Reading Comprehension and Language Understanding

Russ Salakhutdinov

Machine Learning Department
Carnegie Mellon University
Canadian Institute for Advanced Research

Carnegie Mellon University

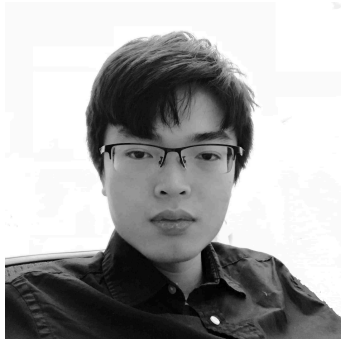
MACHINE LEARNING DEPARTMENT

Talk Outline

- Introduction to Reading Comprehension

- Language Modeling, XLNet and Transformer-XL: Modeling Long-Term Dependencies

Bhuwan Dhingra,
PhD at CMU



Zhilin Yang,
PhD at CMU

Some slides borrowed from Bhuwan Dhingra and Zhiin Yang

Reading Comprehension

Document:
"...arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges ... included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama..."

Query:
"President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat."
Find **X**.

Answer:
Rod Blagojevich

Who-Did-What Dataset (Onishi, Wang, Bansal, Gimpel, McAllester, EMNLP, 2016)

Reading Comprehension

TASK:
Given a document query pair (d, q) find $a \in A$ which answers q .

- d is a document
- q is a question over the contents of that document
- a is the answer to this query

- The answer comes from a fixed vocabulary A .
  - A might consist of all tokens / spans of tokens in the document d (Extractive Question Answering)

- Question Answering / Information Extraction

Approach -- Supervised Learning

$$\mathcal{D} = \{(d, q, a)\}_{i=1}^{N}$$

Dataset

$$Pr(c|d, q) = f_\theta(d, q, c) \quad \forall \quad c \in A$$

Model
(A neural network)

$$\mathcal{L}(\theta) = \sum_{(d,q,a)\in\mathcal{D}} -\log Pr(a|d, q)$$

Loss

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta)$$

Training

What architectural biases can we build into the model?

Architectural Bias

Designing the connectivity pattern of a Neural Network to reflect the nature of the problem being solved.

- CNNs, RNNs have architectural biases towards images / sequences
- For reading comprehension what biases can we build to reflect **linguistic phenomena?**
  - Alignment, Paraphrasing, Aggregation **(This Lecture)**
  - Coreference, Syntactic and Semantic Dependencies **(Tomorrow)**

Text Phenomena

Document:
"...arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges ... included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama."

Query:
"President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat."
Find **X**.

Alignment

Text Phenomena

Document:
"...arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges... included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama..."

Query:
"President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat."
Find **X**.

Alignment

Paraphrasing

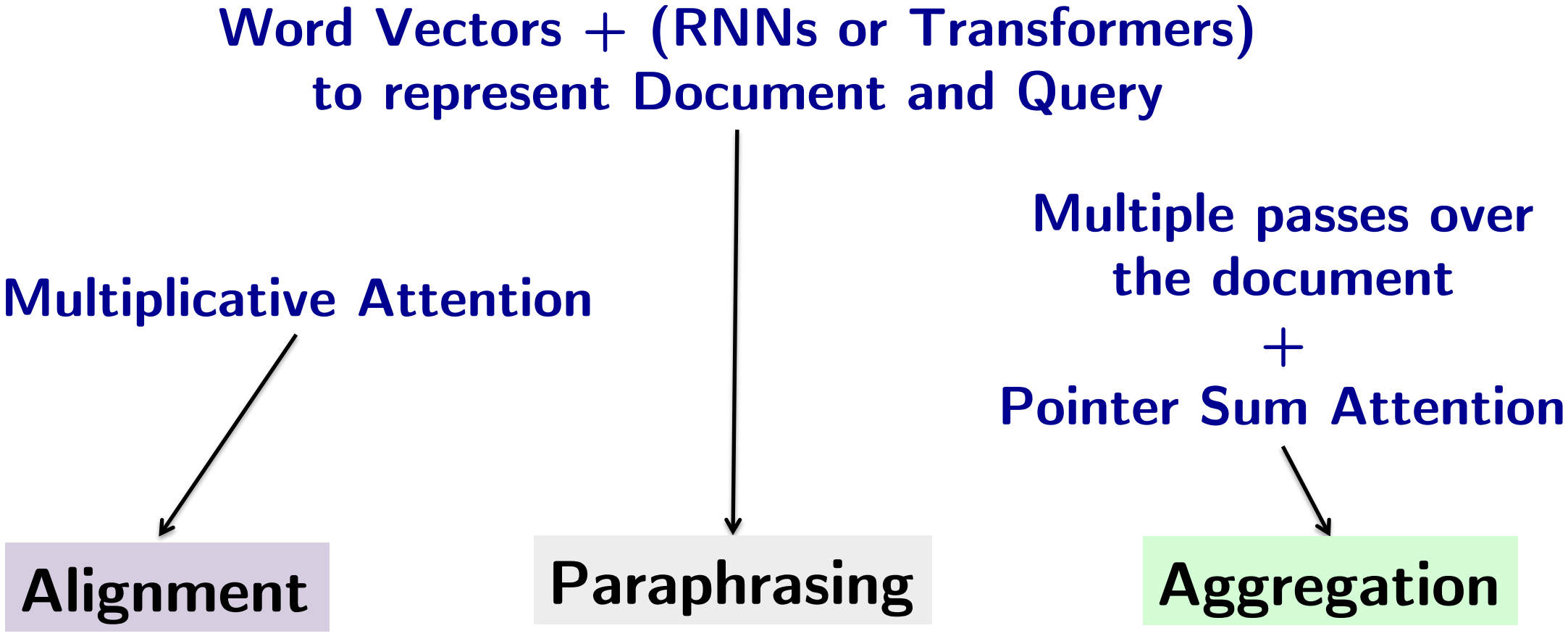
Text Phenomena

Document:
"...arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges ... included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama..."

Query:
"President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat."
Find **X**.

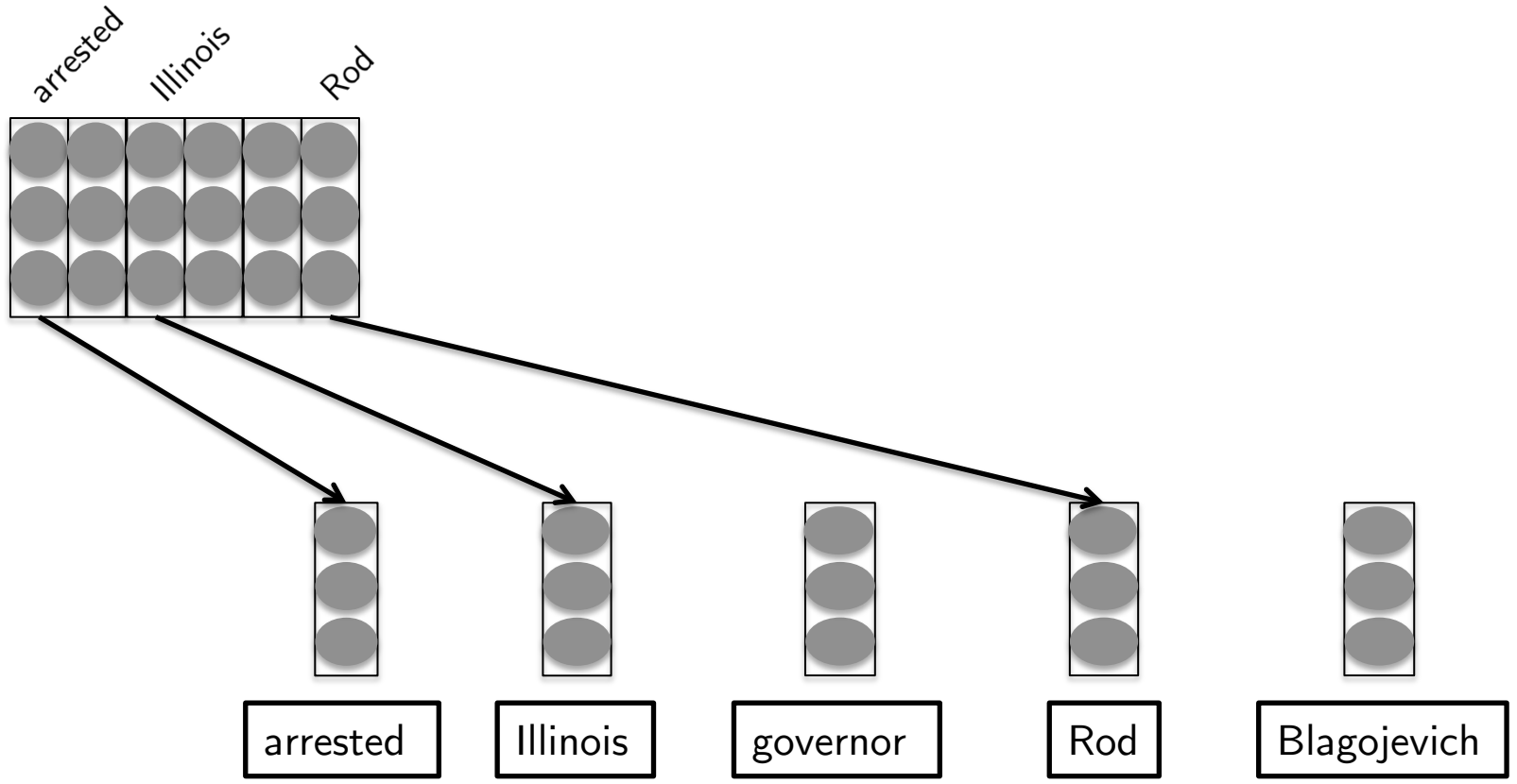
Alignment **Paraphrasing** **Aggregation**

Biases


Word Vectors + (RNNs or Transformers)
to represent Document and Query
Multiplicative Attention
Multiple passes over
the document
+
Pointer Sum Attention
Alignment
Paraphrasing
Aggregation

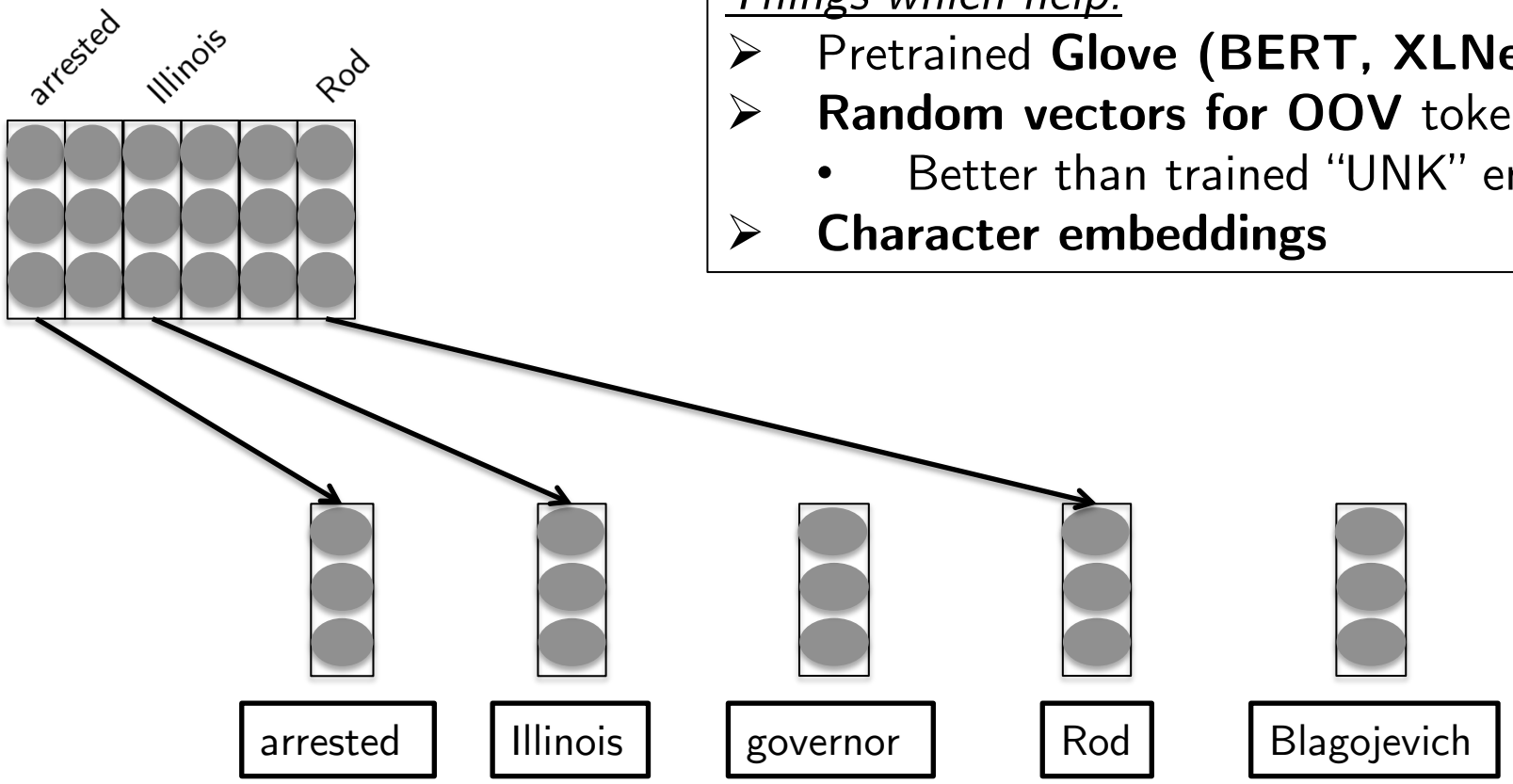
Representing Document / Query

- As compositions of word vectors



Representing Document / Query

- As compositions of word vectors



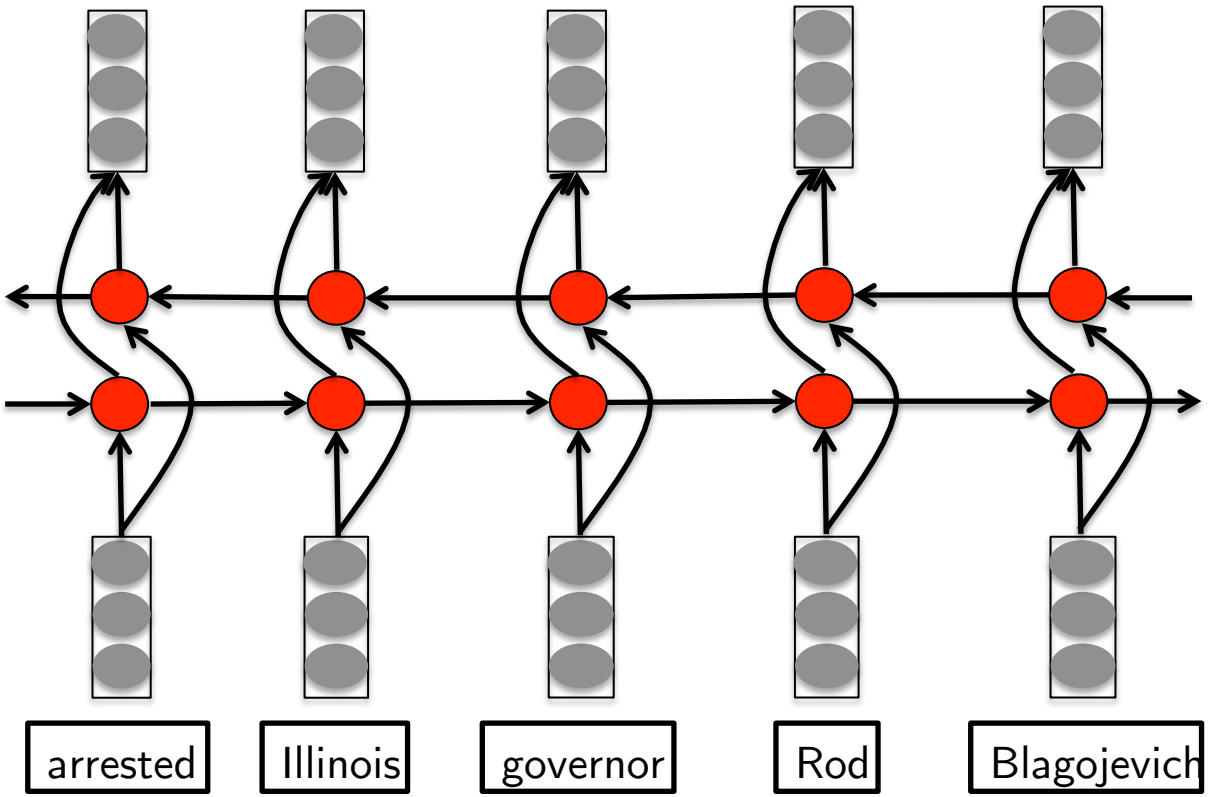
Things which help:

- Pretrained **Glove (BERT, XLNet) embeddings**
- **Random vectors for OOV** tokens at test time.
  - Better than trained "UNK" embedding.
- **Character embeddings**

Representing Document / Query

- **Bidirectional Gated Recurrent Units** process the tokens from left to right and right to left

$$d_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$



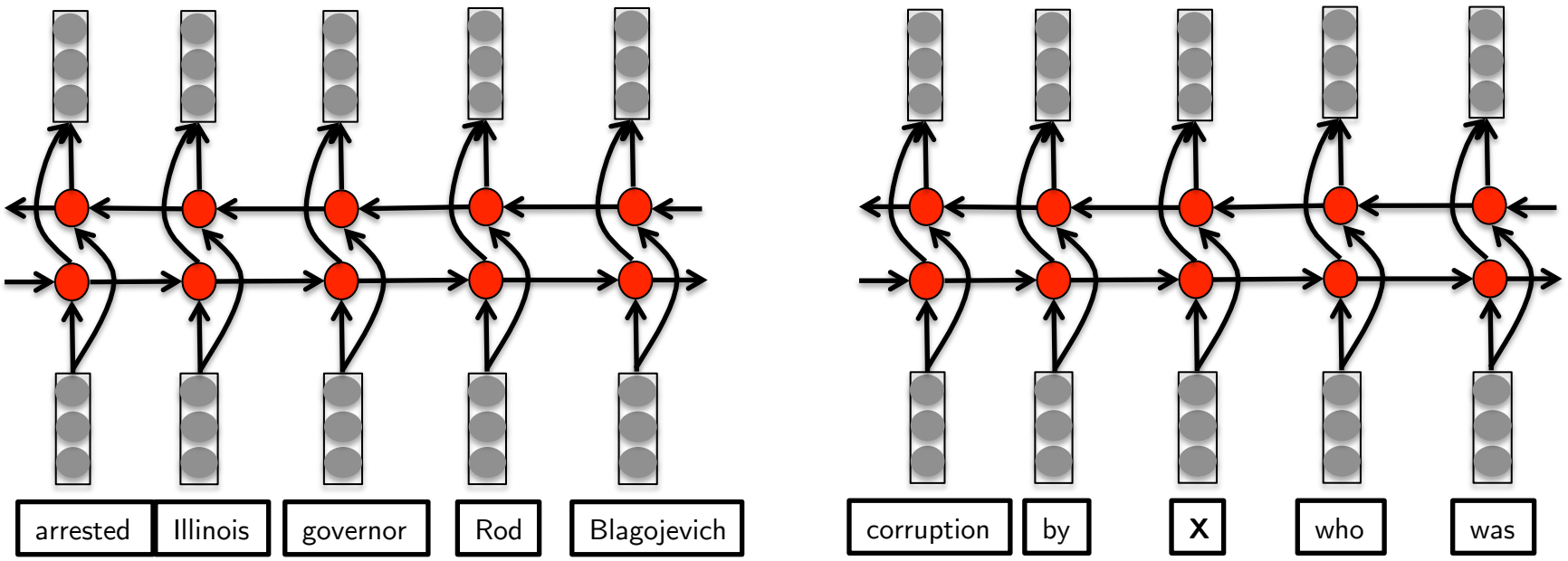
- Later, we will see generalized autoregressive models for representing documents/queries

Representing Document / Query

- Both document and query are represented as matrices

$$D \in \mathbb{R}^{2h \times |D|}$$

$$Q \in \mathbb{R}^{2h \times |Q|}$$



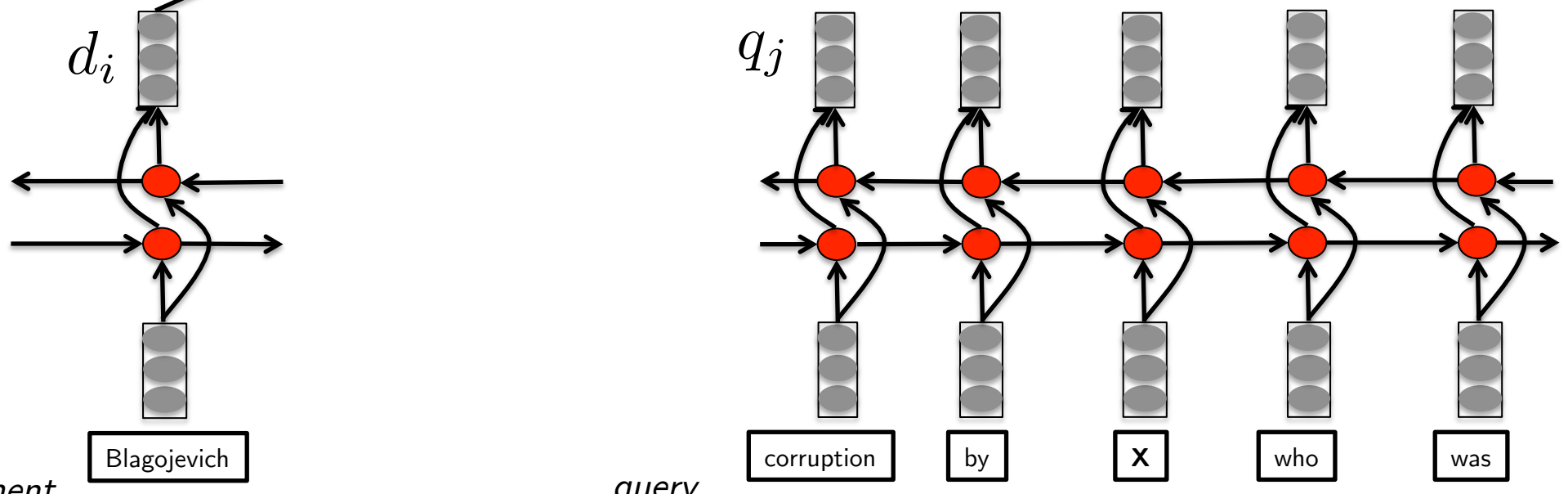
h – State size of each GRU

Gated Attention Mechanism

- For each token in D, we form a **token specific representation** of the query

$$\alpha_{ij} = \text{softmax}(q_j^T d_i)$$

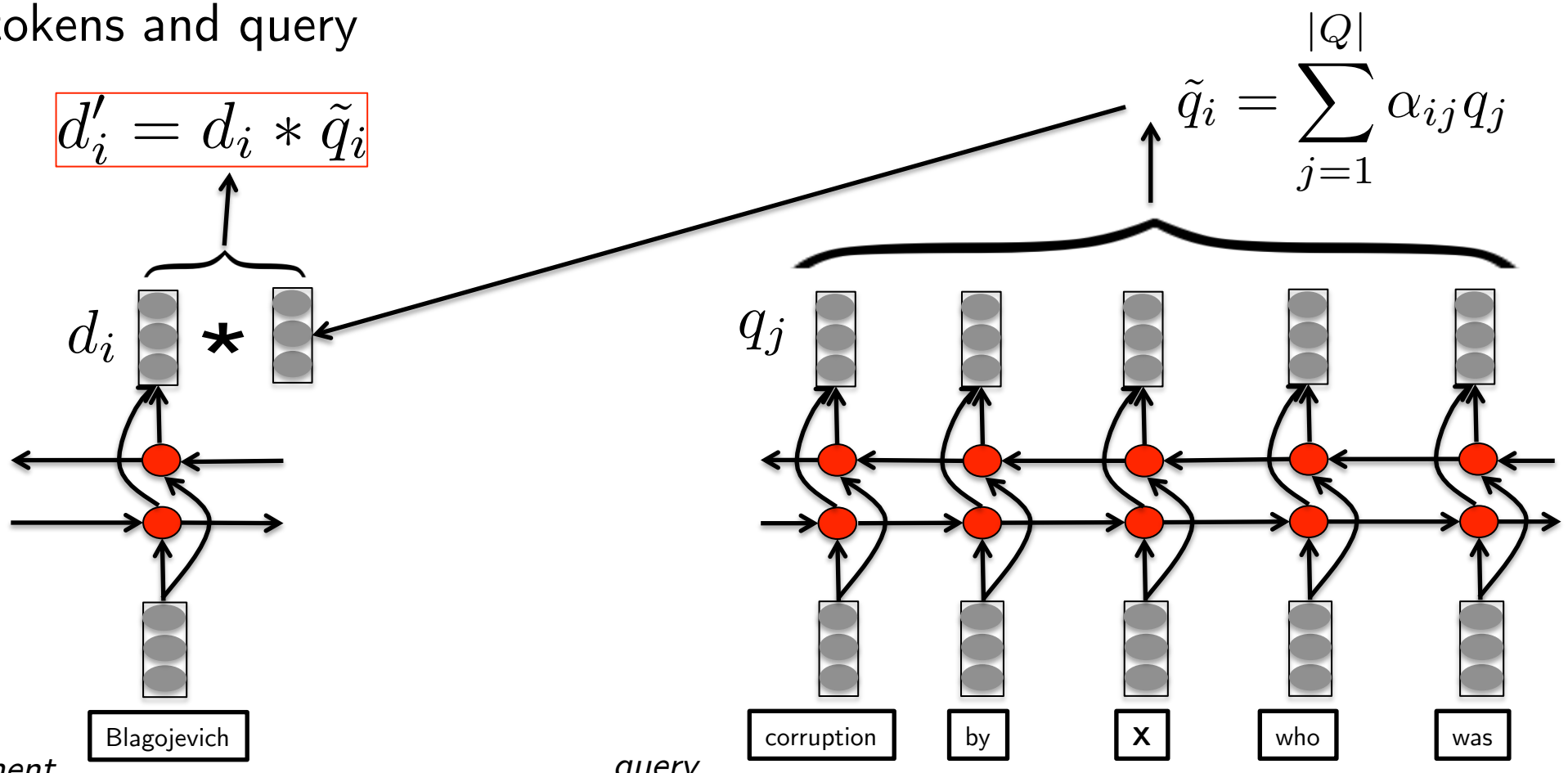
$$\tilde{q}_i = \sum_{j=1}^{|Q|} \alpha_{ij} q_j$$



Dhingra et al (ACL 2017)

Gated Attention Mechanism

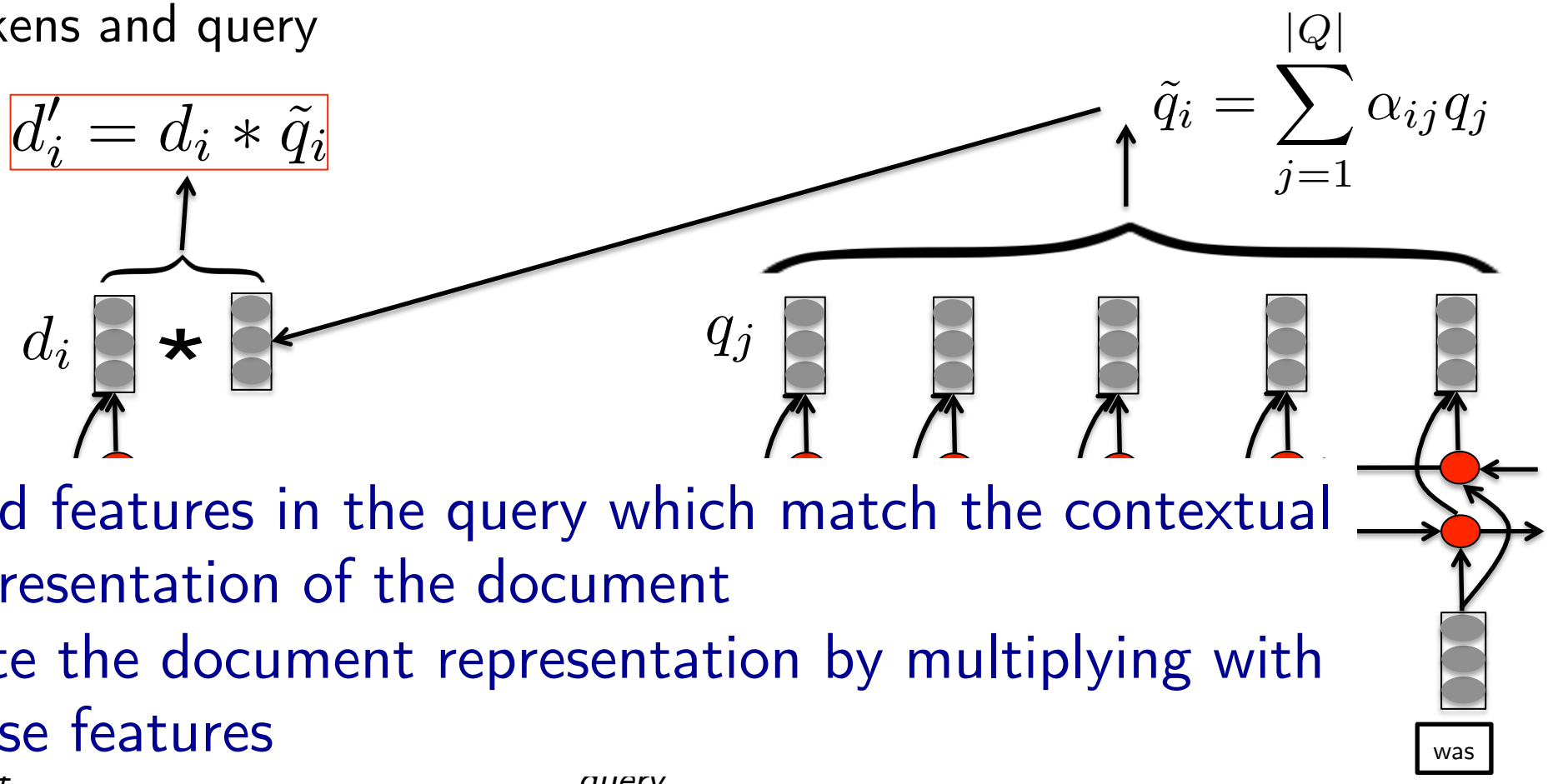
- Use **element-wise multiplication** to gate the interaction between document tokens and query



Dhingra et al (ACL 2017)

Gated Attention Mechanism

- Use **element-wise multiplication** to gate the interaction between document tokens and query

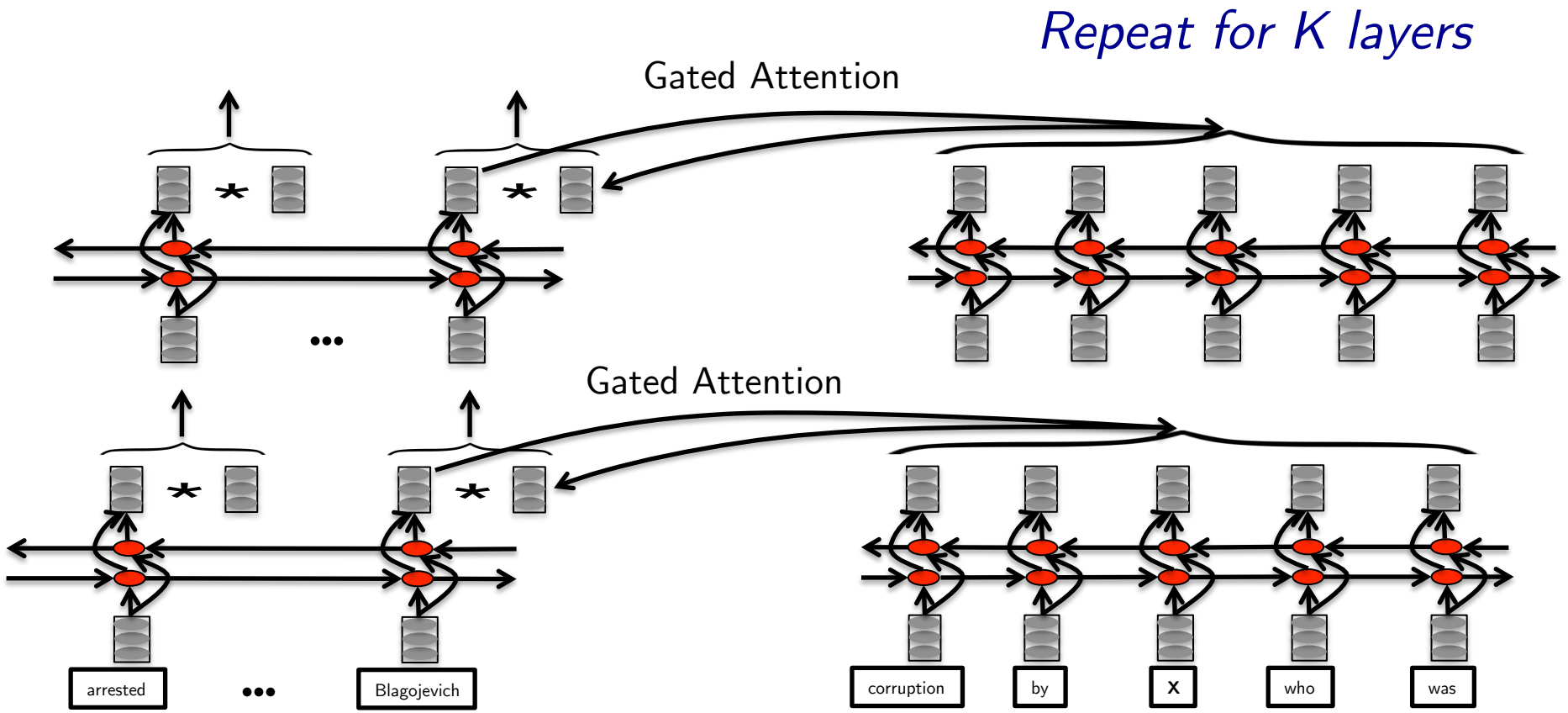


1. Find features in the query which match the contextual representation of the document
2. Gate the document representation by multiplying with these features

Dhingra et al (ACL 2017)

Multi Hop Architecture

- Perform several passes over the document
  - Allow model to combine evidence from multiple sentences



Dhingra et al (ACL 2017)

Output Model

- Probability that a particular token in the document answers the query:
  - Take an **inner product** between the query embedding and the output of the last layer:

$$s_i = \frac{\exp\left(\langle q^{(K)}, d_i^{(K)} \rangle\right)}{\sum_{i'} \exp\left(\langle q^{(K)}, d_{i'}^{(K)} \rangle\right)}, \quad i = 1, \ldots, |D|$$

- The probability of a particular candidate $c \in \mathcal{A}$ is then aggregated over all document tokens which appear in c:

$$P(c|d,q) \propto \sum_{i \in \mathbb{I}(c,d)} s_i$$

set of positions where a token in c appears in the document d.

Pointer Sum Attention of Kadlec et al., 2016

Output Model

- The probability of a particular candidate $c \in \mathcal{A}$ is then aggregated over all document tokens which appear in c:

$$P(c|d,q) \propto \sum_{i \in \mathbb{I}(c,d)} s_i$$

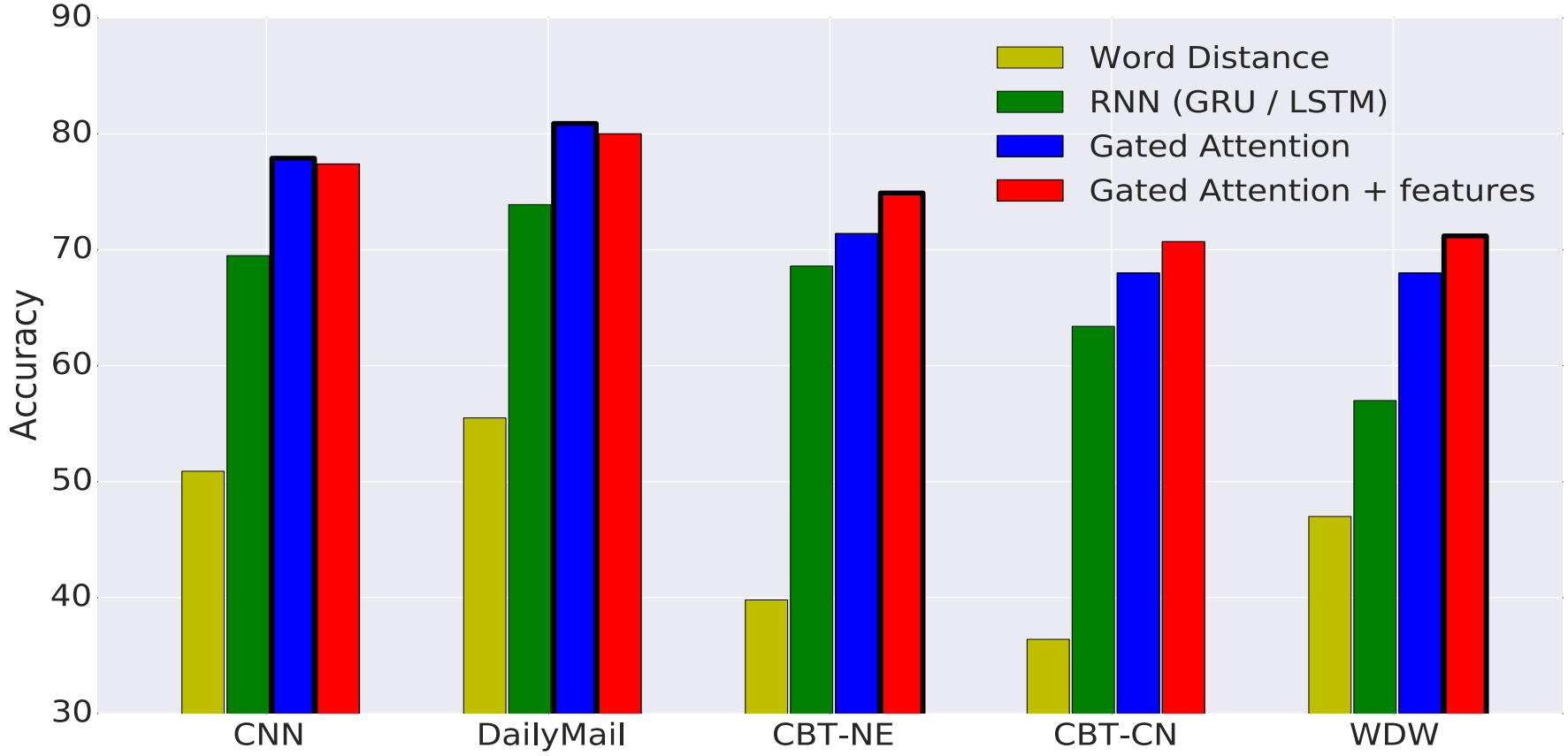
- The candidate with **maximum probability** is selected as the predicted answer:

$$a^* = \text{argmax}_{c \in \mathcal{C}} \; P(c|d,q)$$

- Use cross-entropy loss between the predicted probabilities and the true answers.

Results

- 5 datasets

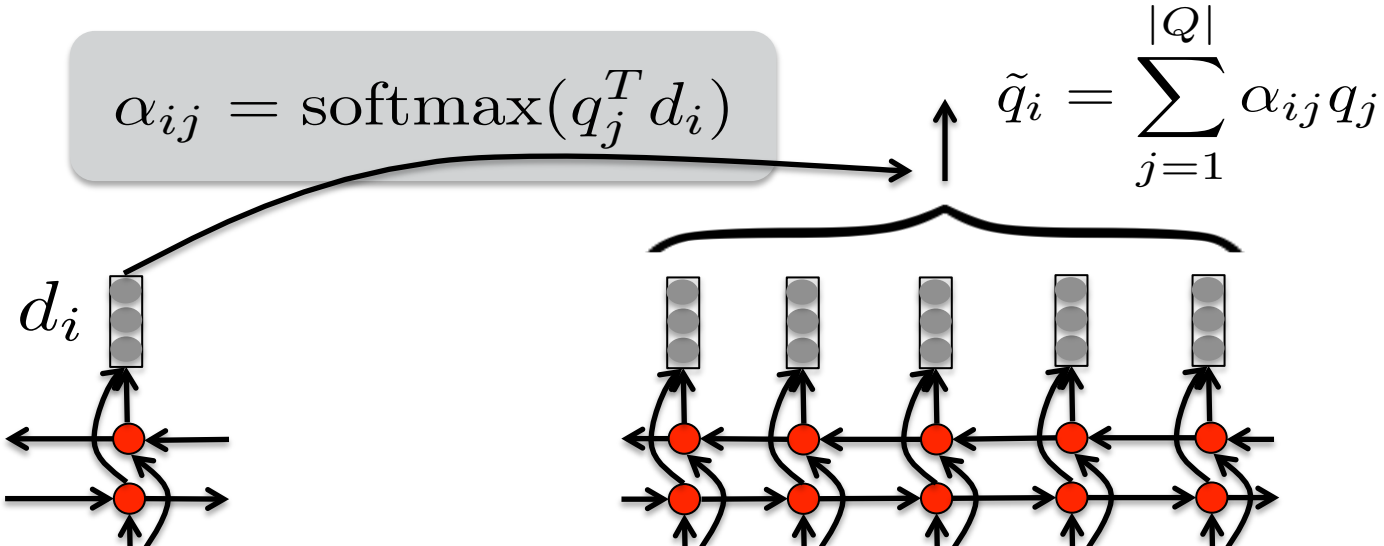


Model	CNN		Daily Mail		CBT-NE		CBT-CN	
	Val	Test	Val	Test	Val	Test	Val	Test
Humans (query) †	–	–	–	–	–	52.0	–	64.4
Humans (context + query) †	–	–	–	–	–	81.6	–	81.6
LSTMs (context + query) †	–	–	–	–	51.2	41.8	62.6	56.0
Deep LSTM Reader †	55.0	57.0	63.3	62.2	–	–	–	–
Attentive Reader †	61.6	63.0	70.5	69.0	–	–	–	–
Impatient Reader †	61.8	63.8	69.0	68.0	–	–	–	–
MemNets †	63.4	66.8	–	–	70.4	66.6	64.2	63.0
AS Reader †	68.6	69.5	75.0	73.9	73.8	68.6	68.8	63.4
DER Network †	71.3	72.9	–	–	–	–	–	–
Stanford AR (relabeling) †	73.8	73.6	77.6	76.6	–	–	–	–
Iterative Attentive Reader †	72.6	73.3	–	–	75.2	68.6	72.1	69.2
EpiReader †	73.4	74.0	–	–	75.3	69.7	71.5	67.4
AoA Reader †	73.1	74.4	–	–	77.8	72.0	72.2	69.4
ReasoNet †	72.9	74.7	77.6	76.6	–	–	–	–
NSE †	–	–	–	–	78.2	73.2	74.3	**71.9**
MemNets (ensemble) †	66.2	69.4	–	–	–	–	–	–
AS Reader (ensemble) †	73.9	75.4	78.7	77.7	76.2	71.0	71.1	68.9
Stanford AR (relabeling,ensemble) †	77.2	77.6	80.2	79.2	–	–	–	–
Iterative Attentive Reader (ensemble) †	75.2	76.1	–	–	76.9	72.0	74.1	71.0
EpiReader (ensemble) †	–	–	–	–	76.6	71.8	73.6	70.6
AS Reader (+BookTest) † ‡	–	–	–	–	80.5	76.2	83.2	80.8
AS Reader (+BookTest,ensemble) † ‡	–	–	–	–	*82.3*	*78.4*	*85.7*	*83.7*
GA--	73.0	73.8	76.7	75.7	74.9	69.0	69.0	63.9
GA (update $L(w)$)	**77.9**	**77.9**	**81.5**	**80.9**	76.7	70.1	69.8	67.3
GA (fix $L(w)$)	77.9	77.8	80.4	79.6	77.2	71.4	71.6	68.0
GA Reader (+feature, update $L(w)$)	77.3	76.9	80.7	80.0	77.2	73.3	73.0	69.8

Analysis of Attention

Document:
"...arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges ... included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama..."

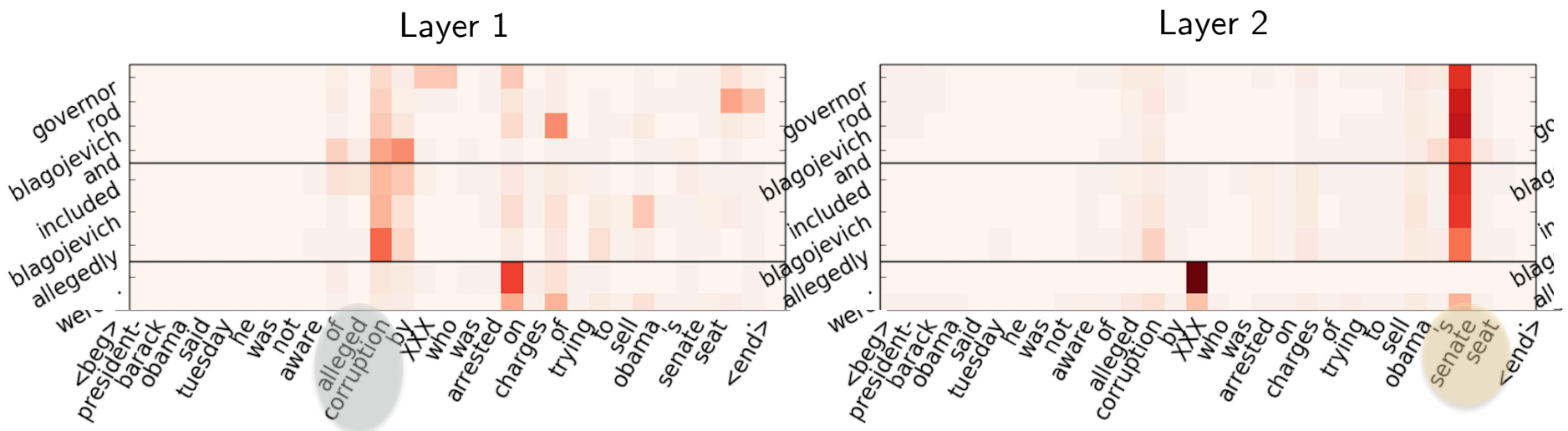
Query:
"President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat."
Find **X**.



Analysis of Attention

Document:
"...arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges ... included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama..."

Query:
"President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat."
Find **X**.



Summary So Far

- Multiplicative attention for document and query alignment
- Multiple layers allow model to focus on different salient aspects of the query

Code + Data: **https://github.com/bdhingra/ga-reader**

"Common" Sense?

Document:
"Eurythmics were a British music duo consisting of members **Annie Lennox** and David A. Stewart"

Query:
"Who was the female member of the 1980's pop music duo Eurythmics?"

Answer:
Annie Lennox

- Where can we find such knowledge?
- Need more datasets for testing this aspect too.

Talk Outline

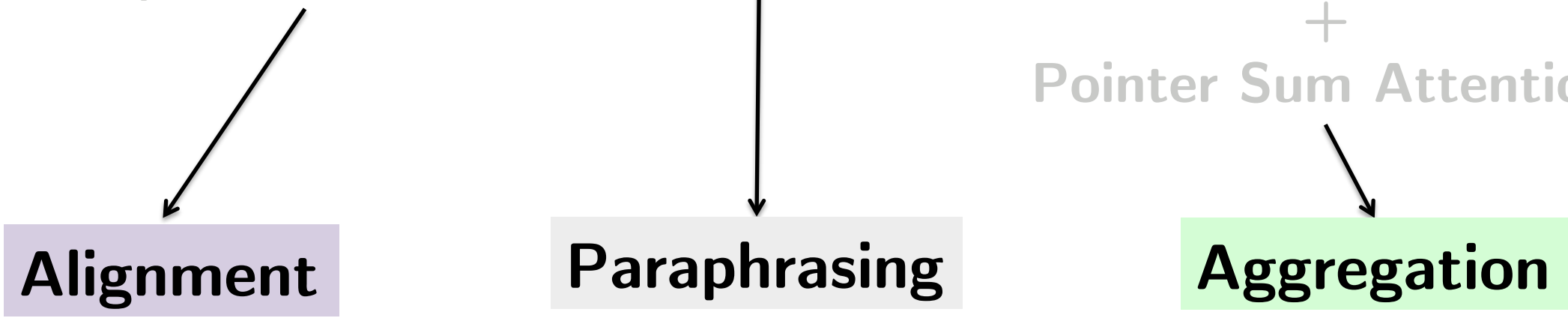
- Introduction to Reading Comprehension
- Language Modeling, XLNet and Transformer-XL: Modeling Long-Term Dependencies

Biases

Word Vectors + (RNNs or Transformers)
to represent Document and Query

Multiplicative Attention

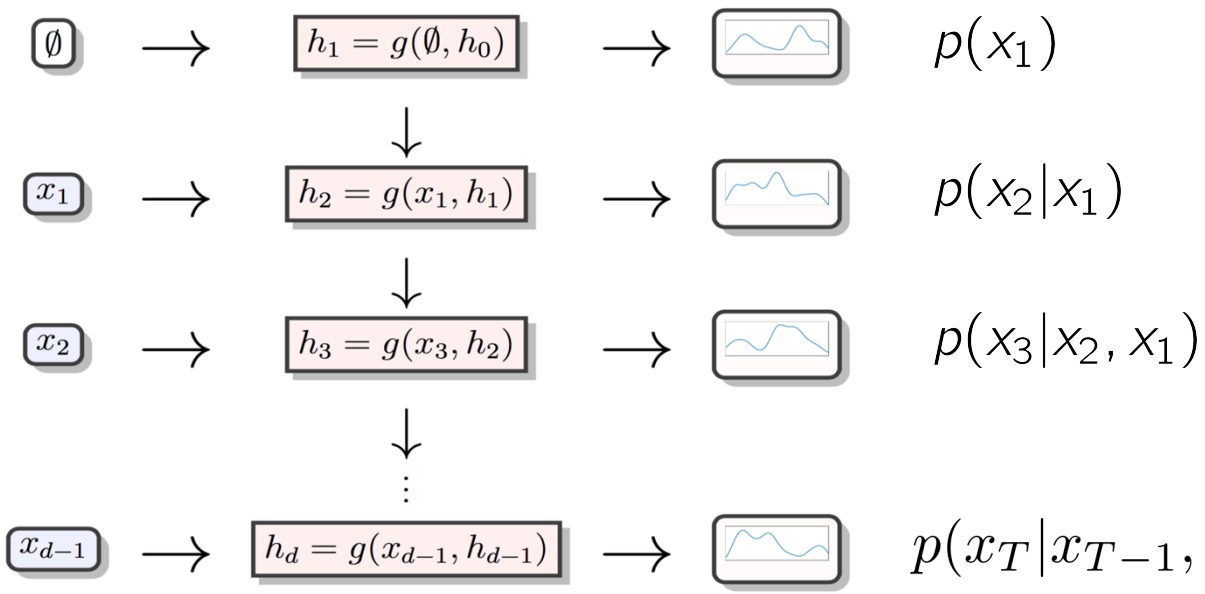
Multiple passes over the document
+
Pointer Sum Attention



Sequence Modelling

- Given a sequence of length T: Density Estimation by Autoregression

$$p(x_1, x_2, ..., x_T) = \prod_{i=1}^{T} p(x_i | x_{i-1}, ..., x_1) \approx \prod_{i=1}^{T} p(x_i | g(x_{i-1}, ..., x_1))$$



Each conditional can be a deep neural network

NADE (Uria 2013), MADE (Germain 2017), PixelCNN (van den Oord, et al, 2016)

Language Modeling

- Given a corpus of T sequential tokens (words) $\mathbf{x} = [x_1, x_2, ..., x_T]$, we model:

$$
\begin{aligned}
P_\theta(\mathbf{x}) &= \prod_{t=1}^{T} P_\theta(x_t \mid \mathbf{x}_{<t}) \\
&= \prod_{t=1}^{T} P(x_t \mid f_\theta(\mathbf{x}_{<t})) \\
&= \prod_{t=1}^{T} \frac{\exp\left(f_\theta(\mathbf{x}_{<t})^\top e_{x_t}\right)}{\sum_{x \in \mathcal{X}} \exp\left(f_\theta(\mathbf{x}_{<t})^\top e_x\right)}
\end{aligned}
$$

  - Here, we will focus on the choice f_θ

Context:

Context:
<unk> undergo <unk> reactions with <unk> to afford a number of unique five membered <unk>, as depicted in the figure below. This reactivity is due to the strained three membered ring and weak N-O bond.
= Battle of Dürenstein =
The Battle of Dürenstein (also known as the Battle of <unk>, Battle of <unk> and Battle of <unk>; German: <unk> bei <unk>), on 11 November 1805 was an engagement in the Napoleonic Wars during the War of the Third Coalition. Dürenstein (modern <unk>) is located in the <unk> Valley, on the River Danube, 73 kilometers (45 mi) upstream from Vienna, Austria. The river makes a crescent-shaped curve between <unk> and nearby Krems an der Donau and the battle was fought in the flood plain between the river and the mountains. At Dürenstein a combined force of Russian and Austrian troops trapped a French division commanded by Théodore Maxime Gazan. The French division was part of the newly created VIII Corps, the so-called Corps Mortier, under command of Édouard Mortier. In pursuing the Austrian retreat from Bavaria, Mortier had over-extended his three divisions along the north bank of the Danube. Mikhail <unk> Kutuzov, commander of the Coalition force, enticed Mortier to send Gazan's division into a trap and French troops were caught in a valley between two Russian columns. They were rescued by the timely arrival of a second division, under command of Pierre Dupont de l 'Étang. The battle extended well into the night. Both sides claimed victory. The French lost more than a third of their participants, and Gazan's division experienced over 40 percent losses. The Austrians and Russians also had heavy <unk> to 16 <unk> perhaps the most significant was the death in action of Johann Heinrich von Schmitt, one of Austria's most capable chiefs of staff. The battle was fought three weeks after the Austrian capitulation at Ulm and three weeks before the Russo-Austrian defeat at the Battle of Austerlitz. After Austerlitz Austria withdrew from the war. The French demanded a high indemnity and Francis II abdicated as Holy Roman Emperor, releasing the German states from their allegiance to the Holy Roman Empire.
= = Background = =
In a series of conflicts from 1803-15 known as the Napoleonic Wars, various European powers formed five coalitions against the First French Empire. Like the wars sparked by the French Revolution (1789), these further revolutionized the formation, organization and training of European armies and led to an unprecedented militarization, mainly due to mass conscription. Under the leadership of Napoleon, French power rose quickly as the Grande Armée conquered most of Europe, and collapsed rapidly

Generation:

In a series of conflicts from 1803-15 known as the Napoleonic Wars, various European powers formed five coalitions against the First French Empire. Like the wars sparked by the French Revolution (1789), these further revolutionized the formation, organization and training of European armies and led to an unprecedented militarization, mainly due to mass conscription. Under the leadership of Napoleon, French power rose quickly as the Grande Armée conquered most of Europe, and collapsed rapidly

Reference:

after the disastrous invasion of Russia in 1812. Napoleon's empire ultimately suffered complete military defeat in the 1813 – 14 campaigns, resulting in the restoration of the Bourbon monarchy in France. Although Napoleon made a spectacular return in 1815, known as the Hundred Days, his defeat at the Battle of Waterloo, the pursuit of his army and himself, his abdication and banishment to the Island of Saint Helena concluded the Napoleonic Wars.

= = Danube campaign = =

From 1803-06 the Third Coalition fought the First French Empire and its client states (see table at right). Although several naval battles determined control of the seas, the outcome of the war was decided on the continent, predominantly in two major land operations in the Danube valley: the Ulm campaign in the upper Danube and the Vienna campaign, in the middle Danube valley. Political conflicts in Vienna delayed Austria's entry into the Third Coalition until 1805. After hostilities of the War of the Second Coalition ended in 1801, Archduke <unk> emperor's <unk> advantage of the subsequent years of peace to develop a military restructuring plan. He carefully put this plan into effect beginning in 1803 – 04, but implementation was incomplete in 1805 when Karl Mack, Lieutenant Field Marshal and Quartermaster-General of the Army, implemented his own restructuring. Mack bypassed Charles ' methodical approach. Occurring in the field, Mack's plan also undermined the overall command and organizational structure. Regardless, Mack sent an enthusiastic report to Vienna on the military's readiness. Furthermore, after misreading Napoleon's maneuvers in Württemberg, Mack also reported to Vienna on the weakness of French dispositions. His reports convinced the war party advising the emperor, Francis II, to enter the conflict against France, despite Charles ' own advice to the contrary. Responding to the report and rampant anti-French fever in Vienna, Francis dismissed Charles from his

Generation:

In a series of conflicts from 1803-15 known as the Napoleonic Wars, various European powers formed five coalitions against the First French Empire. Like the wars sparked by the French Revolution (1789), these further revolutionized the formation, organization and training of European armies and led to an unprecedented militarization, mainly due to mass conscription. Under the leadership of Napoleon, French power rose quickly as the Grande Armée conquered most of Europe, and collapsed rapidly

XL Generation:

after the French conquest of Italy. This victory marked the beginning of the Second Coalition. Napoleon's rapid advance caused Austria – Russia, Britain and Russia to make peace negotiations. The Russian army under Prince Mikhail Mikhailovich Mikhailovich Chaffee, commander of the Imperial Russian Army in Switzerland, was able to drive off the Austrians in the Battle of Stockach (1805) and to take Pressburg (modern Ö3 Austria) from the Austrians. At an early stage the Austro-Hungarian fleet had been damaged in a storm off Cape Matapan; this allowed the Allies to launch a full-scale invasion of Italy in February 1805. The Battle of Austerlitz was the decisive French victory against Napoleon and the largest naval battle in history, involving more modern-day European navies. The French military situation worsened as Napoleon faced several threats related to his newly formed Army of Europe, particularly Russia itself, which was now on the verge of collapse. The Russian general and friend of Napoleon, Alexander, had been dissatisfied with the conduct of the Austrians; he was still in conflict with Emperor Napoleon, the French Republic's king, who had declared war on Austria during the summer of 1804. With the war over, Napoleon decided to launch a second invasion of Italy in October 1805.

= = Prelude = =

In July 1805, the French 1st Army entered southern Italy. The army, under the command of Marshal Marmont, were reinforced by a few battalions of infantry under Claude General Auguste de Marmont at the town of Philippsburg and another battalion at Belluno. On 17 September 1805, the army marched from Belluno towards Krems. By 29 September, they had reached Belluno and conducted its advance against a small Austrian force. By 31 September, the whole force had been reinforced by a brigade from the Army of Tyrol under the command of Pierre Augereau. The Austrians were now under the command of Marshal Jean Victor Marie Moreau, a member of

Reference:

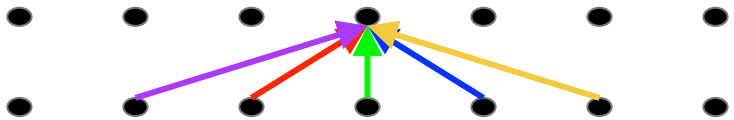
after the disastrous invasion of Russia in 1812. Napoleon's empire ultimately suffered complete military defeat in the 1813 – 14 campaigns, resulting in the restoration of the Bourbon monarchy in France. Although Napoleon made a spectacular return in 1815, known as the Hundred Days, his defeat at the Battle of Waterloo, the pursuit of his army and himself, his abdication and banishment to the Island of Saint Helena concluded the Napoleonic Wars.

= = Danube campaign = =

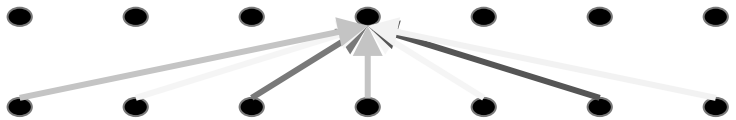
From 1803-06 the Third Coalition fought the First French Empire and its client states (see table at right). Although several naval battles determined control of the seas, the outcome of the war was decided on the continent, predominantly in two major land operations in the Danube valley: the Ulm campaign in the upper Danube and the Vienna campaign, in the middle Danube valley. Political conflicts in Vienna delayed Austria's entry into the Third Coalition until 1805. After hostilities of the War of the Second Coalition ended in 1801, Archduke <unk> emperor's <unk> advantage of the subsequent years of peace to develop a military restructuring plan. He carefully put this plan into effect beginning in 1803 – 04, but implementation was incomplete in 1805 when Karl Mack, Lieutenant Field Marshal and Quartermaster-General of the Army, implemented his own restructuring. Mack bypassed Charles ' methodical approach. Occurring in the field, Mack's plan also undermined the overall command and organizational structure. Regardless, Mack sent an enthusiastic report to Vienna on the military's readiness. Furthermore, after misreading Napoleon's maneuvers in Württemberg, Mack also reported to Vienna on the weakness of French dispositions. His reports convinced the war party advising the emperor, Francis II, to enter the conflict against France, despite Charles ' own advice to the contrary. Responding to the report and rampant anti-French fever in Vienna, Francis dismissed Charles from his

Self Attention

Convolution

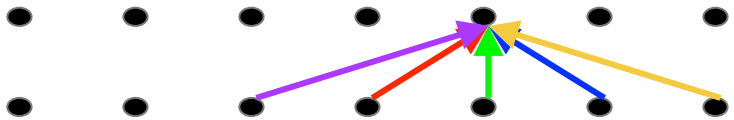


Self-Attention

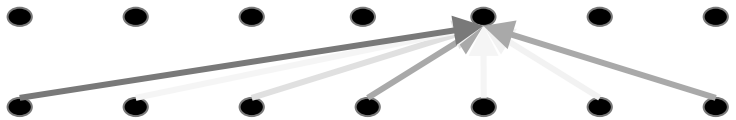


Self Attention

Convolution



Self-Attention



XLNet

- XLNet is a state-of-the-art pretrained model for language understanding
- When to use it?
  - If you are interested in classification, regression, question answering, natural language inference, document ranking, ...
  - Basically any task that maps text to some form of labels/structures
- How to use it?
  - Collect a dataset of interest
  - Finetune XLNet on the dataset
  - See more examples and instructions at https://github.com/zihangdai/xlnet

Yang, Dai, Carbonell, Salakhutdinov, Le. 2019

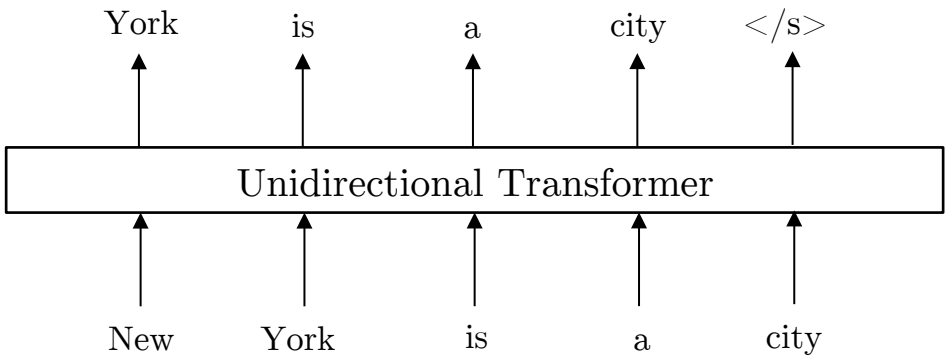
How Pretraining Works

- Pretrain a model on unlabeled data based on language modeling
- Finetune the model or use the model for feature extraction on downstream tasks

word2vec (Mikolov et al.) GloVe (Pennington et al.)
semi-supervised sequence learning (Dai and Le)
ELMo (Peters et al.), CoVe (McCann et al.)
GPT (Radford et al.), BERT (Devlin et al.)

Two Objectives for Pretraining

Auto-regressive (AR) language modeling

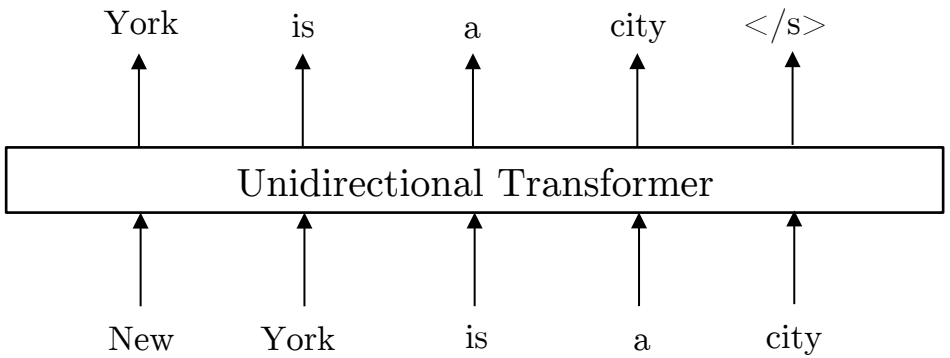


$$\log p(\mathbf{x}) = \sum_{t=1}^{T} \log p(x_t | \mathbf{x}_{<t})$$

Not able to model bidirectional context. ☹

Two Objectives for Pretraining

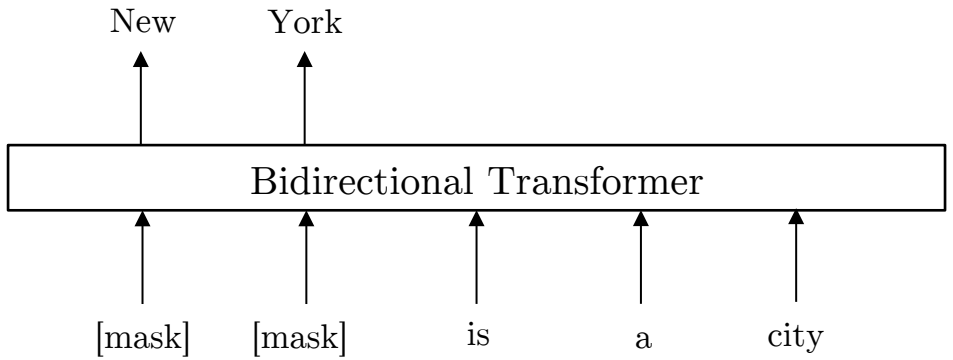
Auto-regressive (AR) language modeling



$$\log p(\mathbf{x}) = \sum_{t=1}^{T} \log p(x_t | \mathbf{x}_{<t})$$

Not able to model bidirectional context. ☹

(Denoising) Auto-encoding (AE)



$$\log p(\bar{\mathbf{x}} | \hat{\mathbf{x}}) = \sum_{t=1}^{T} \text{mask}_t \log p(x_t | \hat{\mathbf{x}})$$

Predicted tokens are independent of each other. ☹

New Objective: Permutation Language Modeling

- Sample a factorization order z
- Determine the attention masks based on the order
- Optimize a standard language modeling objective

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{Z}_T}\left[\sum_{t=1}^{T}\log p(x_{z_t}|\mathbf{x}_{\mathbf{z}<t})\right]$$

- Note that:

$$\begin{aligned} p(a,b) &= p(a)p(b|a) \\ &= p(b)p(a|b) \end{aligned}$$

Example

- Sentence: New York is a city
- Factorization order: New York is a city

$$\begin{aligned}&\text{P(New York is a city)}\\&\quad = \text{P(New) * P(York | New) * P(is | New York) * P(a | New York is) * P(city | New York is a)}\end{aligned}$$

Example

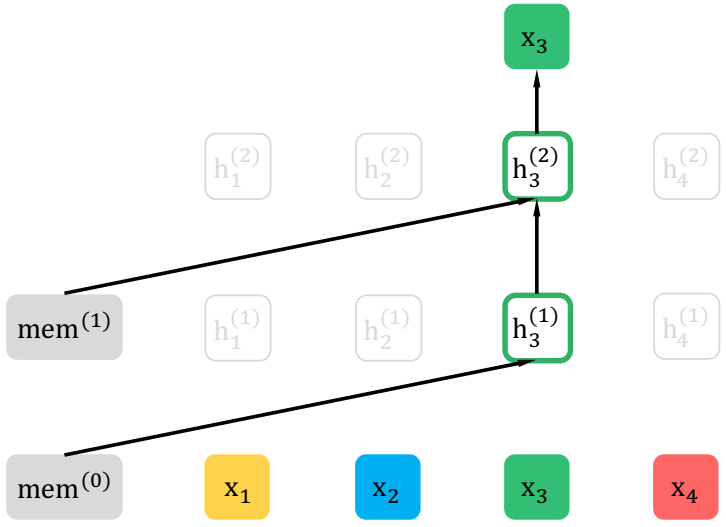
- Sentence: New York is a city
- Factorization order: New York is a city

$$\begin{aligned}&\text{P(New York is a city)}\\&\quad = \text{P(New) * P(York | New) * P(is | New York) * P(a | New York is) * P(city | New York is a)}\end{aligned}$$

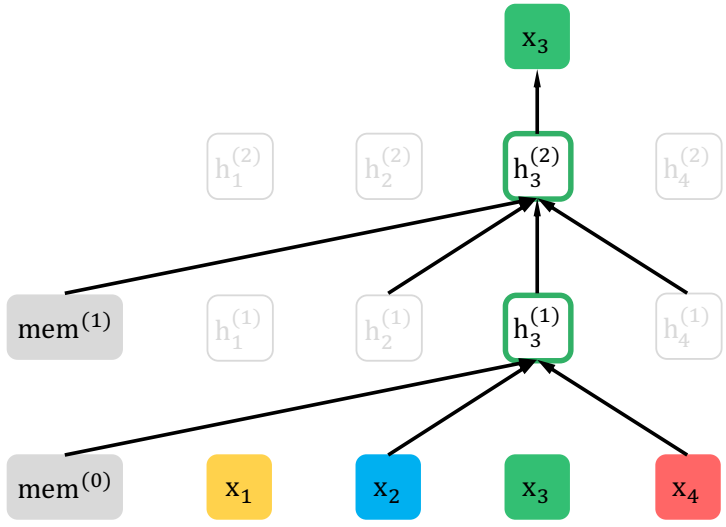
- Factorization order: city a is New York

$$\begin{aligned}&\text{P(New York is a city)}\\&\quad = \text{P(city) * P(a | city) * P(is | city a) * P(New | city a is) * P(York | city a is New)}\end{aligned}$$

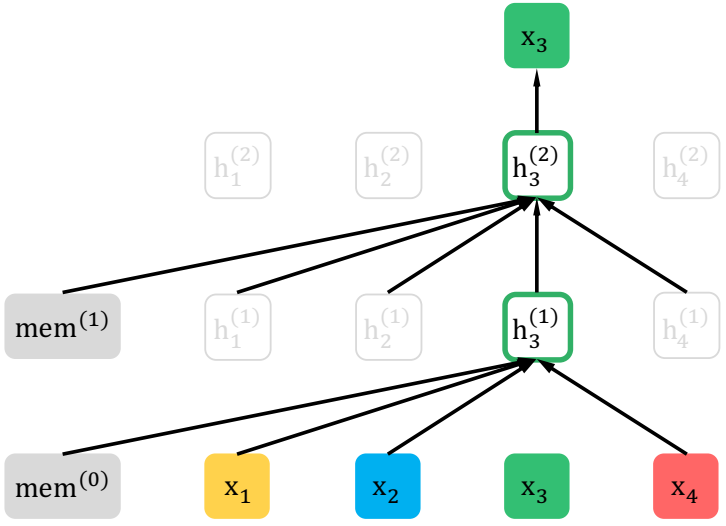
- **Sequence order is not shuffled**
- Attention masks are changed to reflect factorization order



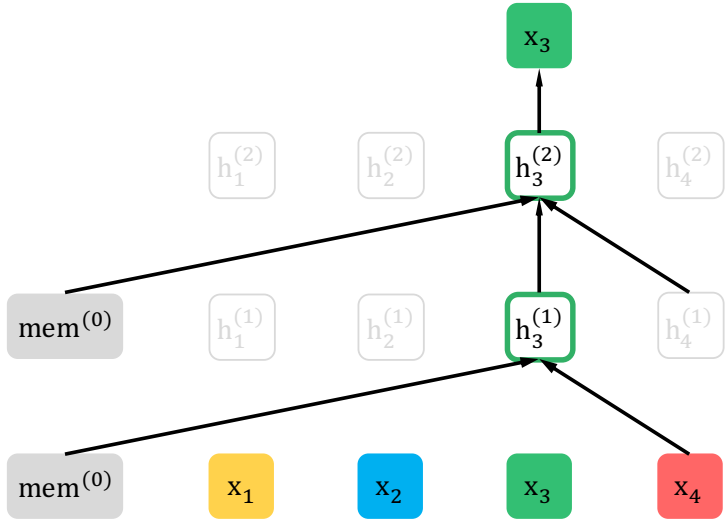
Factorization order: 3 → 2 → 4 → 1



Factorization order: 2 → 4 → 3 → 1



Factorization order: 1 → 4 → 2 → 3



Factorization order: 4 → 3 → 1 → 2

Comparing XLNet and BERT Objectives

- BERT objective (auto-encoding)

$$\mathcal{J}_{\text{BERT}} = \log p(\text{New} \mid \text{is a city}) + \log p(\text{York} \mid \text{is a city})$$

New and York are independent ☹

Comparing XLNet and BERT Objectives

- BERT objective (auto-encoding)

$$\mathcal{J}_{\text{BERT}} = \log p(\text{New} \mid \text{is a city}) + \log p(\text{York} \mid \text{is a city})$$

New and York are independent ☹

- XLNet objective (auto-regressive)

$$\mathcal{J}_{\text{XLNet}} = \log p(\text{New} \mid \text{is a city}) + \log p(\text{York} \mid \text{New}, \text{is a city})$$

$$\mathcal{J}_{\text{XLNet}} = \log p(\text{New} \mid \text{York}, \text{is a city}) + \log p(\text{York} \mid \text{is a city})$$

Able to model dependency between New and York ☺
Able to model bidirectional context ☺
Factorize the joint probability using a product rule that holds universally

Reparameterization

- Standard Parameterization

$$p_\theta(X_{z_t} = x \mid \mathbf{x}_{\mathbf{z}_{<t}}) = \frac{e(x)^\top h_\theta(\mathbf{x}_{\mathbf{z}_{<t}})}{\sum_{x'} e(x')^\top h_\theta(\mathbf{x}_{\mathbf{z}_{<t}})}$$

  - h does not contain the position of the target. Reduced to predicting a bag of words:

$$p(X_3 = \text{is} \mid \text{New York}) = p(X_4 = \text{is} \mid \text{New York}) = p(X_5 = \text{is} \mid \text{New York})$$

Reparameterization

- Standard Parameterization

$$p_\theta(X_{z_t} = x \mid \mathbf{x}_{\mathbf{z}_{<t}}) = \frac{e(x)^\top h_\theta(\mathbf{x}_{\mathbf{z}_{<t}})}{\sum_{x'} e(x')^\top h_\theta(\mathbf{x}_{\mathbf{z}_{<t}})}$$

  - h does not contain the position of the target. Reduced to predicting a bag of words:

$$p(X_3 = \text{is} \mid \text{New York}) = p(X_4 = \text{is} \mid \text{New York}) = p(X_5 = \text{is} \mid \text{New York})$$

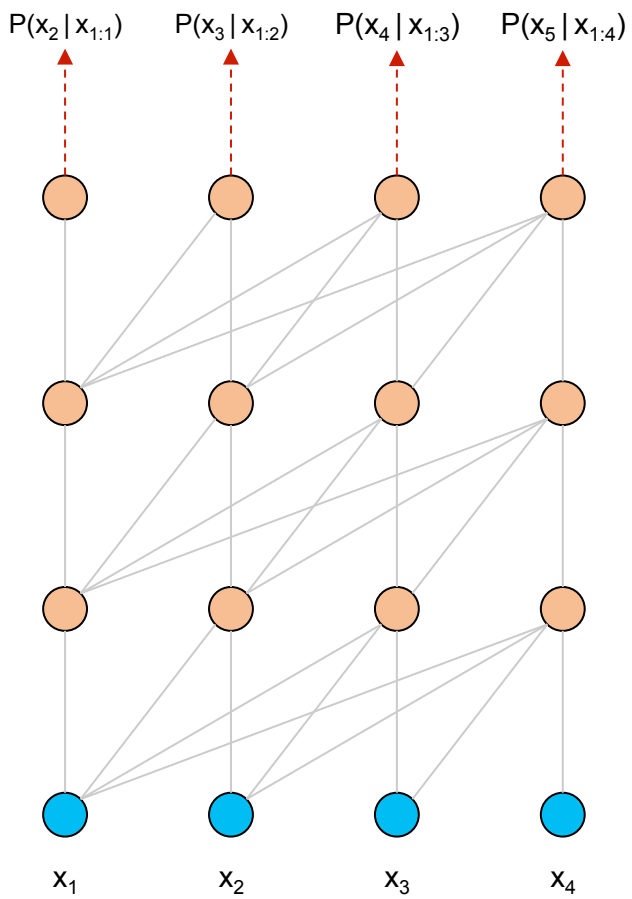
- Solution: condition the distribution on the position.

$$p_\theta(X_{z_t} = x \mid \mathbf{x}_{\mathbf{z}_{<t}}) = \frac{e(x)^\top g_\theta(\mathbf{x}_{\mathbf{z}_{<t}}, z_t)}{\sum_{x'} e(x')^\top g_\theta(\mathbf{x}_{\mathbf{z}_{<t}}, z_t)}$$

XLNet

- Language Modeling (LM) has a deep root in density estimation
  - Rapidly developing
- With XLNet, pretraining is reduced to a LM problem
- Model longer-term dependencies with Transformer-XL

Vanilla Transformer Language Models



Forward Pass

Step 1: break the corpus into segments

$$\mathbf{x} = \Big[\underbrace{(x_1, x_2, \ldots, x_L)}_{\text{segment } 1}, \cdots, \underbrace{(x_{(\tau-1)L+1}, x_{(\tau-1)L+2}, \ldots, x_{\tau T})}_{\text{segment } \tau}, \cdots\Big]$$

$$= \Big[\underbrace{(x_{1,1}, x_{1,2}, \ldots, x_{1,L})}_{\mathbf{s}_1}, \cdots, \underbrace{(x_{\tau,1}, x_{\tau,2}, \ldots, x_{\tau,L})}_{\mathbf{s}_\tau}, \cdots\Big]$$

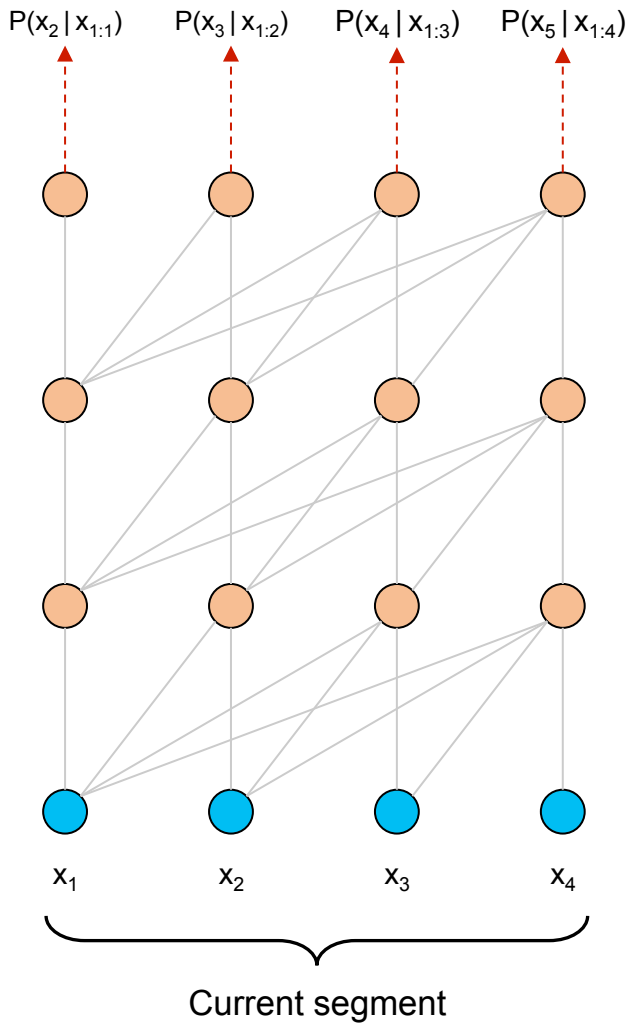
Step 2: Model each segment **independently** (limited memory)

$$P(\mathbf{x}) = \prod_{\tau} P(\mathbf{s}_\tau \mid \mathbf{s}_{<\tau}) \approx \prod_{\tau} P(\mathbf{s}_\tau) \qquad \text{(independence assumption)}$$

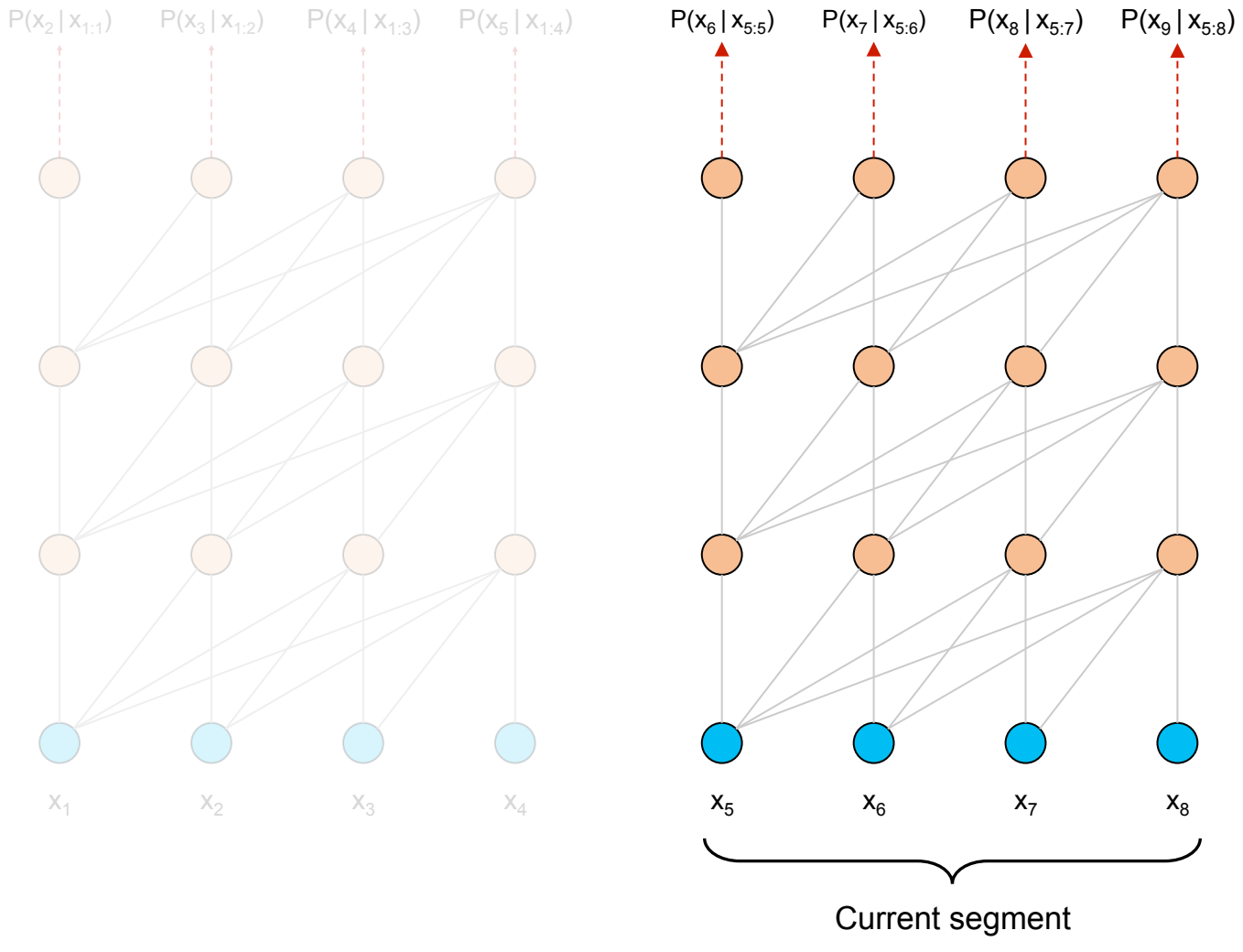
$$= \prod_{\tau} \prod_{i=1}^{L} P(x_{\tau,i} \mid \mathbf{x}_{\tau,<i}) = \prod_{\tau} \prod_{I=1}^{L} P(x_{\tau,i} \mid f(\mathbf{x}_{\tau,<I}))$$

Al-Rfou et al., 2016

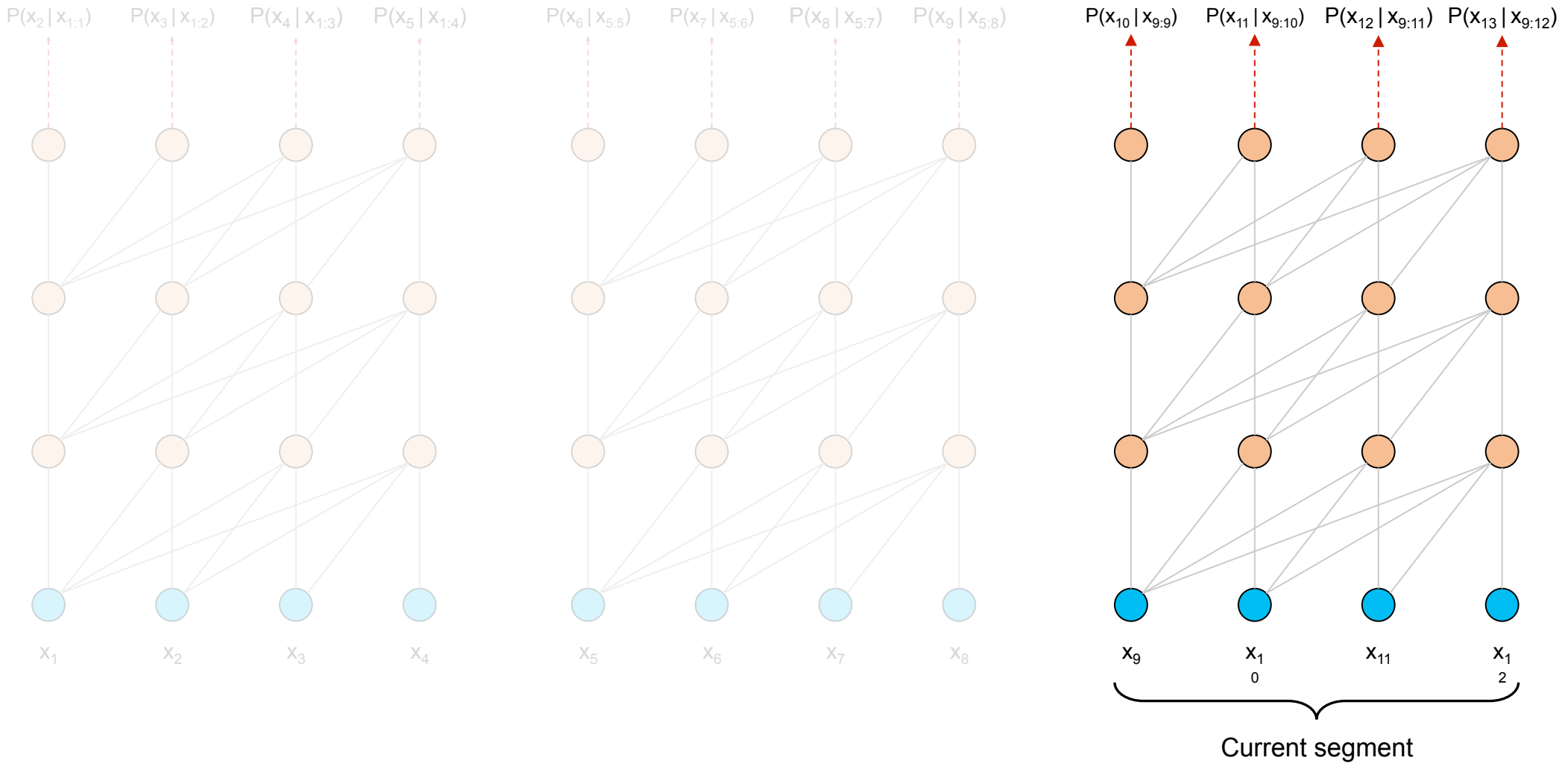
Training with the Vanilla Model



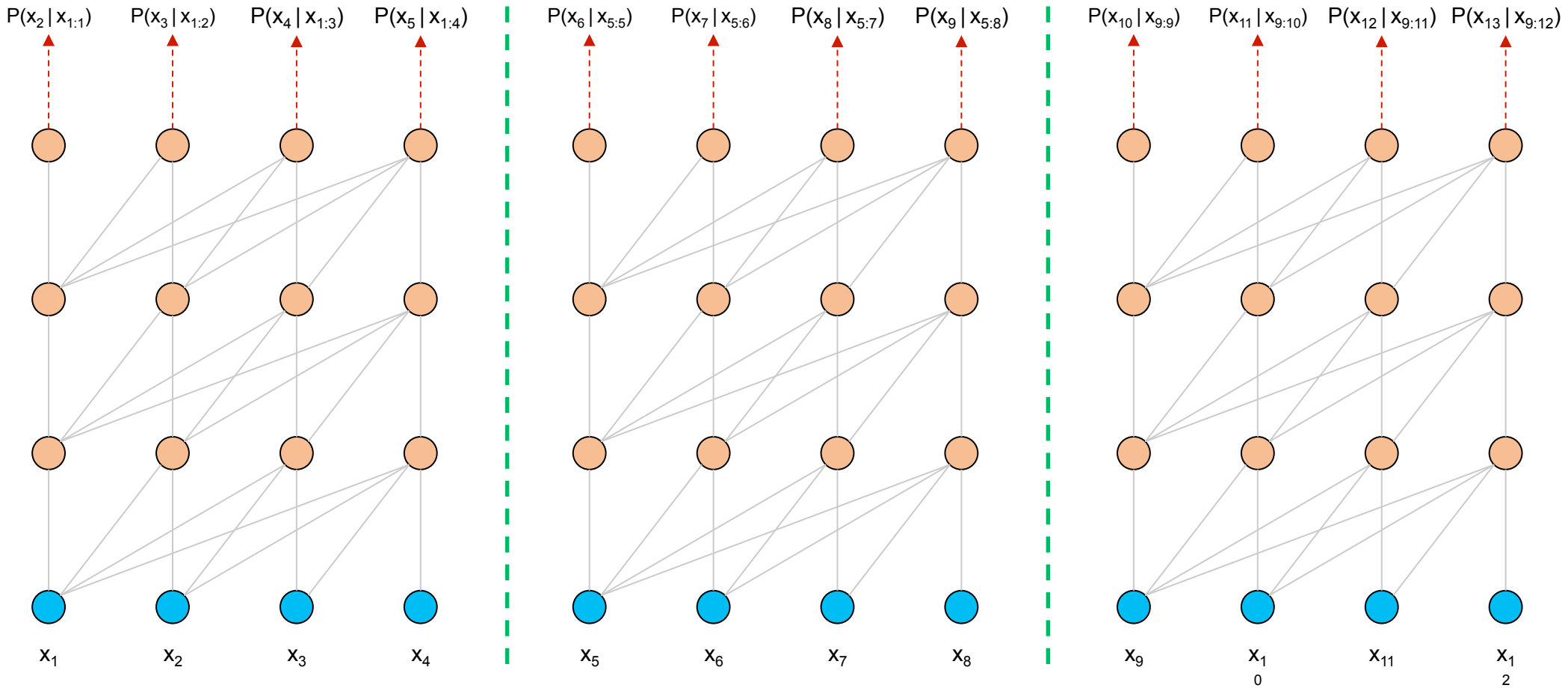
Training with the Vanilla Model



Training with the Vanilla Model



Training with the Vanilla Model



Transformer-XL for Language Modeling

Key Ideas

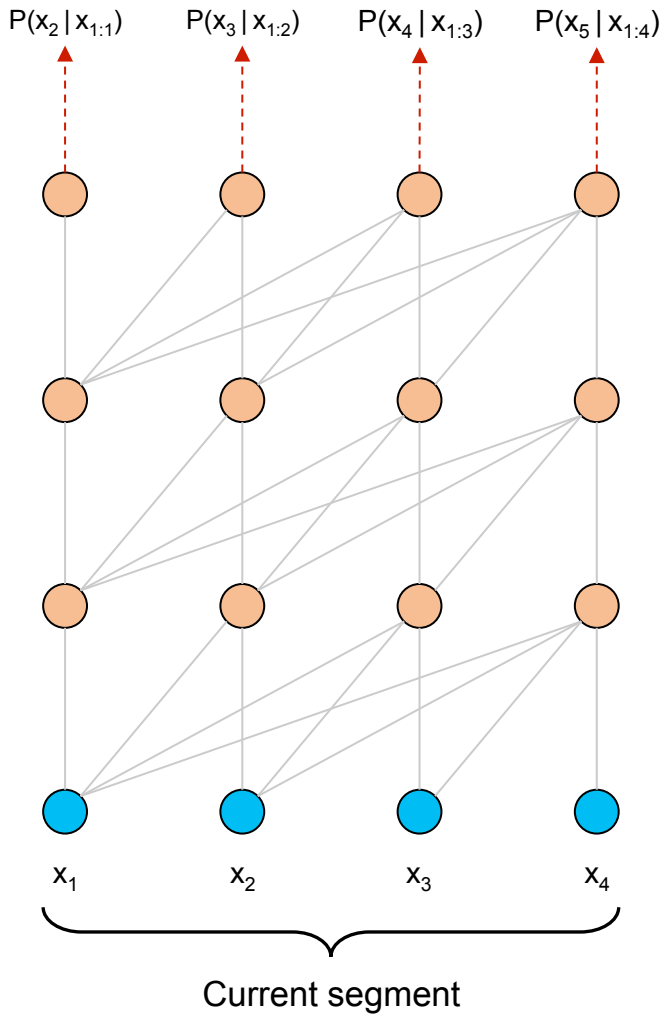
(1) Cache & Reuse previously computed hidden states

- Analogous to Truncated BPTT for RNN: pass the last hidden state to the next segment as the initial hidden
- **Segment-Level Recurrence**

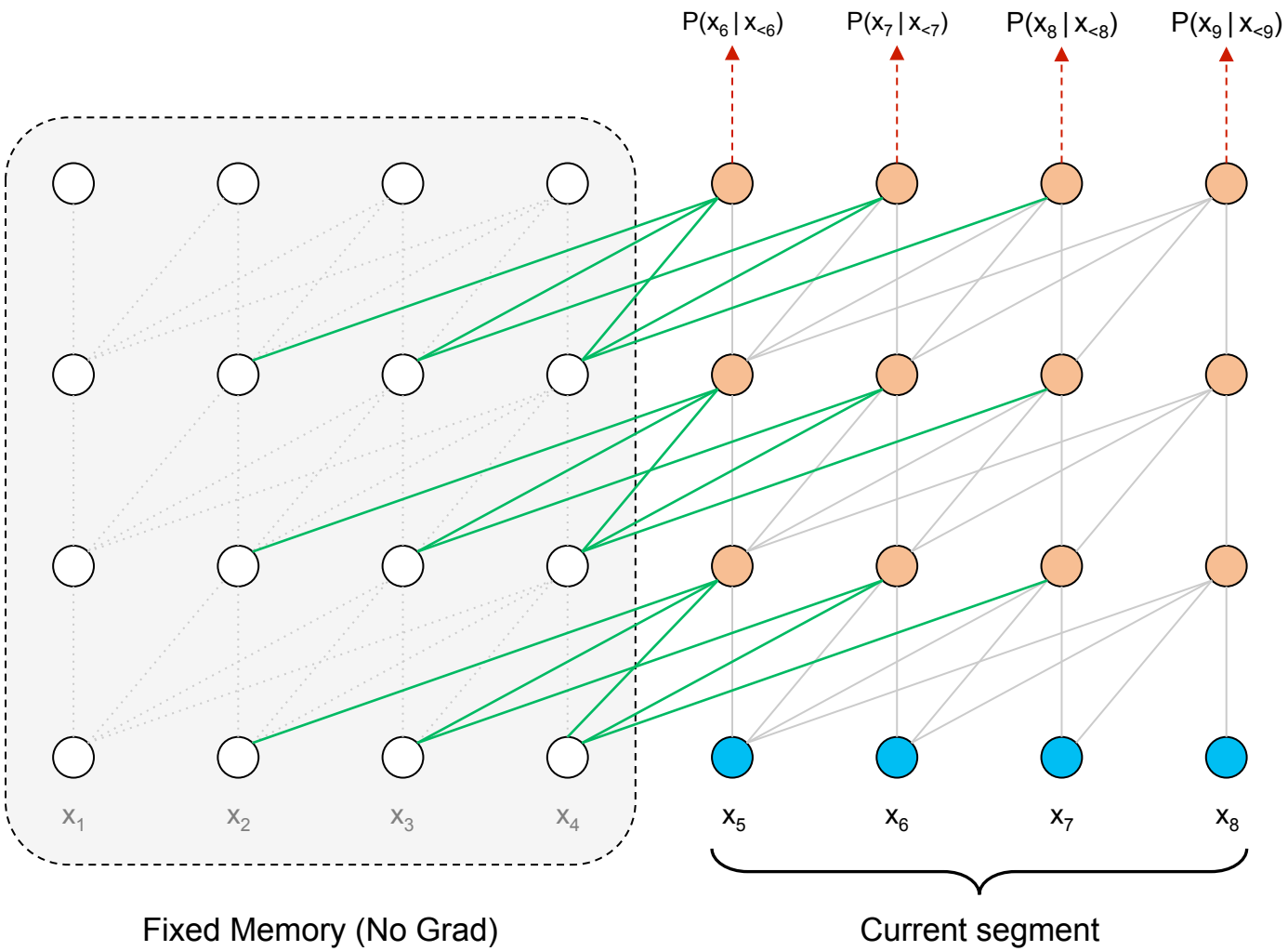
(2) Keep temporal information **"coherent"**

Dai, Yang, Yang, Carbonell, Le, Salakhutdinov. 2019

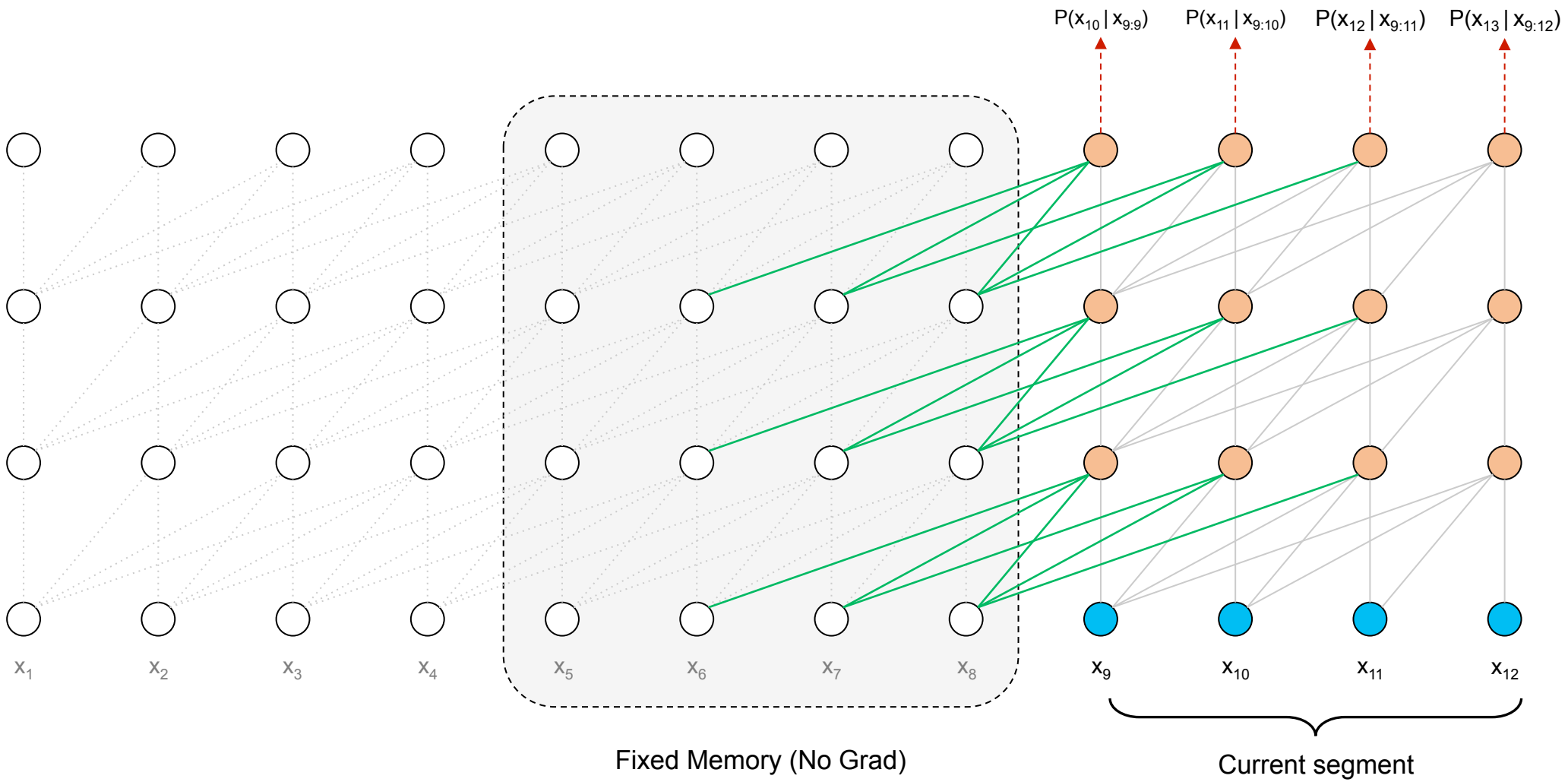
Training with Transformer-XL



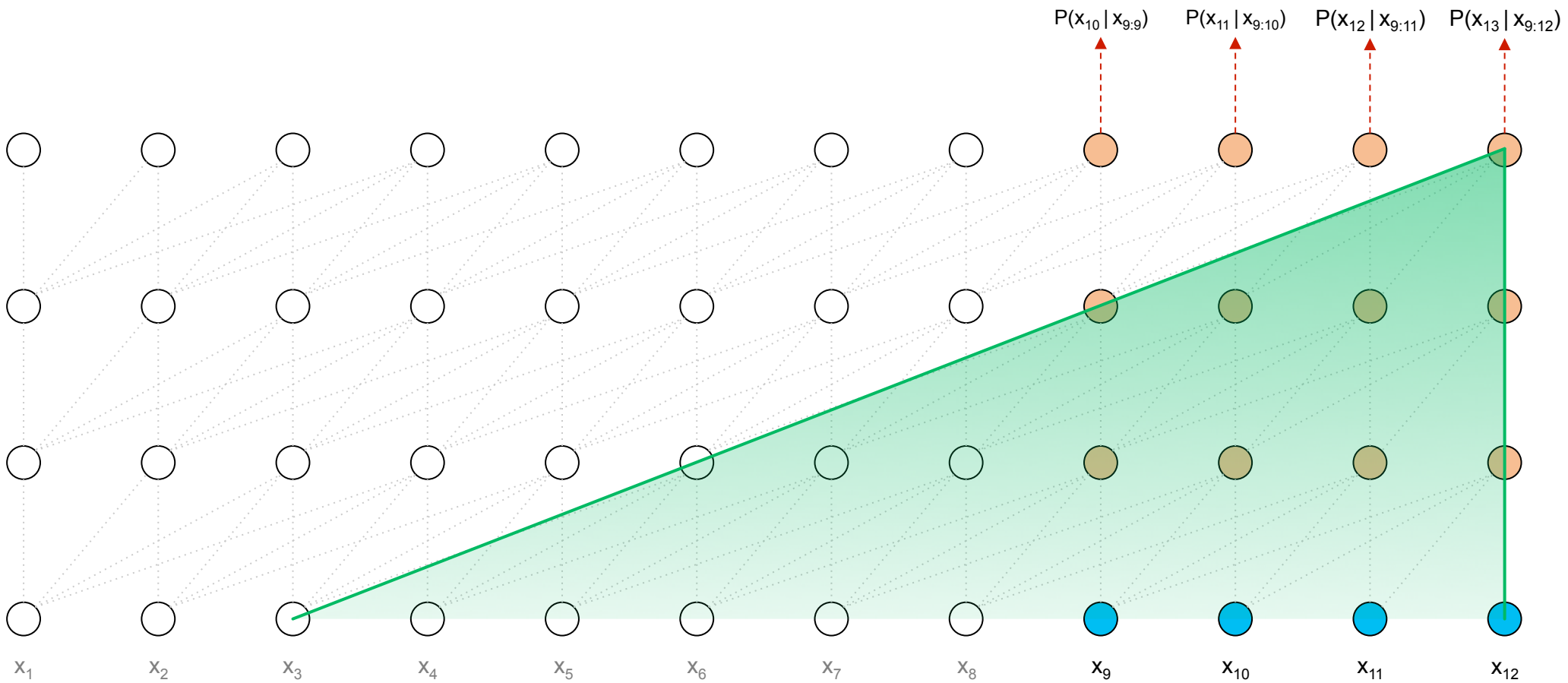
Training with Transformer-XL



Training with Transformer-XL

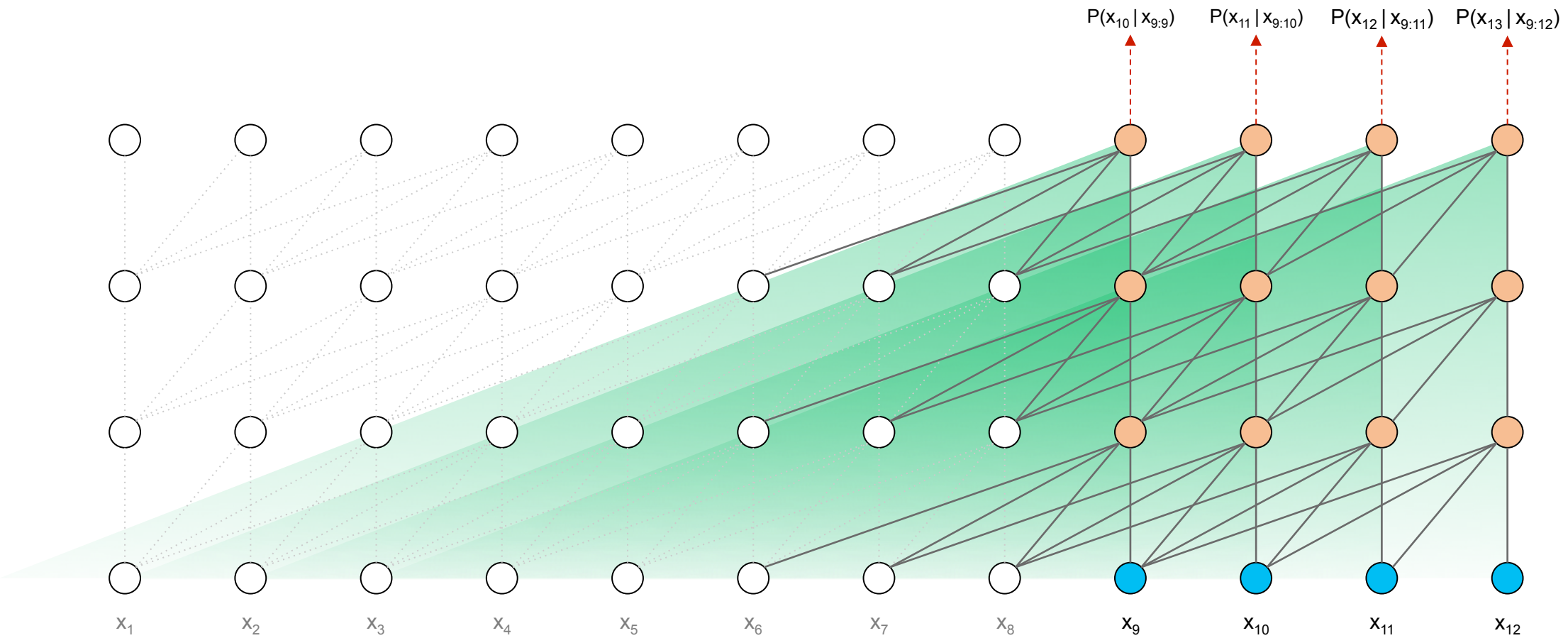


Modeling Much Longer Context



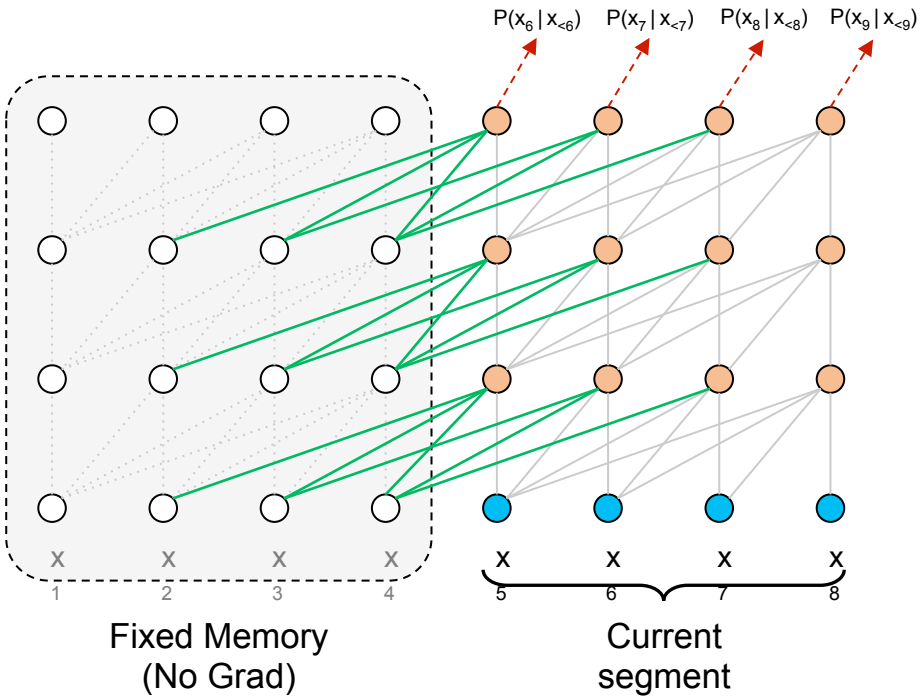
Extra-Long context span: linearly increasing w.r.t. both **segment length** and **number of layers**

Evaluation with Transformer-XL



Process the tokens in a segment in **one forward pass**, without any recomputing → 1800X faster

Transformer-XL



$$\widetilde{\mathbf{h}}_{\tau}^{n-1} = \left[\mathrm{SG}(\mathbf{m}_{\tau}^{n-1}) \circ \mathbf{h}_{\tau}^{n-1}\right]$$

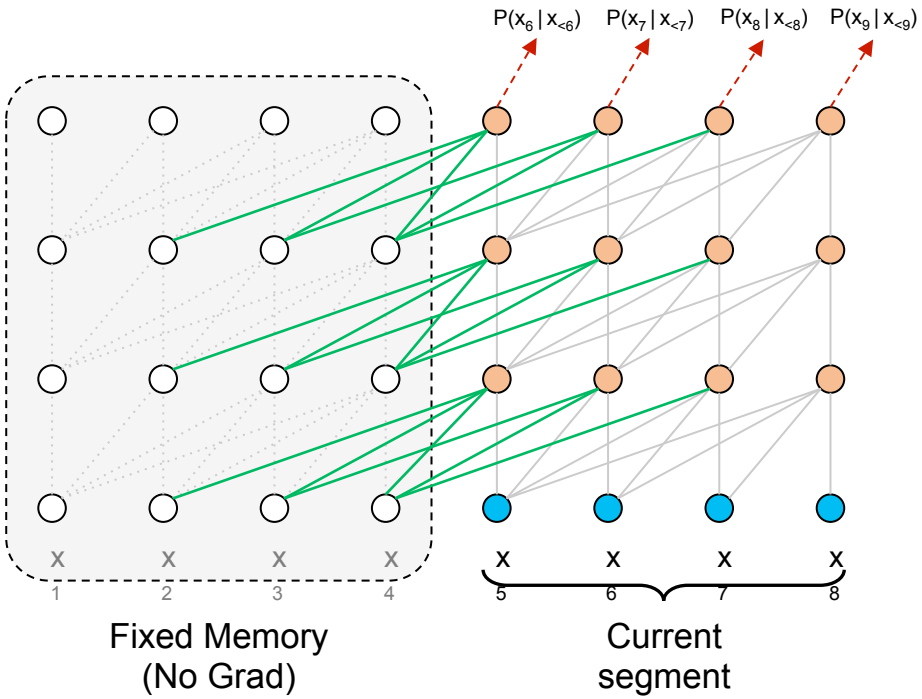
- $\mathbf{h}_{\tau}^{n} \in \mathbb{R}^{L \times d}$ n-th layer hidden state sequence produced for sequence τ
- $\mathbf{m}_{\tau}^{n-1}$ is memory cashed before segment τ
- SG stands for stop gradient
- $[\cdot \circ \cdot]$ stands for concatenation
- Incorporate extended context

$$\mathbf{q}_{\tau}^{n} = \mathbf{h}_{\tau}^{n-1}\mathbf{W}_{q}$$
$$\mathbf{k}_{\tau}^{n} = \tilde{\mathbf{h}}_{\tau}^{n-1}\mathbf{W}_{k}$$
$$\mathbf{v}_{\tau}^{n} = \tilde{\mathbf{h}}_{\tau}^{n-1}\mathbf{W}_{v}$$

Model parameters

$$\mathbf{h}_{\tau}^{n} = \text{Transformer-Layer}\left(\mathbf{q}_{\tau}^{n}, \mathbf{k}_{\tau}^{n}, \mathbf{v}_{\tau}^{n}\right)$$

Transformer-XL



$$\widetilde{\mathbf{h}}_{\tau}^{n-1} = \left[\mathrm{SG}(\mathbf{m}_{\tau}^{n-1}) \circ \mathbf{h}_{\tau}^{n-1}\right]$$

- $\mathbf{h}_{\tau}^{n} \in \mathbb{R}^{L \times d}$ n-th layer hidden state sequence produced for sequence τ
- $\mathbf{m}_{\tau}^{n-1}$ is memory cashed before segment τ
- SG stands for stop gradient
- $[\cdot \circ \cdot]$ stands for concatenation
- Incorporate extended context

$$\mathbf{q}_{\tau}^{n} = \mathbf{h}_{\tau}^{n-1}\mathbf{W}_{q}$$
$$\mathbf{k}_{\tau}^{n} = \widetilde{\mathbf{h}}_{\tau}^{n-1}\mathbf{W}_{k}$$
$$\mathbf{v}_{\tau}^{n} = \widetilde{\mathbf{h}}_{\tau}^{n-1}\mathbf{W}_{v}$$

Model parameters

Extended context at layer n-1

$$\mathbf{h}_{\tau}^{n} = \text{Transformer-Layer}\left(\mathbf{q}_{\tau}^{n}, \mathbf{k}_{\tau}^{n}, \mathbf{v}_{\tau}^{n}\right)$$

Dai, Yang, Yang, Carbonell, Le, Salakhutdinov. 2019

Results

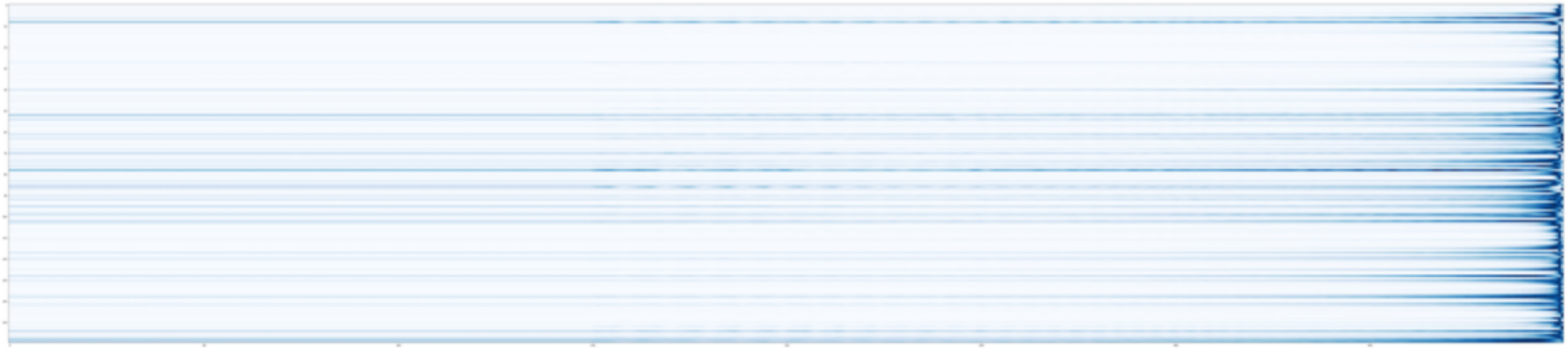
- WikiText-103 Test Corpus:
  - 103M tokens from 28K articles
  - Average length is 3.6K per article
- Training
  - Training segment length 400
  - Test segment length 1600
  - 16 layers

Model	#Param	PPL
Grave et al. (2016b) - LSTM	-	48.7
Bai et al. (2018) - TCN	-	45.2
Dauphin et al. (2016) - GCNN-8	-	44.9
Grave et al. (2016b) - LSTM + Neural cache	-	40.8
Dauphin et al. (2016) - GCNN-14	-	37.2
Merity et al. (2018) - QRNN	151M	33.0
Rae et al. (2018) - Hebbian + Cache	-	29.9
Ours - Transformer-XL Standard	151M	**24.0**
Baevski and Auli (2018) - Adaptive Input$^{\diamond}$	247M	20.5
Ours - Transformer-XL Large	257M	**18.3**

- Achieves State-of-the-Art on 5 publicly available datasets

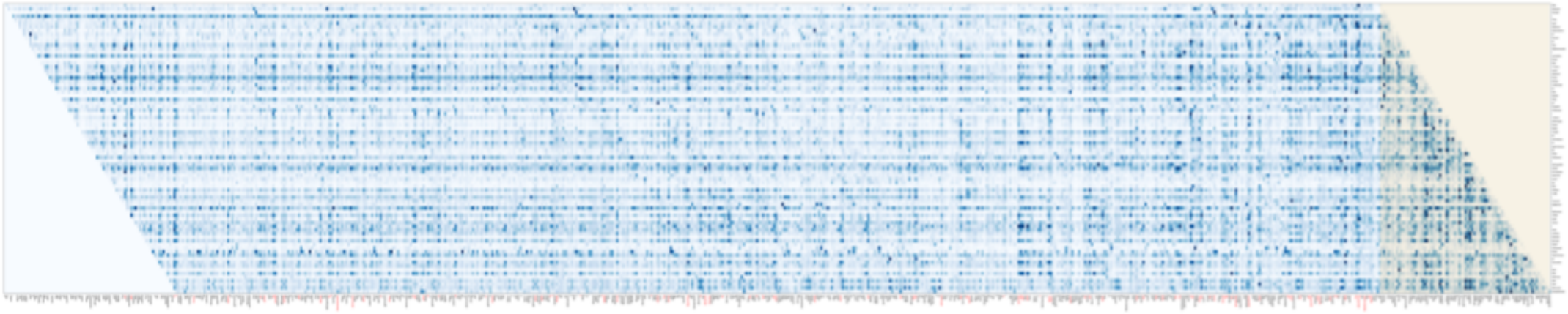
Dai, Yang, Yang, Carbonell, Le, Salakhutdinov. 2019

Visualization of Attention:



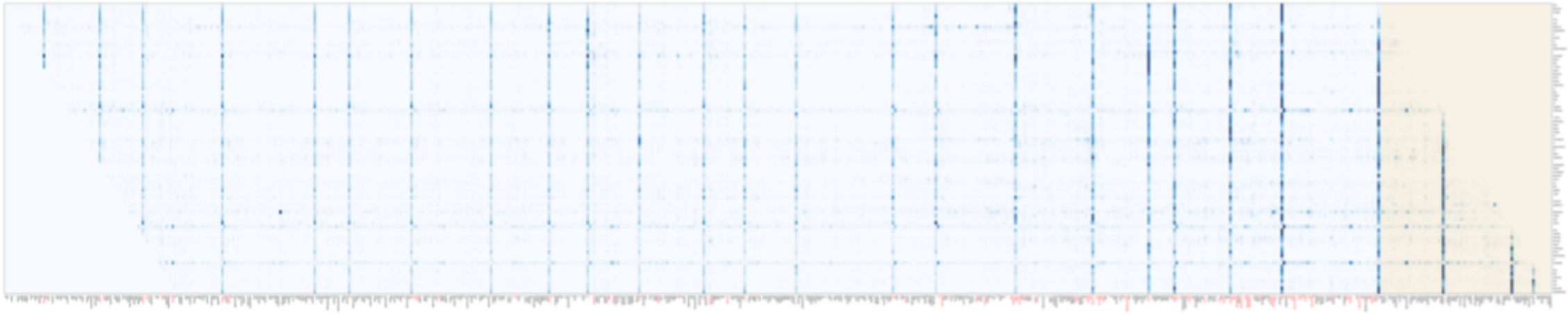
- Average attention over the previous 640 tokens,
- There are totally 160 attention weights across 16 layers

Visualization of Attention:



- Average attention over the previous 640 tokens,
- There are totally 160 attention weights across 16 layers

Visualization of Attention:



- Average attention over the previous 640 tokens,
- There are totally 160 attention weights across 16 layers

Generation:

In a series of conflicts from 1803-15 known as the Napoleonic Wars, various European powers formed five coalitions against the First French Empire. Like the wars sparked by the French Revolution (1789), these further revolutionized the formation, organization and training of European armies and led to an unprecedented militarization, mainly due to mass conscription. Under the leadership of Napoleon, French power rose quickly as the Grande Armée conquered most of Europe, and collapsed rapidly

XL Generation:

after the French conquest of Italy. This victory marked the beginning of the Second Coalition. Napoleon's rapid advance caused Austria – Russia, Britain and Russia to make peace negotiations. The Russian army under Prince Mikhail Mikhailovich Mikhailovich Chaffee, commander of the Imperial Russian Army in Switzerland, was able to drive off the Austrians in the Battle of Stockach (1805) and to take Pressburg (modern Ö3 Austria) from the Austrians. At an early stage the Austro-Hungarian fleet had been damaged in a storm off Cape Matapan; this allowed the Allies to launch a full-scale invasion of Italy in February 1805. The Battle of Austerlitz was the decisive French victory against Napoleon and the largest naval battle in history, involving more modern-day European navies. The French military situation worsened as Napoleon faced several threats related to his newly formed Army of Europe, particularly Russia itself, which was now on the verge of collapse. The Russian general and friend of Napoleon, Alexander, had been dissatisfied with the conduct of the Austrians; he was still in conflict with Emperor Napoleon, the French Republic's king, who had declared war on Austria during the summer of 1804. With the war over, Napoleon decided to launch a second invasion of Italy in October 1805.

= = Prelude = =

In July 1805, the French 1st Army entered southern Italy. The army, under the command of Marshal Marmont, were reinforced by a few battalions of infantry under Claude General Auguste de Marmont at the town of Philippsburg and another battalion at Belluno. On 17 September 1805, the army marched from Belluno towards Krems. By 29 September, they had reached Belluno and conducted its advance against a small Austrian force. By 31 September, the whole force had been reinforced by a brigade from the Army of Tyrol under the command of Pierre Augereau. The Austrians were now under the command of Marshal Jean Victor Marie Moreau, a member of

Reference:

after the disastrous invasion of Russia in 1812. Napoleon's empire ultimately suffered complete military defeat in the 1813 – 14 campaigns, resulting in the restoration of the Bourbon monarchy in France. Although Napoleon made a spectacular return in 1815, known as the Hundred Days, his defeat at the Battle of Waterloo, the pursuit of his army and himself, his abdication and banishment to the Island of Saint Helena concluded the Napoleonic Wars.

= = Danube campaign = =

From 1803-06 the Third Coalition fought the First French Empire and its client states (see table at right). Although several naval battles determined control of the seas, the outcome of the war was decided on the continent, predominantly in two major land operations in the Danube valley: the Ulm campaign in the upper Danube and the Vienna campaign, in the middle Danube valley. Political conflicts in Vienna delayed Austria's entry into the Third Coalition until 1805. After hostilities of the War of the Second Coalition ended in 1801, Archduke <unk> emperor's <unk> advantage of the subsequent years of peace to develop a military restructuring plan. He carefully put this plan into effect beginning in 1803 – 04, but implementation was incomplete in 1805 when Karl Mack, Lieutenant Field Marshal and Quartermaster-General of the Army, implemented his own restructuring. Mack bypassed Charles ' methodical approach. Occurring in the field, Mack's plan also undermined the overall command and organizational structure. Regardless, Mack sent an enthusiastic report to Vienna on the military's readiness. Furthermore, after misreading Napoleon's maneuvers in Württemberg, Mack also reported to Vienna on the weakness of French dispositions. His reports convinced the war party advising the emperor, Francis II, to enter the conflict against France, despite Charles ' own advice to the contrary. Responding to the report and rampant anti-French fever in Vienna, Francis dismissed Charles from his

Thank you

Transformer Networks [Vaswani et al. 2017],

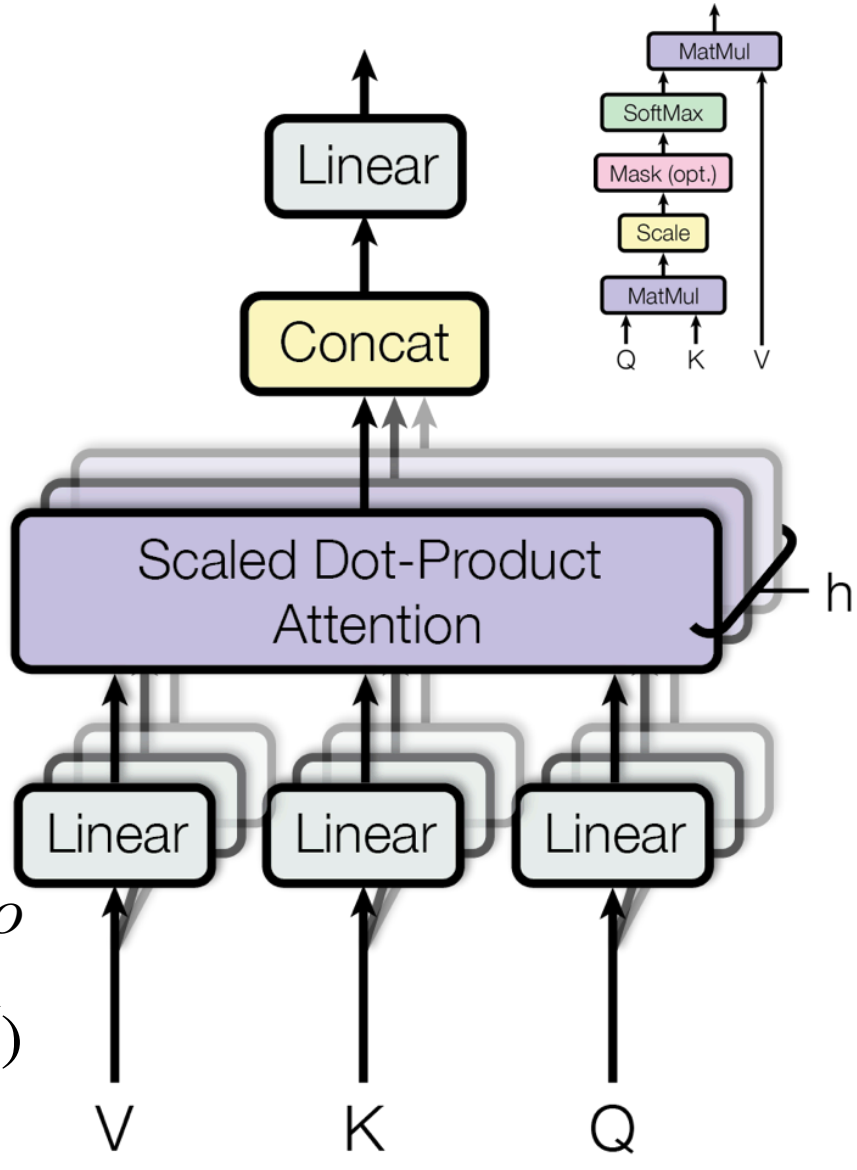
- Core: Scaled Dot-Product Attention Mechanism
  - Also called Single-Head Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}})V$$

- Multi-Head Attention
  - Consider multiple attention hypothesis

$$\text{MutliHead}(Q, K, V) = \text{Concat}(\text{head}_1, \cdots, \text{head}_h)W^O$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



Transformer Networks [Vaswani et al. 2017],

- Originally designed for Neural Machine Translation
- Input/Output Embedding Layer:
  - Lookup table from discrete tokens to continuous word representations
- Positional Encoding
  - Adding temporal information into sequences
- Encoder/ Decoder
  - Performing Sequence-to-Sequence Modeling
  - Core: Scaled Dot-Product Attention Mechanism
- Output Probability Layer
  - Lookup table from continuous word representations to discrete tokens