Deep Learning for Natural Language Processing

Yoshimasa Tsuruoka

The University of Tokyo

Outline

- Introduction
- Word embeddings
  - Word2vec, Skip-gram, fastText
- Recurrent Neural Networks
  - LSTM, GRU
- Neural Machine Translation
  - Encoder-decoder model
  - Transformer
  - Unusupervised NMT
- Pretraining methods
  - ELMo, GPT, BERT, UNI-LM
  - Sentiment Analysis, Textual Entailment, Question Answering, Summarization

Story generation

- GPT-2 [Radford et al., 2019]

SYSTEM PROMPT (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.*

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	The scientist named the population, after their distinctive horn,

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.”

https://blog.opena

Computer Science > Computation and Language

On Extractive and Abstractive Neural Document Summarization with Transformer Language Models

Sandeep Subramanian, Raymond Li, Jonathan Pilault, Christopher Pal

(Submitted on 7 Sep 2019)

We present a method to produce abstractive summaries of long documents that exceed several thousand words via neural abstractive summarization. We perform a simple extractive step before generating a summary, which is then used to condition the transformer language model on relevant information before being tasked with generating a summary. We show that this extractive step significantly improves summarization results. We also show that this approach produces more abstractive summaries compared to prior work that employs a copy mechanism while still achieving higher rouge scores. Note: The abstract above was not written by the authors, it was generated by one of the models presented in this paper.

Subjects: **Computation and Language (cs.CL)**
Cite as: arXiv:1909.03186 **[cs.CL]**
(or **arXiv:1909.03186v1 [cs.CL]** for this version)

Bibliographic data

[Enable Bibex (What is Bibex?)]

Submission history

From: Sandeep Subramanian [view email]
[v1] Sat, 7 Sep 2019 04:33:26 UTC (3,014 KB)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Question Answering

A prime number (or a prime) is a natural number greater than 1 that has no positive divisors other than 1 and itself. A natural number greater than 1 that is not a prime number is called a composite number. For example, 5 is prime because 1 and 5 are its only positive integer factors, whereas 6 is composite because it has the divisors 2 and 3 in addition to 1 and 6. The fundamental theorem of arithmetic establishes the central role of primes in number theory: any integer greater than 1 can be expressed as a product of primes that is unique up to ordering. The uniqueness in this theorem requires excluding 1 as a prime because one can include arbitrarily many instances of 1 in any factorization, e.g., 3, $1 \cdot 3$, $1 \cdot 1 \cdot 3$, etc. are all valid factorizations of 3.

What is the only divisor besides 1 that a prime number can have?

What theorem defines the main role of primes in number theory?

https://rajpurkar.github.io/SQuAD-explorer/

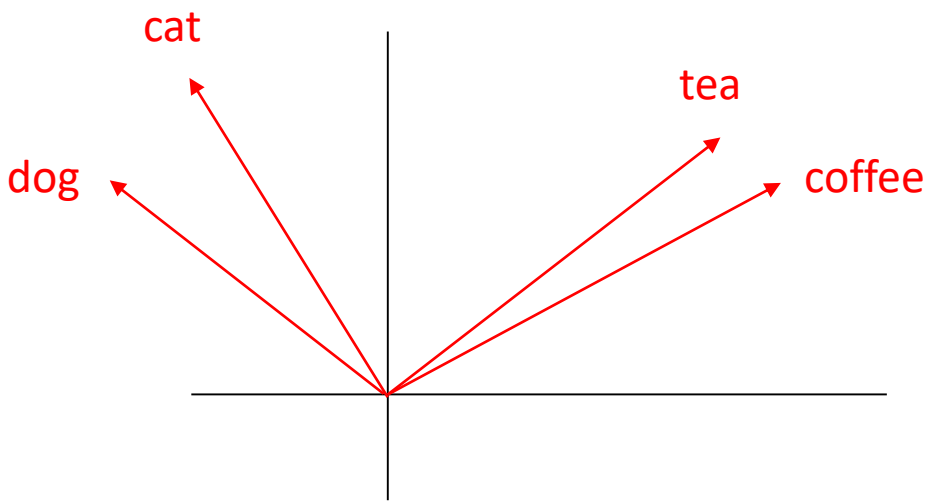
Now machines are better than human!?

Leaderboard

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**90.002**	**92.425**
2 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
3 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859

Word representation

- Vector representation
  - A word is represented as a real-valued vector
    - Dimensionality: 50 ~ 1000
  - Similar words are treated similarly
    - Alleviates the problem of data sparsity

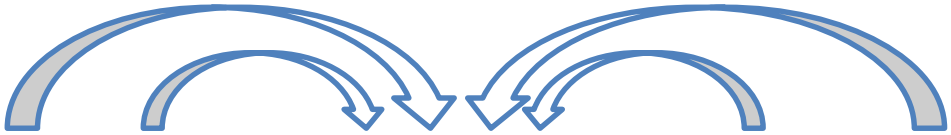


Word2Vec [Mikolov et al., 2013]

- How do you learn good word vectors?
- Optimize the vectors so that they can predict well the occurrence of the words in a document
- Two approaches:
  - Skip-grams
  - Continuous Bag of Words (CBOW)

Continuous Bag of Words (CBOW)

- Predict the center word



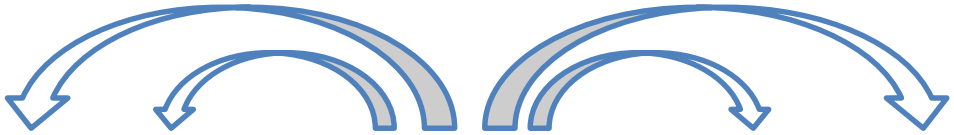
… to prevent another financial crisis such as …

… to prevent another financial crisis such as …

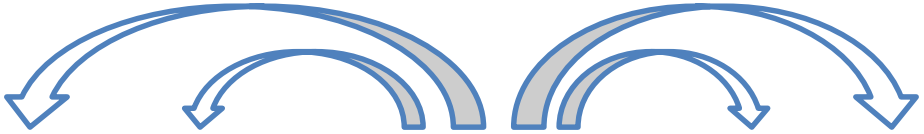
… to prevent another financial crisis such as …

Skip-gram

- Predict each context word



… to prevent another financial crisis such as …



… to prevent another financial crisis such as …

… to prevent another financial crisis such as …

Skip-gram

- Probability of word o given the center word c

$$p(o|c) = \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w=1}^{V} \exp(\mathbf{u}_w^T \mathbf{v}_c)}$$

- Each word in the vocabulary has two vectors
  - $\mathbf{u}$: vector used when the word appears as a context (outside) word
  - $\mathbf{v}$: vector used when the word appears as the center word

Skip-gram

- Objective for training
  - Maximize the likelihood

$$J'(\theta) = \prod_{t=1}^{T} \prod_{-m \le j \le m, j \neq 0} p\left(w_{t+j} \middle| w_t; \theta\right)$$

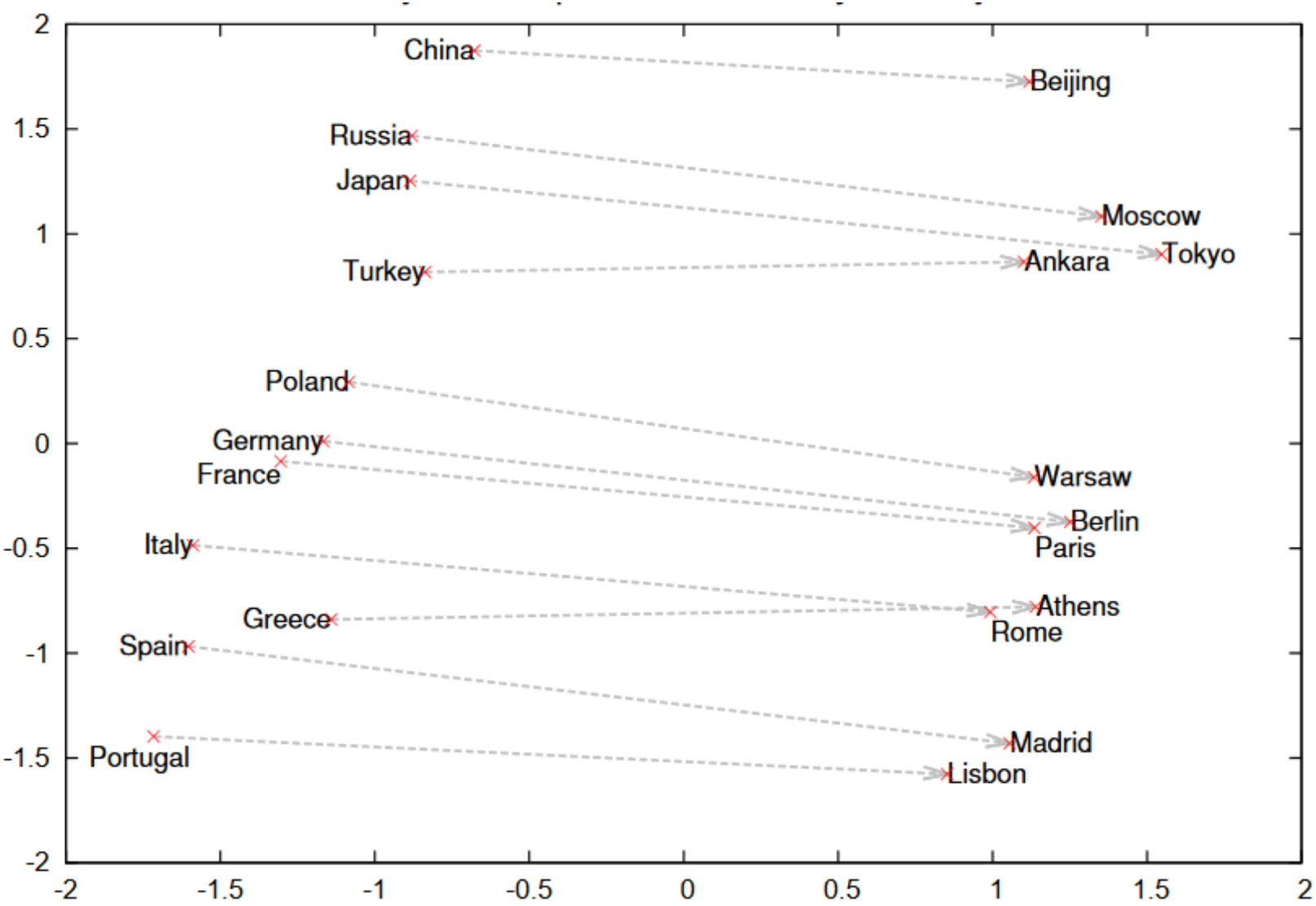
  - Minimize the negative log likelihood

$$J(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} \log p\left(w_{t+j} \middle| w_t; \theta\right)$$

Skip gram with negative sampling

- Skip gram
  - Needs to compute the summation for all words in the vocabulary
  - Computational cost is large
- Skip gram with negative sampling
  - Logistic regression problems
  - Generate negative examples by random sampling

$$\log \sigma(c \cdot w) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim p(w)} \left[\log \sigma(-w_i \cdot w) \right]$$



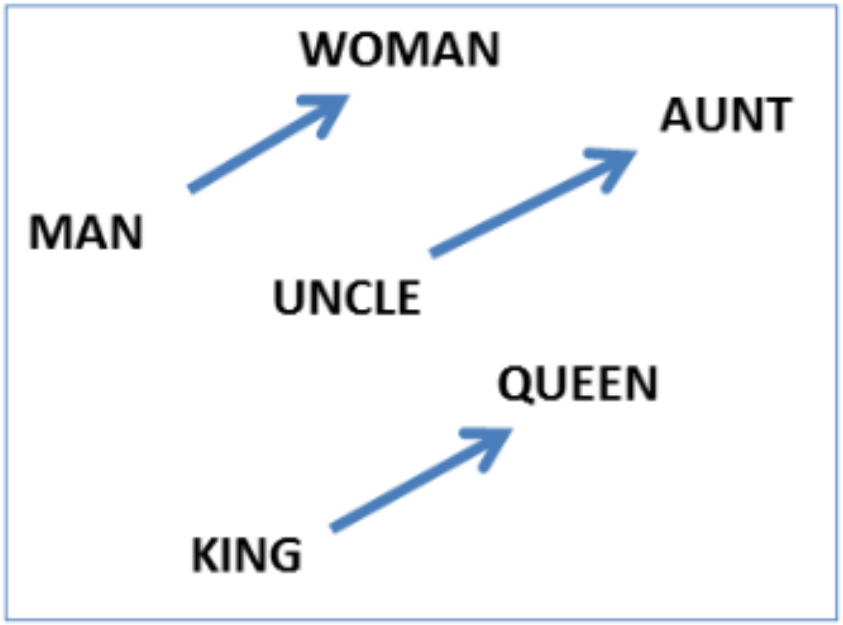
1000-dimensional vectors learned with Skip-gram are converted to two-dimensional vectors by PCA

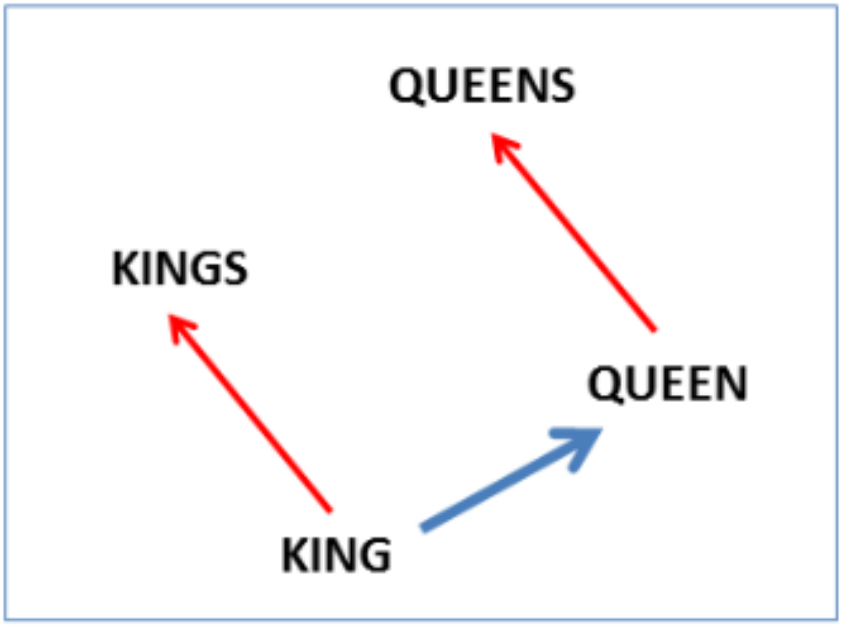
Mikolov, et al., Distributed Representations of Words and Phrases and their Compositionality, NIPS 2013

Analogical reasoning

- What is the female equivalent of a king?
- Analogical reasoning by arithmetic operation of word vectors

vec(king) – vec(man) + vec(woman) ≈ vec(queen)





(Mikolov et al., 2013)

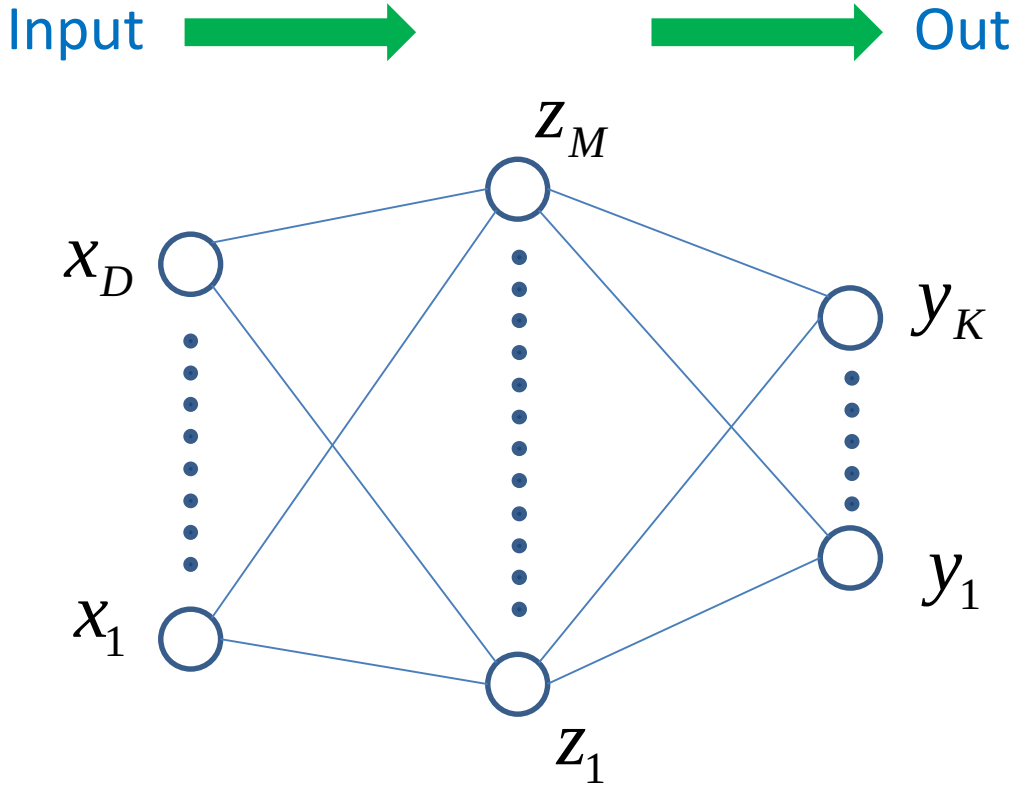
fastText [Bojanowski et al., 2017]

- Address the rare word problem
  - A word is represented as a bag of character *n*-grams
  - Each character *n*-gram has a vector representation
  - Each word is represented as the sum of character *n*-gram vectors
  - Example) The word “interlink”
    - 3-gram: <in, int, nte, ter, erl, rli, lin, ink, nk>
    - 4-gram: <int, inte, nter, terl, erli, rlin, link, ink>
    - 5-gram: <inte, inter, nterl, terli, erlin, rlink, link>
    - 6-gram: <inter, interl, nterli, terlin, erlink, rlink>
    - Whole word: <interlink>

Bojanowski et al., Enriching Word Vectors with Subword Information, TACL 2017

Neural Network

- Feed-forward Neural Network



$$\mathbf{z} = \sigma\left(\mathbf{W}^{\mathrm{zx}}\mathbf{x}\right)$$

$$\mathbf{y} = \mathbf{W}^{\mathrm{yz}}\mathbf{z}$$

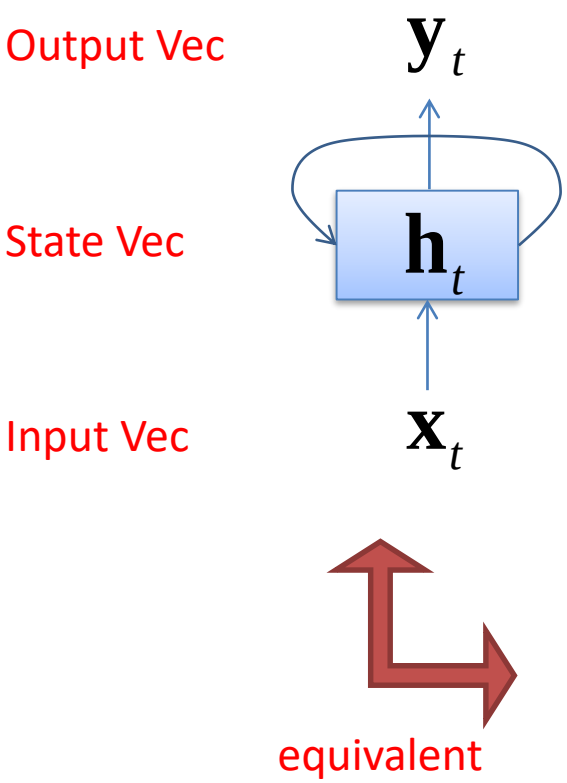
Sigmoid function

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

The sizes of the input and output vectors are fixed

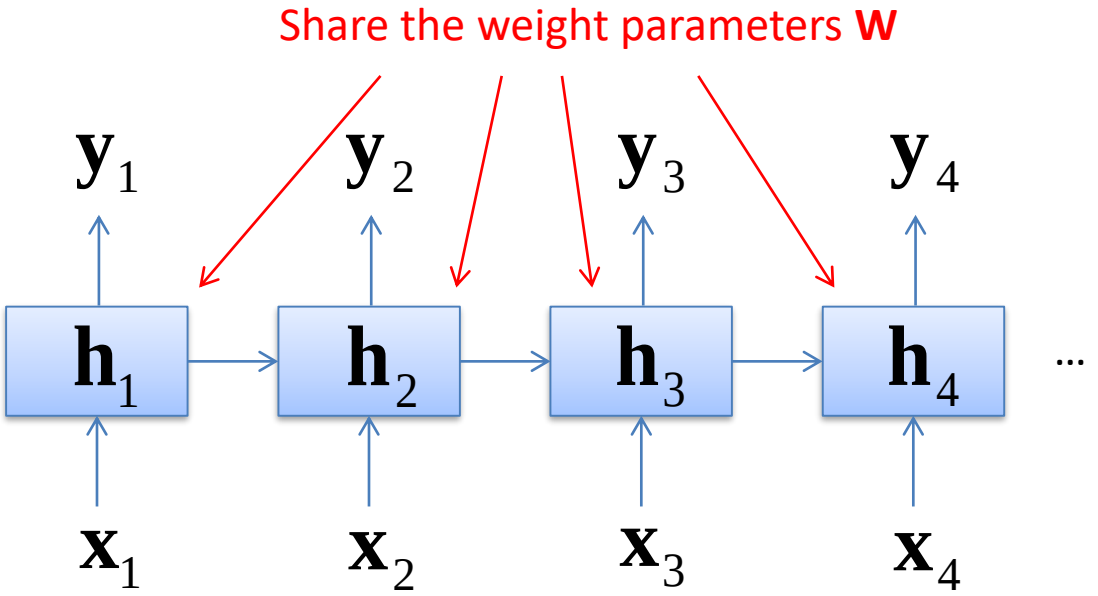
Recurrent Neural Network (RNN)

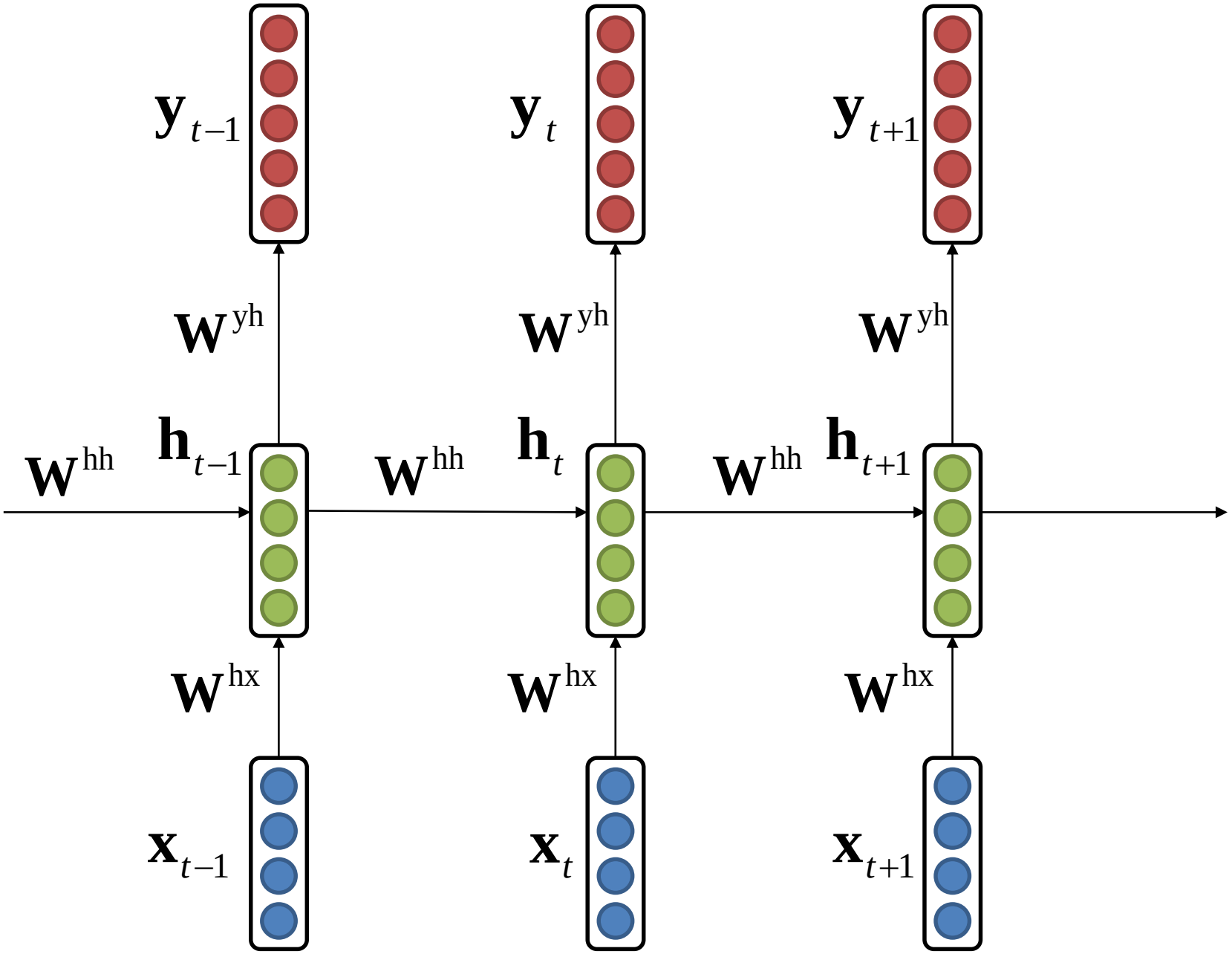
- Can process a sequence of any length



$$\mathbf{h}_t = \sigma\left(\mathbf{W}^{\text{hx}}\mathbf{x}_t + \mathbf{W}^{\text{hh}}\mathbf{h}_{t-1}\right)$$

$$\mathbf{y}_t = \mathbf{W}^{\text{yh}}\mathbf{h}_t$$

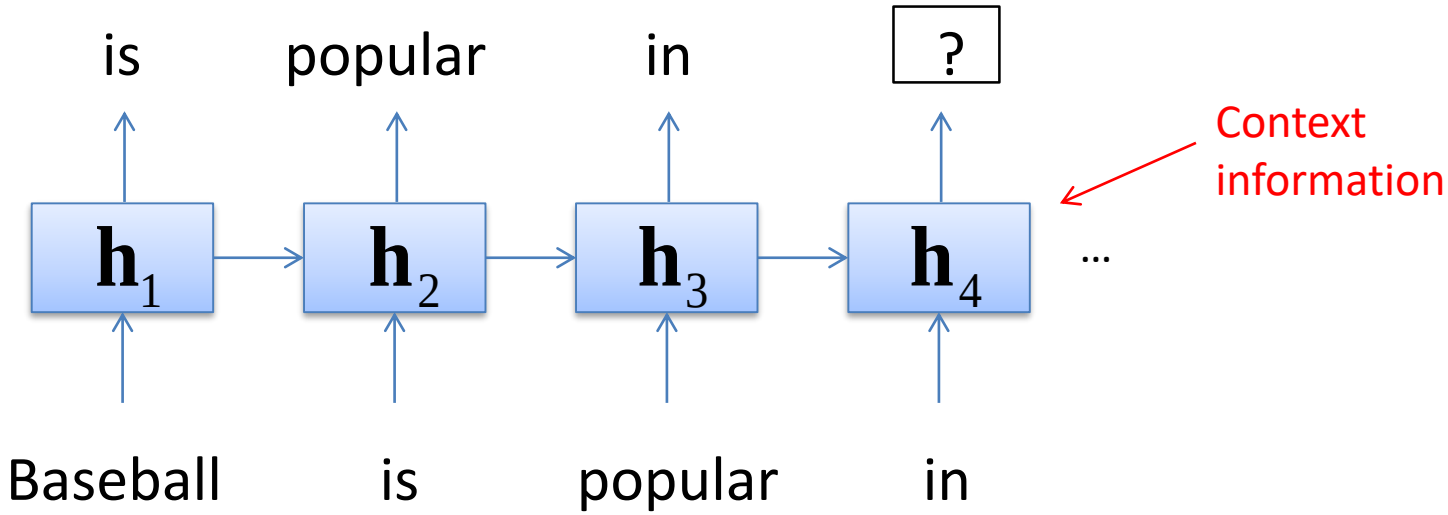



$\mathbf{y}_{t-1}$
$\mathbf{y}_t$
$\mathbf{y}_{t+1}$
$\mathbf{W}^{yh}$
$\mathbf{W}^{yh}$
$\mathbf{W}^{yh}$
$\mathbf{W}^{hh}$
$\mathbf{h}_{t-1}$
$\mathbf{W}^{hh}$
$\mathbf{h}_t$
$\mathbf{W}^{hh}$
$\mathbf{h}_{t+1}$
$\mathbf{W}^{hx}$
$\mathbf{W}^{hx}$
$\mathbf{W}^{hx}$
$\mathbf{x}_{t-1}$
$\mathbf{x}_t$
$\mathbf{x}_{t+1}$

RNN and NLP

- In NLP, we process sequences of words and characters
  - Language modeling, part-of-speech tagging, machine translation, etc.

- E.g., Language modeling
  - Predict the next word



RNN language model

- Words: $w_1, w_2, \ldots, w_{t-1}, w_t, w_{t+1}, \ldots, w_T$
- Word vectors: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_T$
- Hidden states

$$\mathbf{h}_t = \sigma\left(\mathbf{W}^{\mathrm{hx}}\mathbf{x}_t + \mathbf{W}^{\mathrm{hh}}\mathbf{h}_{t-1}\right)$$

- Word prediction

$$\hat{\mathbf{y}}_t = \mathrm{softmax}\left(\mathbf{W}^{\mathrm{S}}\mathbf{h}_t\right)$$

$$P\left(w_{t+1} = v_j \middle| w_t, \ldots, w_1\right) = \hat{y}_{t,j}$$

$$\mathbf{x}_t \in \mathbb{R}^d$$

$$\mathbf{h}_t \in \mathbb{R}^{D_h}$$

$$\mathbf{W}^{\mathrm{hx}} \in \mathbb{R}^{D_h \times d}$$

$$\mathbf{W}^{\mathrm{hh}} \in \mathbb{R}^{D_h \times D_h}$$

$$\mathbf{W}^{\mathrm{S}} \in \mathbb{R}^{|V| \times D_h}$$

Softmax function

- Softmax function
  - Convert real-valued scores to a probability distribution

$$\text{softmax}(\mathbf{x}) = \left(\frac{\exp(x_1)}{Z(\mathbf{x})}, \frac{\exp(x_2)}{Z(\mathbf{x})}, \ldots, \frac{\exp(x_n)}{Z(\mathbf{x})} \right)$$

$$Z(\mathbf{x}) = \sum_{i=1}^{n} \exp(x_i)$$

- Example

Add up to 1

$$\begin{bmatrix} 3.5 \\ -1.2 \\ 2.0 \end{bmatrix} \xrightarrow{\text{softmax( )}} \frac{1}{e^{3.5} + e^{-1.2} + e^{2.0}} \begin{bmatrix} e^{3.5} \\ e^{-1.2} \\ e^{2.0} \end{bmatrix} = \begin{bmatrix} 0.812 \\ 0.007 \\ 0.181 \end{bmatrix}$$

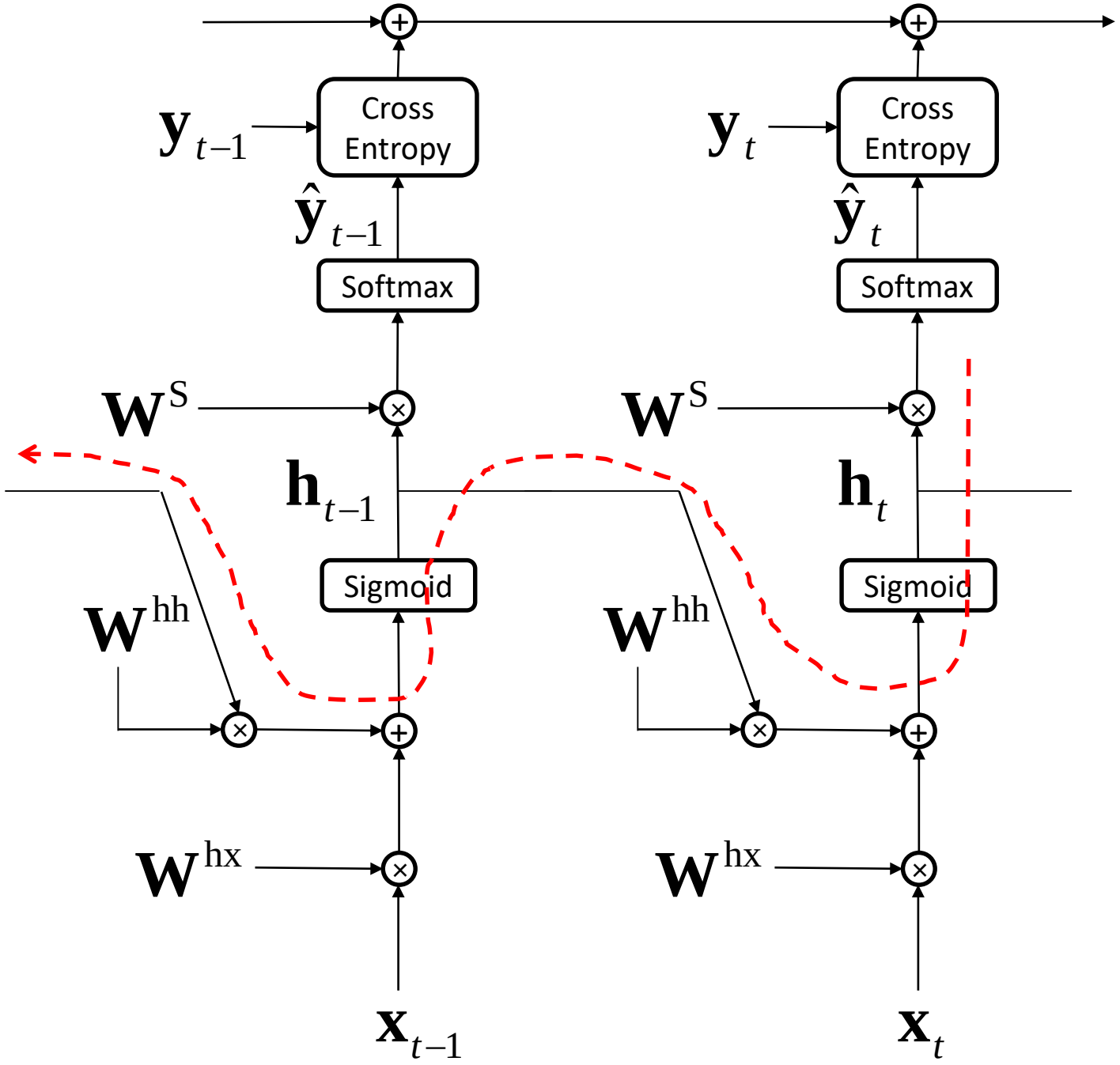
Training

- Objective
  - Cross Entropy Loss

$$J = -\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{|V|} y_{t,j} \log \hat{y}_{t,j}$$

Takes on 1 for the correct word, 0 otherwise

  - Equivalent to negative log-likelihood
    - Correct word should be assigned a high probability


$\mathbf{y}_{t-1}$
Cross Entropy
$\hat{\mathbf{y}}_{t-1}$
Softmax
$\mathbf{W}^{S}$
$\mathbf{h}_{t-1}$
Sigmoid
$\mathbf{W}^{hh}$
$\mathbf{W}^{hx}$
$\mathbf{x}_{t-1}$
$\mathbf{y}_{t}$
Cross Entropy
$\hat{\mathbf{y}}_{t}$
Softmax
$\mathbf{W}^{S}$
$\mathbf{h}_{t}$
Sigmoid
$\mathbf{W}^{hh}$
$\mathbf{W}^{hx}$
$\mathbf{x}_{t}$

Problems with vanilla RNNs

- Vanishing/exploding gradients
- Fail to capture long-distance dependencies

- Solutions
  - Directly connect $\mathbf{h}_t$ and $\mathbf{h}_{t-1}$ depending on context
  - Gated Recurrent Units (GRUs)
  - Long Short-Term Memory (LSTM)

Gated Recurrent Unit (GRU) [Cho et al., 2014]

- Update gate: $\mathbf{z}_t = \sigma\left(\mathbf{W}^{(z)}\mathbf{x}_t + \mathbf{U}^{(z)}\mathbf{h}_{t-1}\right)$
- Reset gate: $\mathbf{r}_t = \sigma\left(\mathbf{W}^{(r)}\mathbf{x}_t + \mathbf{U}^{(r)}\mathbf{h}_{t-1}\right)$
- State update

$$\tilde{\mathbf{h}}_t = \tanh\left(\mathbf{W}\mathbf{x}_t + \mathbf{r}_t \circ \mathbf{U}\mathbf{h}_{t-1}\right)$$

$$\mathbf{h}_t = \mathbf{z}_t \circ \mathbf{h}_{t-1} + \left(1 - \mathbf{z}_t\right) \circ \tilde{\mathbf{h}}_t$$

  - Ignore the previous state if the reset gate is 0
  - Copy the previous state if the update gate is 1

Long Short-Term Memory (LSTM)

[Hochreiter and Schmidhuber, 1997]

- Input gate: $\mathbf{i}_t = \sigma\left(\mathbf{W}^{(i)}\mathbf{x}_t + \mathbf{U}^{(i)}\mathbf{h}_{t-1} + \mathbf{b}^{(i)}\right)$
- Forget gate: $\mathbf{f}_t = \sigma\left(\mathbf{W}^{(f)}\mathbf{x}_t + \mathbf{U}^{(f)}\mathbf{h}_{t-1} + \mathbf{b}^{(f)}\right)$
- Output gate: $\mathbf{o}_t = \sigma\left(\mathbf{W}^{(o)}\mathbf{x}_t + \mathbf{U}^{(o)}\mathbf{h}_{t-1} + \mathbf{b}^{(o)}\right)$
- Memory cell

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{W}^{(\tilde{c})}\mathbf{x}_t + \mathbf{U}^{(\tilde{c})}\mathbf{h}_{t-1} + \mathbf{b}^{(\tilde{c})}\right)$$

$$\mathbf{c}_t = \mathbf{i}_t \circ \tilde{\mathbf{c}}_t + \mathbf{f}_t \circ \mathbf{c}_{t-1}$$

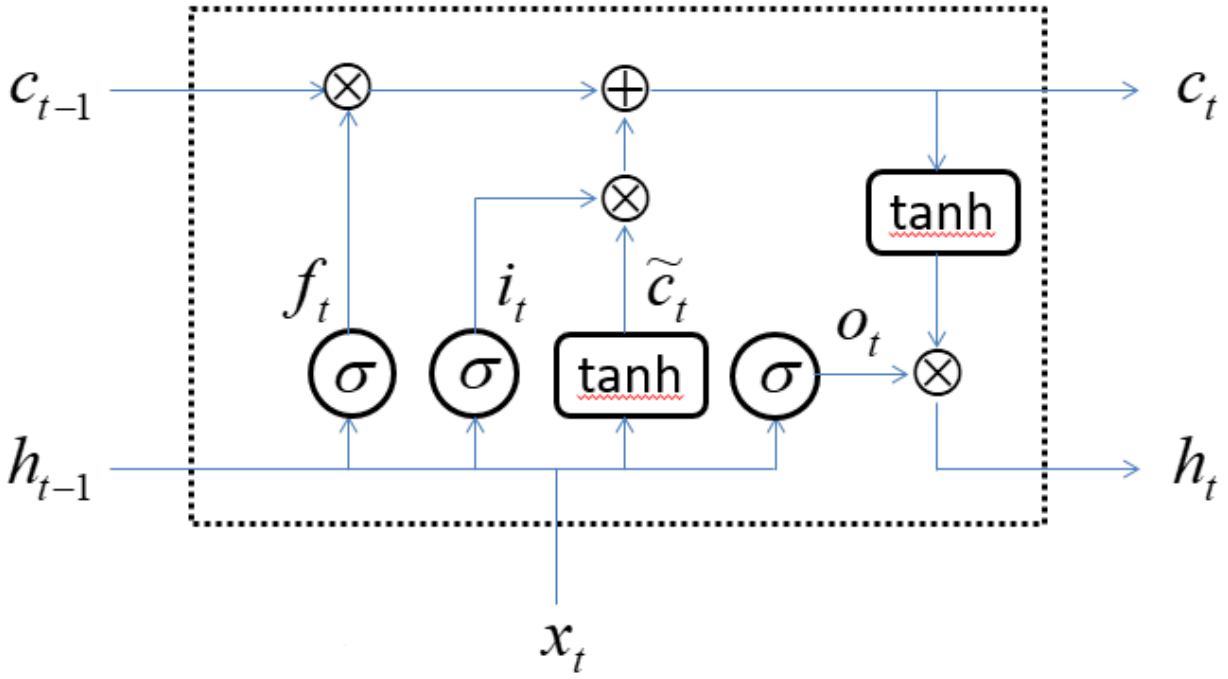
- State update

$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh\left(\mathbf{c}_t\right)$$

LSTM (Long Short-Term Memory)

$$\mathbf{i}_t = \sigma\left(\mathbf{W}^{(i)}\mathbf{x}_t + \mathbf{U}^{(i)}\mathbf{h}_{t-1} + \mathbf{b}^{(i)}\right)$$
$$\mathbf{f}_t = \sigma\left(\mathbf{W}^{(f)}\mathbf{x}_t + \mathbf{U}^{(f)}\mathbf{h}_{t-1} + \mathbf{b}^{(f)}\right)$$
$$\mathbf{o}_t = \sigma\left(\mathbf{W}^{(o)}\mathbf{x}_t + \mathbf{U}^{(o)}\mathbf{h}_{t-1} + \mathbf{b}^{(o)}\right)$$

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{W}^{(\tilde{c})}\mathbf{x}_t + \mathbf{U}^{(\tilde{c})}\mathbf{h}_{t-1} + \mathbf{b}^{(\tilde{c})}\right)$$
$$\mathbf{c}_t = \mathbf{i}_t \circ \tilde{\mathbf{c}}_t + \mathbf{f}_t \circ \mathbf{c}_{t-1}$$
$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh\left(\mathbf{c}_t\right)$$



Performance of RNN language models

	PPL	Size
LSTM-Word-Small	97.6	5 m
LSTM-Char-Small	92.3	5 m
LSTM-Word-Large	85.4	20 m
LSTM-Char-Large	78.9	19 m
KN-5 (Mikolov et al. 2012)	141.2	2 m
RNN† (Mikolov et al. 2012)	124.7	6 m
RNN-LDA† (Mikolov et al. 2012)	113.7	7 m
genCNN† (Wang et al. 2015)	116.4	8 m
FOFE-FNNLM† (Zhang et al. 2015)	108.0	6 m
Deep RNN (Pascanu et al. 2013)	107.5	6 m
Sum-Prod Net† (Cheng et al. 2014)	100.0	5 m
LSTM-1† (Zaremba et al. 2014)	82.7	20 m
LSTM-2† (Zaremba et al. 2014)	78.4	52 m

Kim et al., Character-Aware Neural Language Models, 2015

Perplexity

- Perplexity (PP or PPL)

$$\mathrm{PP}(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}}$$

$$= \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1 \ldots w_{i-1})}}$$

- Average branching factor of word prediction
- The smaller, the better

$$\log \mathrm{PP}(W) = -\frac{1}{N} \sum_{i=1}^{N} \log P(w_i | w_1, \ldots, w_{i-1})$$

→ Cross-entropy loss per word

Examples generated by LSTM

- Trained with LaTeX source (word-level)

Proof. Omitted. □

Lemma 0.1. *Let $\mathcal{C}$ be a set of the construction.*

Let $\mathcal{C}$ be a gerber covering. Let $\mathcal{F}$ be a quasi-coherent sheaves of $\mathcal{O}$-modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

.

Proof. This is an algebraic space with the composition of sheaves $\mathcal{F}$ on $X_{\acute{e}tale}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{morph_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where $\mathcal{G}$ defines an isomorphism $\mathcal{F} \to \mathcal{F}$ of $\mathcal{O}$-modules. □

Lemma 0.2. *This is an integer $\mathcal{Z}$ is injective.*

Proof. See Spaces, Lemma ??. □

Lemma 0.3. *Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $\mathcal{U} \subset \mathcal{X}$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.*

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

$$b : X \to Y' \to Y \to Y \to Y' \times_X Y \to X.$$

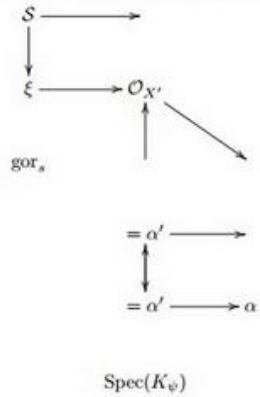
be a morphism of algebraic spaces over S and Y.

Proof. Let X be a nonzero scheme of X. Let X be an algebraic space. Let $\mathcal{F}$ be a quasi-coherent sheaf of $\mathcal{O}_X$-modules. The following are equivalent

(1) $\mathcal{F}$ is an algebraic space over S.
(2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. □

This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram



is a limit. Then $\mathcal{G}$ is a finite type and assume S is a flat and $\mathcal{F}$ and $\mathcal{G}$ is a finite type $f_\star$. This is of finite type diagrams, and

- the composition of $\mathcal{G}$ is a regular sequence,
- $\mathcal{O}_{X'}$ is a sheaf of rings.

□

Proof. We have see that $X = \mathrm{Spec}(R)$ and $\mathcal{F}$ is a finite type representable by algebraic space. The property $\mathcal{F}$ is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U. □

Proof. This is clear that $\mathcal{G}$ is a finite presentation, see Lemmas ??.
A *reduced above* we conclude that U is an open covering of $\mathcal{C}$. The functor $\mathcal{F}$ is a "field

$$\mathcal{O}_{X,x} \longrightarrow \mathcal{F}_{\overline{x}} \quad \text{-}1(\mathcal{O}_{X_{\acute{e}tale}}) \longrightarrow \mathcal{O}_{X_t}^{-1}\mathcal{O}_{X_\lambda}(\mathcal{O}_{X_\eta}^{\overline{v}})$$

is an isomorphism of covering of $\mathcal{O}_{X_t}$. If $\mathcal{F}$ is the unique element of $\mathcal{F}$ such that X is an isomorphism.
The property $\mathcal{F}$ is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme $\mathcal{O}_X$-algebra with $\mathcal{F}$ are opens of finite type over S.
If $\mathcal{F}$ is a scheme theoretic image points. □

If $\mathcal{F}$ is a finite direct sum $\mathcal{O}_{X_\lambda}$ is a closed immersion, see Lemma ??. This is a sequence of $\mathcal{F}$ is a similar morphism.

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

Machine Translation

- Translate one language into another

I'm here on vacation

Je suis là pour les vacances

- Train a translation model with a parallel corpus
  - E.g., WMT'14 English-to-French dataset
    - 12 million sentences from Europarl, News Commentary, etc.
    - 300 million words (English)
    - 350 million words (Frence)

Japanese-English Subtitle Corpus

[Pryzant et al., 2017]

English	Japanese
look, i don't do that shit anymore.	私は卒業した
thank you! you're so sweet	ありがとう
look, his name is cyrus gold.	いいか 彼の名前はサイラス・ゴールド
is that so? i hate to disappoint you.	そうか それは残念だったな。

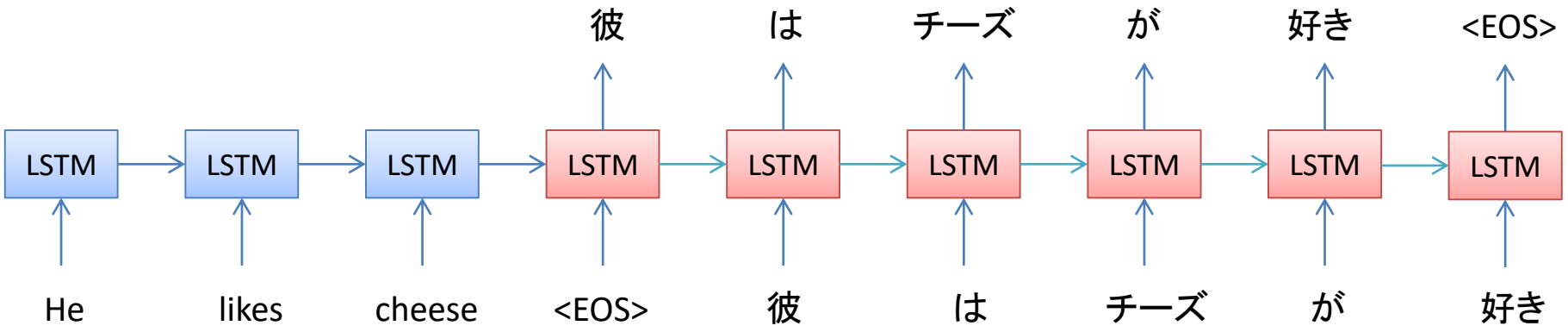
ASPEC corpus [Nakazawa et al., 2016]

DID: G-01A0204677	SID: 1	Sim: 0.137
Ja: リドカイン使用濃度，使用量は０．５～１０％，０．１～１．０ｍｌ（１０～６０ｍｇ）であった。		
En: The use concentration and the amount of lidocaine were 0.5～10% and 0.1～1.0m l(10～60mg) respectively.		

DID: G-93A0370292	SID: 0	Sim: 0.048
Ja: 症例は４３歳の女性で，心臓弁膜症手術後２２日目頃より，発熱と共に全身に紅斑が出現した。		
En: A 43 - year - old female was seen at our clinic with complaints of high fever and erythroderma like skin rashes, which have developed in 3 weeks after her heart operation.		

Neural Machine Translation

- Encoder-decoder model (Sutskever et al., 2014)
  - Encoder RNN
    - Convert the source sentence into a real-valued vector
  - Decoder RNN
    - Generate a sentence in the target language from the vector



Example

Output of the system

Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance .

Reference translation

Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années .

Sutskever et al., Sequence to Sequence Learning with Neural Networks, NIPS 2014

Results

- WMT'14 English to French

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

The BLUE score of the state-of-the-art system was 37.0

Sutskever et al., Sequence to Sequence Learning with Neural Networks, NIPS 2014

BLEU score

- BLEU score [Papineni et al., 2002]
  - Most widely used evaluation measure for MT

$$\text{BLUE} = \text{BP} \cdot \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log p_n \right)$$

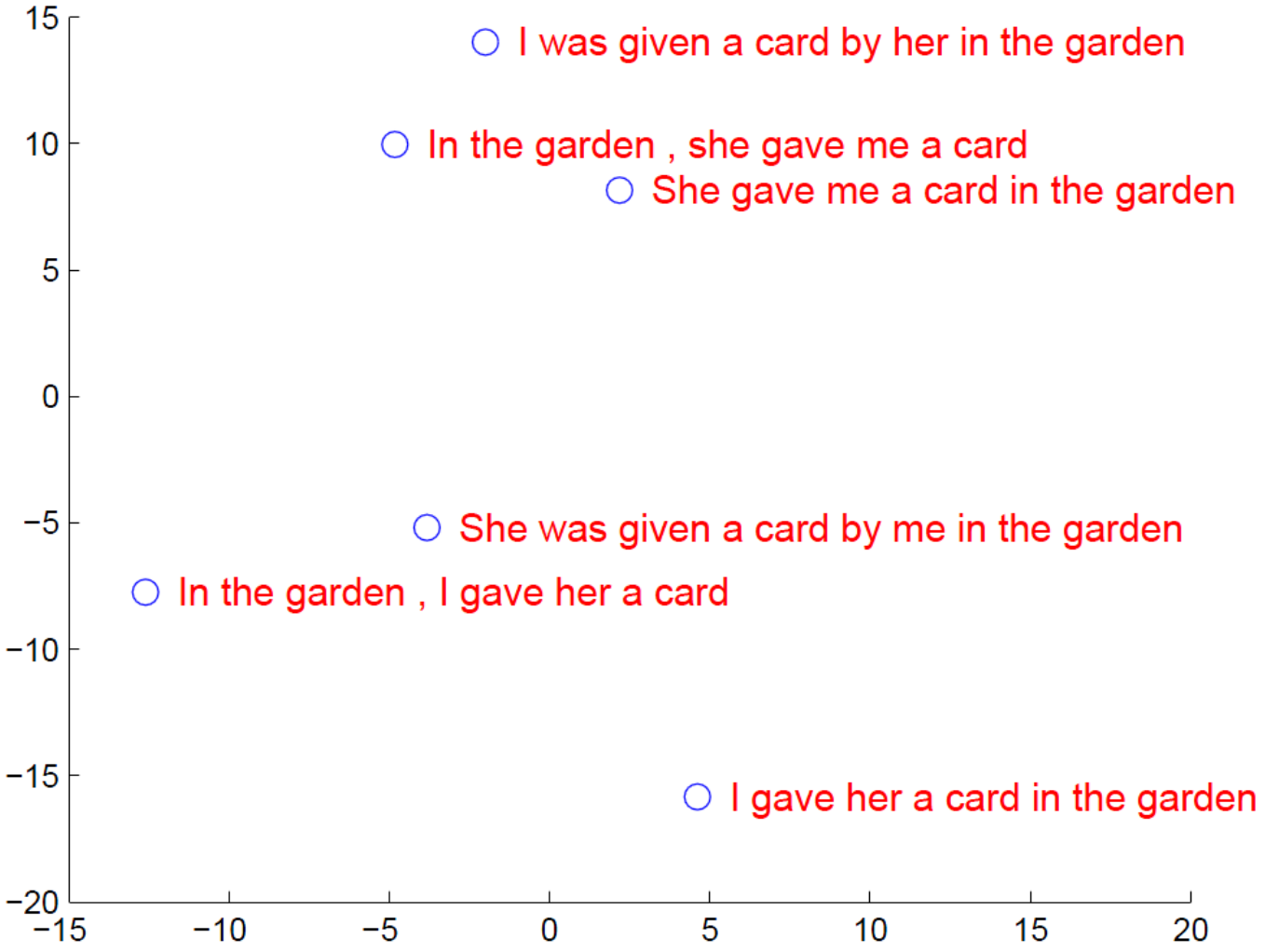
Modified n-gram precision

Brevity Penalty

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{otherwise} \end{cases}$$

c: length of the generated sentence
r: length of the reference sentence

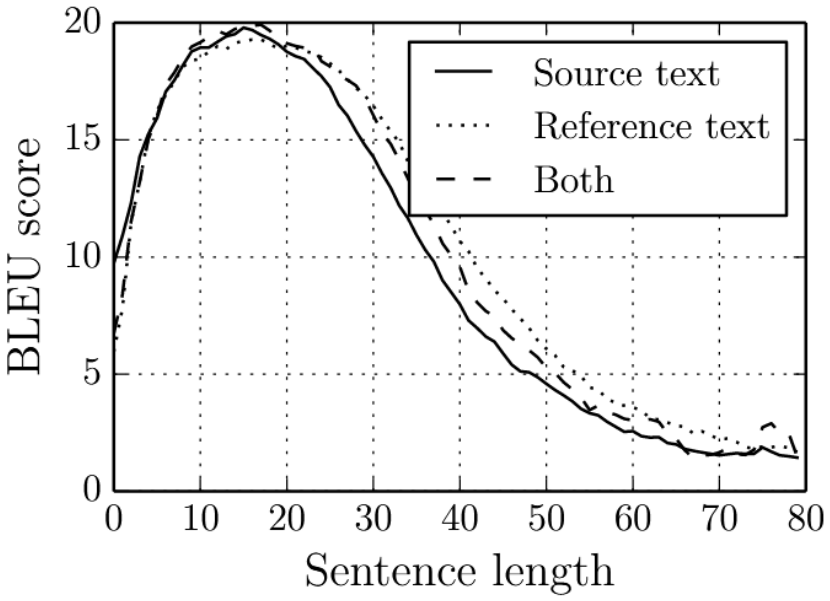
Vector representation of source sentences



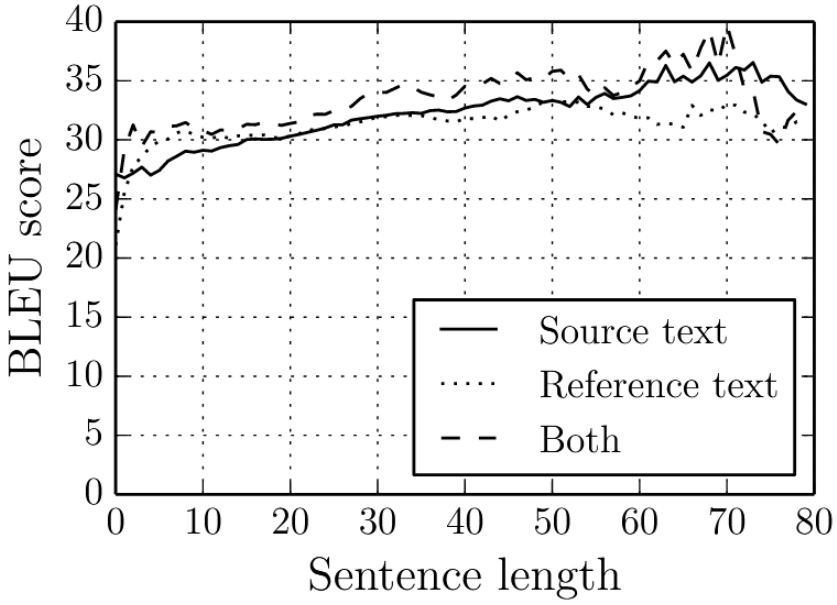
Sutskever et al., Sequence to Sequence Learning with Neural Networks, NIPS 2014

Problems

- Represents the content of the source sentence with a single vector
  - Hard to represent a long sentence
- Sentence length vs translation accuracy



Encoder-decoder model



Statistical Machine Translation

Cho et al., On the Properties of Neural Machine Translation: Encoder–Decoder Approaches, 2014

Attention (Bahdanau et al., 2015)

- Look at the hidden state of each word in the source sentence when updating the states of the decoder

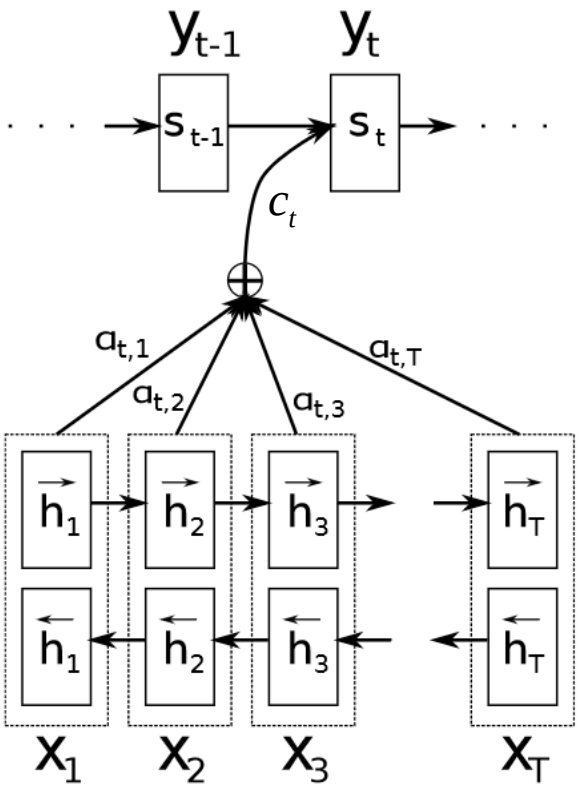
Weighted average of hidden states in the decor

$$\mathbf{c}_i = \sum_{j=1}^{T_x} \alpha_{ij} \mathbf{h}_j$$

Weights (which add up to 1)

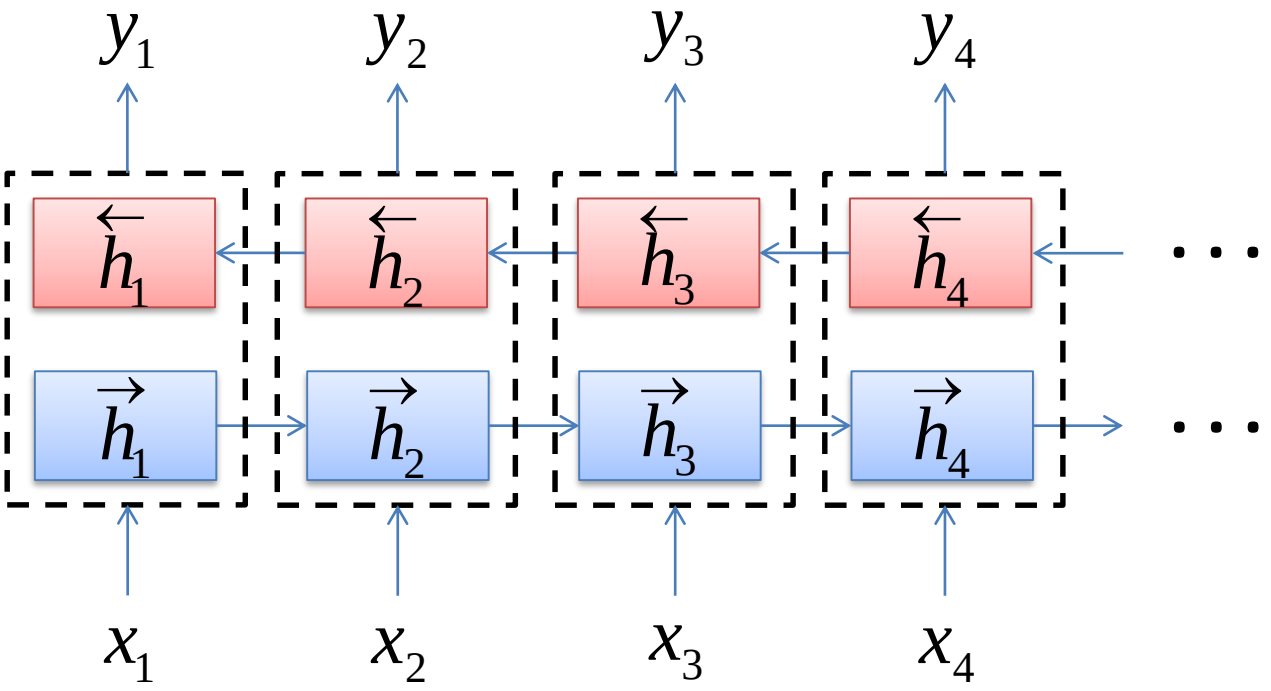
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = FeedForwardNN(\mathbf{s}_{i-1}, \mathbf{h}_j)$$

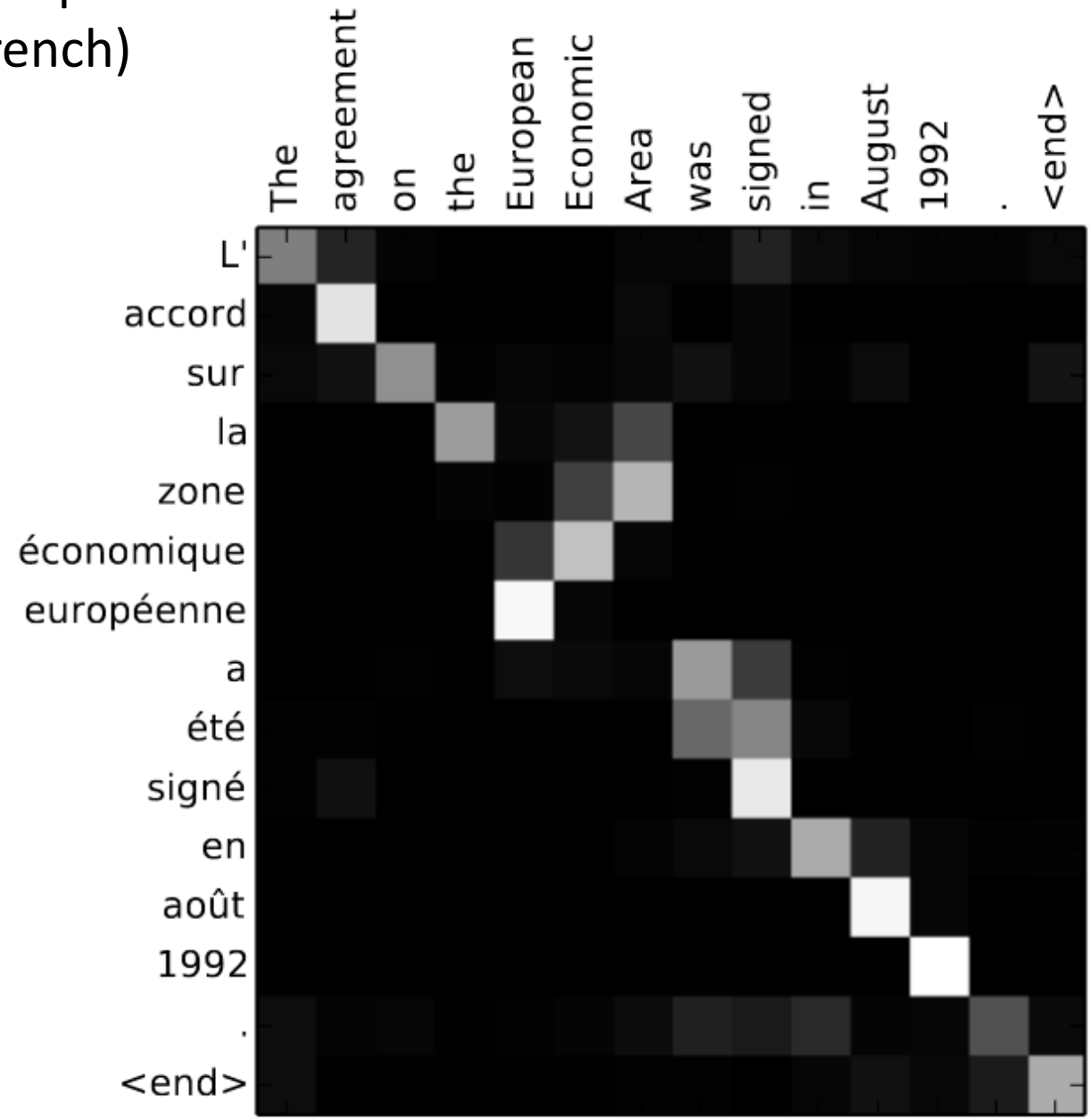


Bidirectional LSTM (BiLSTM)

- Stack two RNNs (forward and backward)
  - Capture left and right context information

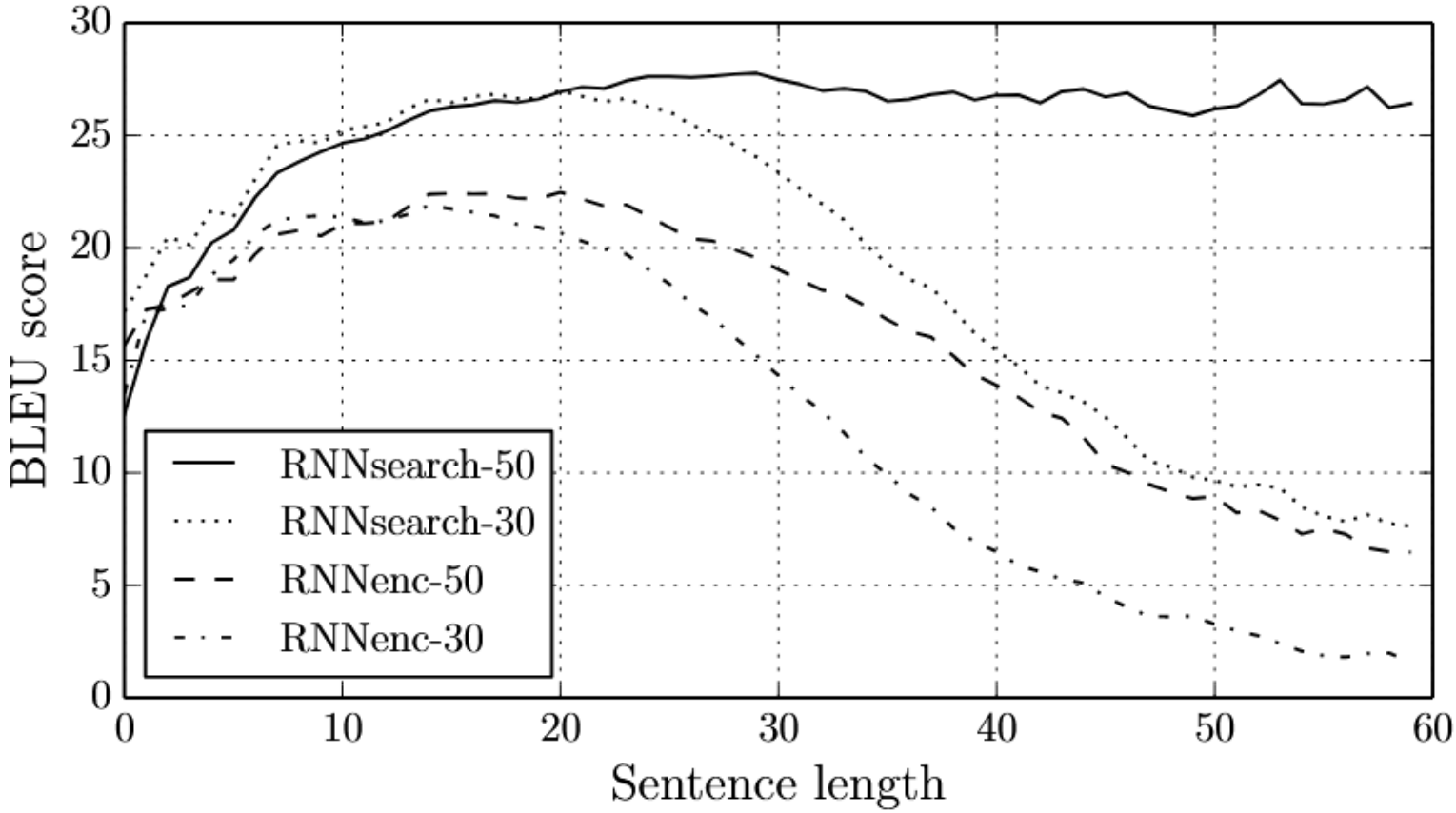


Attention Example (English to French)



(Bahdanau et al., 2015)

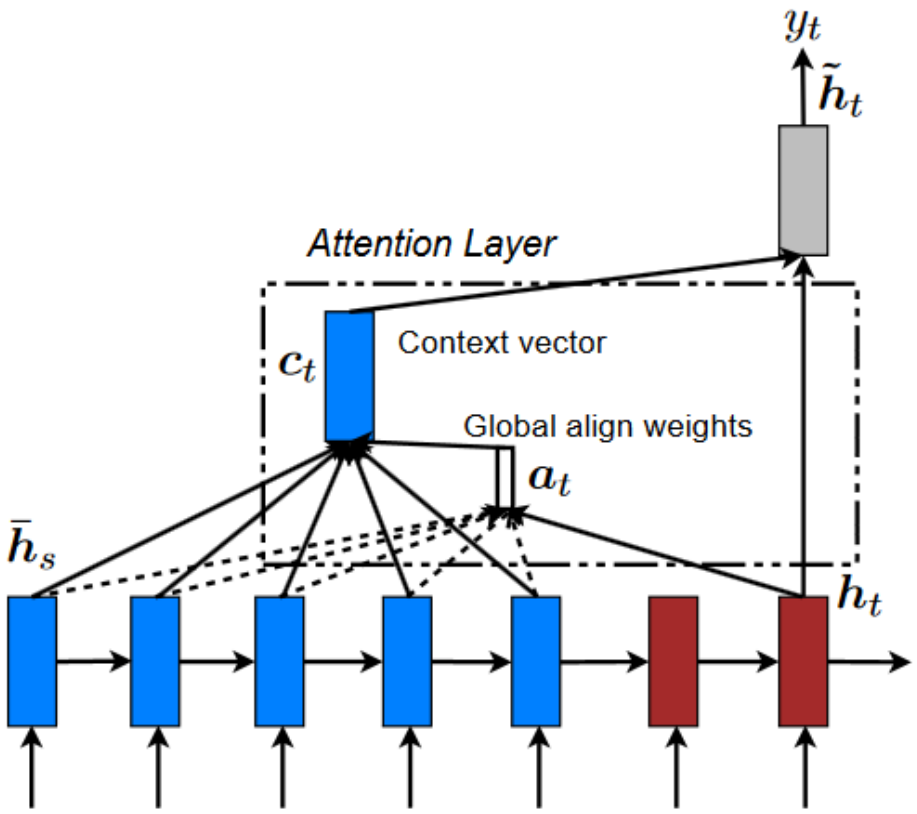
Sentence length and Translation Accuracy



(Bahdanau et al., 2015)

Attention

- Improvements by Luong et al. (2015)



$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W_c}[\boldsymbol{c}_t; \boldsymbol{h}_t])$$

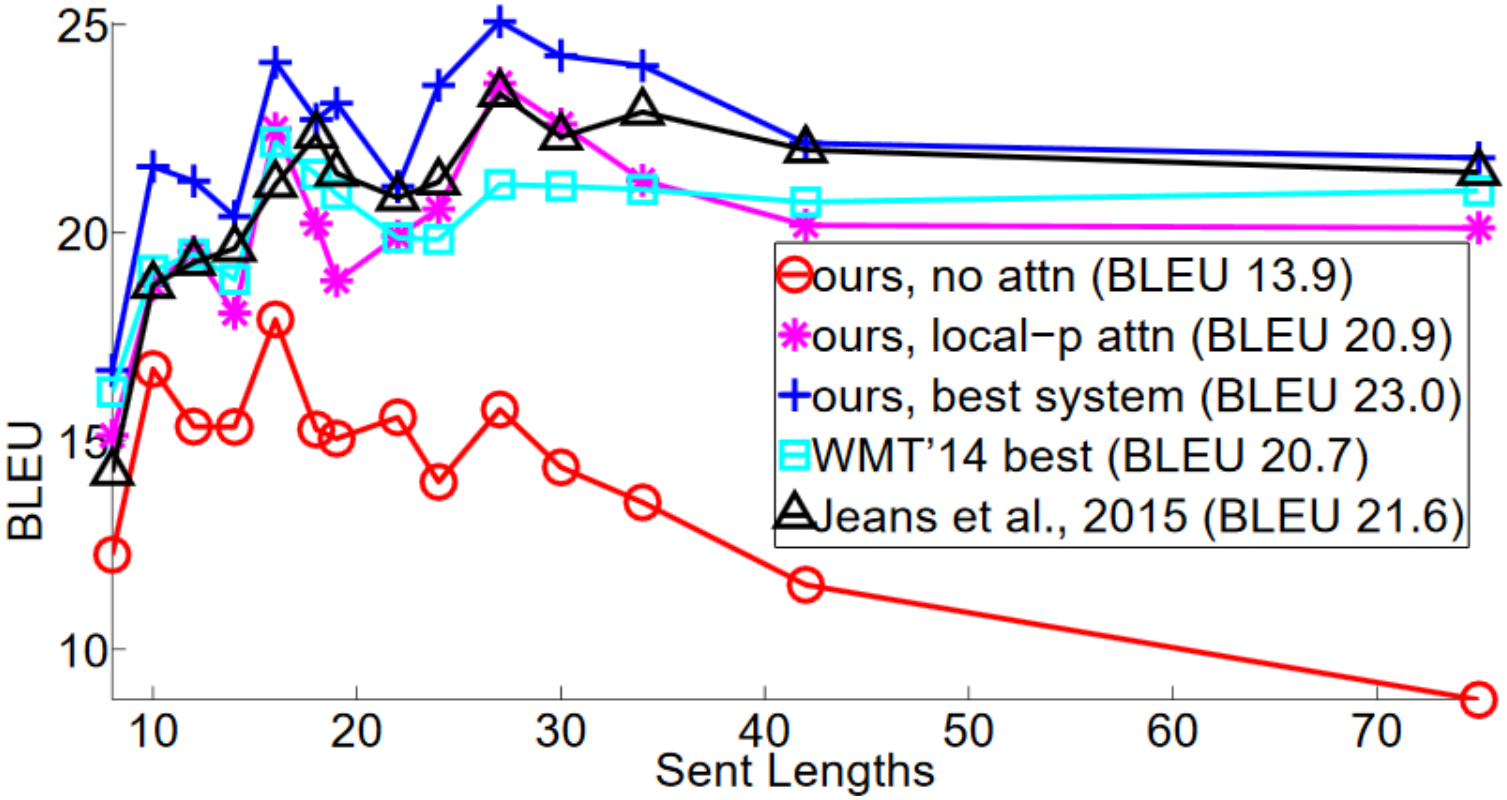
$$p(y_t|y_{<t}, x) = \text{softmax}(\boldsymbol{W_s}\tilde{\boldsymbol{h}}_t)$$

$$\boldsymbol{a}_t(s) = \text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)}$$

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W_a} \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s] & \textit{concat} \end{cases}$$

Translation accuracy

- WMT'14 English-German results
  - 4.5M sentence pairs



Luong et al., (2015)

Examples

German-English translations	
src	In einem Interview sagte Bloom jedoch , dass er und Kerr sich noch immer lieben .
ref	However , in an interview , Bloom has said that he and *Kerr* still love each other .
best	In an interview , however , Bloom said that he and *Kerr* still love .
base	However , in an interview , Bloom said that he and **Tina** were still <unk> .
src	Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
ref	The *austerity imposed by Berlin and the European Central Bank , coupled with the straitjacket* imposed on national economies through adherence to the common currency , has led many people to think Project Europe has gone too far .
best	Because of the strict *austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket* in which the respective national economy is forced to adhere to the common currency , many people believe that the European project has gone too far .
base	Because of the pressure **imposed by the European Central Bank and the Federal Central Bank with the strict austerity** imposed on the national economy in the face of the single currency , many people believe that the European project has gone too far .

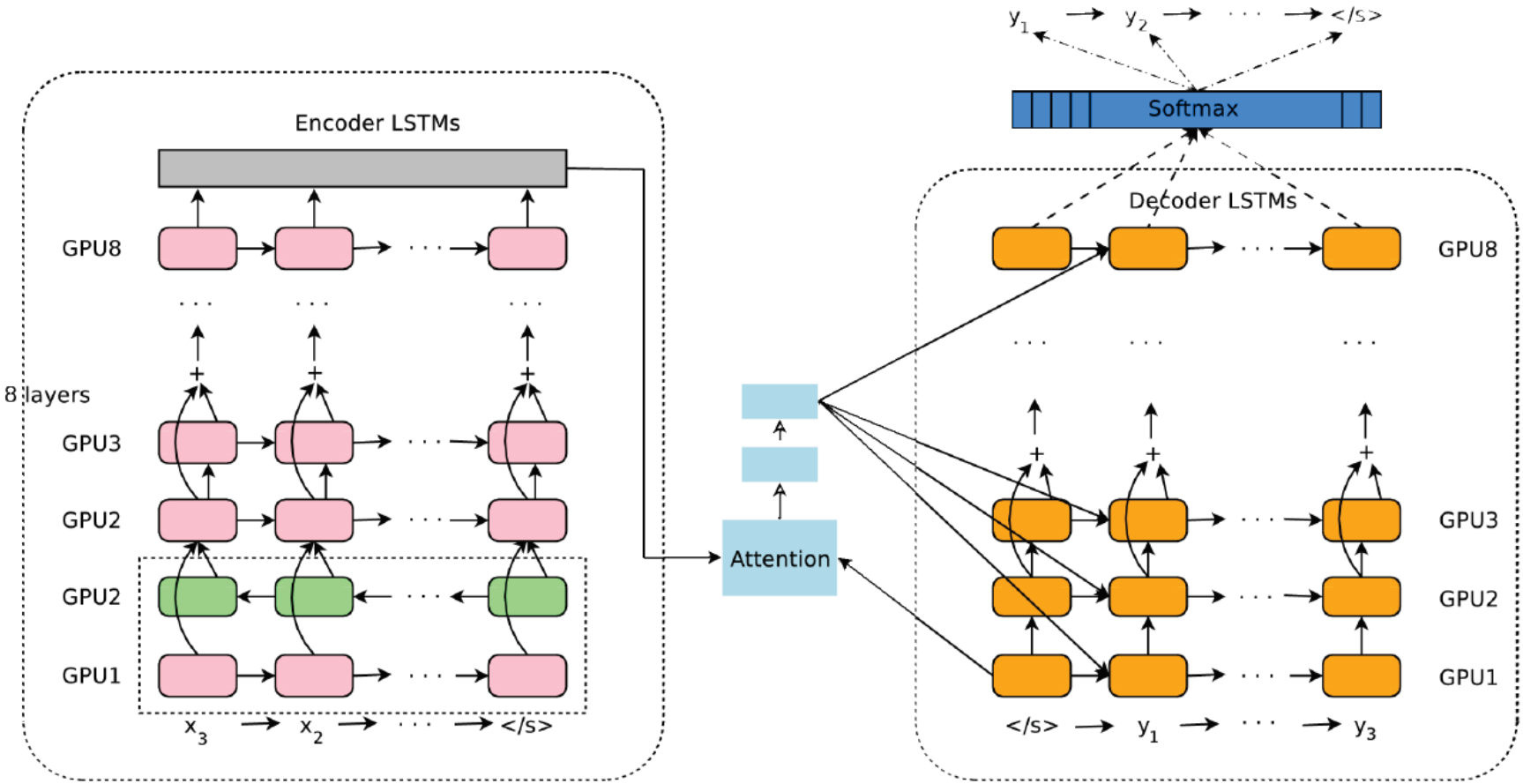
Luong et al., (2015)

Google Neural Machine Translation system (GNMT) (Wu et al., 2016)

- Model
  - Encoder: 8-layer LSTM (the lowest layer is bidirectional)
  - Decoder: 8-layer LSTM
  - Attention: from the bottom layer of the decoder to the top layer of the encoder
  - Unit: Wordpiece model

- Training data
  - For research
    - WMT corpus: En-Fr (36M), En-De (5M), etc.
  - For production
    - Two or three orders of magnitude larger than the WMT corpus

GNMT system

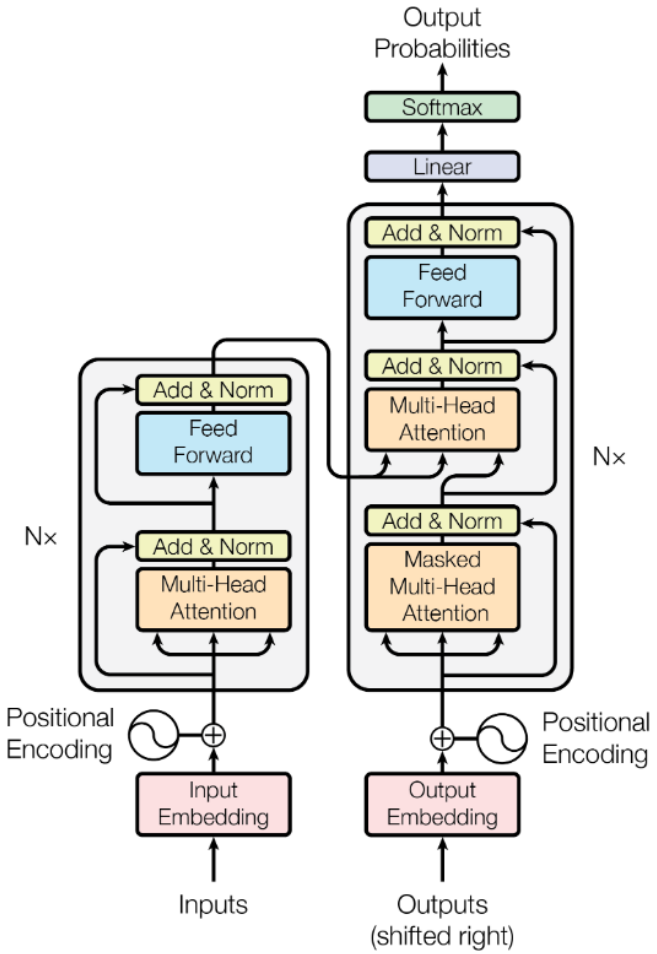


（Wu et al., 2016）

Transformer (Vaswani et al., 2017)

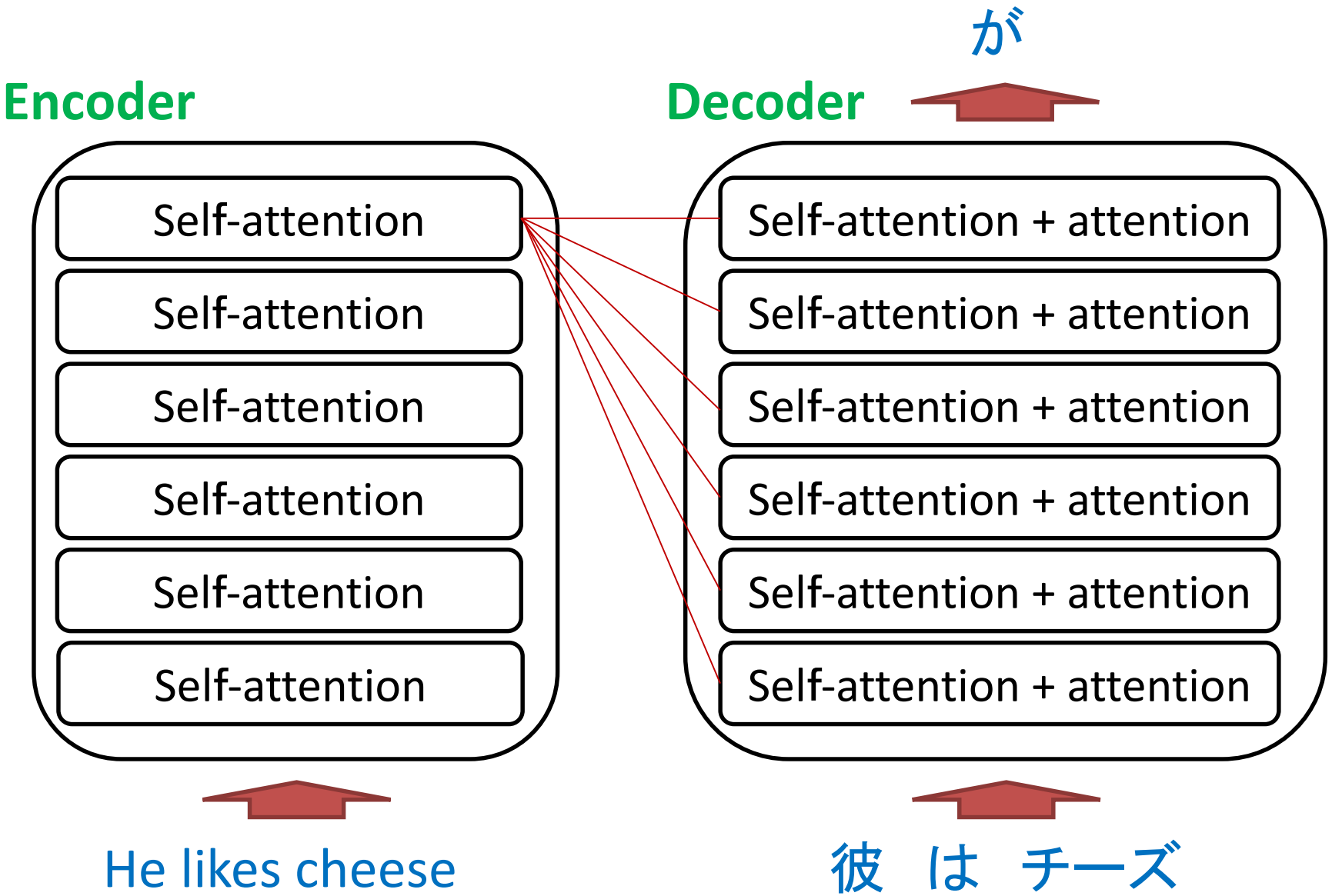
- Published in June of 2017
- No use of RNN

  ``**Attention is all you need**''
- Outperformed GNMT with much less computational cost
  - Training
    - 36M English-French sentence pairs
    - 8 GPUs
    - Completed in 3.5 day

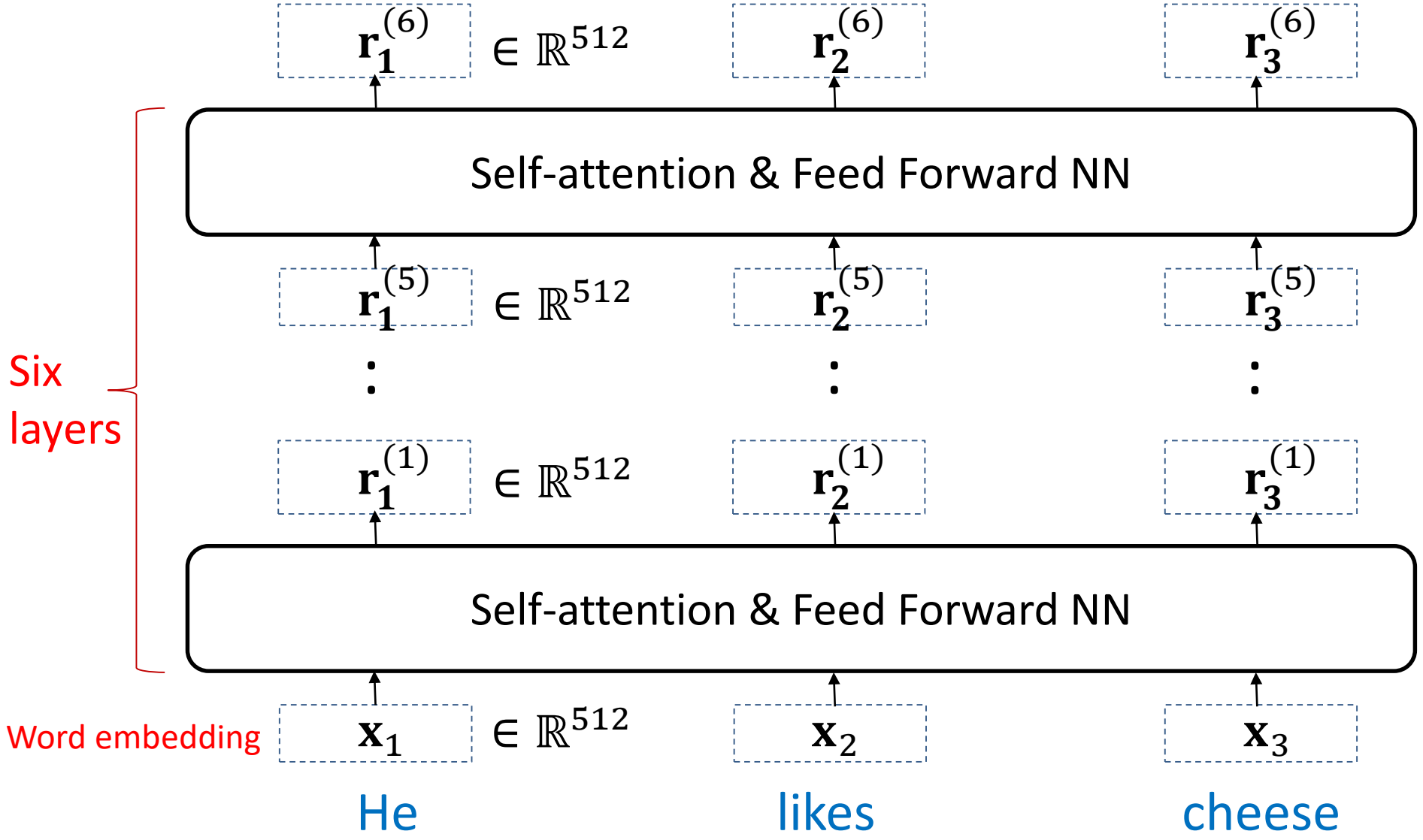


Vaswani et al., Attention Is All You Need, NIPS 2017

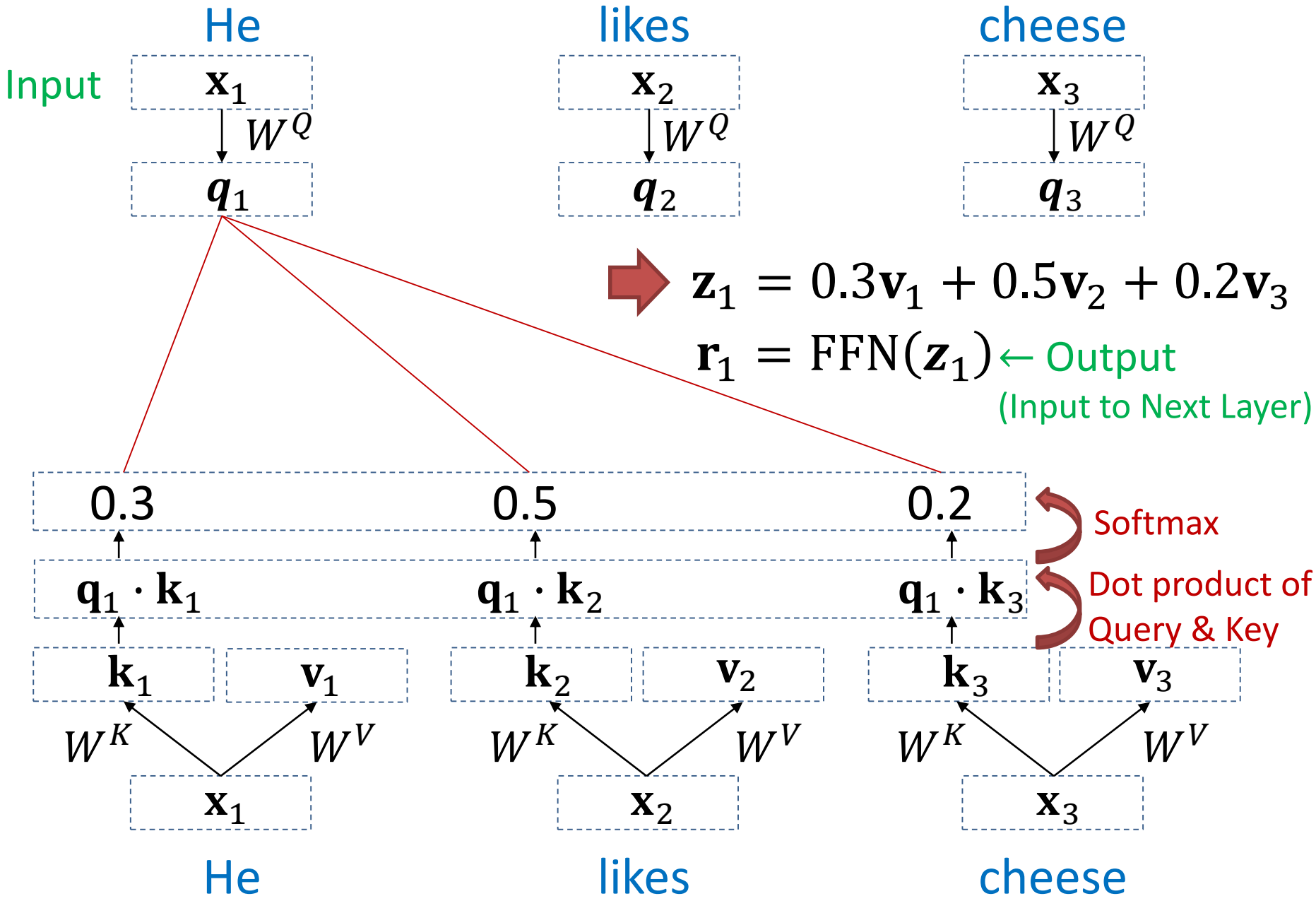
Transformer



- Encoder
  - Compute a contextual embedding for each word



- Encoder self-attention (1st layer)



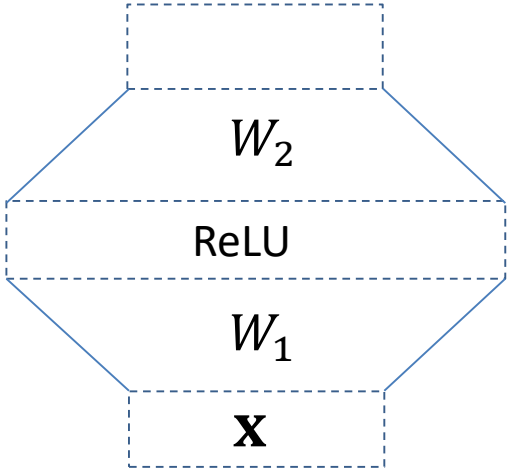
Position-wise Feed-Forward Networks

- Compute the output of each attention layer
  - Two-layer feed-forward neural network
  - Linear transformation -> ReLU -> Linear Transformation

$$\mathrm{FFN}(\mathbf{x}) = \max(0, \mathbf{x}W_1 + \mathbf{b}_1)W_2 + \mathbf{b}_2$$

$W_1 \in \mathbb{R}^{512\times2048}$ $\mathbf{b}_1 \in \mathbb{R}^{2048}$

$W_2 \in \mathbb{R}^{2048\times512}$ $\mathbf{b}_2 \in \mathbb{R}^{512}$



Scaled dot product attention

- Compute the attention for all words at once

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

$$Q \in \mathbb{R}^{n \times d_k}$$
$$K \in \mathbb{R}^{n \times d_k}$$
$$V \in \mathbb{R}^{n \times d_v}$$

scaling

$$\begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \mathbf{q}_3 \end{bmatrix} \begin{bmatrix} \mathbf{k}_1 & \mathbf{k}_2 & \mathbf{k}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1 \cdot \mathbf{k}_1 & \mathbf{q}_1 \cdot \mathbf{k}_2 & \mathbf{q}_1 \cdot \mathbf{k}_3 \\ \mathbf{q}_2 \cdot \mathbf{k}_1 & \mathbf{q}_2 \cdot \mathbf{k}_2 & \mathbf{q}_2 \cdot \mathbf{k}_3 \\ \mathbf{q}_3 \cdot \mathbf{k}_1 & \mathbf{q}_3 \cdot \mathbf{k}_2 & \mathbf{q}_3 \cdot \mathbf{k}_3 \end{bmatrix}$$

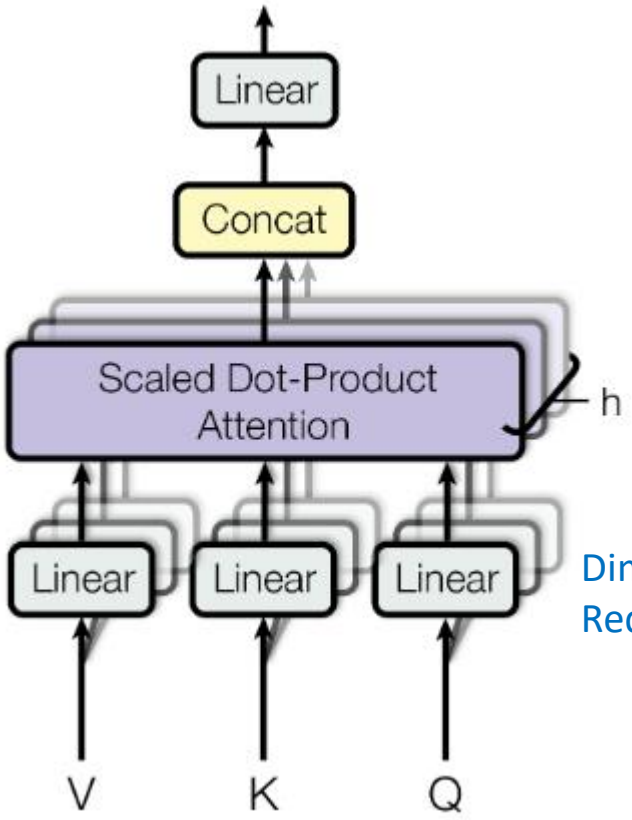
$$\begin{bmatrix} 0.1 & 0.2 & 0.7 \\ 0.4 & 0.3 & 0.3 \\ 0.8 & 0.1 & 0.1 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \end{bmatrix} = \begin{bmatrix} 0.1\mathbf{v}_1 + 0.2\mathbf{v}_2 + 0.7\mathbf{v}_3 \\ 0.4\mathbf{v}_1 + 0.3\mathbf{v}_2 + 0.3\mathbf{v}_3 \\ 0.8\mathbf{v}_1 + 0.1\mathbf{v}_2 + 0.1\mathbf{v}_3 \end{bmatrix}$$

Q K^T $\text{softmax}(\cdot)$ V

Multi-Head Attention

- Use 8 heads with different parameters

$$\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_8) W^O$$



$$W^O \in \mathbb{R}^{512\times512} \quad \text{Multihead}(\cdot) \in \mathbb{R}^{n\times512}$$

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

$$W_i^Q \in \mathbb{R}^{512\times64} \quad W_i^V \in \mathbb{R}^{512\times64}$$

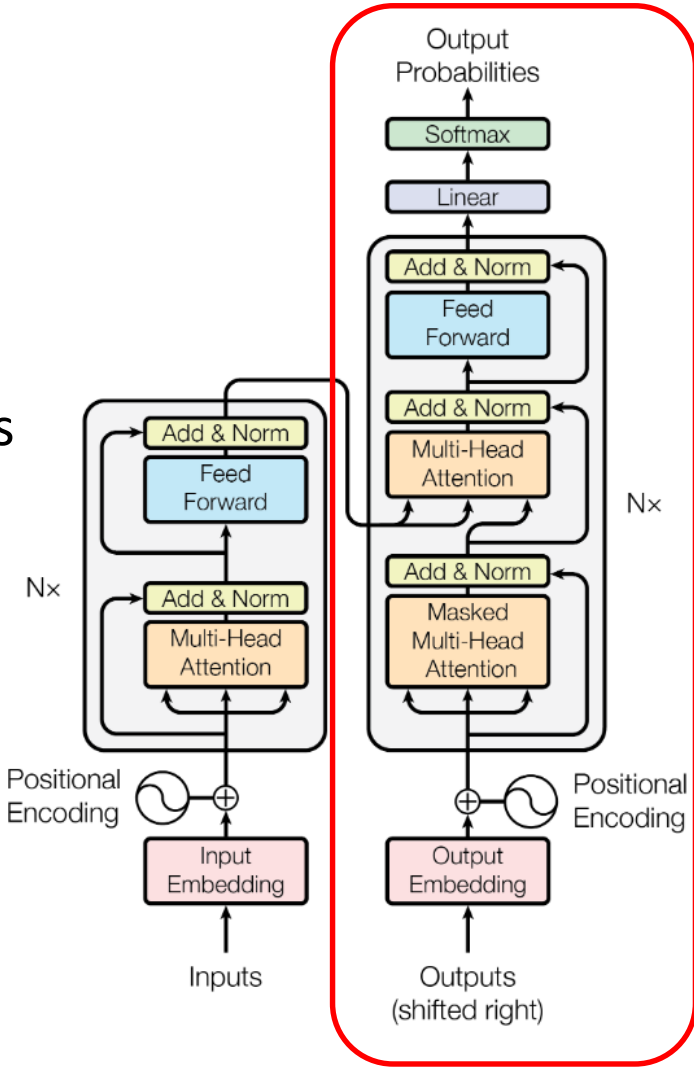
$$W_i^K \in \mathbb{R}^{512\times64} \quad \text{head}_i \in \mathbb{R}^{n\times64}$$

Dimensionality Reduction

Each head collects a different kinds of information

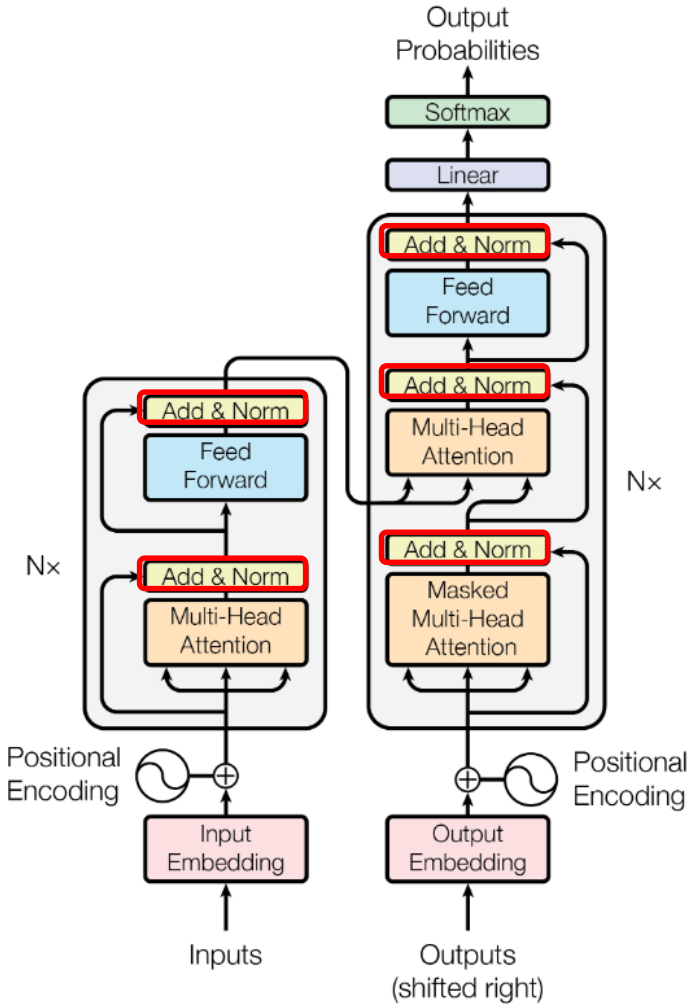
Decoder

- Six layers of attention
  - Each layer has two types of attention mechanism
- Self-attention
  - Collects information on the generated words
    - Attend to the output of the layer below
  - Mask the information about the future
- Encoder-decoder attention
  - Collects information on the source sentence
    - Attend to the output of the encoder



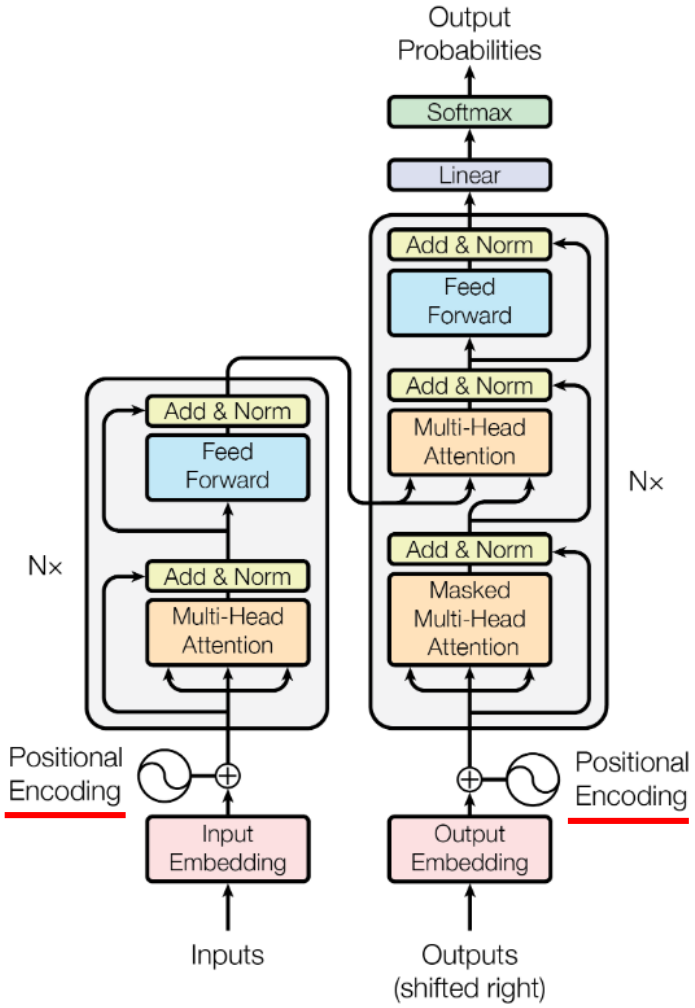
Add & Norm

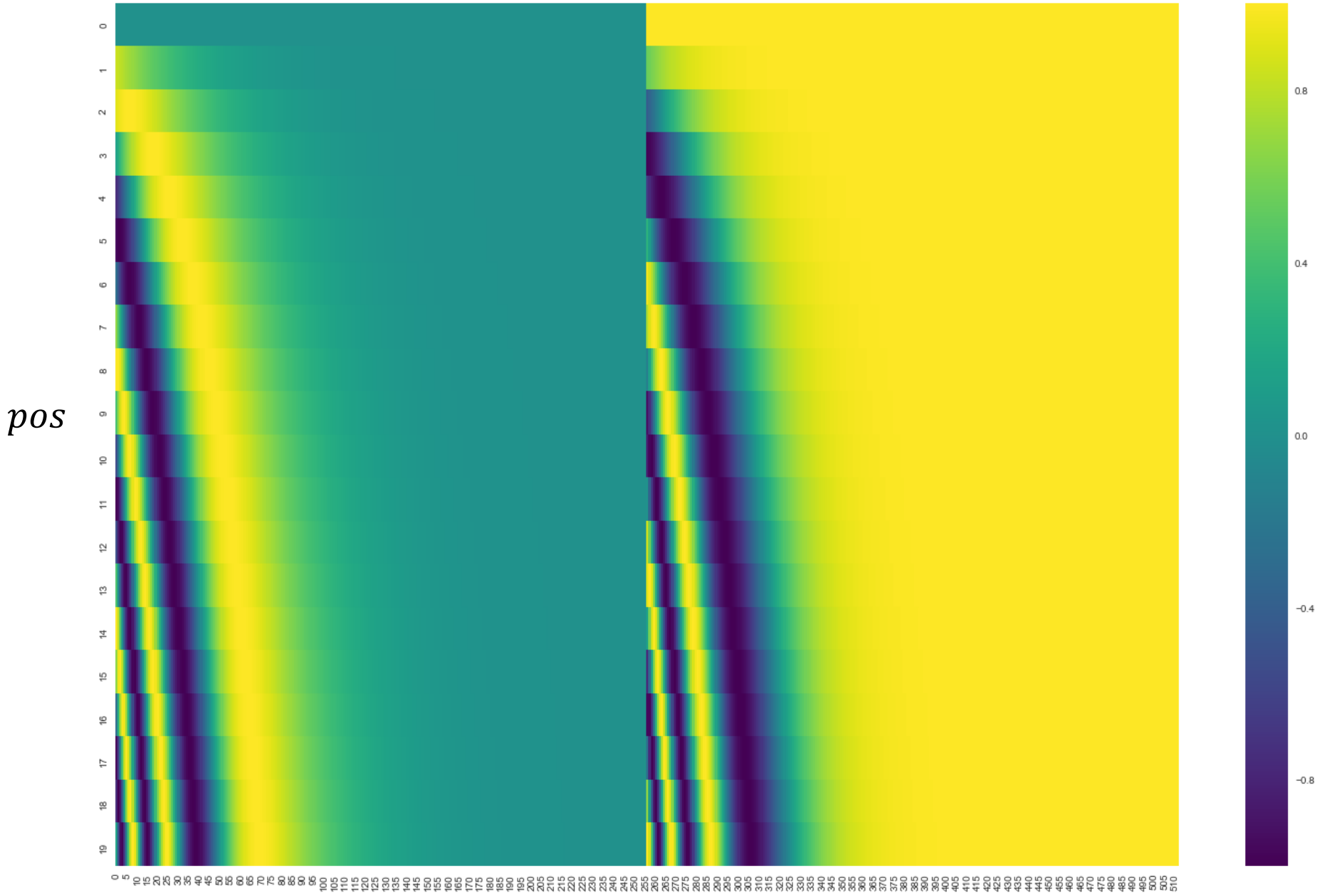
- Add (Residual Connection)
  - Learn the difference between input and output
  - Implementation
    - Simply add input to output
  - Reduces train/test error
- Norm (Layer Normalization)
  - Normalize the output of each layer
    - Mean = 0, Variance = 1
  - Faster convergence



Positional Encoding

- Encode position information
  - $PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/512}}\right)$
  - $PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/512}}\right)$
  - $0 \leq i < 256$
  - Wavelength: $2\pi \sim 10000 \cdot 2\pi$
- Represents each position with soft binary notation





http://jalammar.github.io/illustrated-transformer/

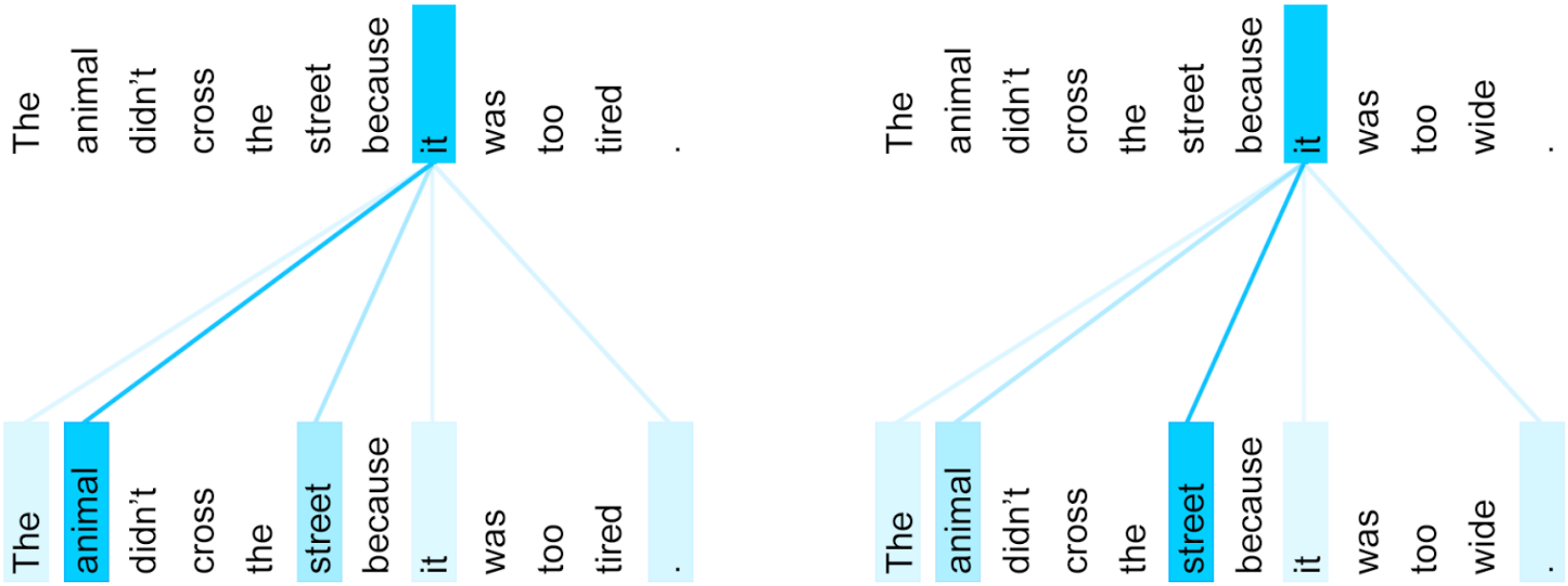
Transformer

I arrived at the

https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Visualizing attention

The animal didn't cross the street because **it** was too **tired**.
The animal didn't cross the street because **it** was too **wide**.



The encoder self-attention distribution for the word "it" from the 5th to the 6th layer of a Transformer trained on English to French translation (one of eight attention heads).

https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Performance

- Translation accuracy and training cost

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Results with different hyper-parameter settings

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65
(A)				1	512	512				5.29	24.9	
				4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16					5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4									5.19	25.3	50
	8									4.88	25.5	80
		256			32	32				5.75	24.5	28
		1024			128	128				4.66	26.0	168
			1024							5.12	25.4	53
			4096							4.75	26.2	90
(D)							0.0			5.77	24.6	
							0.2			4.95	25.5	
								0.0		4.67	25.3	
								0.2		5.47	25.7	
(E)	positional embedding instead of sinusoids									4.92	25.7	
big	6	1024	4096	16			0.3		300K	**4.33**	**26.4**	213

Unsupervised NMT

- Unsupervised NMT [Artetxe et al., 2018; Lample et al., 2018]
  - Build translation models by using only monolingual corpora

- Method
  1. Learn initial translation models
  2. Repeat
     - Generate source and target sentences using the current translation models
     - Train new translation models using the generated sentences

Example

Source	un homme est debout près d' une série de jeux vidéo dans un bar .
Iteration 0	a man is seated near a series of games video in a bar .
Iteration 1	a man is standing near a closeup of other games in a bar .
Iteration 2	a man is standing near a bunch of video video game in a bar .
Iteration 3	a man is standing near a bunch of video games in a bar .
Reference	**a man is standing by a group of video games in a bar .**

Source	une femme aux cheveux roses habillée en noir parle à un homme .
Iteration 0	a woman at hair roses dressed in black speaks to a man .
Iteration 1	a woman at glasses dressed in black talking to a man .
Iteration 2	a woman at pink hair dressed in black speaks to a man .
Iteration 3	a woman with pink hair dressed in black is talking to a man .
Reference	**a woman with pink hair dressed in black talks to a man .**

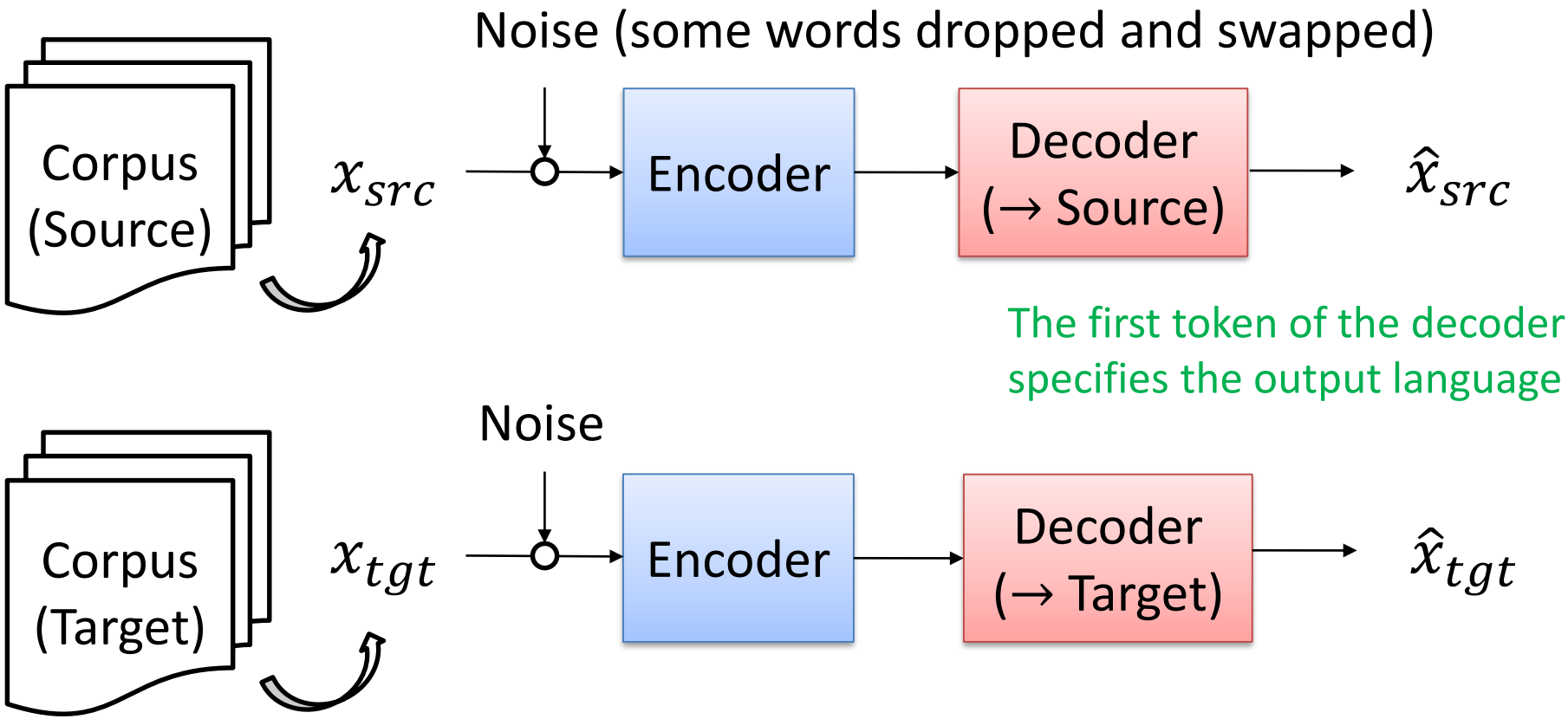
Lample et al., Unsupervised Machine Translation Using Monolingual Corpora Only, ICLR 2018

Initialization

- Approaches
  - Use a bilingual dictionary (Klementiev et al. 2012)
  - Use a dictionary inferred in an unsupervised way (Conneau et al., 2018; Artetxe et al., 2017)
  - Use a shared sub-word vocabulary between two languages (Lample et al., 2018)
    1. Join the monolingual corpora
    2. Apply BPE tokenization (Sennrich et al., 2016) on the resulting corpus
    3. Learn token embeddings by fastText

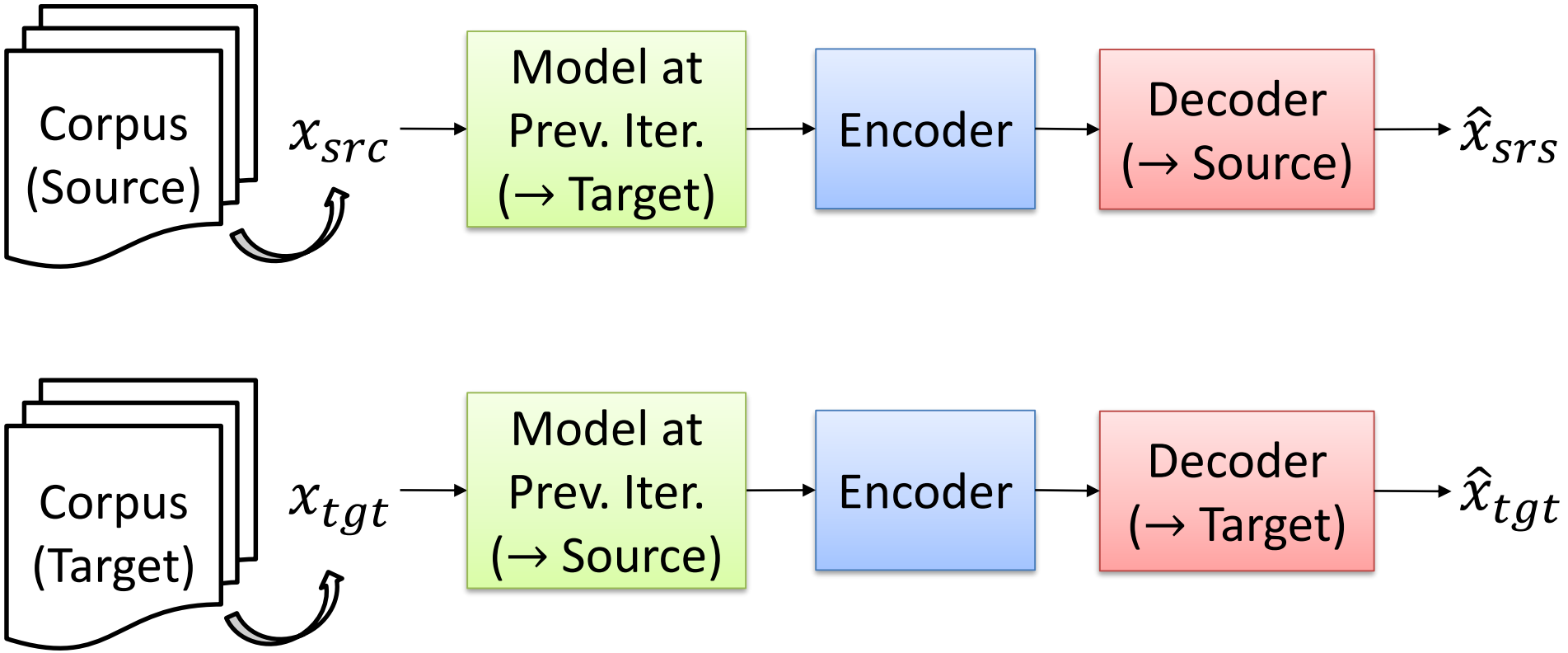
Language modeling (denoising auto-encoding)

$$\mathcal{L}^{lm} = E_{x\sim S}\left[-\log P_{s\to s}\left(x|C(x)\right)\right] + E_{x\sim T}\left[-\log P_{t\to t}\left(x|C(x)\right)\right]$$



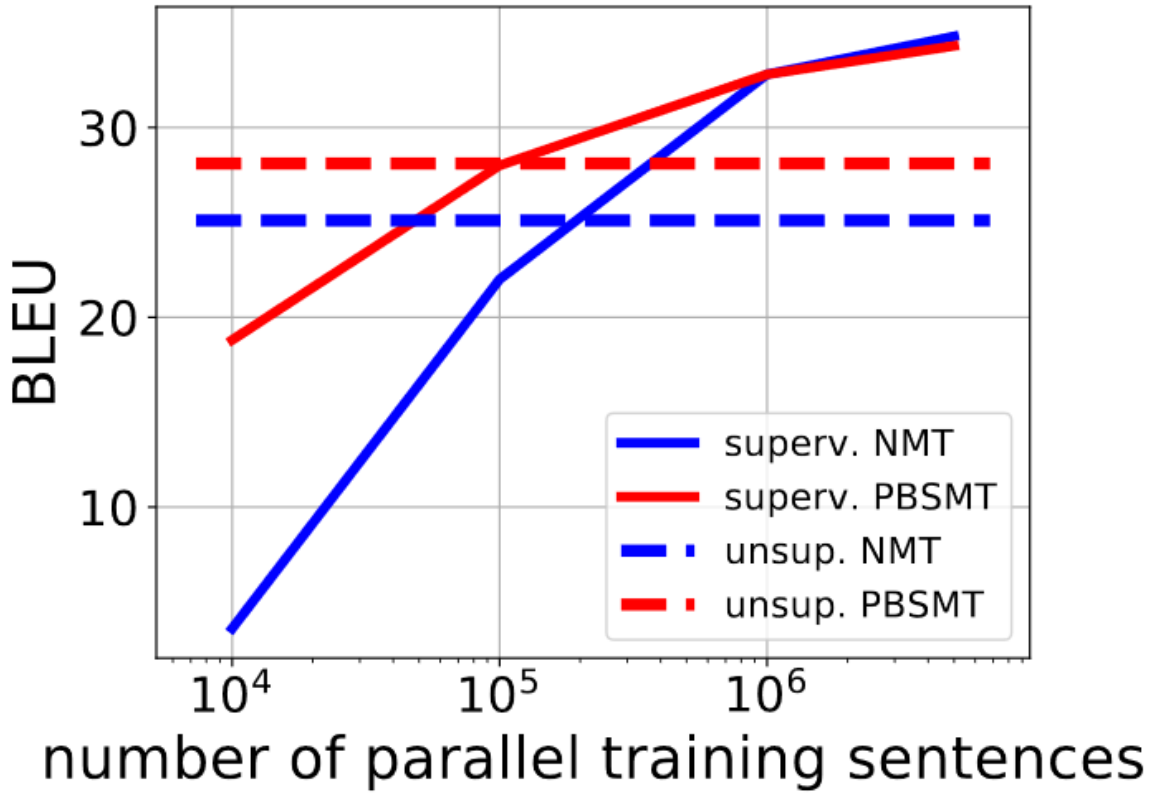
Back-translation

$$\mathcal{L}^{back} = E_{x\sim S}\left[-\log P_{t\to s}\left(x|v^*(x)\right)\right] + E_{x\sim T}\left[-\log P_{s\to t}\left(x|u^*(x)\right)\right]$$



Comparison to supervised MT

- WMT'14 En-Fr



Lample et al., Phrase-Based & Neural Unsupervised Machine Translation, EMNLP 2018

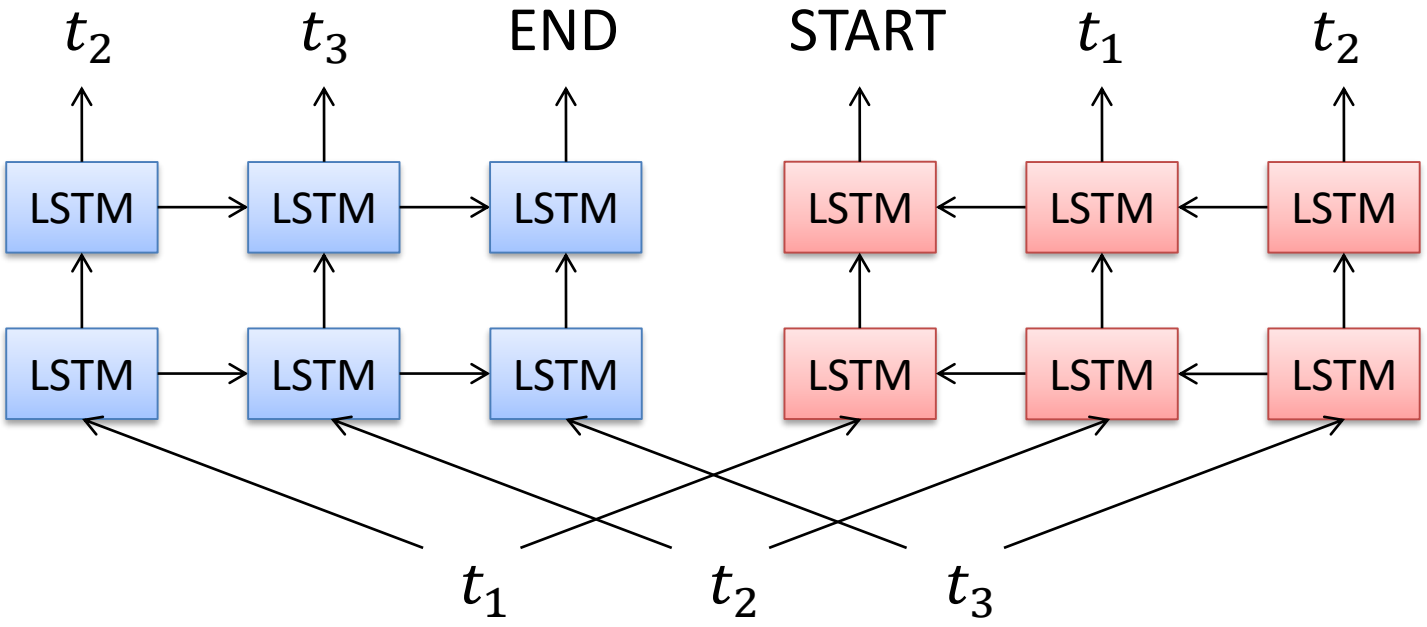
Pretraining methods

- Deep learning models require a large amount of labeled data for training
- How can we build accurate models without using a large amount of labeled data?

- Pretraining methods
  - Train a language model with a large amount of raw (unlabeled) text and then adapt it to various NLP tasks
    - Feature-based
      - ELMo (Peters et al., 2018)
    - Fine-tuning
      - GPT (Radford et al., 2018)
      - BERT (Devlin et al., 2019)

ELMo (Peters et al., 2018)

- Train two language models (left-to-right and right-to-left) on a large raw corpus



- Use the hidden vectors of LSTMs as features for NLP models

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}$$

Peters et al., Deep contextualized word representations, NAACL 2018

ELMo (Peters et al., 2018)

- Accuracy of existing NLP models can be improved by simply adding features produced by ELMo

TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

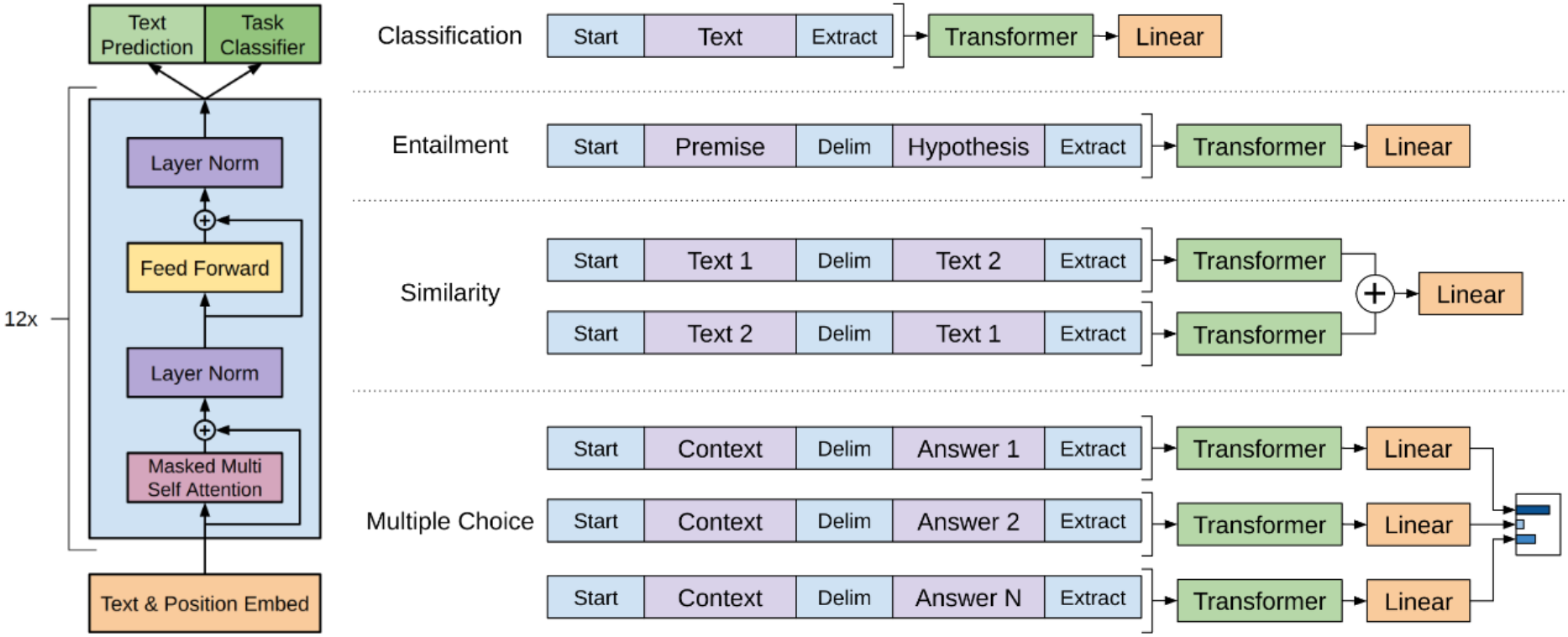
Peters et al., Deep contextualized word representations, NAACL 2018

GPT (Radford et al., 2018)

- GPT (Generative Pre-trained Transformer)
- Method
  - Train a Transformer language model with a large amount of raw text
    - 12-layer decoder-only Transformer
      - sequences of up to 512 tokens.
    - Took one month with 8 GPUs
    - BooksCorpus: 7000 unpublished books (~5GB of text).
  - Add a task-specific layer to the Transformer model and fine-tune it with labeled data
    - 3 epochs of training was sufficient for most cases

Radford et al., Improving language understanding by generative pre-training, 2018

GPT (Radford et al., 2018)



Radford et al., Improving language understanding by generative pre-training, 2018

MNLI (Multi-Genre Natural Language Inference, MultiNLI) [Williams et al., 2018]

- Entailment, Neural, or Contradiction

Met my first girlfriend that way.	FACE-TO-FACE **contradiction** C C N C	I didn't meet my first girlfriend until later.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** N N N N	The 8 million dollars for emergency housing was still not enough to solve the problem.
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** N N N N	All of the children love working in their gardens.
At 8:34, the Boston Center controller received a third transmission from American 11	9/11 **entailment** E E E E	The Boston Center controller got a third transmission from American 11.
I am a lacto-vegetarian.	SLATE **neutral** N N E N	I enjoy eating cheese too much to abstain from dairy.
someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny	TELEPHONE **contradiction** C C C C	No one noticed and it wasn't funny at all.

Williams et al., A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, NAACL-HLT 2018

RACE test (Lai et al., 2017)

In a small village in England about 150 years ago, a mail coach was standing on the street. It didn't come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn't have to pay the postage, while the receiver had to. "Here's a letter for Miss Alice Brown," said the mailman. " I'm Alice Brown," a girl of about 18 said in a low voice. Alice looked at the envelope for a minute, and then handed it back to the mailman. "I'm sorry I can't take it, I don't have enough money to pay it", she said. A gentleman standing around were very sorry for her. Then he came up and paid the postage for her. When the gentleman gave the letter to her, she said with a smile, " Thank you very much, This letter is from Tom. I'm going to marry him. He went to London to look for work. I've waited a long time for this letter, but now I don't need it, there is nothing in it." "Really? How do you know that?" the gentleman said in surprise. "He told me that he would put some signs on the envelope. Look, sir, this cross in the corner means that he is well and this circle means he has found work. That's good news." The gentleman was Sir Rowland Hill. He didn't forgot Alice and her letter. "The postage to be paid by the receiver has to be changed," he said to himself and had a good plan. "The postage has to be much lower, what about a penny? And the person who sends the letter pays the postage. He has to buy a stamp and put it on the envelope." he said . The government accepted his plan. Then the first stamp was put out in 1840. It was called the "Penny Black". It had a picture of the Queen on it.

The girl handed the letter back to the mailman because:

1. she didn't know whose letter it was
2. she had no money to pay the postage
3. she received the letter but she didn't want to open it
4. she had already known what was written in the letter

https://openai.com/blog/language-unsupervised/

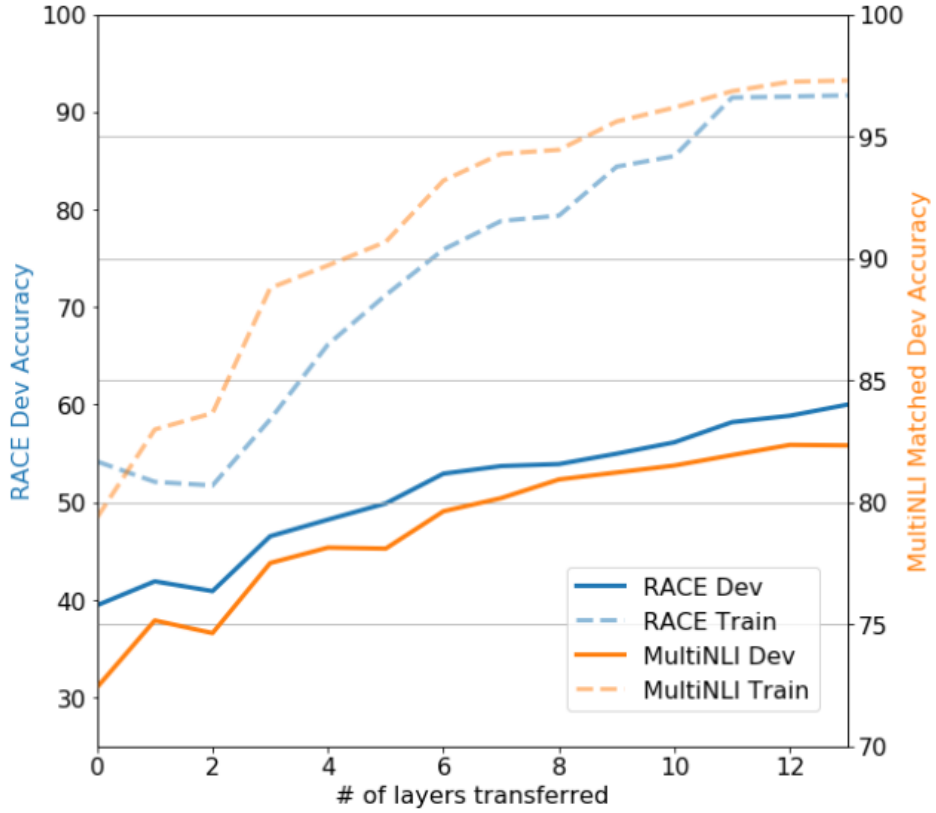
- Results on natural language inference tasks

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

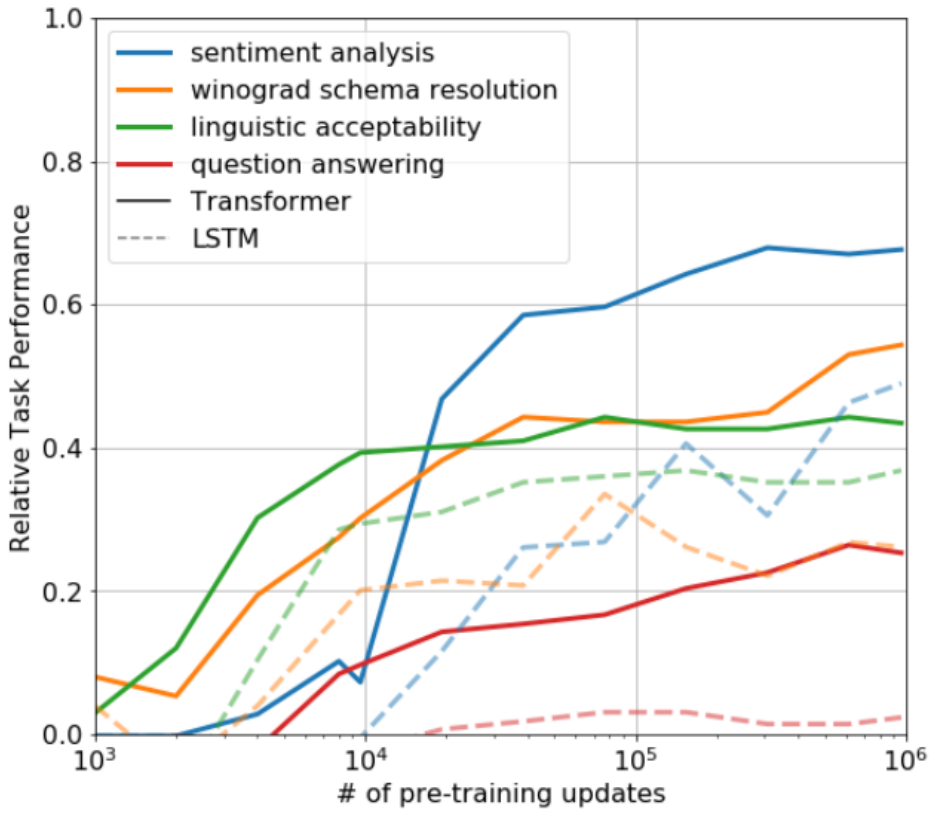
- Results on QA and common sense reasoning tasks

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

GPT (Radford et al., 2018)



Number of layers and accuracy



Zero-shot performance

BERT (Devlin et al., 2019)

- BERT (Bidirectional Encoder Representations from Transformers)
- Method
  1. Pretraining with unlabeled data
     - Learn deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context
  2. Fine-tuning with labeled data
     - Add a task-specific layer to the model
     - Fine-tune all the parameters

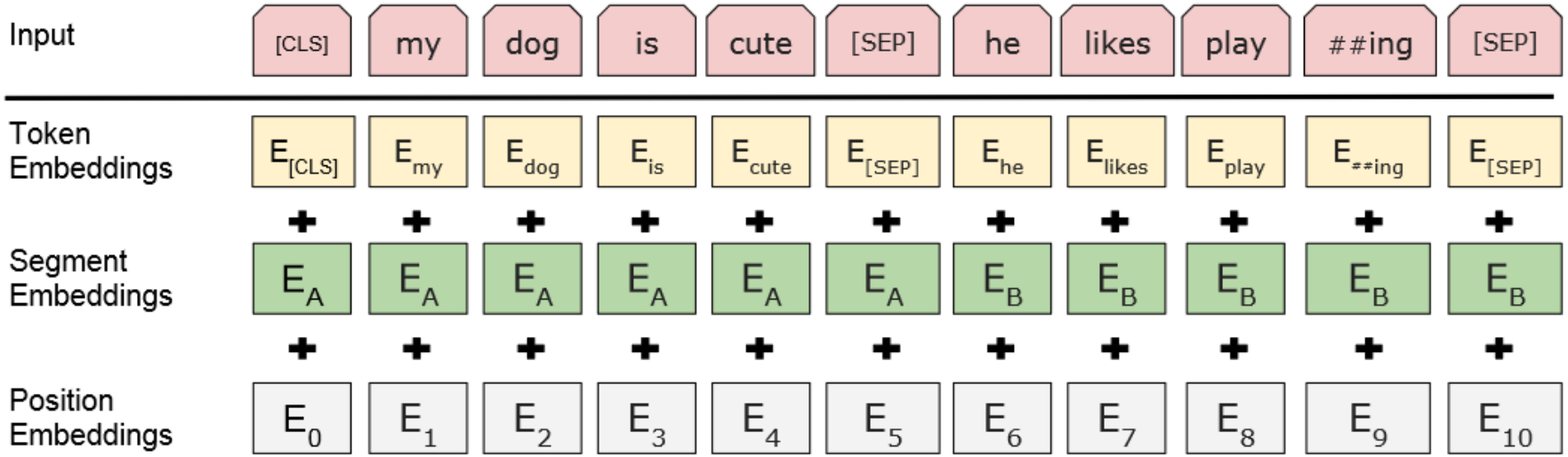
Pretraining

- Masked Language Model (MLM)
  - Some of the tokens are replaced with a special token **[MASK]**
  - The model is trained to predict them

- Next Sentence Prediction (NSP)
  - Predict whether the given two sentences are consecutive sentences in a document
- Data
  - BooksCorpus (800M words)
  - English Wikipedia (2,500M words)

Input representation

- First token is always [CLS] (special token for classification)
- Segments are separated by [SEP]
- Add a segment-specific embedding to each token



Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding, 2019

Pretraining

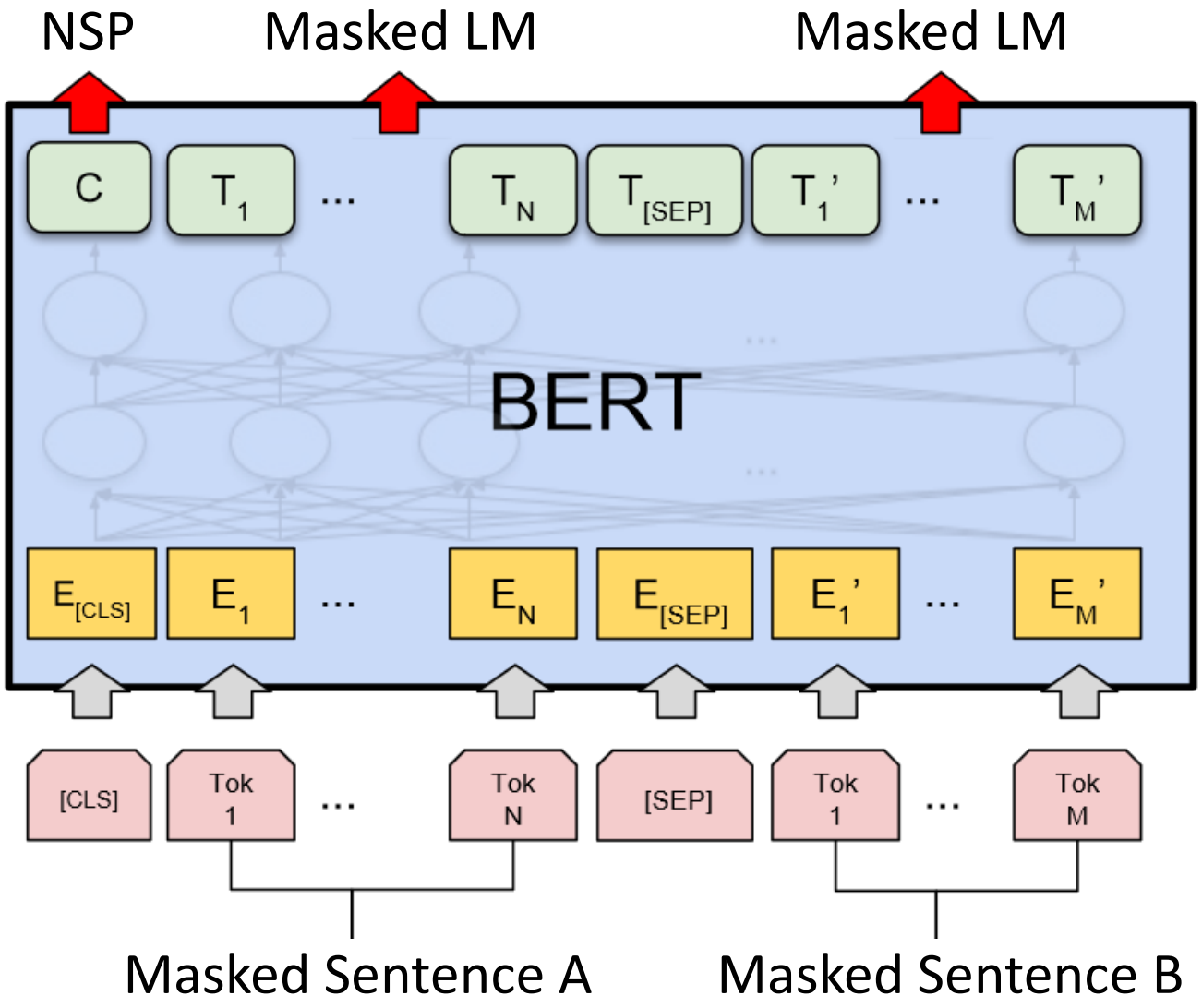
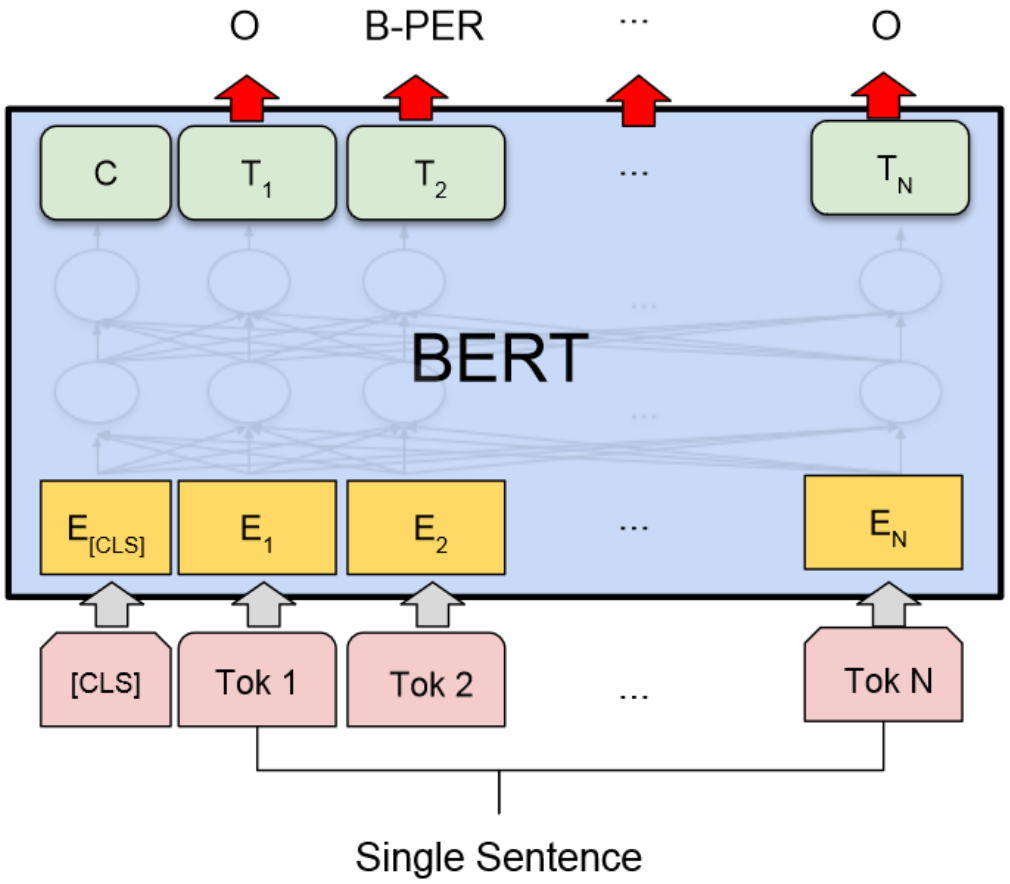


Figure adapted from Devlin et al. (2019)

Fine-tuning for Single Sentence Tagging Tasks

- CoNLL-2003 NER



Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding, 2019

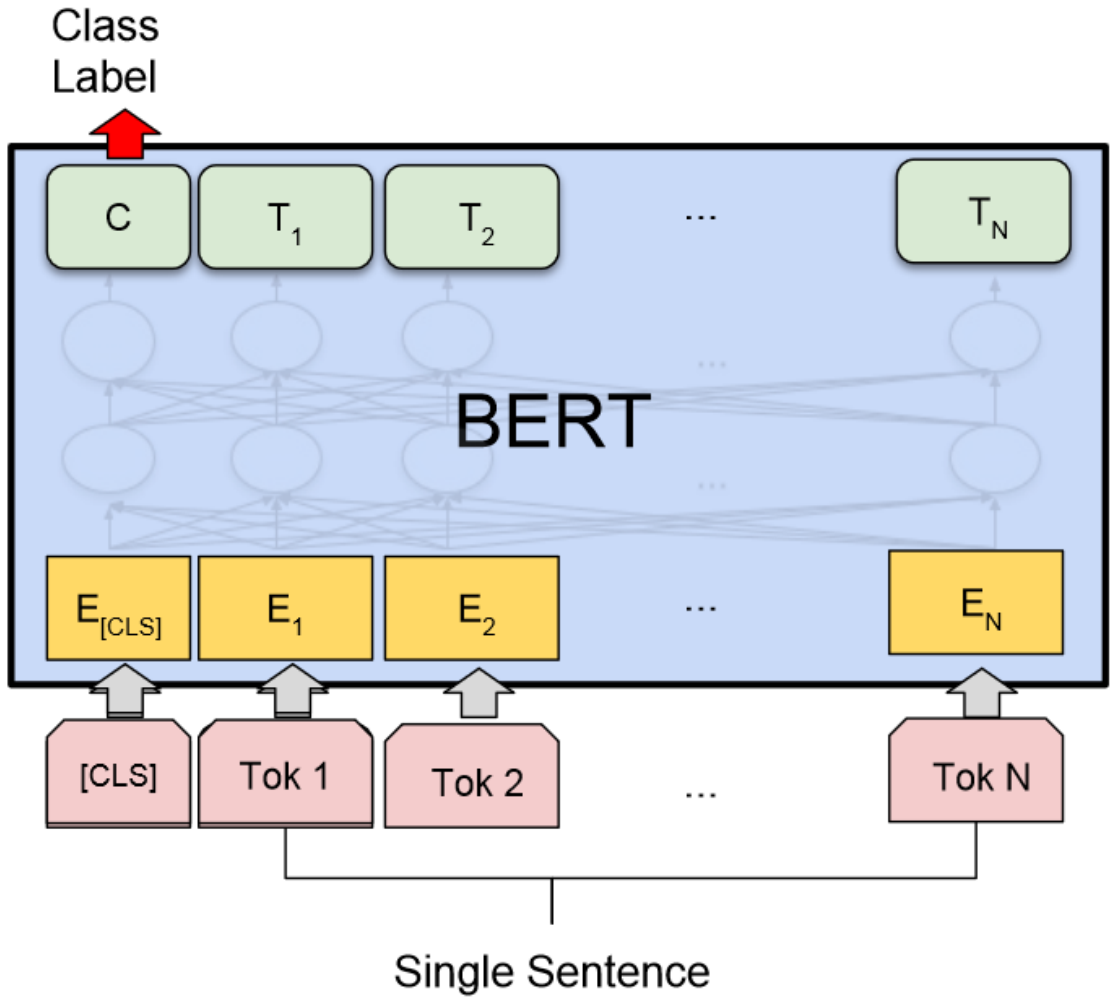
CoNLL-2003 NER
[Tjong Kim Sang, 2003]

- Named Entity Recognition

B-ORG	O	B-PER	O	O	B-LOC
U.N.	official	Ekeus	heads	for	Baghdad

Fine-turning for Single Sentence Classification tasks

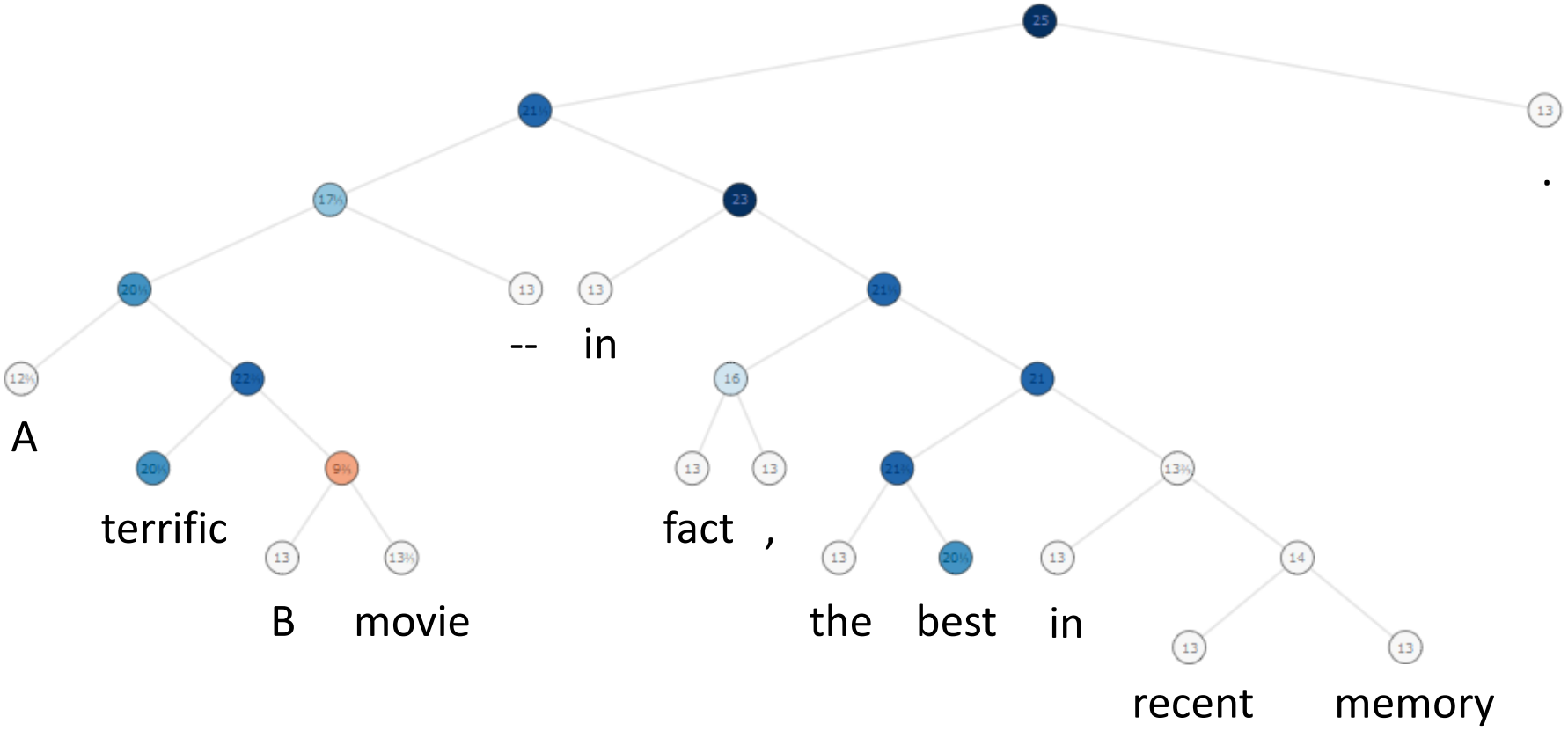
- SST-2, CoLA



Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding, 2019

Stanford Sentiment Treebank (SST)

[Socher et al., 2013]



https://nlp.stanford.edu/sentiment/treebank.html

CoLA (The Corpus of Linguistic Acceptability)

[Warstadtet al., 2018]

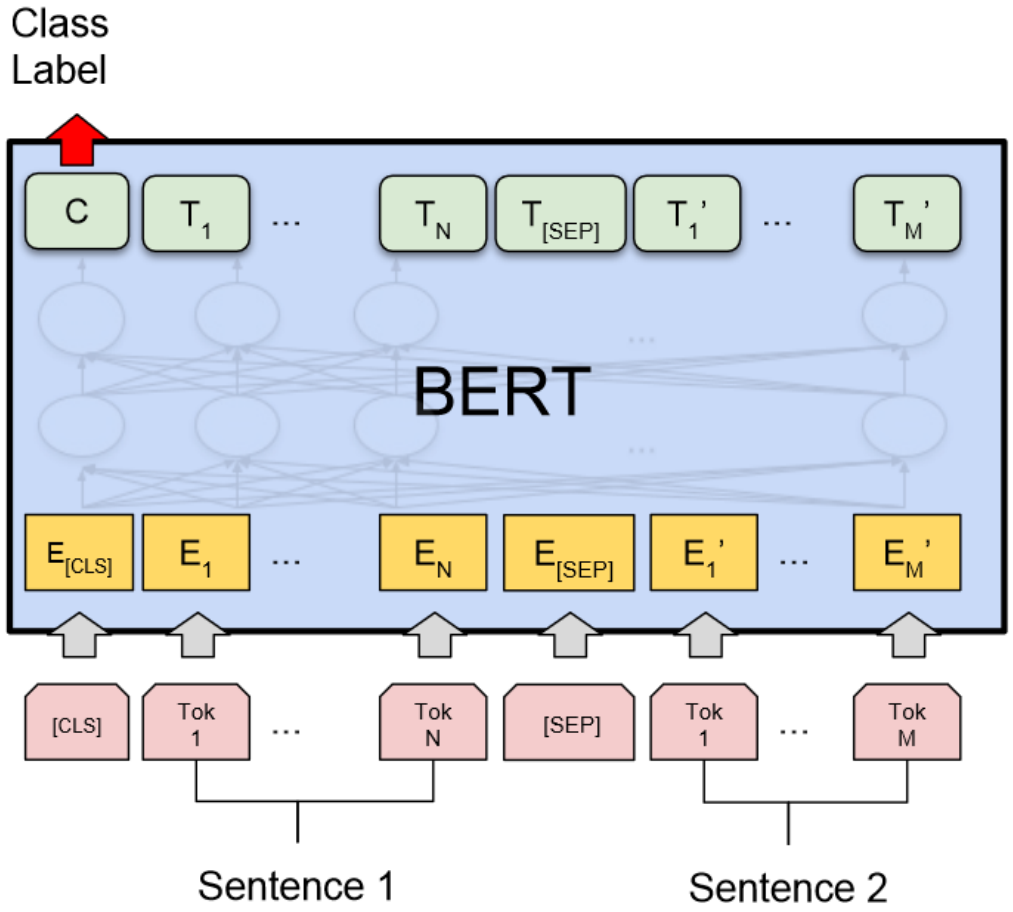
- Grammatical or ungrammatical

Label	Sentence
*	The more books I ask to whom he will give, the more he reads.
✓	I said that my father, he was tight as a hoot-owl.
✓	The jeweller inscribed the ring with the name.
*	many evidence was provided.
✓	They can sing.
✓	The men would have been all working.
*	Who do you think that will question Seamus first?
*	Usually, any lion is majestic.
✓	The gardener planted roses in the garden.
✓	I wrote Blair a letter, but I tore it up before I sent it.

(✓= acceptable, *=unacceptable)

Fine-tuning for Sentence Pair Classification tasks

- MNLI, QQP, QNLI, STS-B, MPRC, RTE, SWAG



Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding, 2019

Textual Entailment datasets

DATASET	EXAMPLE	LABEL
SNLI	1. A black race car starts up in front of a crowd of people. 2. A man is driving down a lonely road.	Contra.
MNLI	1. At the other end of Pennsylvania Avenue, people began to line up for a White House tour. 2. People formed a line at the end of Pennsylvania Avenue.	Entails
SciTail	1. Because type 1 diabetes is a relatively rare disease, you may wish to focus on prevention only if you know your child is at special risk for the disease. 2. Diabetes is unpreventable in the type one form but may be prevented by diet if it is of the second type.	Neutral
QNLI	Context: In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. Statement: What causes precipitation to fall?	Entails
RTE	1. Passions surrounding Germany's final match turned violent when a woman stabbed her partner because she didn't want to watch the game. 2. A woman passionately wanted to watch the game.	Contra.

https://openai.com/blog/language-unsupervised/

Evaluation

- GLUE test results
  - BERT outperformed GPT on all tasks

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

  - $\text{BERT}_{\text{BASE}}$
    - L=12, H=768, A=12, Total Param-eters=110M
    - Roughly the same size as OpenAI GPT
  - $\text{BERT}_{\text{LARGE}}$
    - L=24, H=1024, A=16, Total Parameters=340M

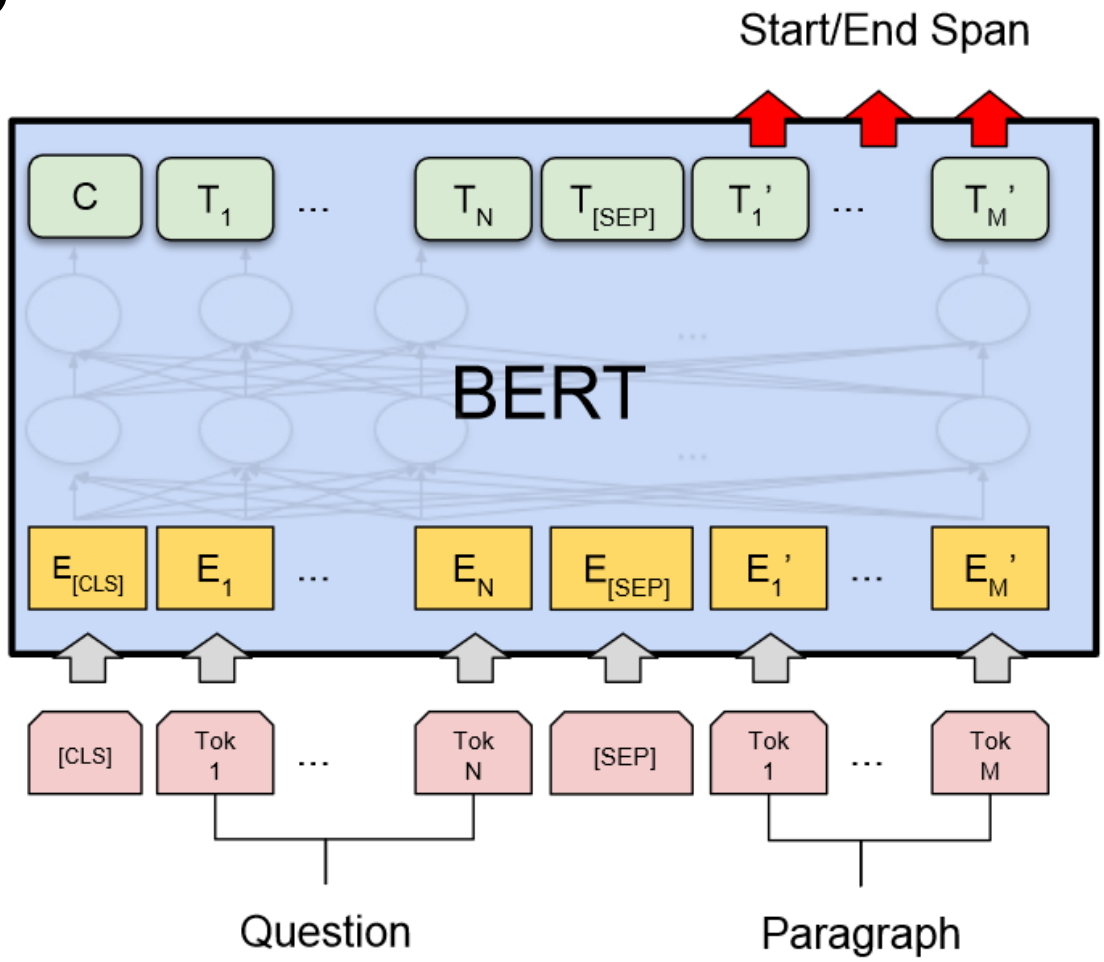
Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding, 2019

GLUE (General Language Understanding Evaluation) benchmark [Wang et al., 2019]

- Nine tasks on natural language understanding
  - Single-Sentence Tasks
    - CoLA (The Corpus of Linguistic Acceptability) [Warstadtet al., 2018]
    - SST-2 (The Stanford Sentiment Treebank) [Socheret al., 2013]
  - Similarity and Paraphrase Tasks
    - MRPC (Microsoft Research Paraphrase Corpus) [Dolan and Brockett, 2005]
    - STS-B (The Semantic Textual Similarity Benchmark) [Cer et al., 2017]
    - QQP (Quora Question Pairs) [Chen et al., 2018]
  - Inference Tasks
    - MNLI (Multi-Genre Natural Language Inference) [Williams et al., 2018]
    - QNLI (Question Natural Language Inference) [Wanget al., 2018]
    - RTE (Recognizing Textual Entailment) [Bentivogli et al., 2009]
    - WNLI (Winograd NLI) [Levesque et al., 2011]

Fine-tuning for extractive QA

- SQuAD



Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding, 2019

SQuAD [Rajpurkar et al., 2016]

A prime number (or a prime) is a natural number greater than 1 that has no positive divisors other than 1 and itself. A natural number greater than 1 that is not a prime number is called a composite number. For example, 5 is prime because 1 and 5 are its only positive integer factors, whereas 6 is composite because it has the divisors 2 and 3 in addition to 1 and 6. The fundamental theorem of arithmetic establishes the central role of primes in number theory: any integer greater than 1 can be expressed as a product of primes that is unique up to ordering. The uniqueness in this theorem requires excluding 1 as a prime because one can include arbitrarily many instances of 1 in any factorization, e.g., 3, $1 \cdot 3$, $1 \cdot 1 \cdot 3$, etc. are all valid factorizations of 3.

What is the only divisor besides 1 that a prime number can have?

What theorem defines the main role of primes in number theory?

Evaluation

- Results on SQuAD 1.1

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

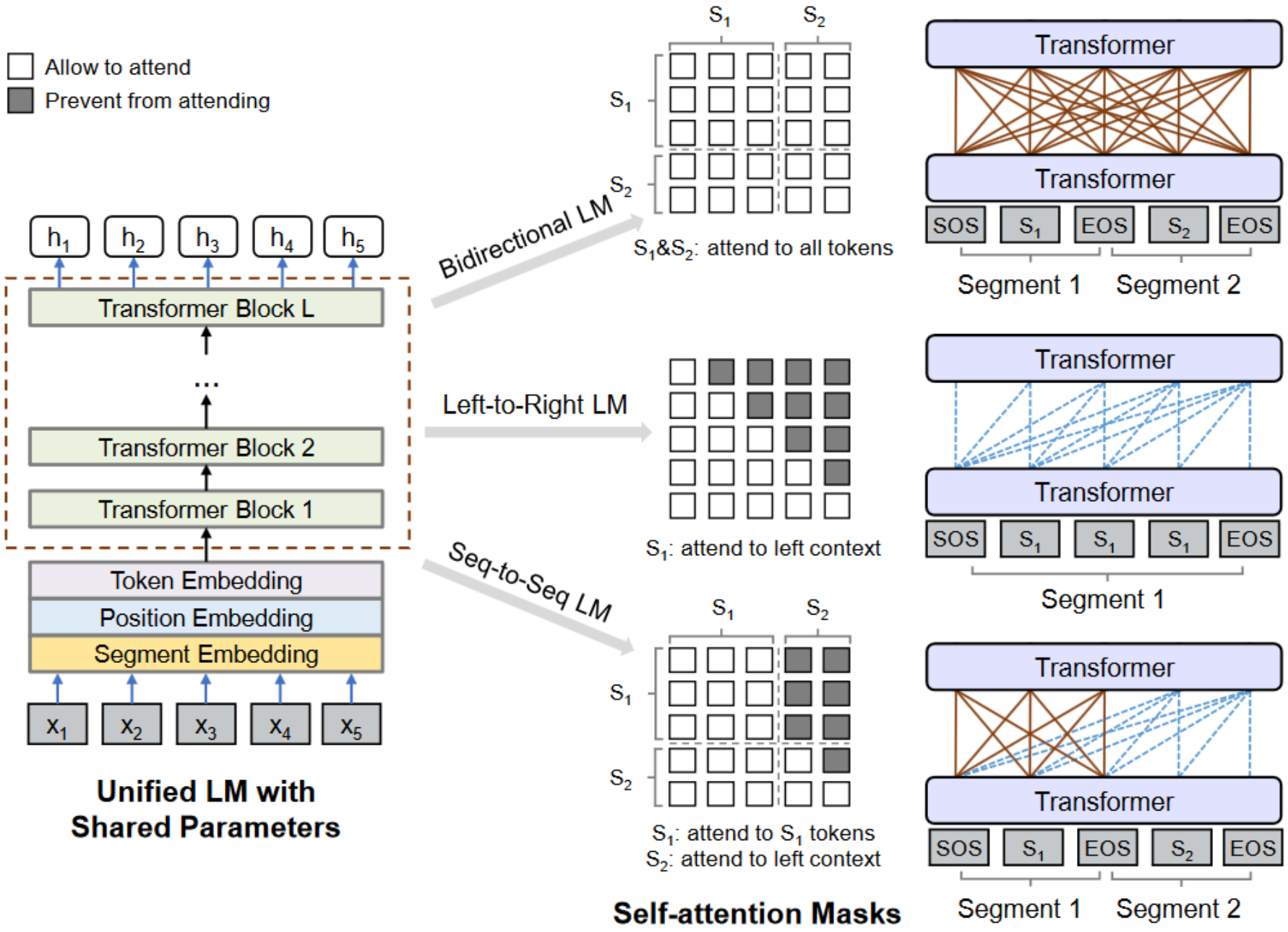
Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding, 2019

UNI-LM (Dong et al., 2019)

- UNI-LM (Unified Pre-trained Language Model)
  - Model that can be fine-tuned for both NLU and NLG

- Pre-training
  - Three kinds of language models
    - Unidirectional LM (left-to-right / right-to-left)
    - Bidirectional LM
    - Sequence-to-sequence LM
  - Data
    - English Wikipedia & BooksCorpus
- Fine-tuning

Dong et al., Unified Language Model Pre-training for Natural Language Understanding and Generation, NeurIPS 2019



Dong et al., Unified Language Model Pre-training for Natural Language Understanding and Generation, NeurIPS 2019

Examples from Gigaword corpus

S: norway delivered a diplomatic protest to russia on monday after three norwegian fisheries research expeditions were barred from russian waters . the norwegian research ships were to continue an annual program of charting fish resources shared by the two countries in the barents sea region .
T: norway protests russia barring fisheries research ships

S: volume of transactions at the nigerian stock exchange has continued its decline since last week , a nse official said thursday . the latest statistics showed that a total of ##.### million shares valued at ###.### million naira -lrb- about #.### million us dollars -rrb- were traded on wednesday in , deals .
T: transactions dip at nigerian stock exchange

Nallapati et al., Abstractive Text Summarization using Sequence-to-sequence RNNs andBeyond, 2016

Example from CNN/DailyMail corpus

Article (truncated): lagos , nigeria (cnn) a day after winning nigeria 's presidency , *muhammadu buhari* told cnn 's christiane amanpour that he plans to aggressively fight corruption that has long plagued nigeria and go after the root of the nation 's unrest . *buhari* said he 'll " rapidly give attention " to curbing violence in the northeast part of nigeria , where the terrorist group boko haram operates . by cooperating with neighboring nations chad , cameroon and niger , he said his administration is confident it will be able to thwart criminals and others contributing to nigeria 's instability . for the first time in nigeria 's history , the opposition defeated the ruling party in democratic elections . *buhari* defeated incumbent goodluck jonathan by about 2 million votes , according to nigeria 's independent national electoral commission . the win comes after a long history of military rule , coups and botched attempts at democracy in africa 's most populous nation .

Reference Summary:

muhammadu buhari tells cnn 's christiane amanpour that he will fight corruption in nigeria .
nigeria is the most populous country in africa and is grappling with violent boko haram extremists .
nigeria is also africa 's biggest economy , but up to 70 % of nigerians live on less than a dollar a day .

See et al., Get To The Point: Summarization with Pointer-Generator Networks, 2016

Evaluation

- CNN/DailyMail

	RG-1	RG-2	RG-L
Extractive Summarization			
LEAD-3	40.42	17.62	36.67
Best Extractive [27]	43.25	**20.24**	39.63
Abstractive Summarization			
PGNet [37]	39.53	17.28	37.98
Bottom-Up [16]	41.22	18.68	38.34
S2S-ELMo [13]	41.56	18.94	38.47
UNILM	**43.33**	20.21	**40.51**

- Gigaword

	RG-1	RG-2	RG-L
10K Training Examples			
Transformer [43]	10.97	2.23	10.42
MASS [39]	25.03	9.48	23.48
UNILM	**32.96**	**14.68**	**30.56**
Full Training Set			
OpenNMT [23]	36.73	17.86	33.68
Re3Sum [4]	37.04	19.03	34.46
MASS [39]	37.66	18.53	34.89
UNILM	**38.45**	**19.45**	**35.75**

Dong et al., Unified Language Model Pre-training for Natural Language Understanding and Generation, NeurIPS 2019

Summary

- Word embeddings
  - Word2vec, Skip-gram, fastText
- Recurrent Neural Networks
  - LSTM, GRU
- Neural Machine Translation
  - Encoder-decoder model
  - Transformer
  - Unusupervised NMT
- Pretraining methods
  - ELMo, GPT, BERT, UNI-LM
  - Sentiment Analysis, Textual Entailment, Question Answering, Summarization