DTU Course 02456 Deep learning
3 Recurrent neural networks
2017 Updates

Ole Winther

Dept for Applied Mathematics and Computer Science
Technical University of Denmark (DTU)

September 18, 2017

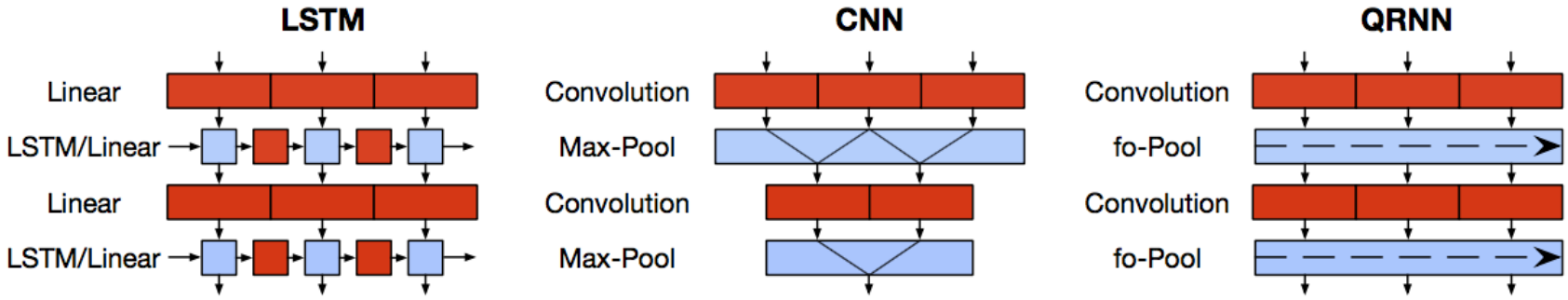
Objectives of RNN 2017

- P1: Quasi RNN (QRNN)
- how to perform sequence modeling with CNNs
- P2: Non-recurrent sequence to sequence models
- P3: Text summarization
- A cool application of encoder-decoder models.

Part 1:
Quasi RNN (QRNN)

Quasi RNN

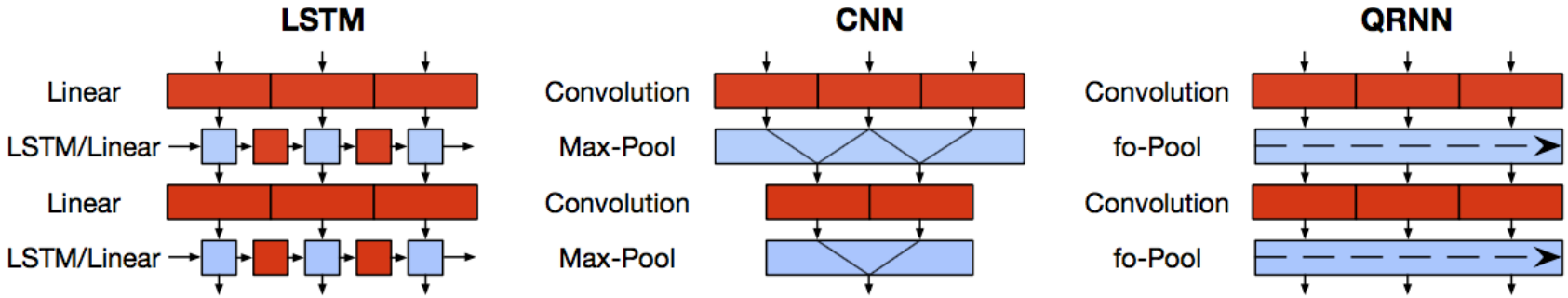
- Sequential computations in RNN $\rightarrow$ RNN relatively slow



- Red: conv/matmul and blue: element-wise operations

Quasi RNN

- Sequential computations in RNN $\rightarrow$ RNN relatively slow



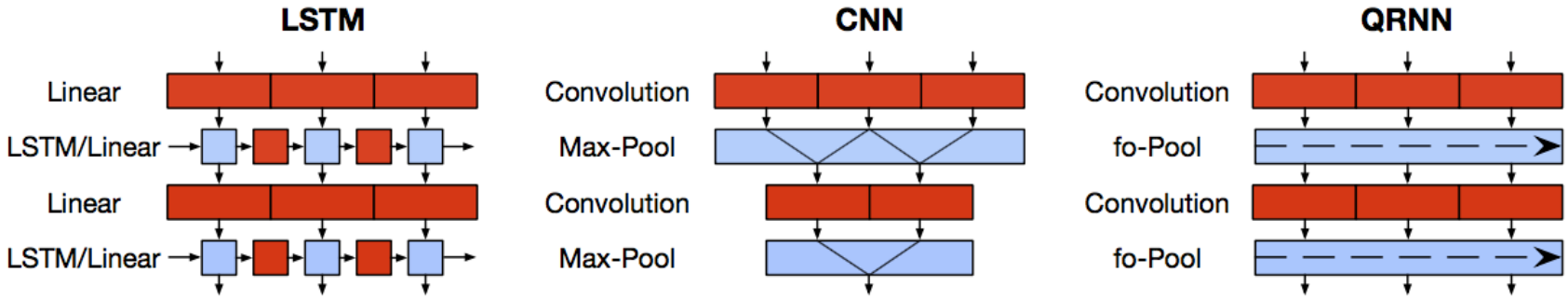
- Red: conv/matmul and blue: element-wise operations
- QRNN: replace recurrent step with masked convolutions
- width k: z_t depends upon $x_{t-k+1}, \ldots, x_{t-1}$.

$$Z = \tanh(W_z * X)$$

- $W_z \in \mathbb{R}^{k \times n \times m}$

Quasi RNN

- Sequential computations in RNN $\rightarrow$ RNN relatively slow



- Red: conv/matmul and blue: element-wise operations
- QRNN: replace recurrent step with masked convolutions
- width k: z_t depends upon $x_{t-k+1}, \ldots, x_{t-1}$.

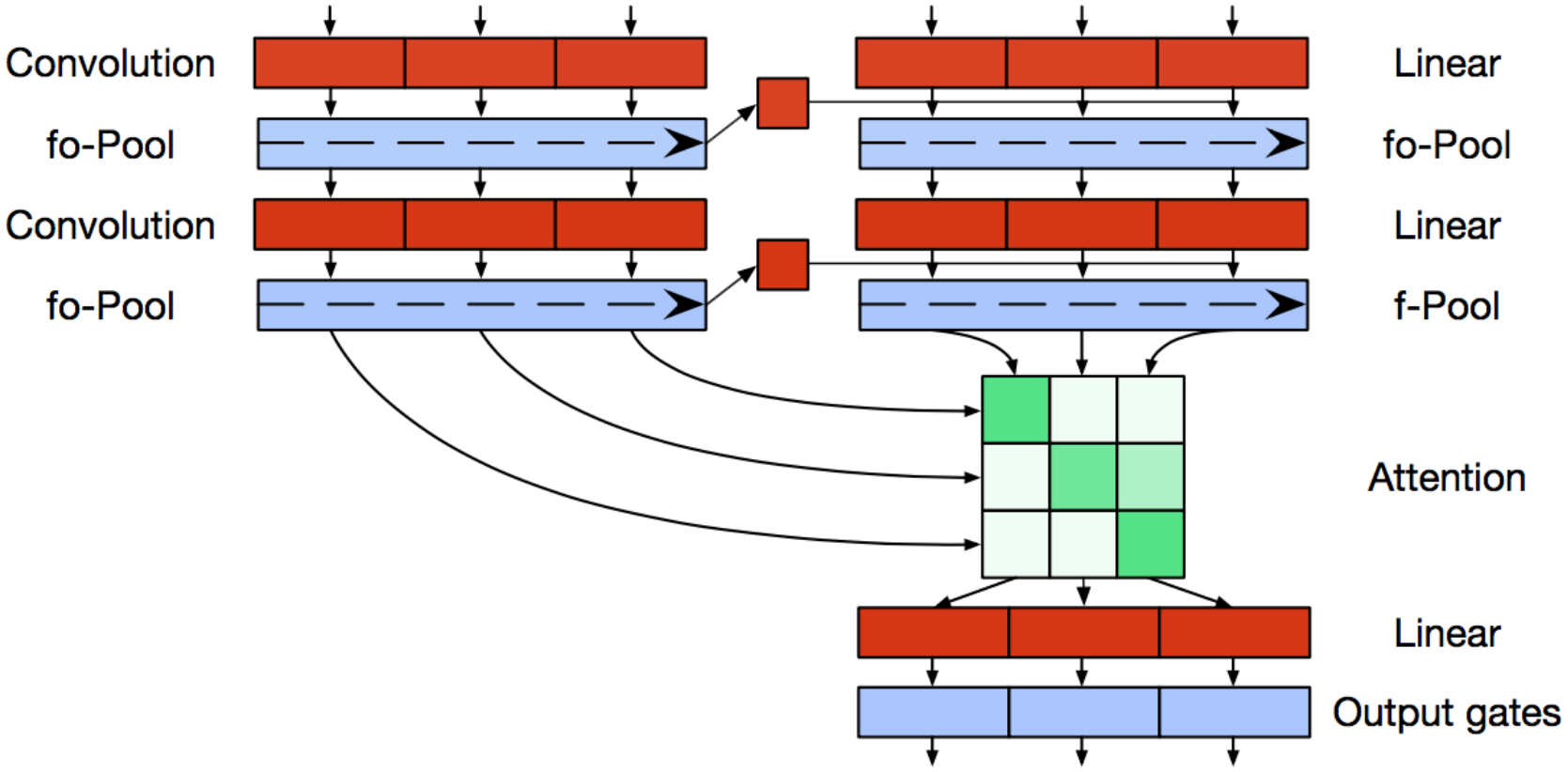
$$Z = \tanh(W_z * X)$$

- $W_z \in \mathbb{R}^{k \times n \times m}$
- f-pooling - forget gate $F = \sigma(W_f * X)$

$$h_t = f_t \odot h_{t-1} + (1 - f_t) \odot z_t$$

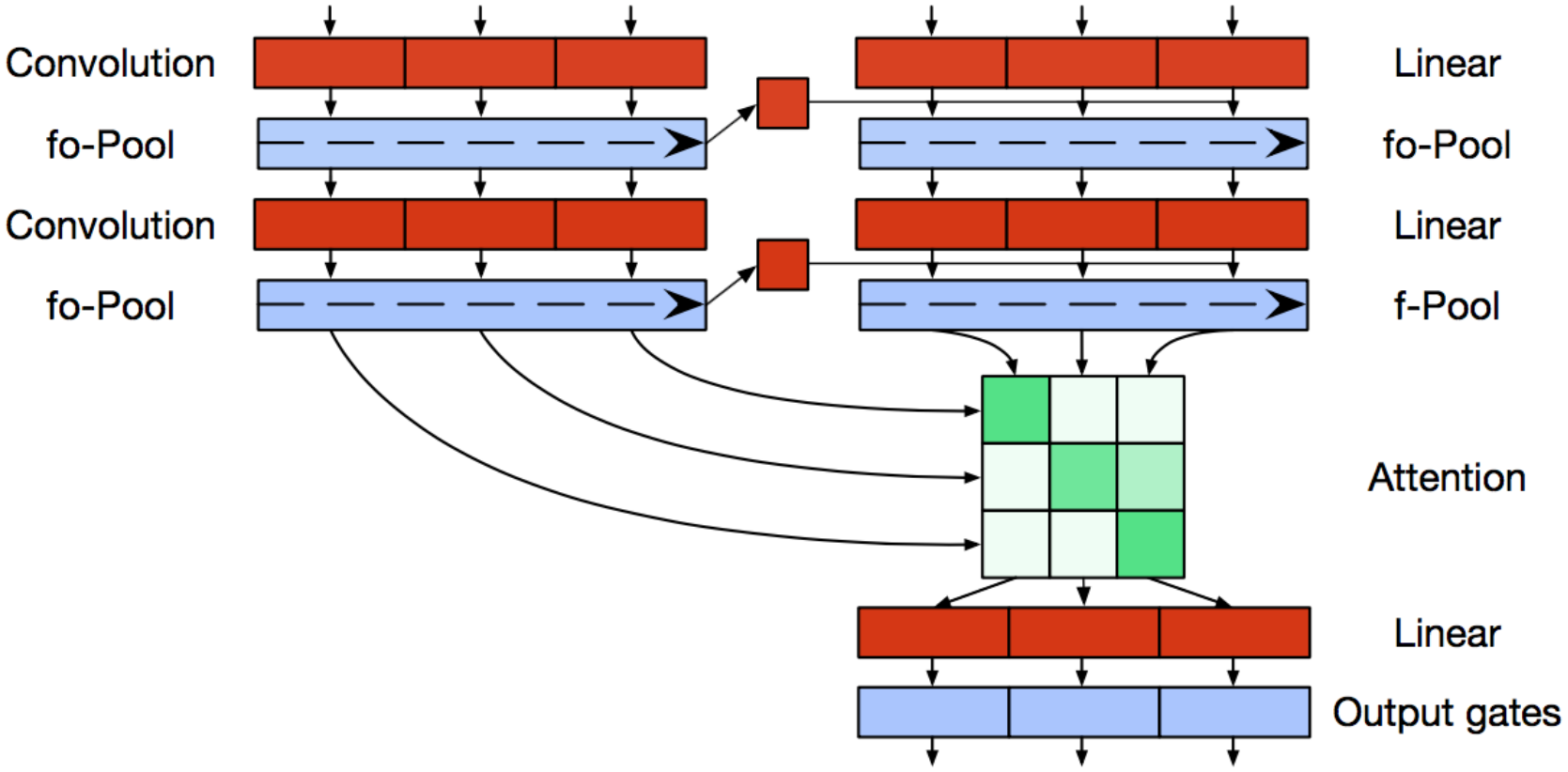
- Sequential operation - but fast because it is element-wise

QRNN for machine translation



- fo-Pool = forget and output gate pooling

QRNN for machine translation



- fo-Pool = forget and output gate pooling
- Close to state-of-the-art performance
- Big savings in the recurrent part of model
- Additional tweaks: regularization and densenet layers.

Part 2:

Non-recurrent sequence to sequence models

Attention instead of recurrent connections!

- Attention recap!

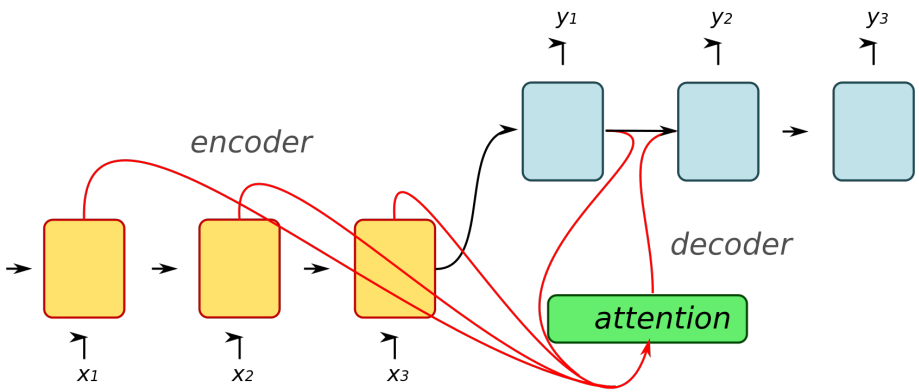
Attention instead of recurrent connections!

- Attention recap!
- H = hidden state of encoder
- Attention encoder-decoder:
  - Query vector Q from decoder
  - Key vector K from encoder

$$K = W_K H$$

  - Value vector V from encoder

$$V = W_V H$$

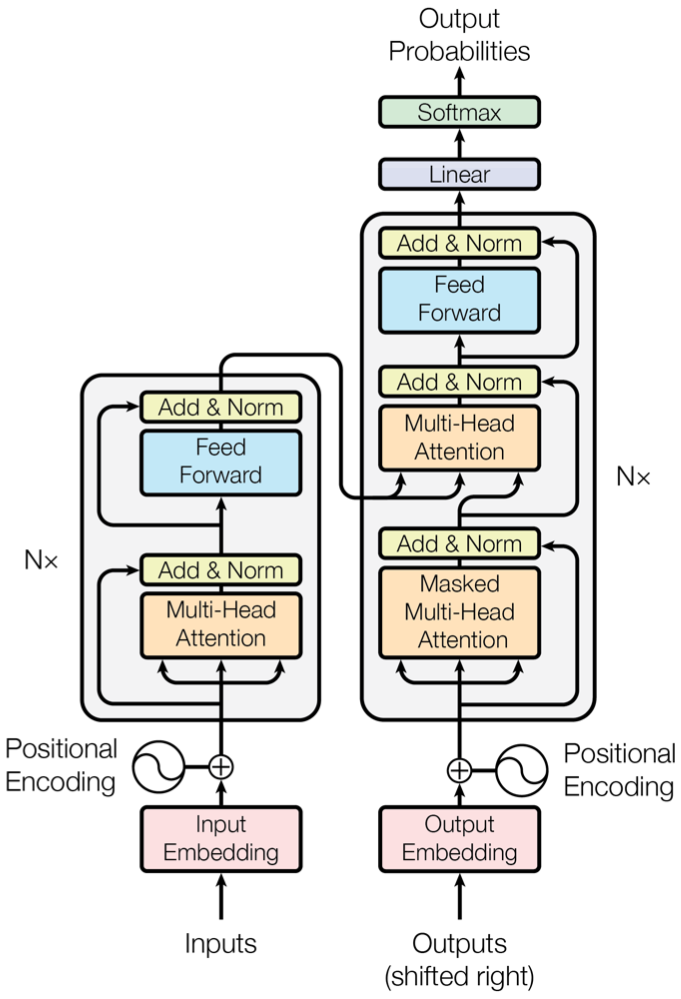


- Output softmax weighted combination of V:

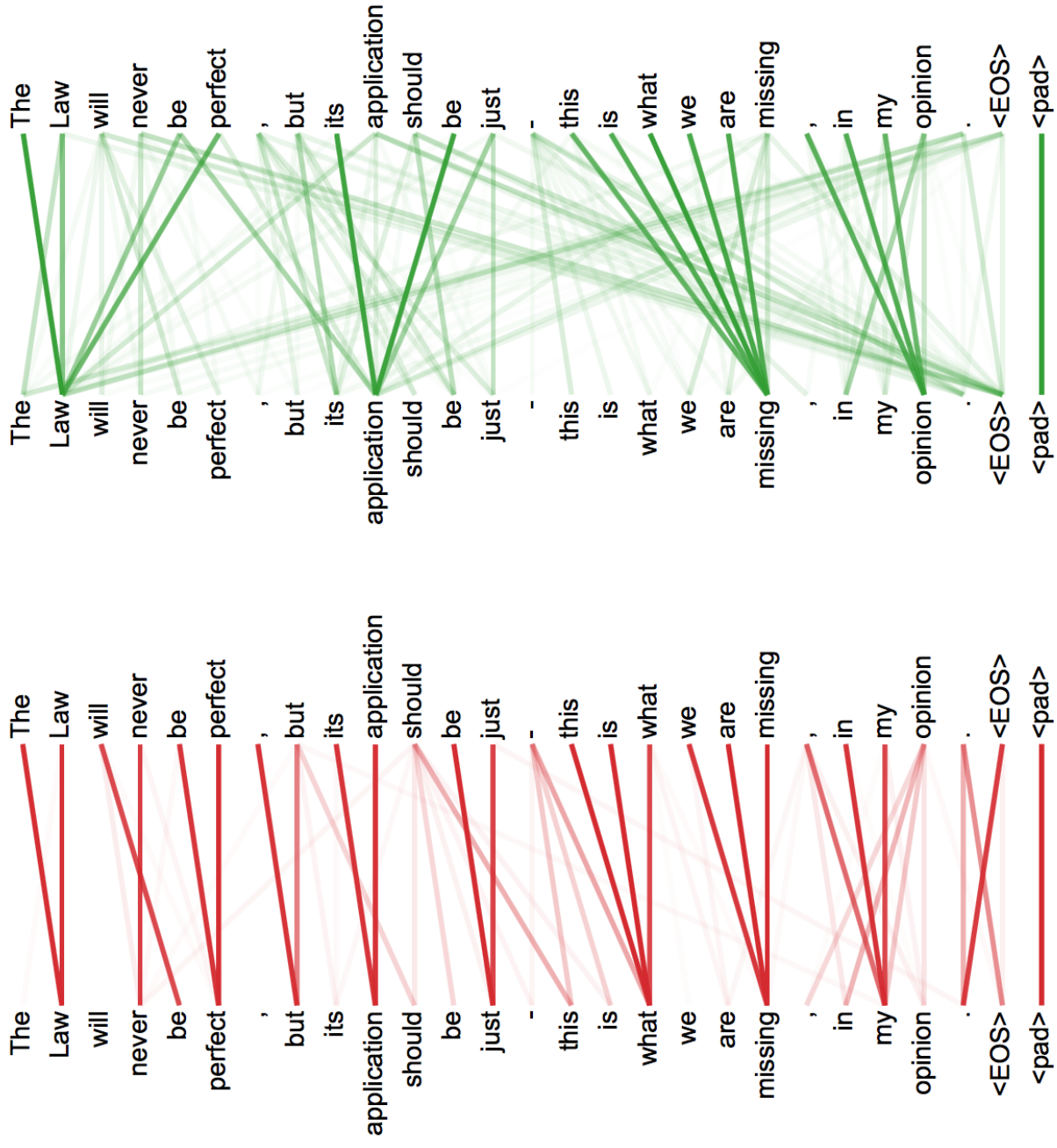
$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T)V$$

Extension to self-attention

- Encoder self-attention layer:
  - Q, K, V are functions of previous encoder layer
- Decoder self-attention layer:
  - The same but we need to mask out the “future”.
- Gehring et. al. similar but with CNN.



Multi-head self-attention



Position embedding

- Model is not really aware of word position in sequence!

Position embedding

- Model is not really aware of word position in sequence!
- Input sequence $x = (x_1, \ldots, x_m)$
- Embedded input sequence $w = (w_1, \ldots, w_m) = \mathcal{D}x$
- Add position encoding $p = (p_1, \ldots, p_m)$:

$$\text{embedding} = e = (e_1, \ldots, e_m) = (w_1 + p_1, \ldots, w_m + p_m)$$

Position embedding

- Model is not really aware of word position in sequence!
- Input sequence $x = (x_1, \ldots, x_m)$
- Embedded input sequence $w = (w_1, \ldots, w_m) = \mathcal{D}x$
- Add position encoding $p = (p_1, \ldots, p_m)$:

$$\text{embedding} = e = (e_1, \ldots, e_m) = (w_1 + p_1, \ldots, w_m + p_m)$$

- Position embedding p can be learned or
- chosen e.g. to sine and cosine basis functions.

Position embedding

- Model is not really aware of word position in sequence!
- Input sequence $x = (x_1, \ldots, x_m)$
- Embedded input sequence $w = (w_1, \ldots, w_m) = \mathcal{D}x$
- Add position encoding $p = (p_1, \ldots, p_m)$:

$$\text{embedding} = e = (e_1, \ldots, e_m) = (w_1 + p_1, \ldots, w_m + p_m)$$

- Position embedding p can be learned or
- chosen e.g. to sine and cosine basis functions.
- Models works pretty well without position encoding!

Part 3:
Text summarization

Text summarization model in action

The bottleneck is no longer access to information; now it's our ability to keep up.
AI can be trained on a variety of different types of texts and summary lengths.
A model that can generate long, coherent, and meaningful summaries remains an open research problem.

The last few decades have witnessed a fundamental change in the challenge of taking in new information. The bottleneck is no longer access to information; now it's our ability to keep up. We all have to read more and more to keep up-to-date with our jobs, the news, and social media. We've looked at how AI can improve people's work by helping with this information deluge and one potential answer is to have algorithms automatically summarize longer texts. Training a model that can generate long, coherent, and meaningful summaries remains an open research problem. In fact, generating any kind of longer text is hard for even the most advanced deep learning algorithms. In order to make summarization successful, we introduce two separate improvements: a more contextual word generation model and a new way of training summarization models via reinforcement learning (RL). The combination of the two training methods enables the system to create relevant and highly readable multi-sentence summaries of long text, such as news articles, significantly improving on previous results. Our algorithm can be trained on a variety of different types of texts and summary lengths. In this blog post, we present the main contributions of our model and an overview of the natural language challenges specific to text summarization.

Text summarization input

Source document
Jenson Button was denied his 100th race for McLaren after an ERS prevented him from making it to the start-line. It capped a miserable weekend for the Briton; his time in Bahrain plagued by reliability issues. Button spent much of the race on Twitter delivering his verdict as the action unfolded. 'Kimi is the man to watch,' and 'loving the sparks', were among his pearls of wisdom, but the tweet which courted the most attention was a rather mischievous one: 'Ooh is Lewis backing his team mate into Vettel?' he quizzed after Rosberg accused Hamilton of pulling off such a manoeuvre in China. Jenson Button waves to the crowd ahead of the Bahrain Grand Prix which he failed to start Perhaps a career in the media beckons Lewis Hamilton has out-qualified and finished ahead of Nico Rosberg at every race this season. Indeed Rosberg has now beaten his Mercedes team-mate only once in the 11 races since the pair infamously collided in Belgium last year. Hamilton secured the 36th win of his career in Bahrain and his 21st from pole position. Only Michael Schumacher (40), Ayrton Senna (29) and Sebastian Vettel (27) have more. He also became only the sixth F1 driver to lead 2,000 laps. Nico Rosberg has been left in the shade by Lewis Hamilton who celebrates winning his third race of the year Kimi Raikkonen secured a record seventh podium finish in Bahrain following his superb late salvo, although the Ferrari driver has never won in the Gulf Kingdom. It was the Finn's first trip to the rostrum since the 2013 Korean Grand Prix, but his triumph brought a typically deadpan response: 'You're never happy when you finish second... I'm a bit pleased to get a result.' Sparks fly off the back of Kimi Raikkonen's Ferrari en route to finishing second in Bahrain Bernie Ecclestone was in the Bahrain paddock this weekend. He denied trying to engineer a deal for Hamilton, out of contract at the end of the season, to join Ferrari despite earlier insisting that such a move would be 'great' for the sport. The 84-year-old also confirmed that F1 would be in Azerbaijan for the first time next year, even with concerns surrounding the countrys human rights record. 'I think everybody seems to be happy,' Ecclestone said. 'There doesn't seem to be any big problem there. There's no question of it not being on the calendar. It's going to be another good race. Formula One supremo Bernie Ecclestone speaks to Nico Rosberg ahead of the Bahrain Grand Prix

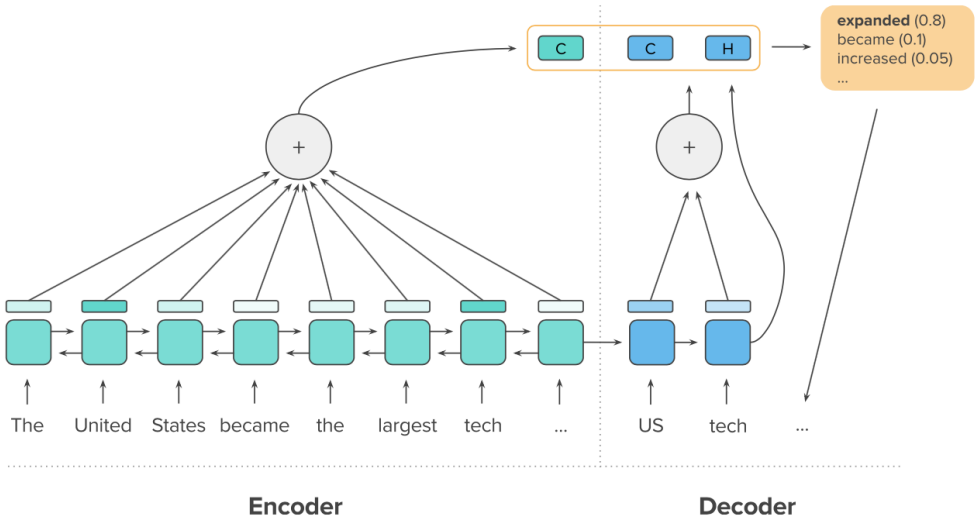
Datasets: CNN/Daily Mail and New York TImes

Text summarization ground truth and predictions

Ground truth summary
Button denied 100th race start for McLaren after ERS failure. Button then spent much of the Bahrain Grand Prix on Twitter delivering his verdict on the action as it unfolded. Lewis Hamilton has out-qualified and finished ahead of Mercedes team-mate Nico Rosberg at every race this season. Bernie Ecclestone confirms F1 will make its bow in Azerbaijan next season.
ML, with intra-attention (ROUGE-1 41.58)
Button was denied his 100th race for McLaren. ERS prevented him from making it to the start-line. The Briton. He quizzed after Nico Rosberg accused Lewis Hamilton of pulling off such a manoeuvre in China. Button has been in Azerbaijan for the first time since 2013.
RL, with intra-attention (ROUGE-1 50.00)
Button was denied his 100th race for McLaren after an ERS prevented him from making it to the start-line. It capped a miserable weekend for the Briton. Button has out-qualified. Finished ahead of Nico Rosberg at Bahrain. Lewis Hamilton has. In 11 races. . The race. To lead 2,000 laps. . In. . . And. .
ML+RL, with intra-attention (ROUGE-1 44.00)
Button was denied his 100th race for McLaren. The ERS prevented him from making it to the start-line. Button was his team mate in the 11 races in Bahrain. He quizzed after Nico Rosberg accused Lewis Hamilton of pulling off such a manoeuvre in China.

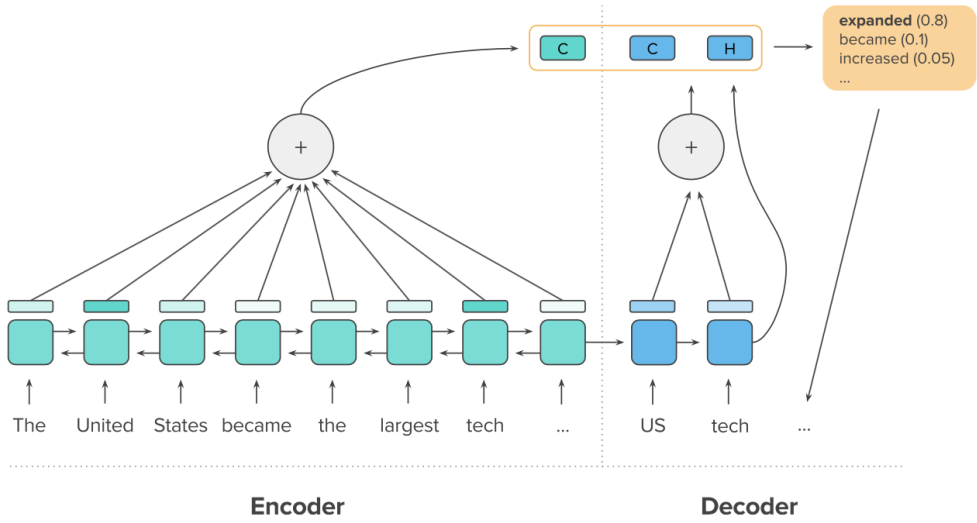
Main ingredients

- Intra-temporal attention and intra decoder attention:
  - penalize inputs that have high attention in previous steps
  - use previous generated outputs as input to attention function



Main ingredients

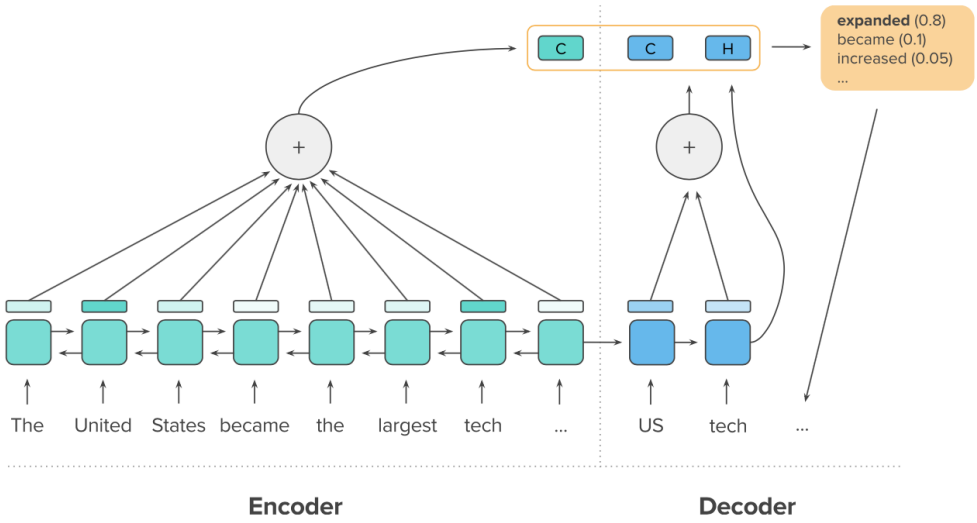
- Intra-temporal attention and intra decoder attention:
  - penalize inputs that have high attention in previous steps
  - use previous generated outputs as input to attention function



- Token generation and pointer mechanism to rare input

Main ingredients

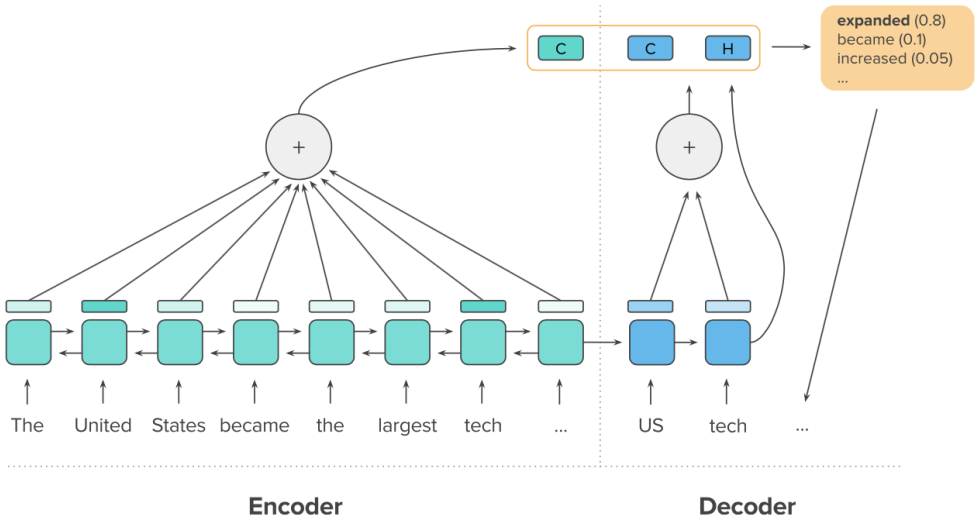
- Intra-temporal attention and intra decoder attention:
  - penalize inputs that have high attention in previous steps
  - use previous generated outputs as input to attention function



- Token generation and pointer mechanism to rare input
- Heuristic: Avoid repetition at test time.

Main ingredients

- Intra-temporal attention and intra decoder attention:
  - penalize inputs that have high attention in previous steps
  - use previous generated outputs as input to attention function



- Token generation and pointer mechanism to rare input
- Heuristic: Avoid repetition at test time.
- Objective - combination of
  - word level softmax (maximum likelihood) and
  - sentence level reinforcement learning (ROUGH score)

References

- J. Bradbury, S. Merit, C. Xiong and R Socher, Quasi-recurrent neural networks,
- A. Vaswani et. al., Attention is all you need
- J. Gehring et. al., Convolutional Sequence to Sequence Learning
- R. Paulus, C. Xiong and R. Socher, A Deep Reinforced Model for Abstractive Summarization

Thanks!
Ole Winther