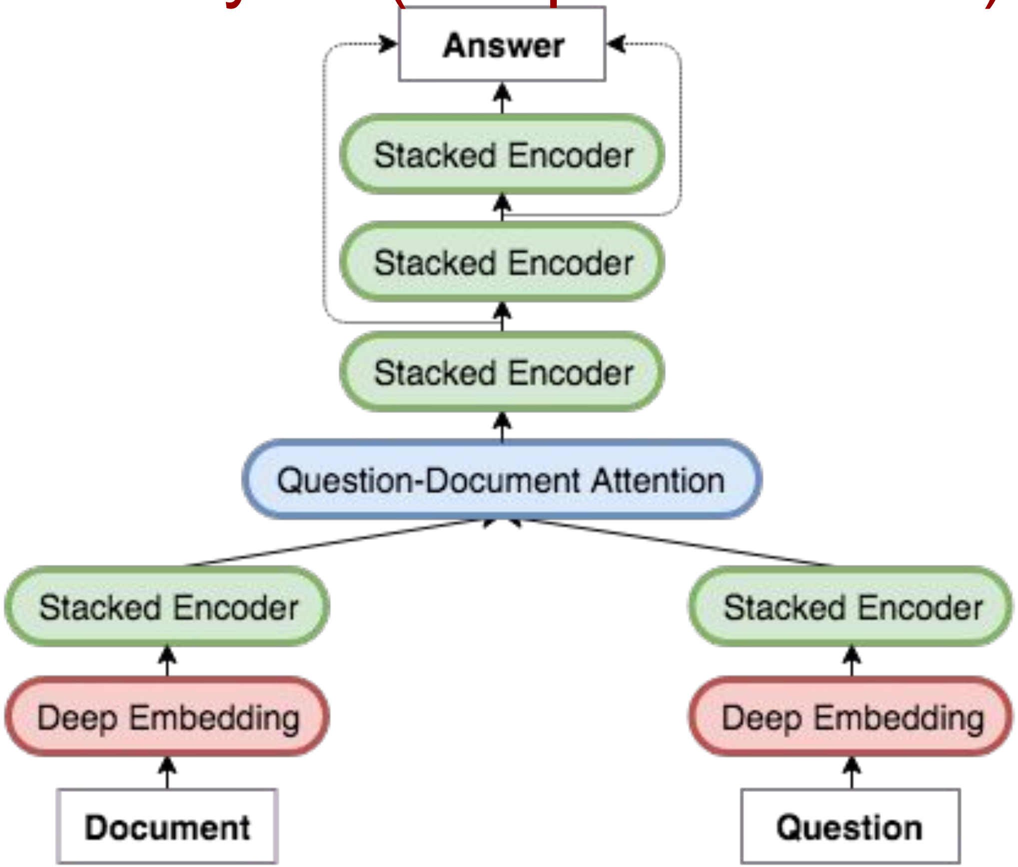
[Yu et al., ICLR'18]

Base Model (*BiDAF*) → QANet

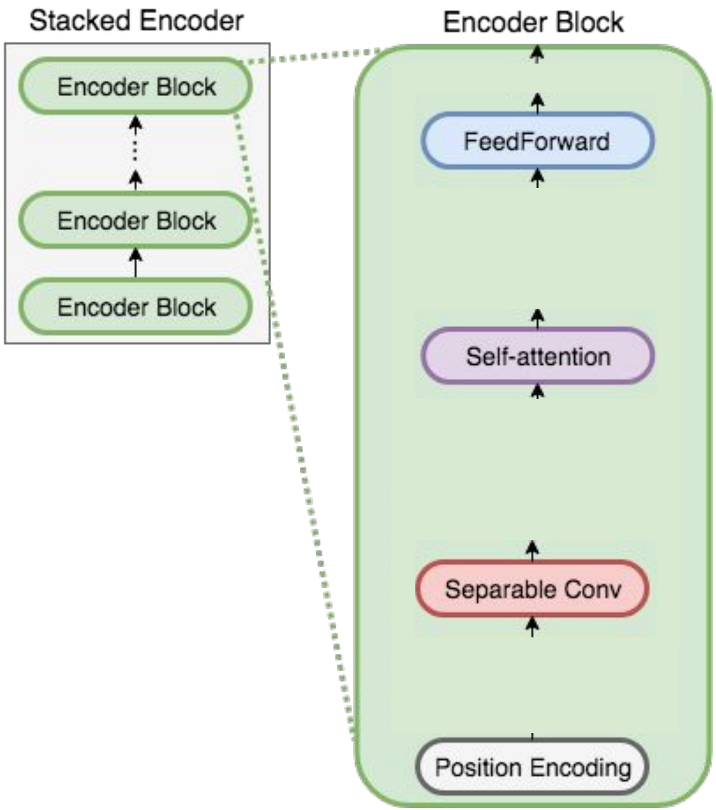


QANet – 130+ layers (Deepest NLP NN)

130 layers

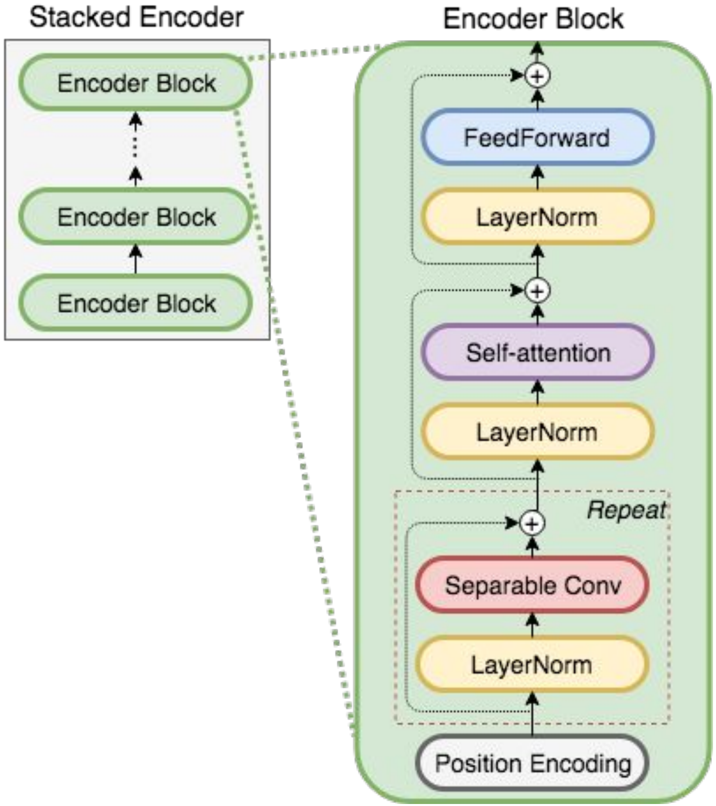


QANet – First QA system with *No Recurrence*



- **Very fast!**
  - Training: 3x - 13x
  - Inference: 4x - 9x

QANet – 130+ layers (Deepest NLP NN)

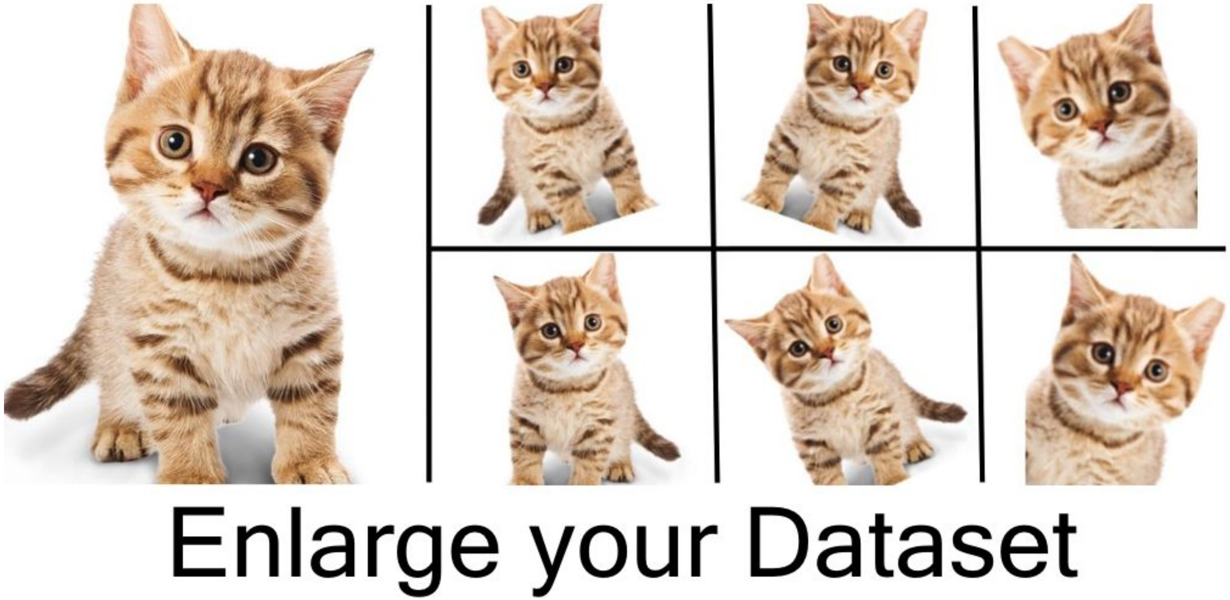


- Layer normalization
- Residual connections
- L_2 regularization
- Stochastic Depth
- Squeeze and Excitation
- ...

Roadmap

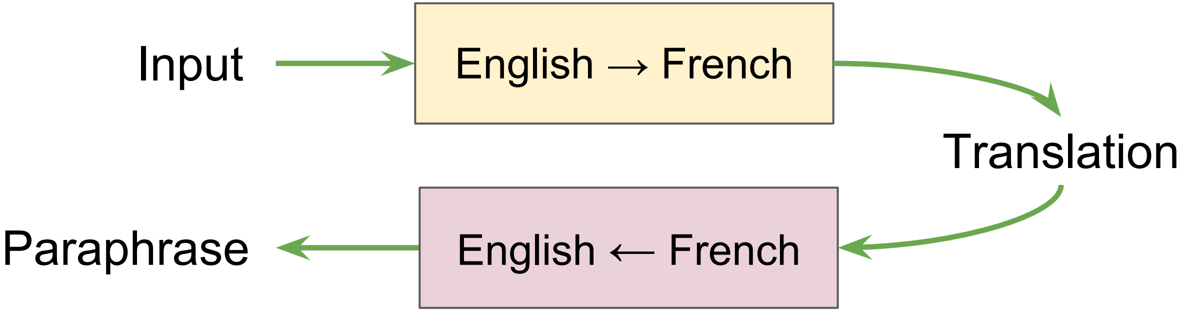
- Models for text
- General neural structures for QA
- Building blocks for QANet
  - Fully parallel (CNN + Self-attention)
  - data augmentation via back-translation
  - transfer learning from unsupervised tasks

Data augmentation: popular in vision & speech



More data with NMT back-translation

Previously, tea had been used primarily for Buddhist monks to stay awake during meditation.

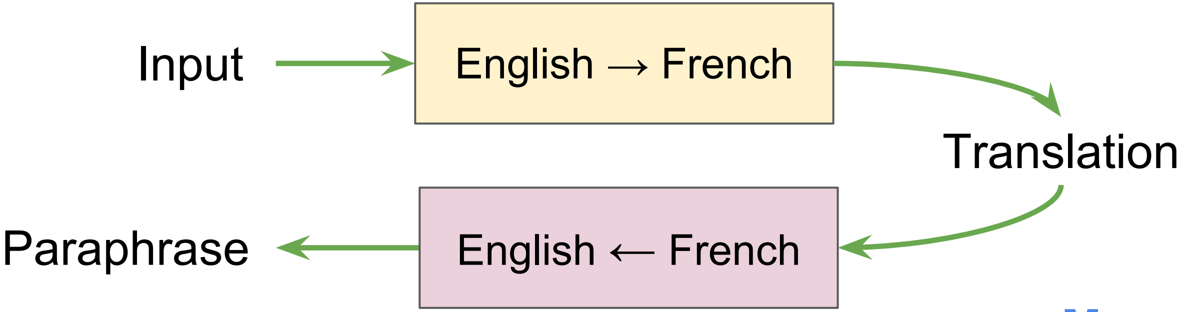


Autrefois, le thé avait été utilisé surtout pour les moines bouddhistes pour rester éveillé pendant la méditation.

In the past, tea was used mostly for Buddhist monks to stay awake during the meditation.

More data with NMT back-translation

Previously, tea had been used primarily for Buddhist monks to stay awake during meditation.

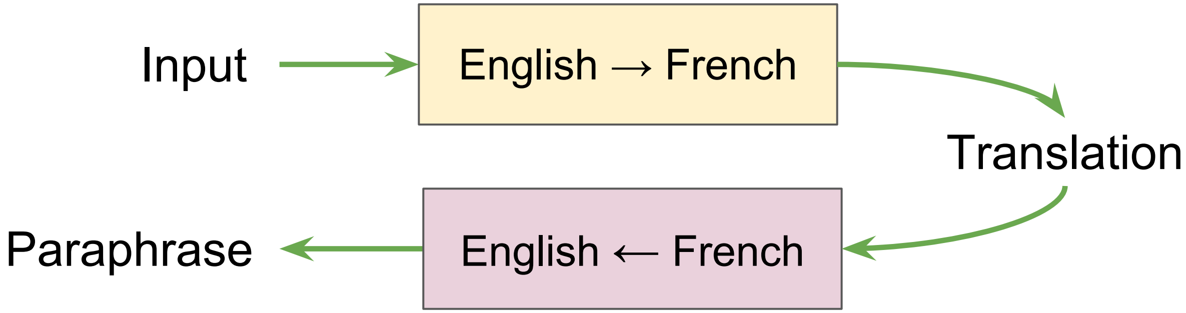


In the past, tea was used mostly for Buddhist monks to stay awake during the meditation.

- **More data**
  - (Input, *label*)
  - (Paraphrase, *label*)

Applicable to virtually any NLP tasks!

QANet augmentation



Use 2 language pairs: *English-French, English-German*. **3x data**.

Improvement: +1.1 F1

Roadmap

- Models for text
- General neural structures for QA
- Building blocks for QANet
    - Fully parallel (CNN + Self-attention)
    - data augmentation via back-translation
    - transfer learning from unsupervised tasks

ELMo

Deep contextualized word representations

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner,
Christopher Clark, Kent Zettlemoyer.

Universal Language Model Fine-tuning for Text Classification

Jeremy Howard*
fast.ai
University of San Francisco
j@fast.ai

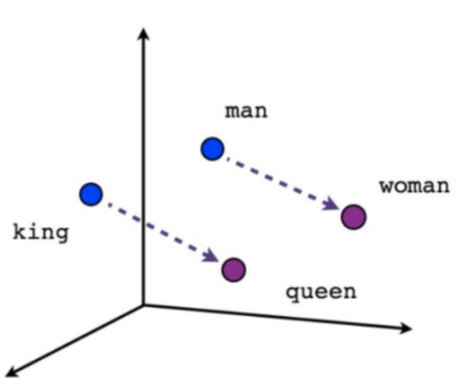
Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

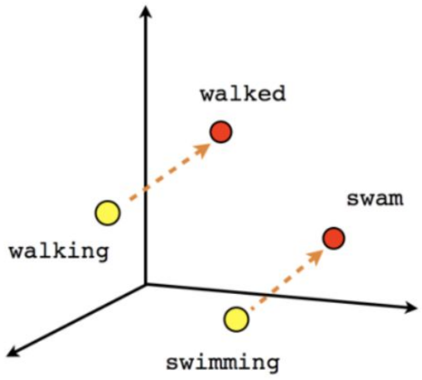
Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

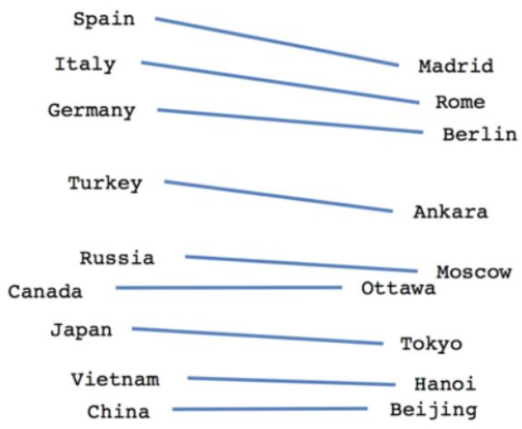
Ilya Sutskever
OpenAI
ilyasu@openai.com



Male-Female

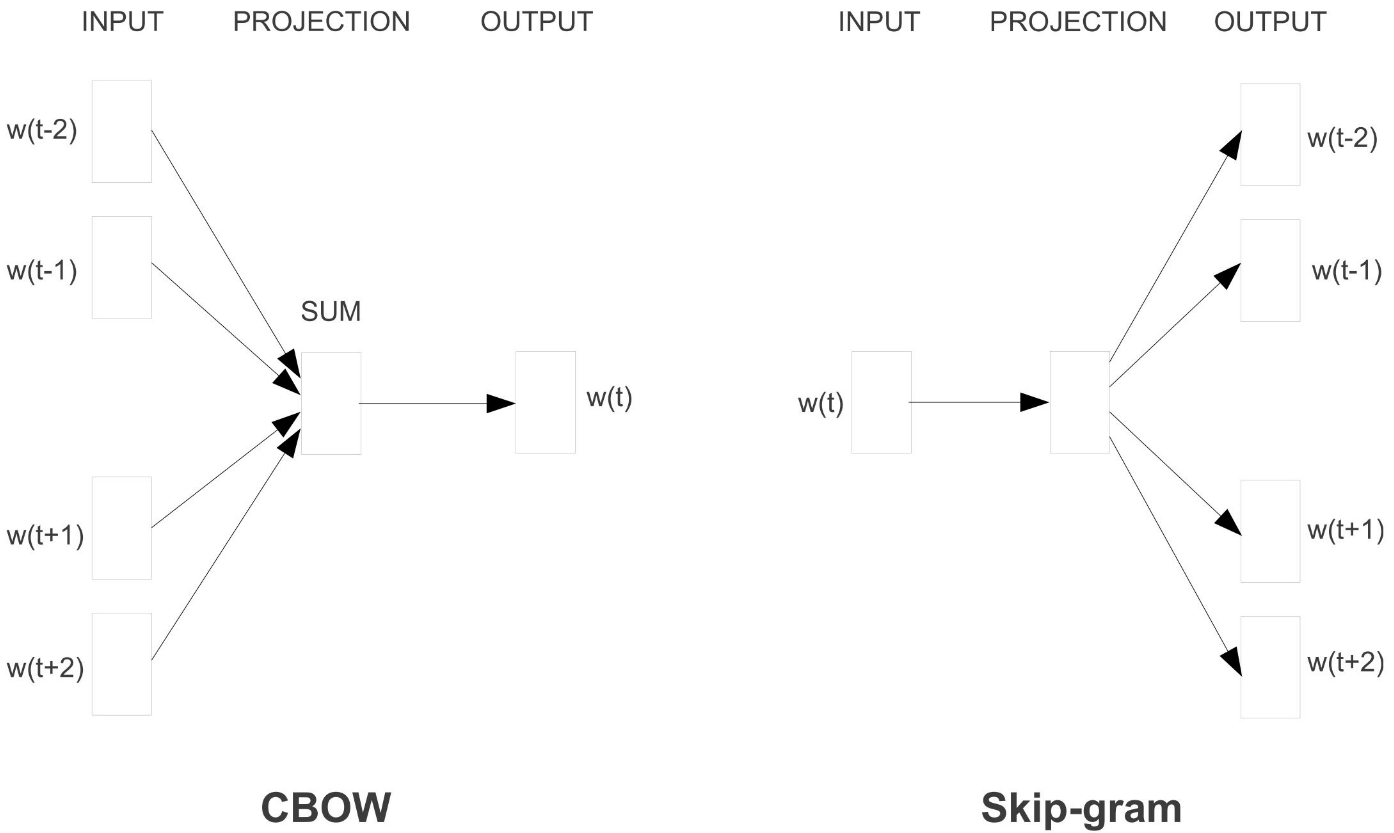


Verb tense

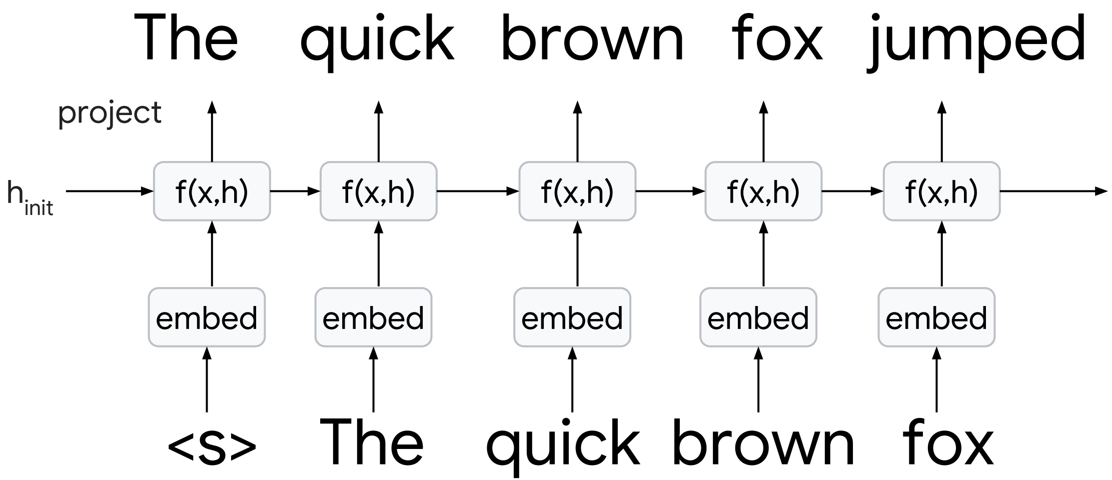


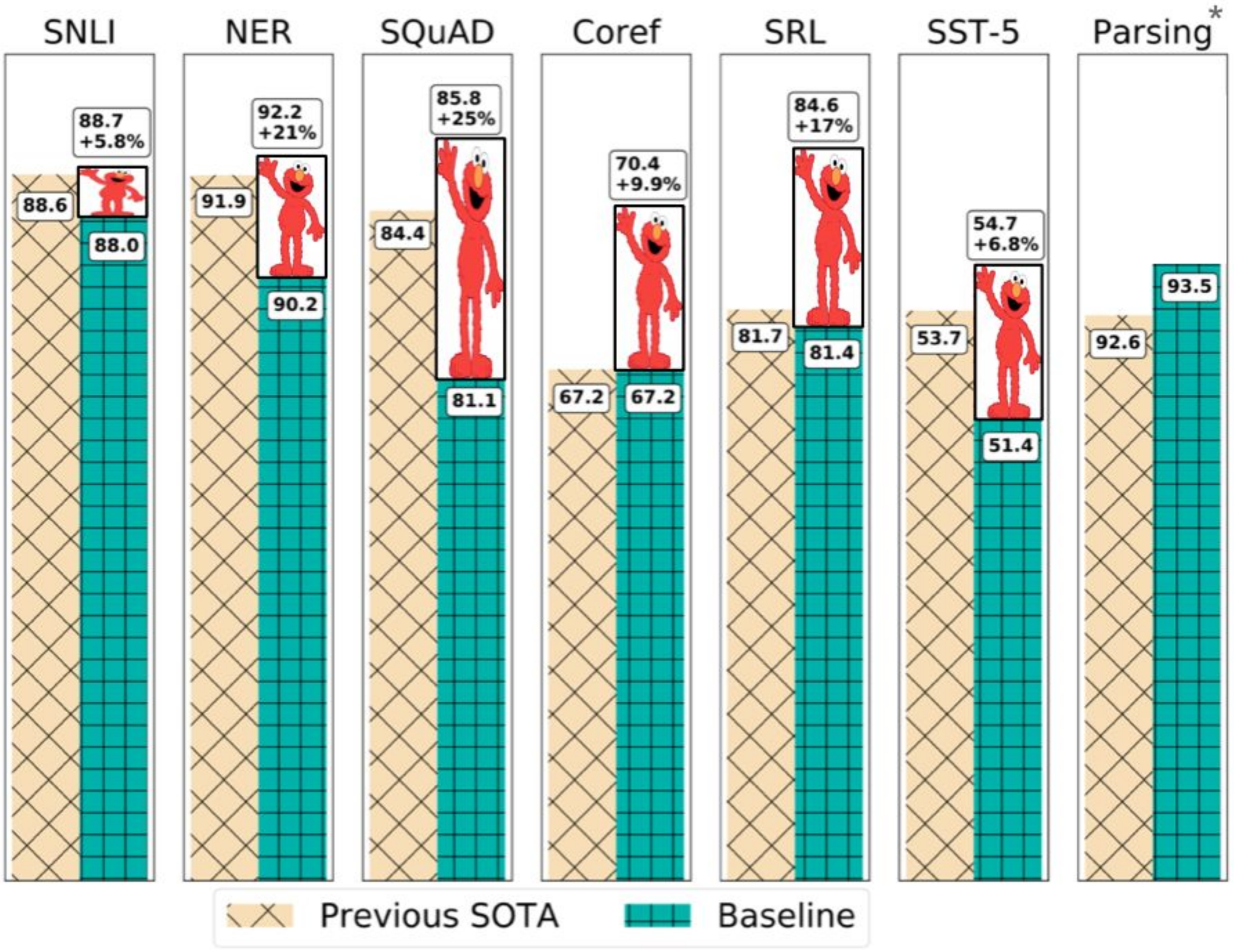
Country-Capital

Transfer learning for richer presentation



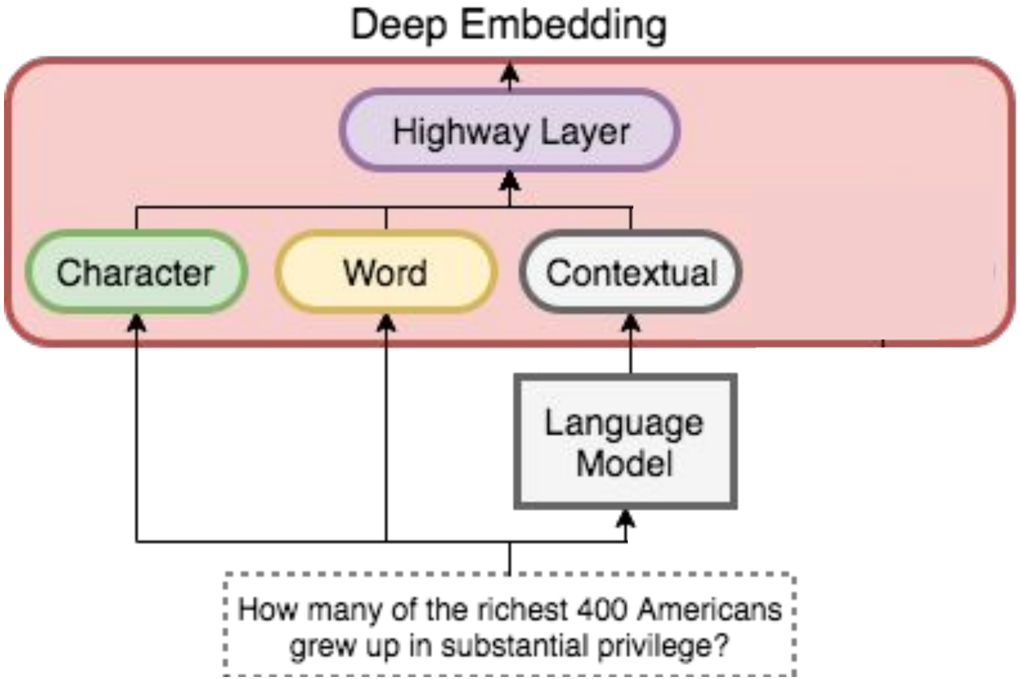
Language Models




SNLI
NER
SQuAD
Coref
SRL
SST-5
Parsing*
88.7
+5.8%
88.6
88.0
92.2
+21%
91.9
90.2
85.8
+25%
84.4
81.1
70.4
+9.9%
67.2
67.2
84.6
+17%
81.7
81.4
54.7
+6.8%
53.7
51.4
93.5
92.6
Previous SOTA
Baseline

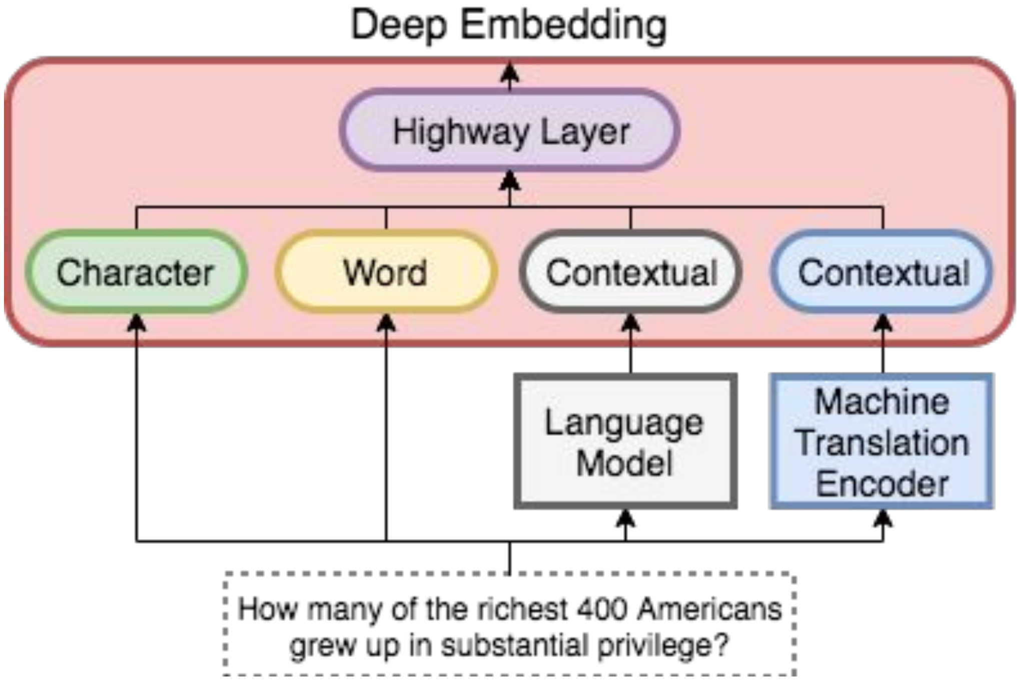
Transfer learning for richer presentation

- Pretrained language model (ELMo, [Peters et al., NAACL'18])
    - + 4.0 F1



Transfer learning for richer presentation

- Pretrained language model (ELMo, [Peters et al., NAACL'18])
    - + 4.0 F1
- Pretrained machine translation model (CoVe [McCann, NIPS'17])
    - + 0.3 F1



QANet – *3 key ideas*

- Deep Architecture without RNN
  - 130-layer (Deepest in NLP)
- Transfer Learning
  - leverage unlabeled data
- Data Augmentation
  - with back-translation

#1 on SQuAD (Mar-Aug 2018)

SQuAD1.1 Leaderboard

Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	**85.954**	**91.677**
2 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
3 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
4 Sep 09, 2018	nlnet (single model) *Microsoft Research Asia*	83.468	90.133
4 Jun 20, 2018	MARS (ensemble) *YUANFUDAO research NLP*	83.982	89.796
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737
6 Sep 01, 2018	MARS (single model) *YUANFUDAO research NLP*	83.185	89.547
7 Jun 20, 2018	QANet (single) *Google Brain & CMU*	82.471	89.306
7 May 09, 2018	MARS (single model) *YUANFUDAO research NLP*	82.587	88.880

QA is not Solved!!

QA is not Solved!!

Thank you!