Lecture 7: NLP highlights and future

Radoslav Neychev
Anastasia Ianina

Harbour.Space University
16.07.2019, Barcelona, Spain

Loss functions

Two general problems: classification and regression.

Classification:

- Cross-entropy loss
  - Difference between distributions
  - See it again soon

$$H(p, q) = -\sum_{x} p(x) \log q(x)$$

Regression:

- MSE
- MAE
- Any stuff you love

Word representations

Word vectors are simply vectors of numbers that represent the meaning of a word

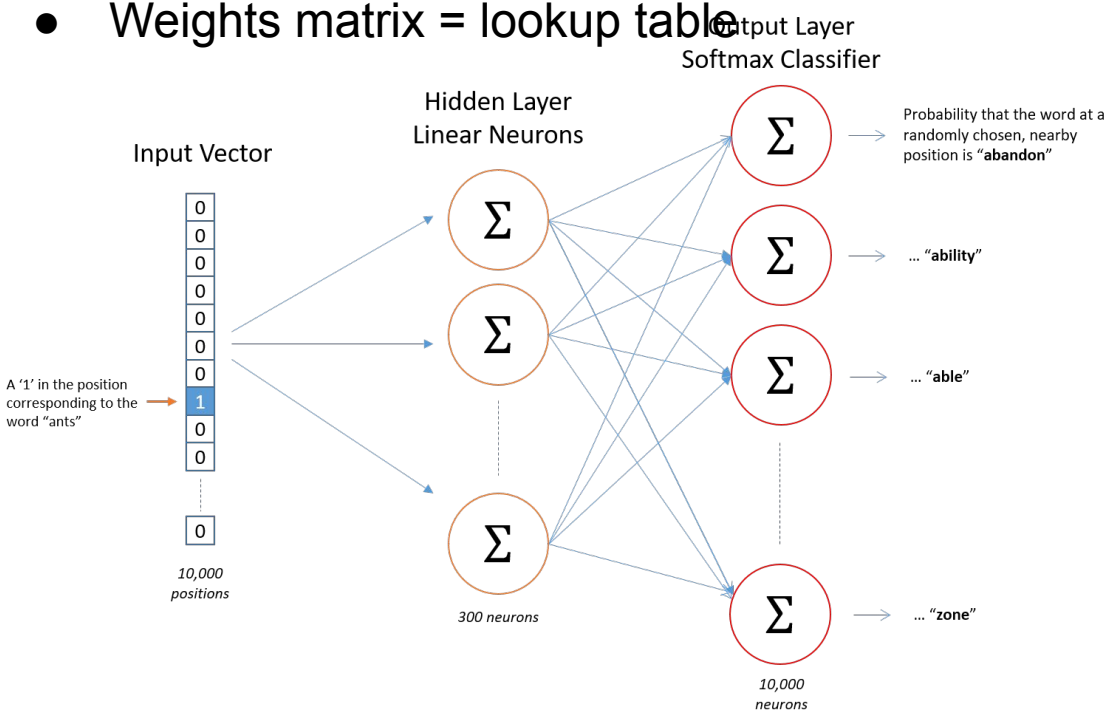
Approaches:

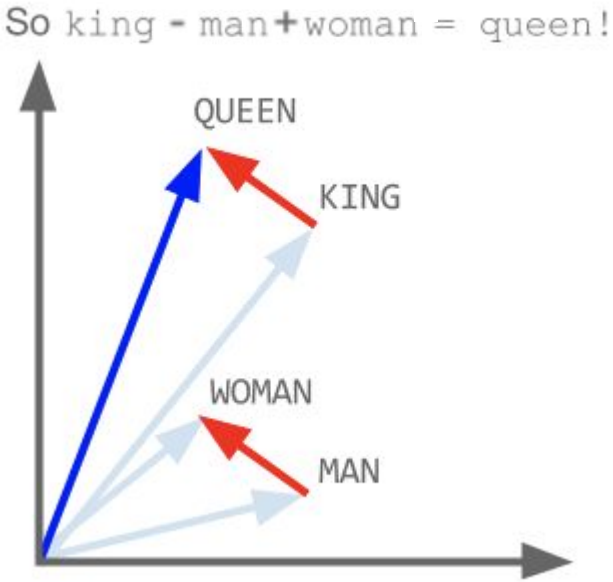
- One-hot encoding
- Bag-of-words models
- Counts of word / context co-occurrences
- TF-IDF
- Word Embeddings (e.g. word2vec, GloVe):
  - “You shall know a word by the company it keeps” (Firth, 1957)

Embeddings

- Learned from *unlabeled* data
- Predict context by the word (skip-gram) or word by its context (CBOW)
- Weights matrix = lookup table

- Solves matrix factorization problem
- Engineering tricks are important as well
  - Negative sampling, subsampling, etc.
- Provides word alignments





RNNs generating...

Shakespeare

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

Algebraic Geometry (Latex)

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.

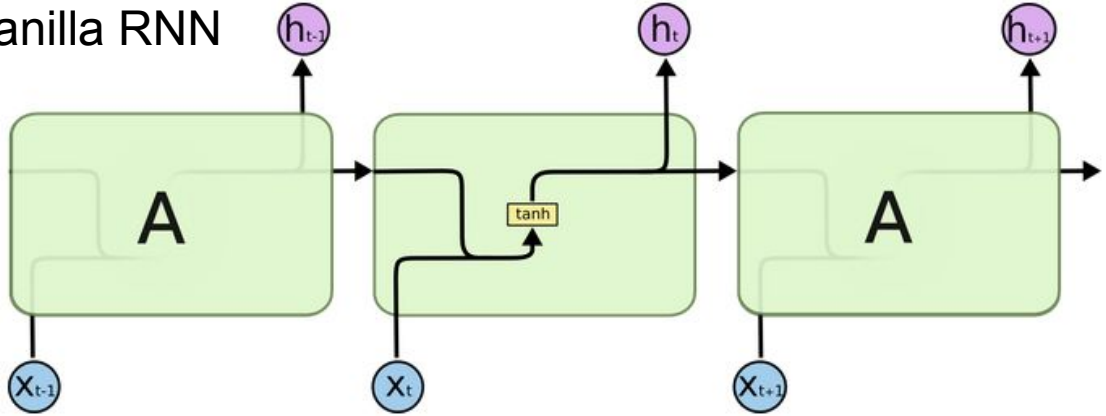
Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. □

Linux kernel (source code)

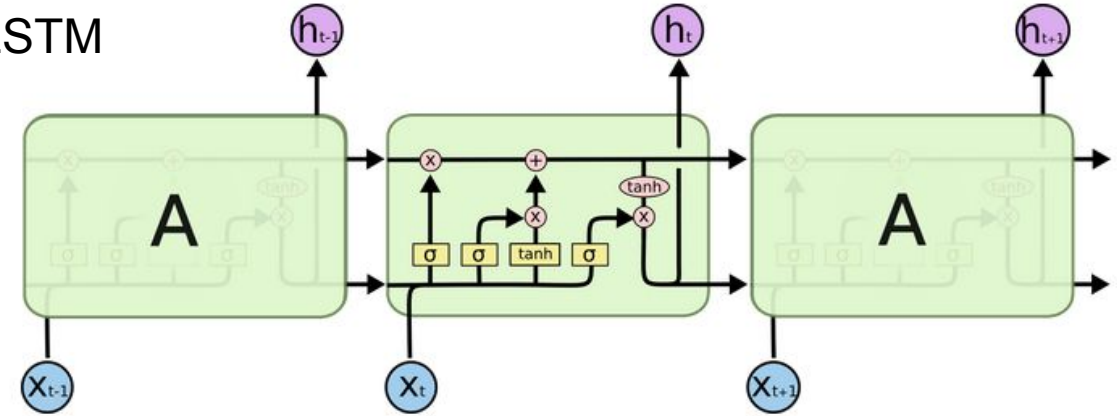
```
/*
 * If this error is set, we will need anything right after that BSD.
 */
static void action_new_function(struct s_stat_info *wb)
{
  unsigned long flags;
  int lel_idx_bit = e->edd, *sys & ~((unsigned long) *FIRST_COMPAT);
  buf[0] = 0xFFFFFFFF & (bit << 4);
  min(inc, slist->bytes);
  printk(KERN_WARNING "Memory allocated %02x/%02x, "
    "original MLL instead\n"),
    min(min(multi_run - s->len, max) * num_data_in),
    frame_pos, sz + first_seg);
  div_u64_w(val, inb_p);
  spin_unlock(&disk->queue_lock);
  mutex_unlock(&s->sock->mutex);
  mutex_unlock(&func->mutex);
  return disassemble(info->pending_bh);
}
```

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

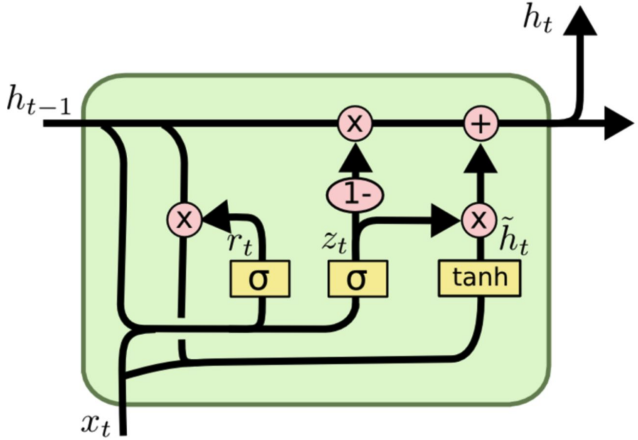
Vanilla RNN



LSTM



GRU



Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM vs GRU

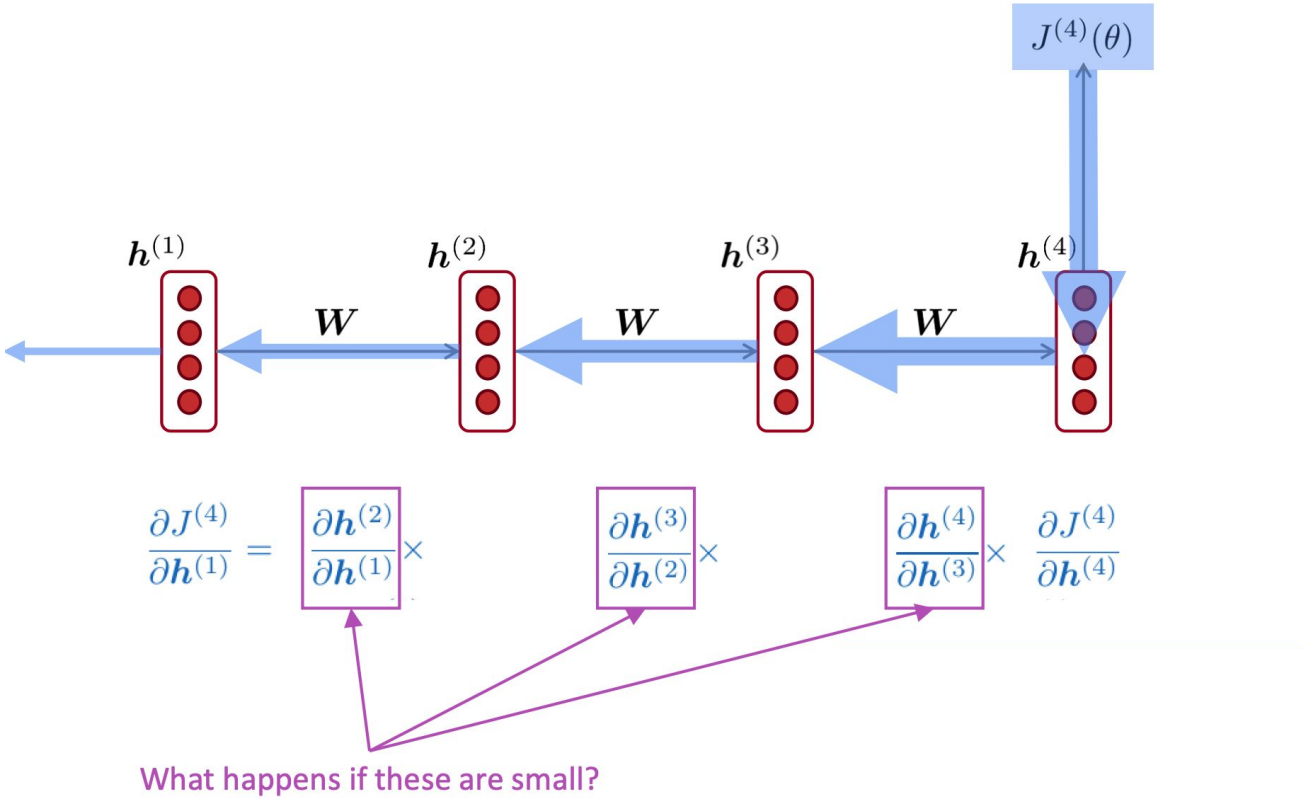
- LSTM and GRU are both great
  - GRU is quicker to compute and has fewer parameters than LSTM
  - There is no conclusive evidence that one consistently performs better than the other
  - LSTM is a good default choice (especially if your data has particularly long dependencies, or you have lots of training data)

Rule of thumb: start with LSTM, but switch to GRU if you want something more efficient

Vanishing gradient

Vanishing gradient problem:

When the derivatives are small, the gradient signal gets smaller and smaller as it backpropagates further



More info: "On the difficulty of training recurrent neural networks", Pascanu et al, 2013
http://proceedings.mlr.press/v28/pascanu13.pdf

Vanishing gradient in non-RNN

Vanishing gradient is present in **all** deep neural network architectures.

- Due to chain rule / choice of nonlinearity function, gradient can become vanishingly small during backpropagation
- Lower levels are hard to train and are trained slower
- **Potential solution**: or skip-connections/dense-connections/other shortcuts

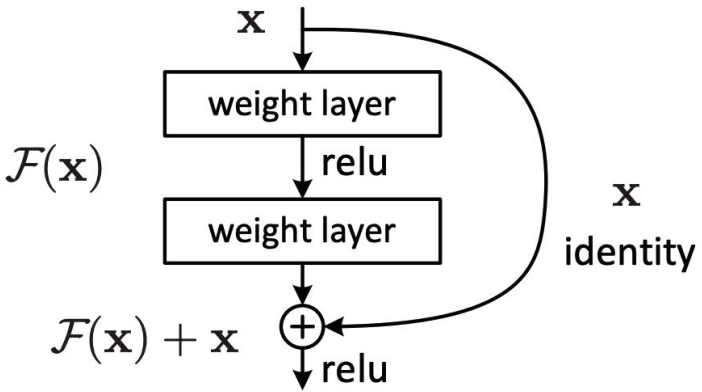
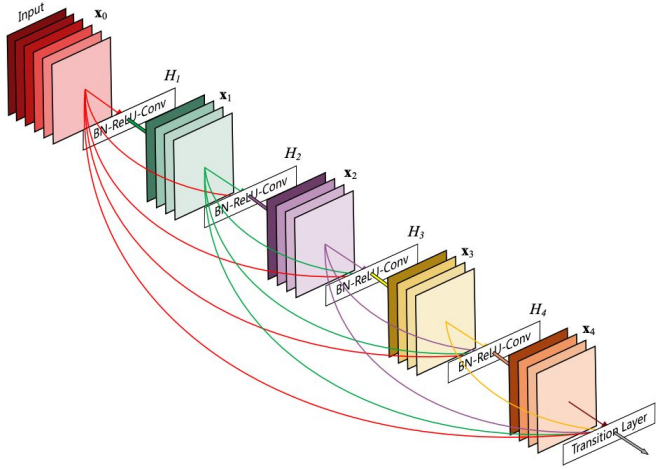


Figure 2. Residual learning: a building block.



Source: "Deep Residual Learning for Image Recognition", He et al, 2015. https://arxiv.org/pdf/1512.03385.pdf

Exploding gradient problem

- If the gradient becomes too big, then the SGD update step becomes too big:

$$\theta^{new} = \theta^{old} - \underbrace{\overbrace{\alpha}^{\text{learning rate}}}_{\text{gradient}} \nabla_\theta J(\theta)$$

- This can cause bad updates: we take too large a step and reach a bad parameter configuration (with large loss)

- In the worst case, this will result in Inf or NaN in your network (then you have to restart training from an earlier checkpoint)

Based on: Lecture by Abigail See, CS224n Lecture 7

Exploding gradient solution

- Gradient clipping: if the norm of the gradient is greater than some threshold, scale it down before applying SGD update

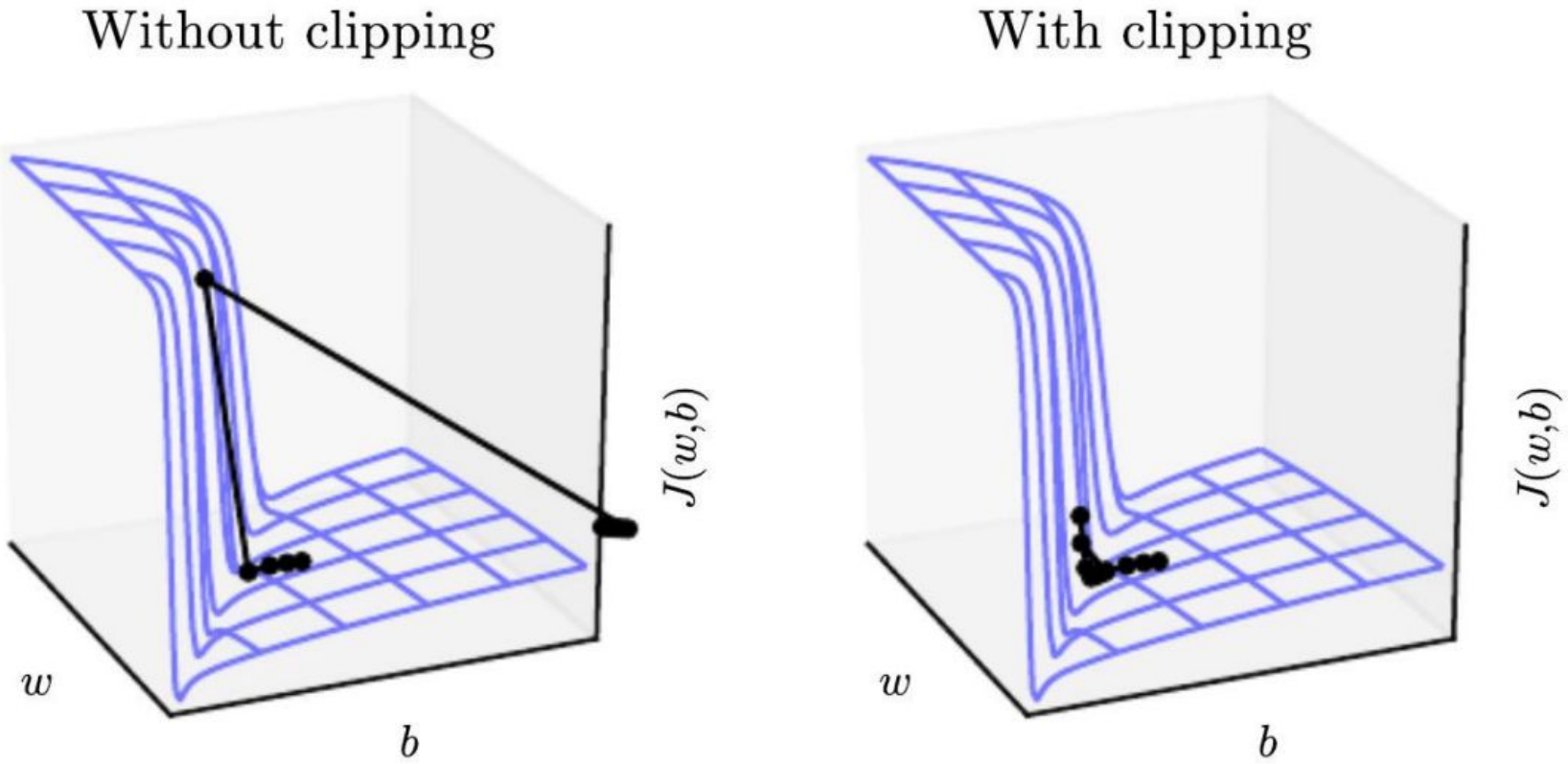
Algorithm 1 Pseudo-code for norm clipping

$\hat{\mathbf{g}} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$
if $\|\hat{\mathbf{g}}\| \geq threshold$ **then**
  $\hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|}\hat{\mathbf{g}}$
end if

- Intuition: take a step in the same direction, but a smaller step

Based on: Lecture by Abigail See, CS224n Lecture 7

Exploding gradient solution



Based on: Lecture by Abigail See, CS224n Lecture 7

From RNN to CNN

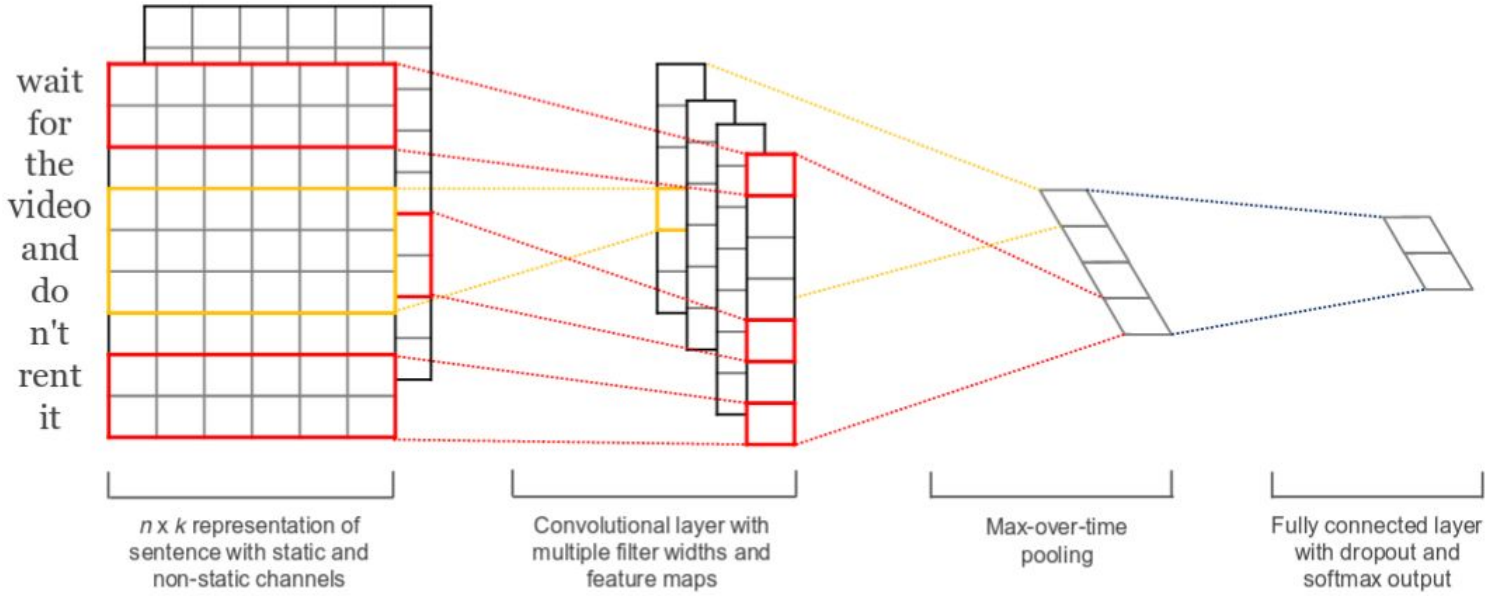
- RNN: Get compositional vectors for grammatical phrases only

- CNN: What if we compute vectors for every possible phrase?
  - Example: *“the country of my birth”* computes vectors for:
    - *the country, country of, of my, my birth, the country of, country of my, of my birth, the country of my, country of my birth*

- Regardless of whether it is grammatical
- Wouldn’t need parser
- Not very linguistically or cognitively plausible

Based on: Lecture by Richard Socher 5/12/16, http://cs224d.stanford.edu

Example from Kim (2014) paper



Main highlights

- Vanishing gradient is present not only in RNNs
  - Use some kind of memory or skip-connections
- LSTM and GRU are both great
  - GRU is quicker, LSTM catch more complex dependencies
- Clip your gradients
- Combining RNN and CNN worlds? Why not ;)

Optimizers

There are much more optimizers:

- Momentum
- Adagrad
- Adadelta
- RMSprop
- Adam

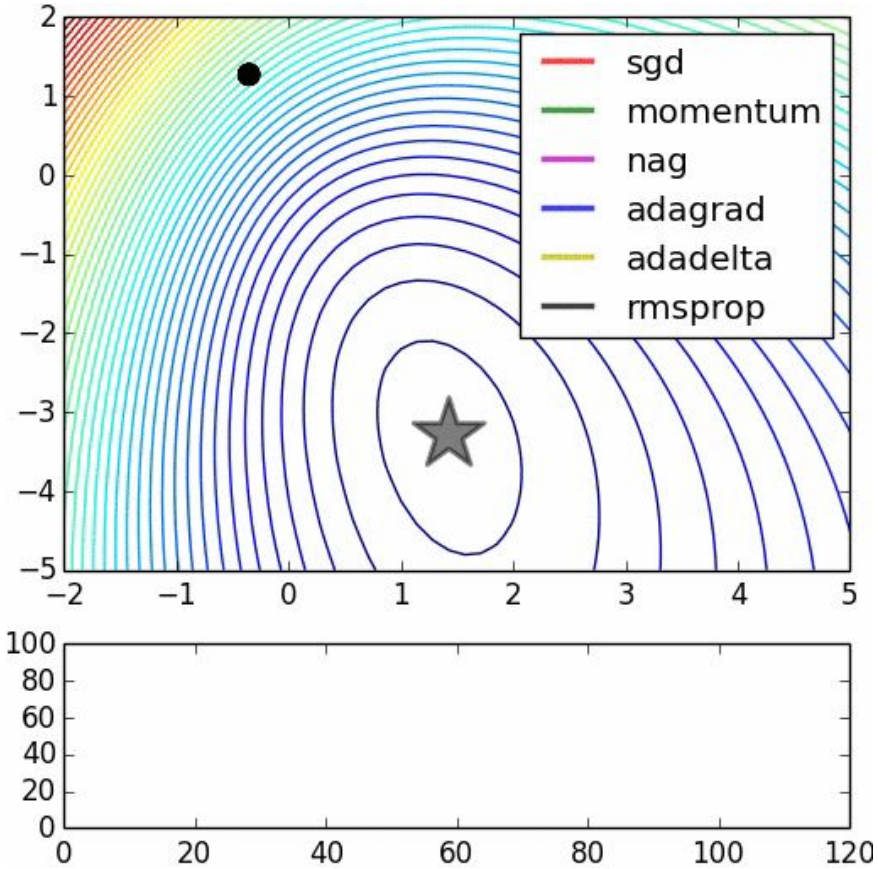
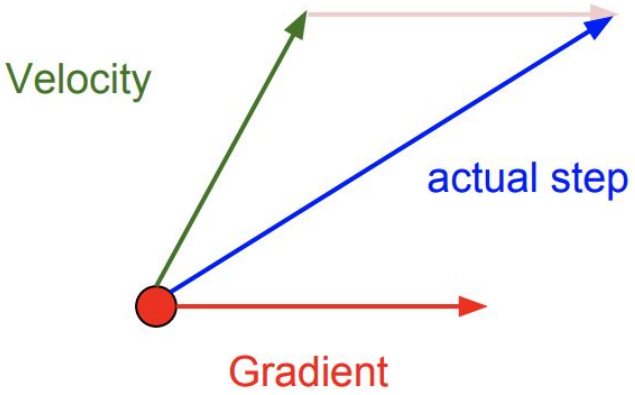


Image credits: Alec Radford

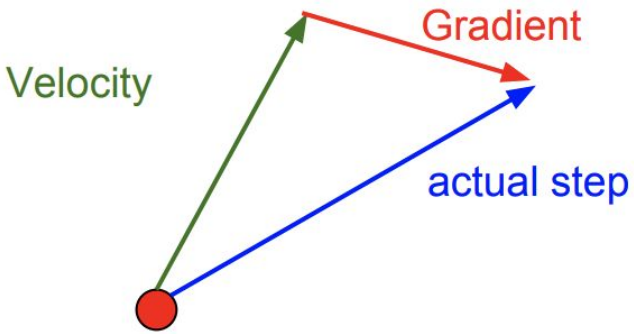
Nesterov momentum

Momentum update:



$$v_{t+1} = \rho v_t + \nabla f(x_t)$$
$$x_{t+1} = x_t - \alpha v_{t+1}$$

Nesterov Momentum



$$v_{t+1} = \rho v_t - \alpha \nabla f(x_t + \rho v_t)$$
$$x_{t+1} = x_t + v_{t+1}$$

source: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture7.pdf

Second idea: different dimensions are different

Adagrad: SGD with cache

$$\text{cache}_{t+1} = \text{cache}_t + (\nabla f(x_t))^2$$

$$x_{t+1} = x_t - \alpha \frac{\nabla f(x_t)}{\text{cache}_{t+1} + \varepsilon}$$

RMSProp: SGD with cache with exp. Smoothing

$$\text{cache}_{t+1} = \beta \text{cache}_t + (1-\beta)(\nabla f(x_t))^2$$

$$x_{t+1} = x_t - \alpha \frac{\nabla f(x_t)}{\text{cache}_{t+1} + \varepsilon}$$

Slide 29 Lecture 6 of Geoff Hinton's Coursera class
http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

Adam

Let's combine the momentum idea and RMSProp normalization:

$$v_{t+1} = \gamma v_t + (1-\gamma)\nabla f(x_t)$$

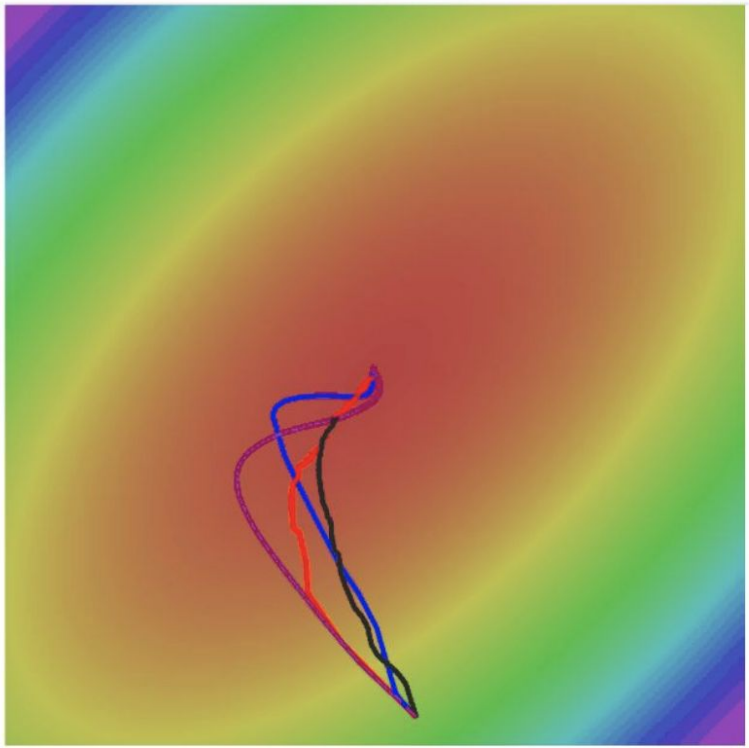
$$\text{cache}_{t+1} = \beta \text{cache}_t + (1-\beta)(\nabla f(x_t))^2$$

$$x_{t+1} = x_t - \alpha \frac{v_{t+1}}{\text{cache}_{t+1} + \varepsilon}$$

Actually, that's not quite Adam.

Adam full form involves bias correction term. See http://cs231n.github.io/neural-networks-3/ for more info.

Comparing optimizers



SGD

SGD+Momentum

RMSProp

Adam

source: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture7.pdf

Regularization

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq y_i}\max(0, f(x_i; W)_j - f(x_i; W)_{y_i} + 1) + \boxed{\lambda R(W)}$$

Adding some extra term to the loss function.

Common cases:

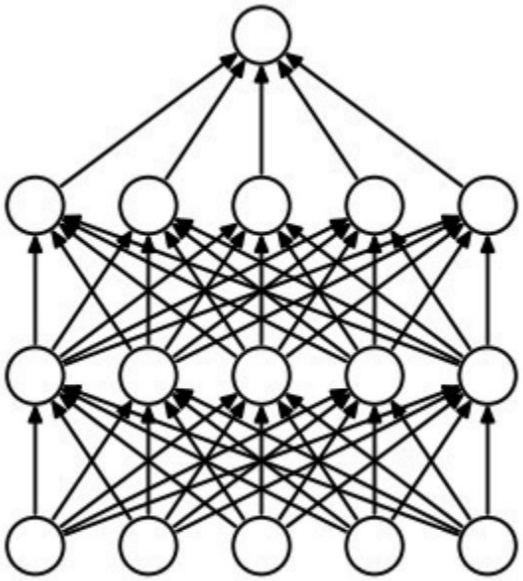
- L2 regularization: $R(W) = \|W\|_2^2$
- L1 regularization: $R(W) = \|W\|_1$
- Elastic Net (L1 + L2): $R(W) = \beta\|W\|_2^2 + \|W\|_1$

source: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture7.pdf

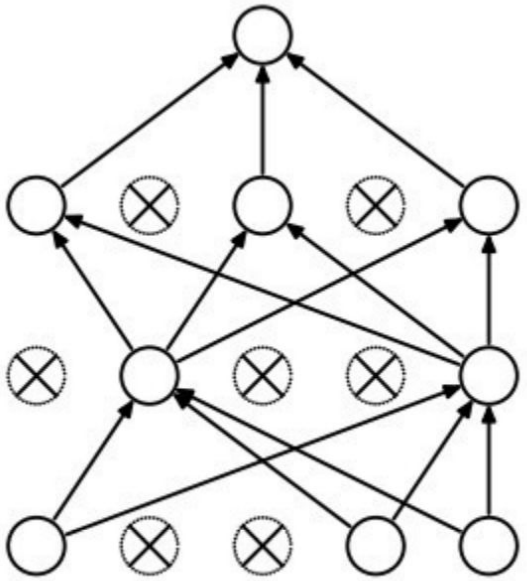
Regularization: Dropout

Some neurons are “dropped” during training.

Prevents overfitting.



(a) Standard Neural Net

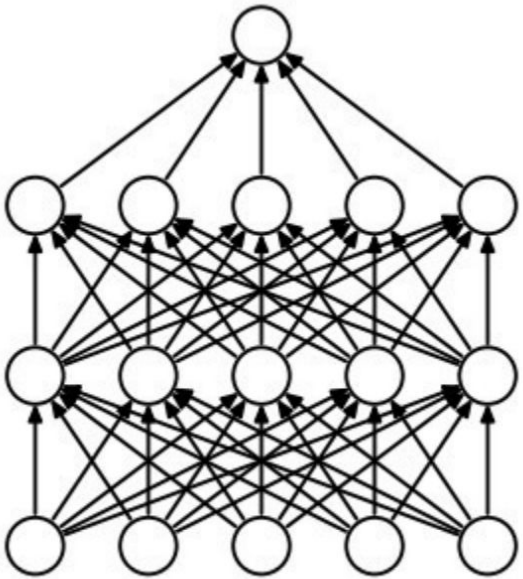


(b) After applying dropout.

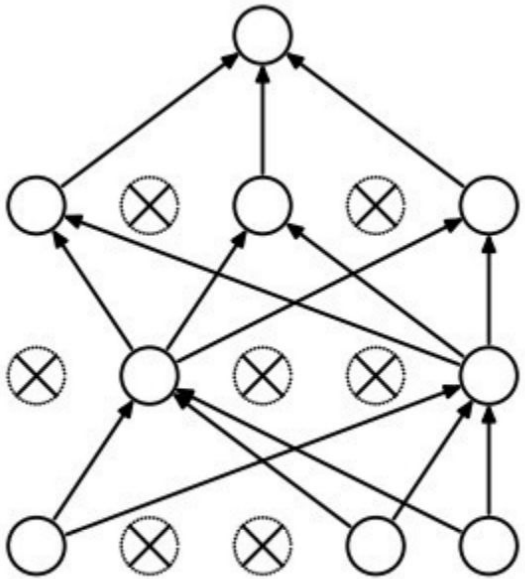
Regularization: Dropout

Some neurons are “dropped” during training.

Prevents overfitting.



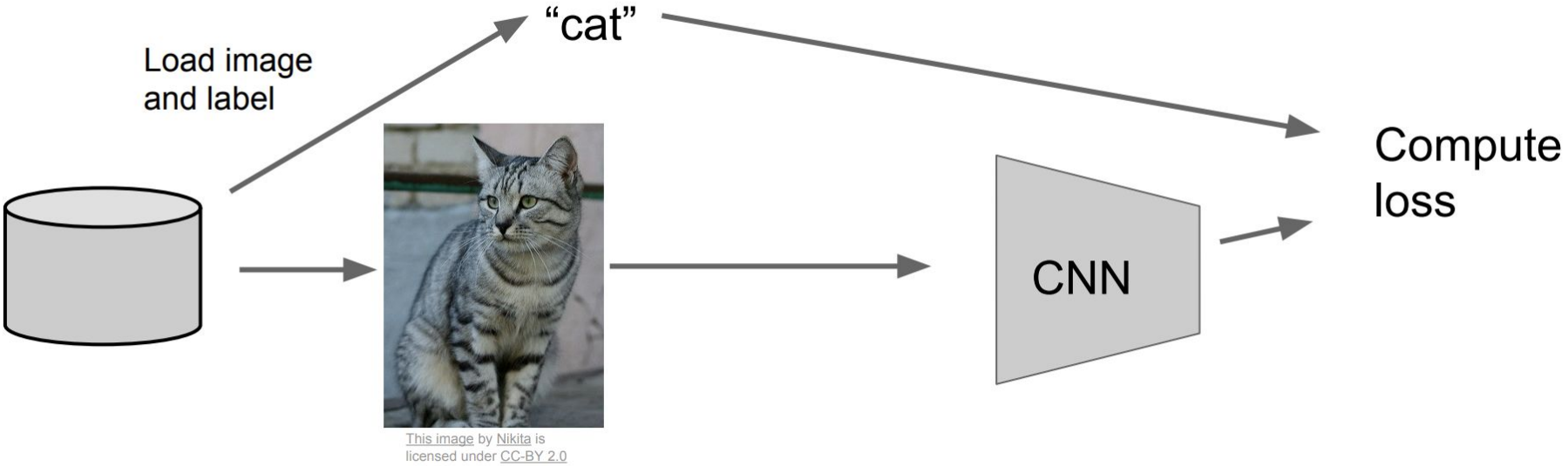
(a) Standard Neural Net



(b) After applying dropout.

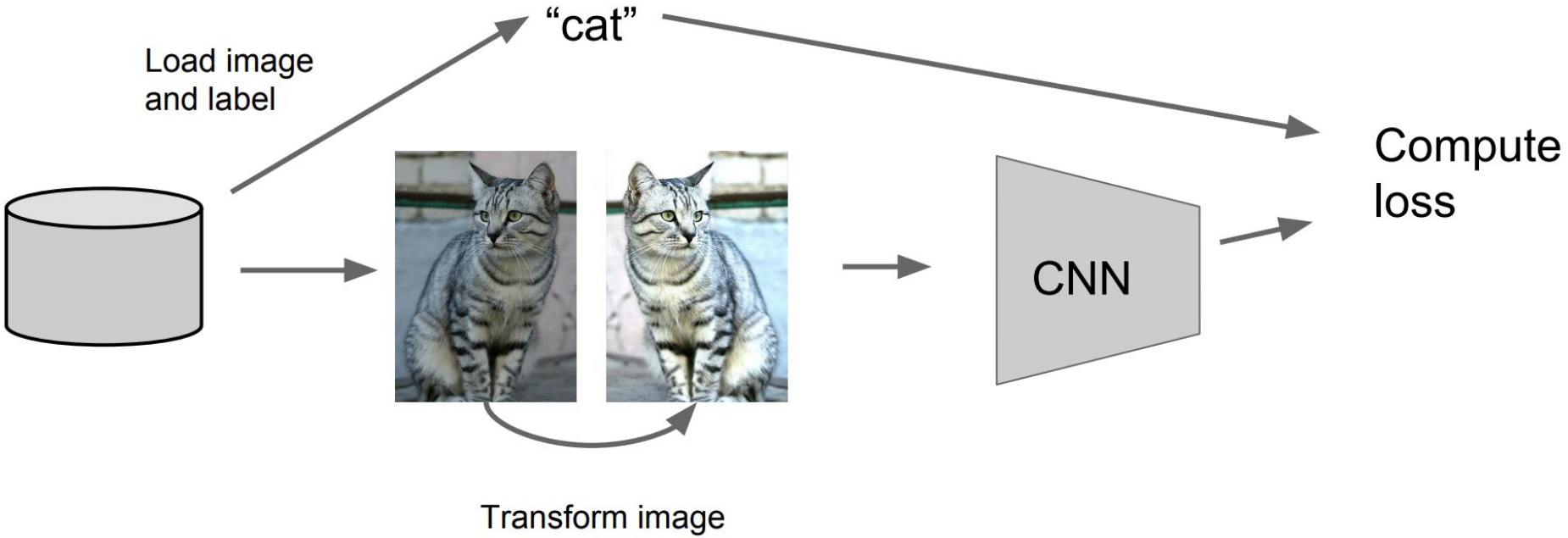
Actually, on test case output should be normalized. See sources for more info.

Regularization: data augmentation



source: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture7.pdf

Regularization: data augmentation



source: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture7.pdf

Main highlights

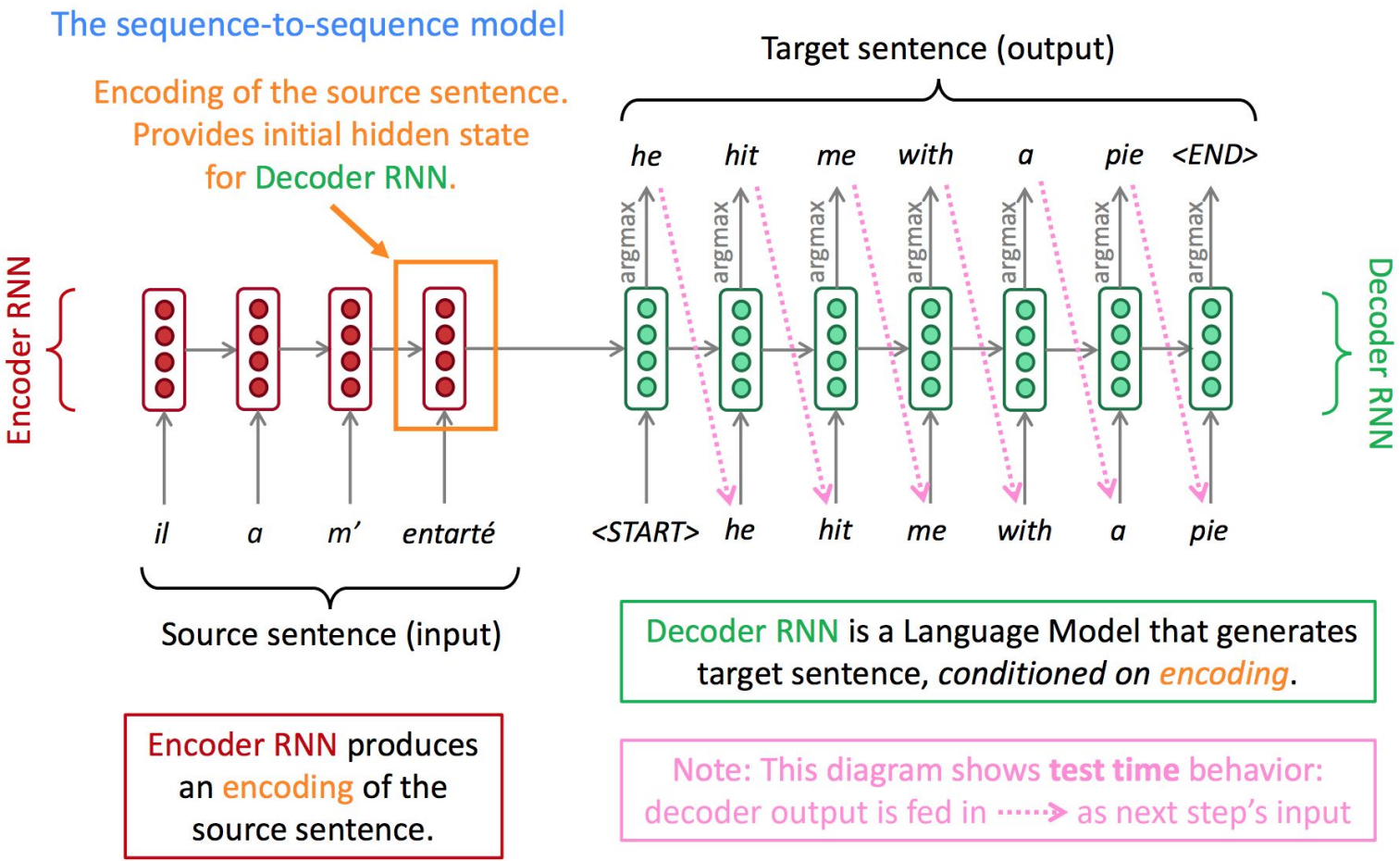
Optimization:

- Adam is great to start
  - Initial learning rate 3e-4
- Momentum is great
- Remember the learning rate decay

Regularization:

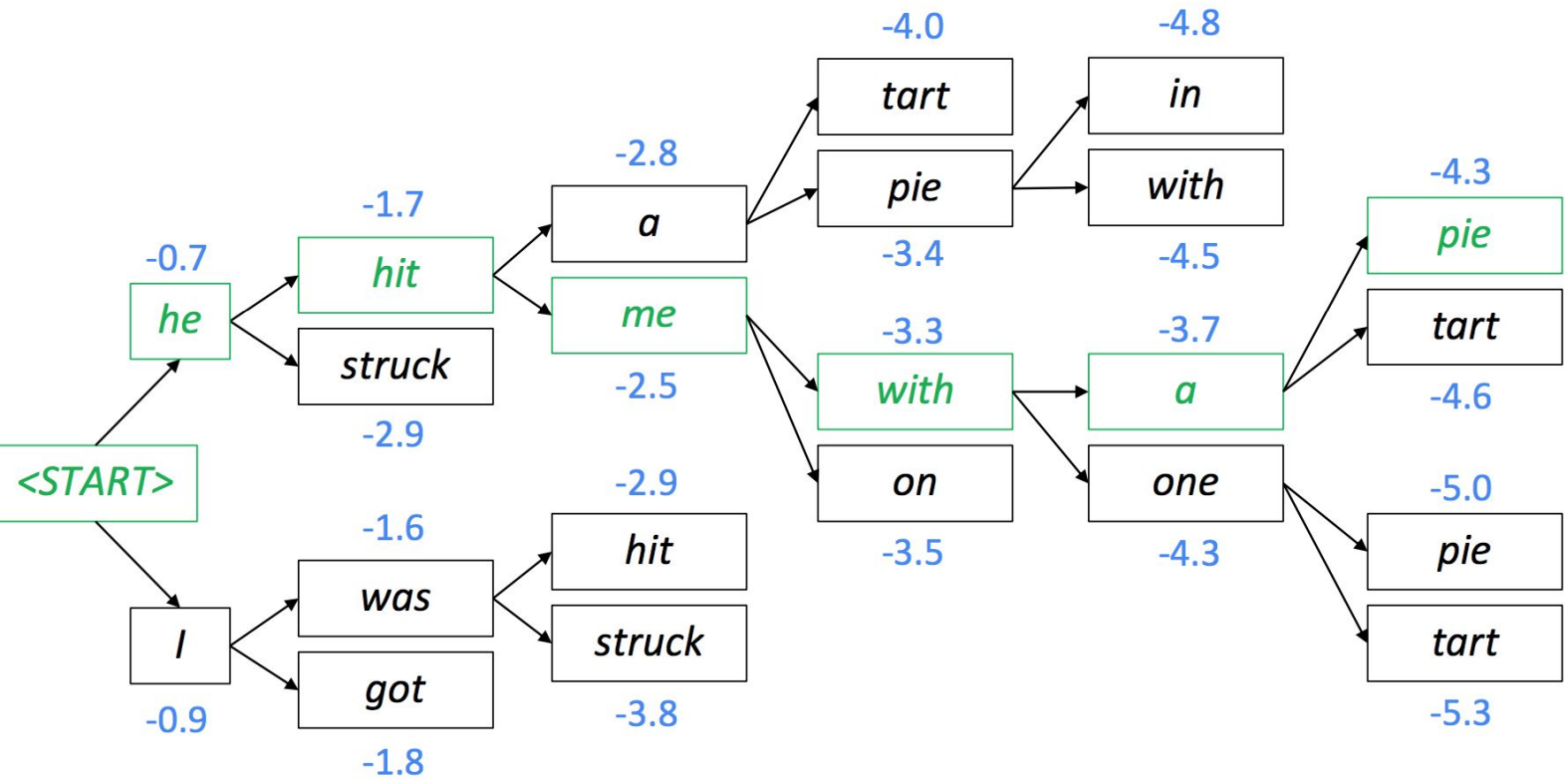
- Add some weight constraints
- Add some random noise during train and marginalize it during test
- Add some prior information in appropriate form

Seq2Seq approach



Beam Search

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$



Main highlights

Seq2Seq:

- Let’s train models to convert sequences from one domain (e.g. sentences in English) to sequences in another domain (e.g. the same sentences in French)
- Encoder-Decoder architecture

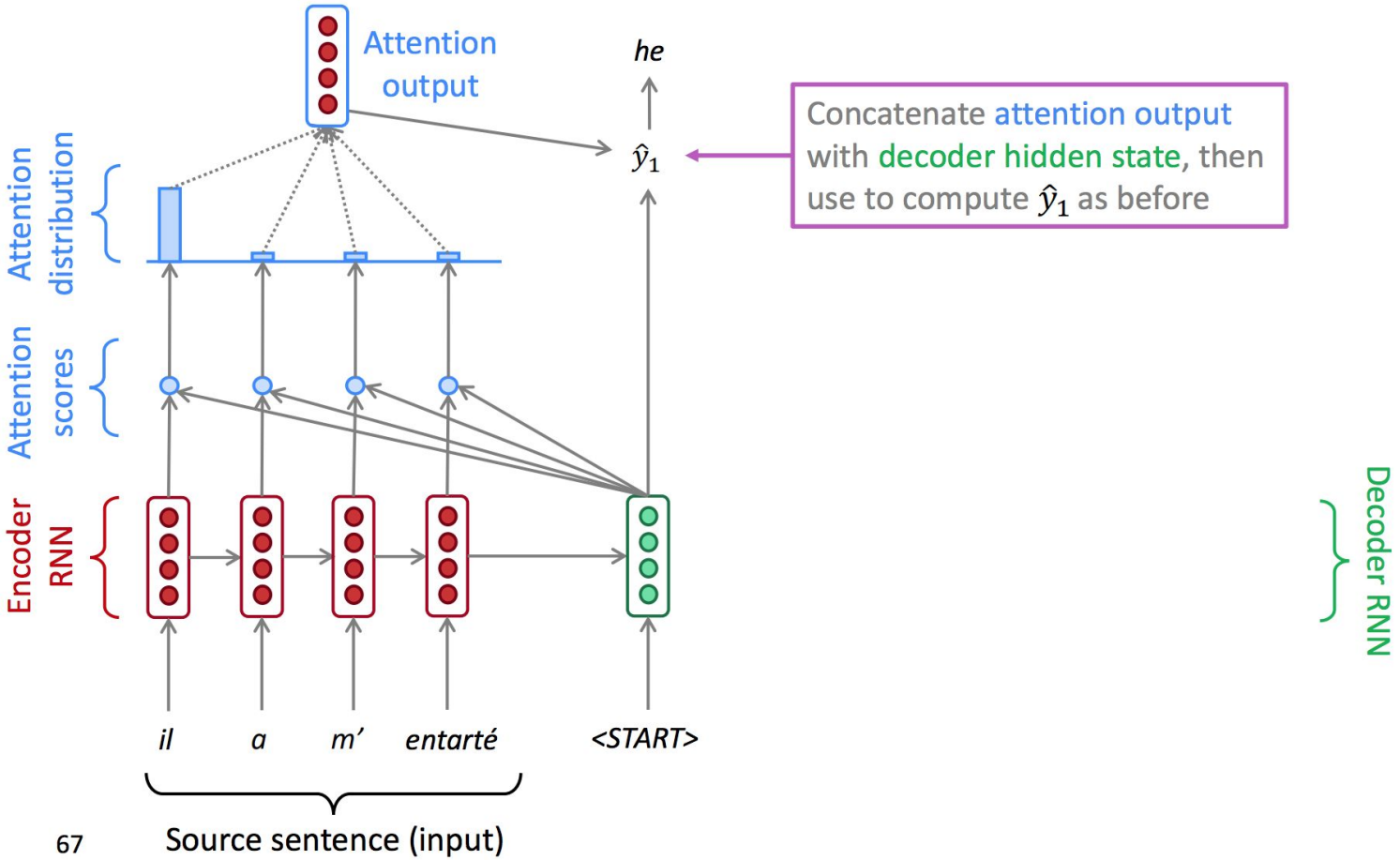
Beam Search:

- Much better than greedy decoding or exhaustive search
- May work with hypotheses of different lengths
- Use this formula to calculate score for a hypothesis (y - target sentence, x - source sentence):

$$\frac{1}{t}\sum_{i=1}^{t}\log P_{\mathrm{LM}}(y_i|y_1,\ldots,y_{i-1},x)$$

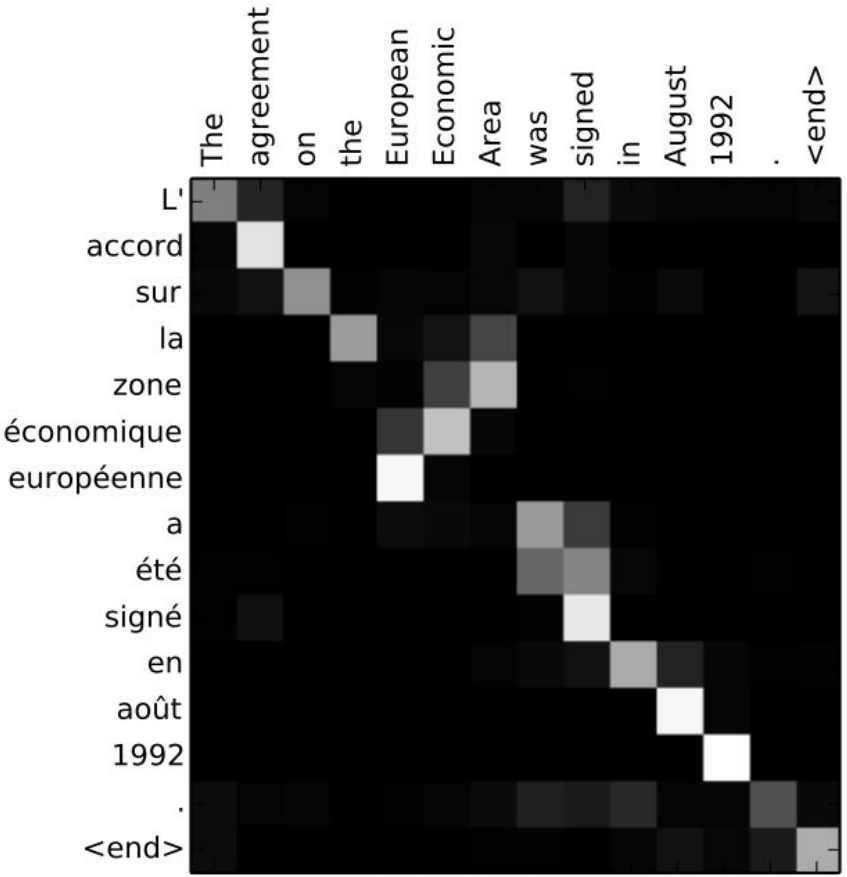
- Don’t forget to normalize the score by length of the sentence!

Attention

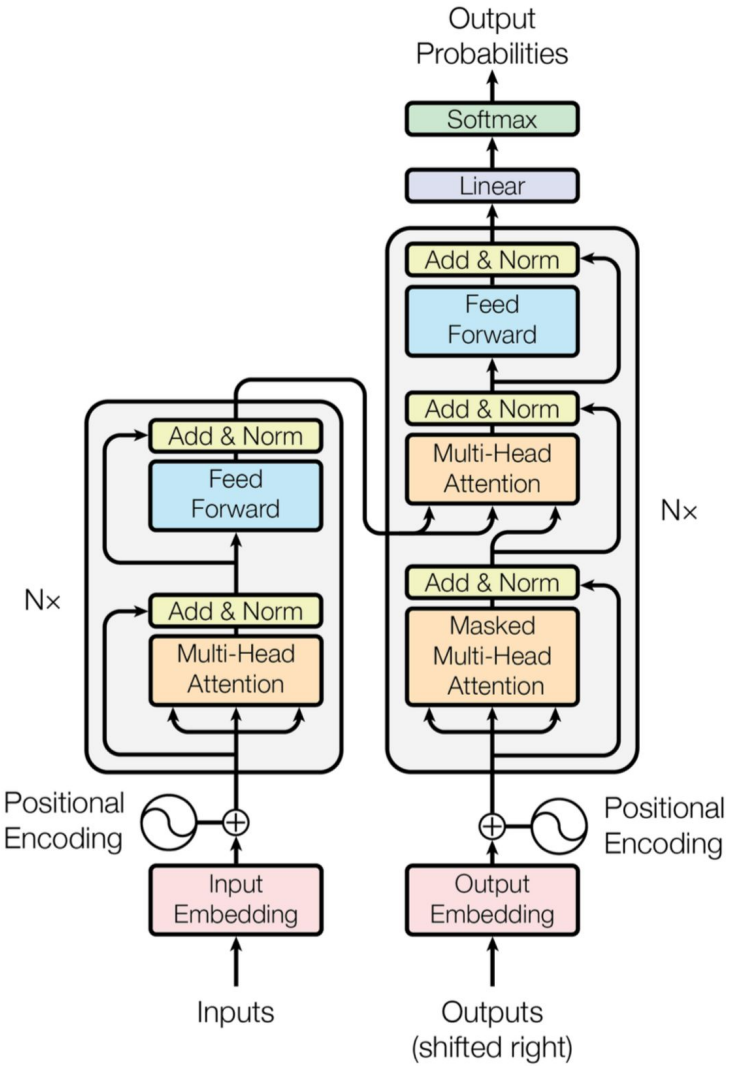

Attention output
he
$\hat{y}_1$
Concatenate attention output with decoder hidden state, then use to compute $\hat{y}_1$ as before
Attention distribution
Attention scores
Encoder RNN
Decoder RNN
il
a
m'
entarté
<START>
Source sentence (input)

Attention: free-lunch word alignments

- We may see what the decoder was focusing on
- We get word alignment for free!

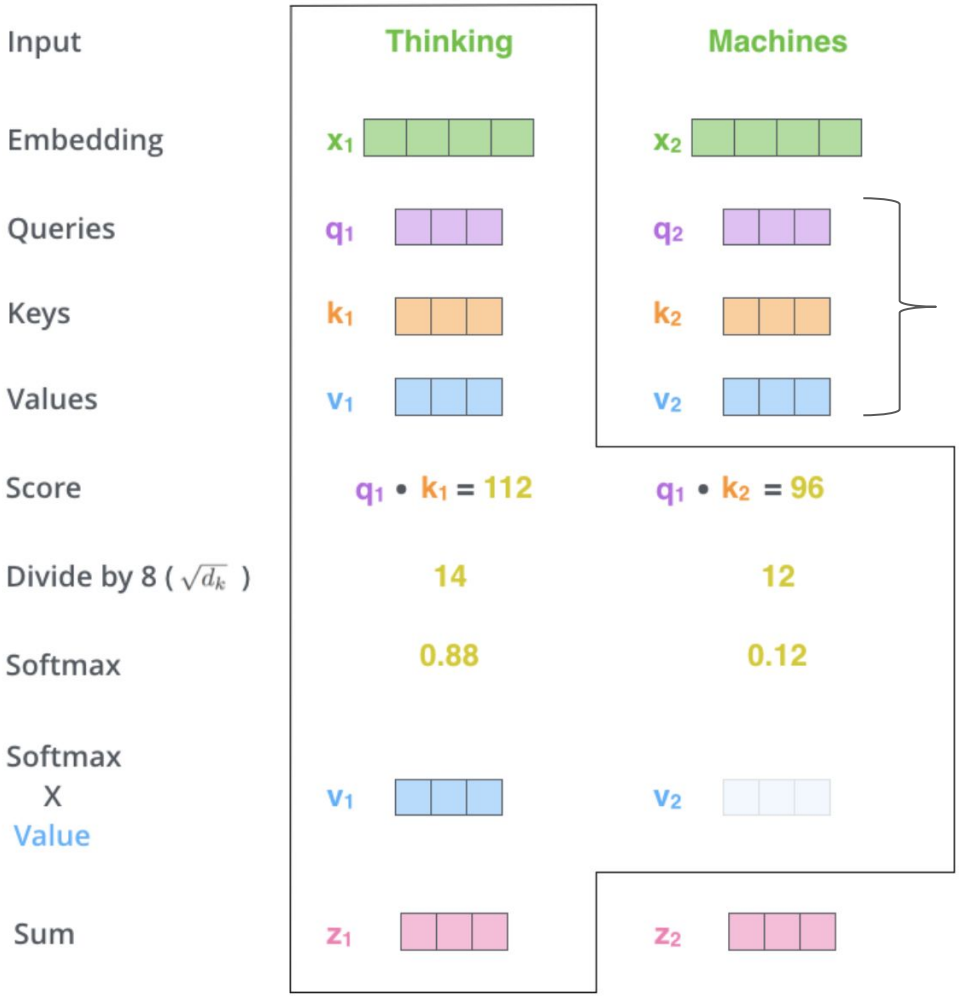


The Transformer



- Proposed in the paper “Attention is All You Need” (Ashish Vaswani et al.)
- No recurrent or convolutional neural networks -> just attention
- Uses Multi-Head **self-attention** concept

Self-Attention



STEP 1: create Query, Key, Value

STEP 2: calculate scores

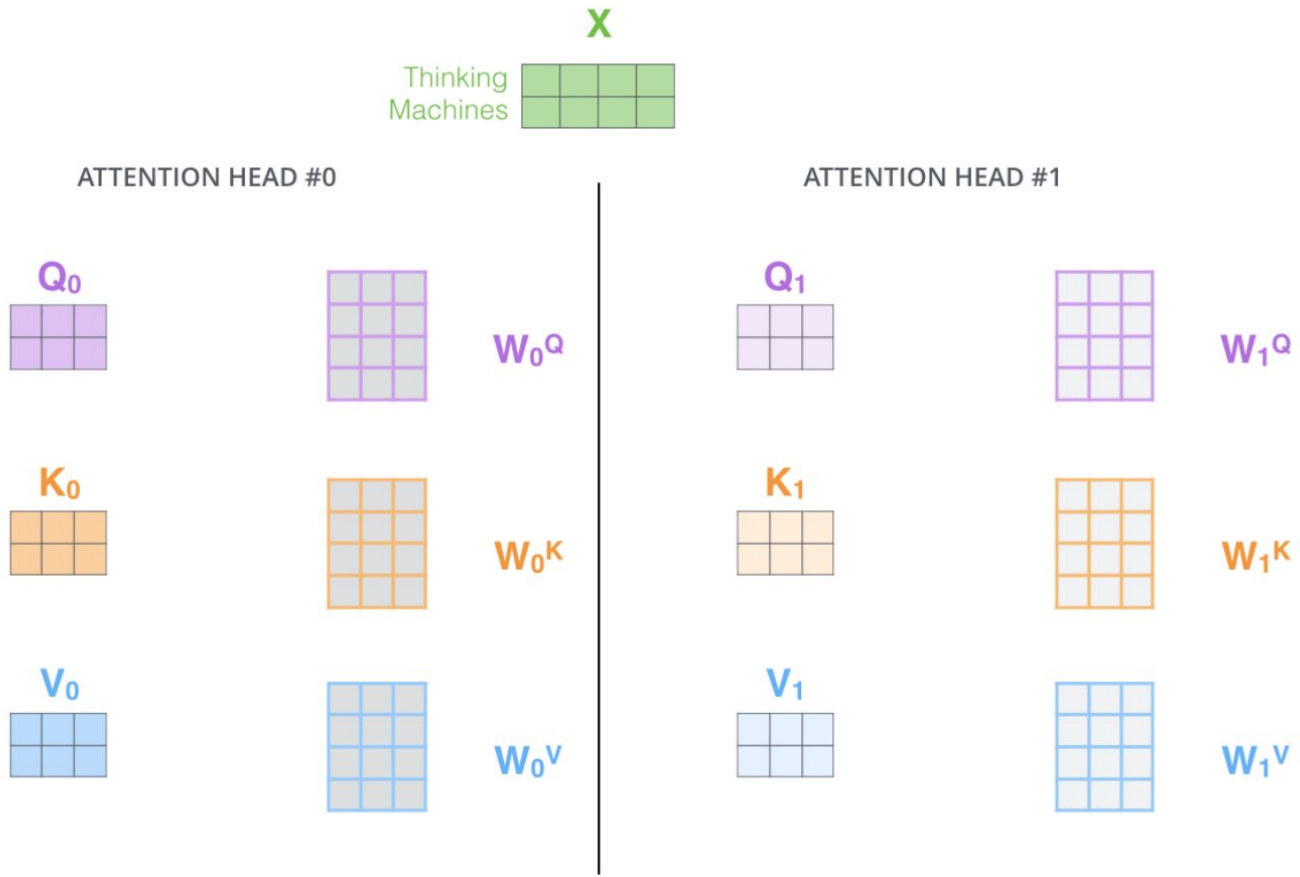
STEP 3: divide by $\sqrt{d_k}$

STEP 4: softmax

STEP 5: multiply each value vector by the softmax score

STEP 6: sum up the weighted value vectors

Multi-Head Attention



1) This is our input sentence*

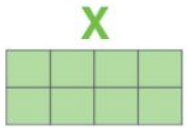
2) We embed each word*

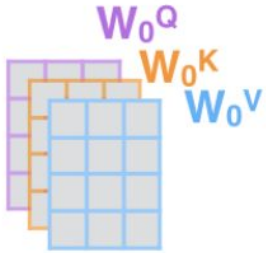
3) Split into 8 heads. We multiply X or R with weight matrices

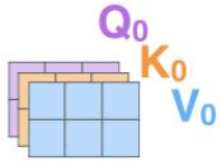
4) Calculate attention using the resulting Q/K/V matrices

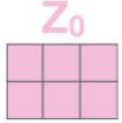
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

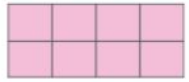
Thinking Machines

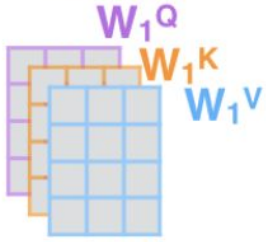


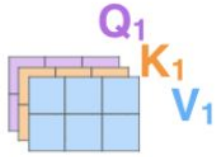


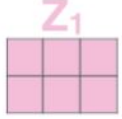




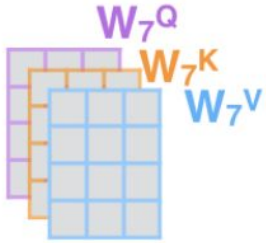


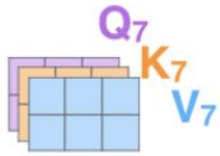


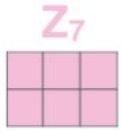




* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one







Main highlights

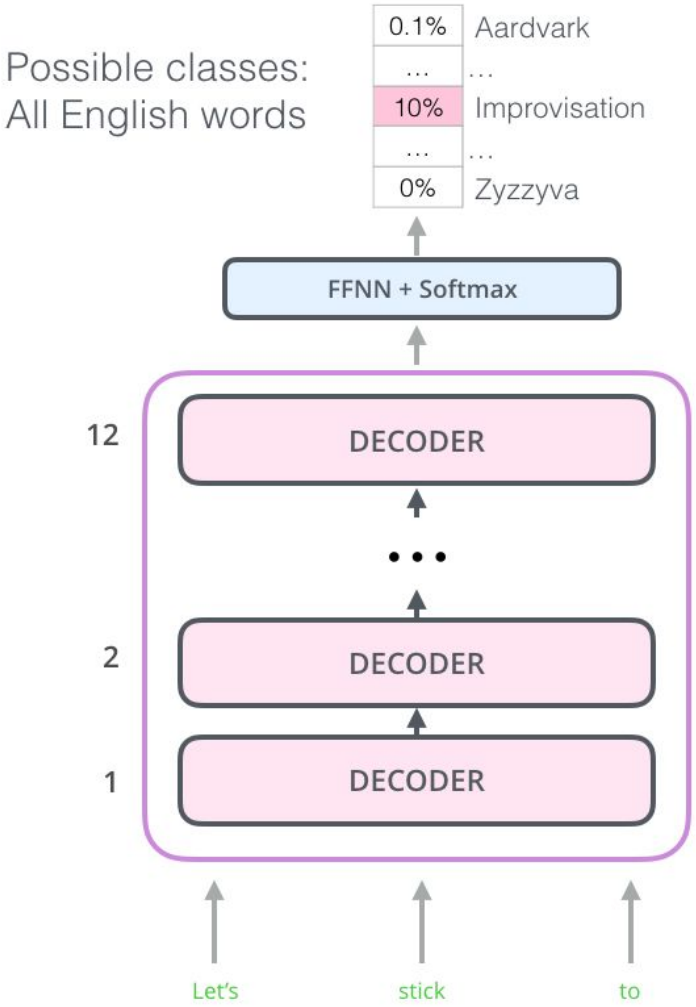
Attention:

- each time the model predicts an output word, it uses the most relevant parts of an input instead of an entire sentence
- Solves “bottleneck” problem
- Provides word alignments for free

Self-attention:

- Forget about RNNs or CNNs: Attention is all you need!
- Query, Key, Value vectors
- $\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) V$

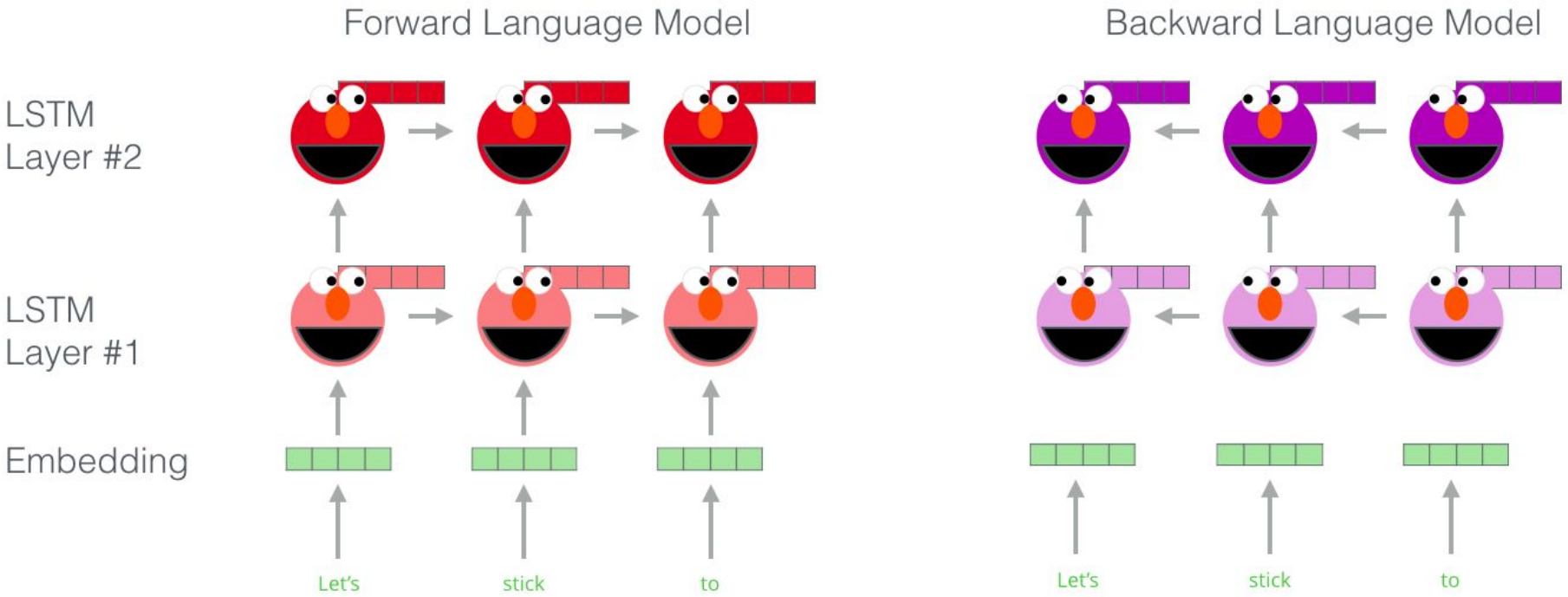
OpenAI Transformer



Differences from vanilla Transformer:

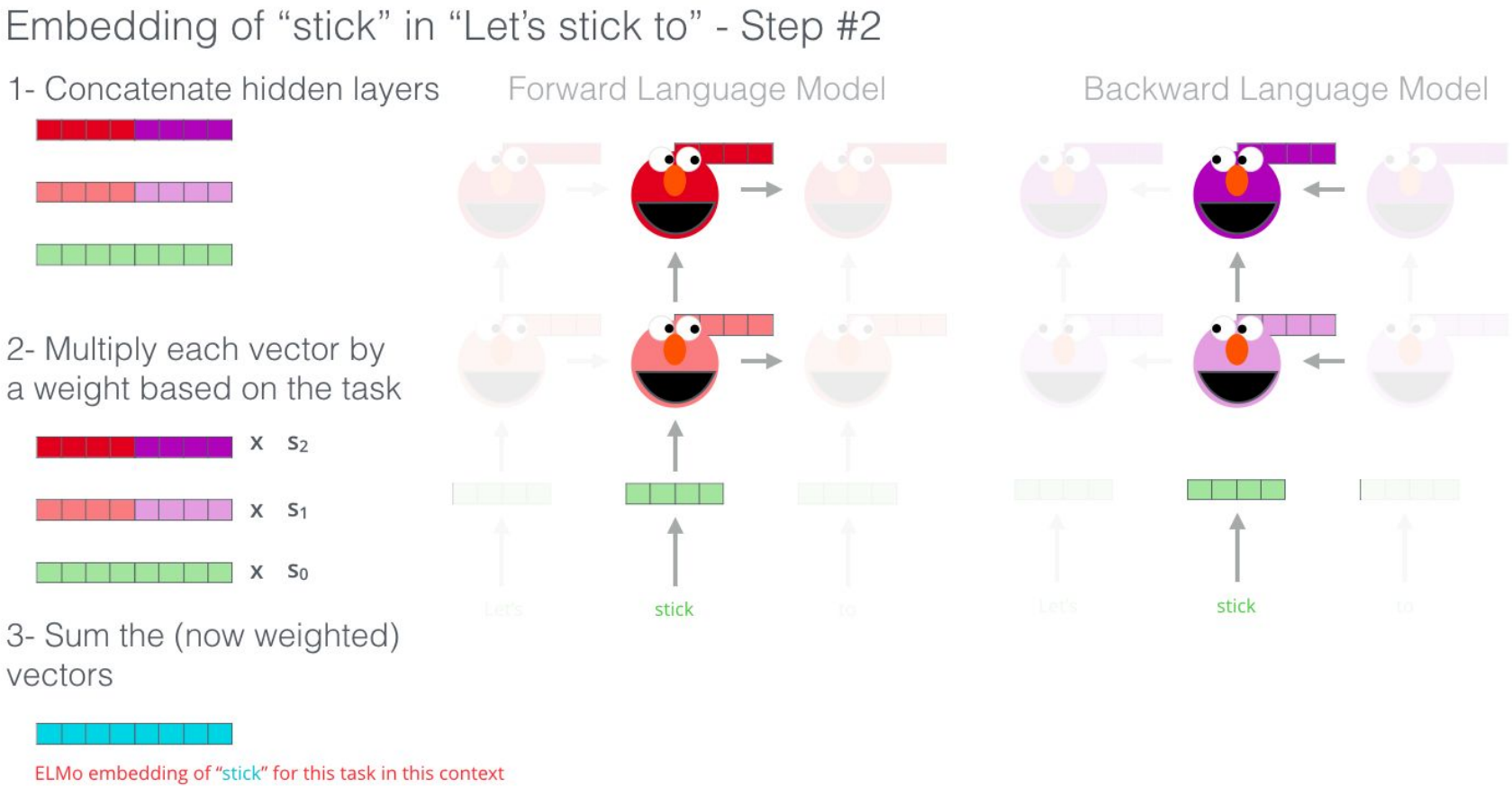
- no encoder
- decoder layers would not have the encoder-decoder attention sublayer
- Pre-train the model on predicting the next word using massive (unlabeled) datasets (language modeling)

ELMo: let's use biLMs

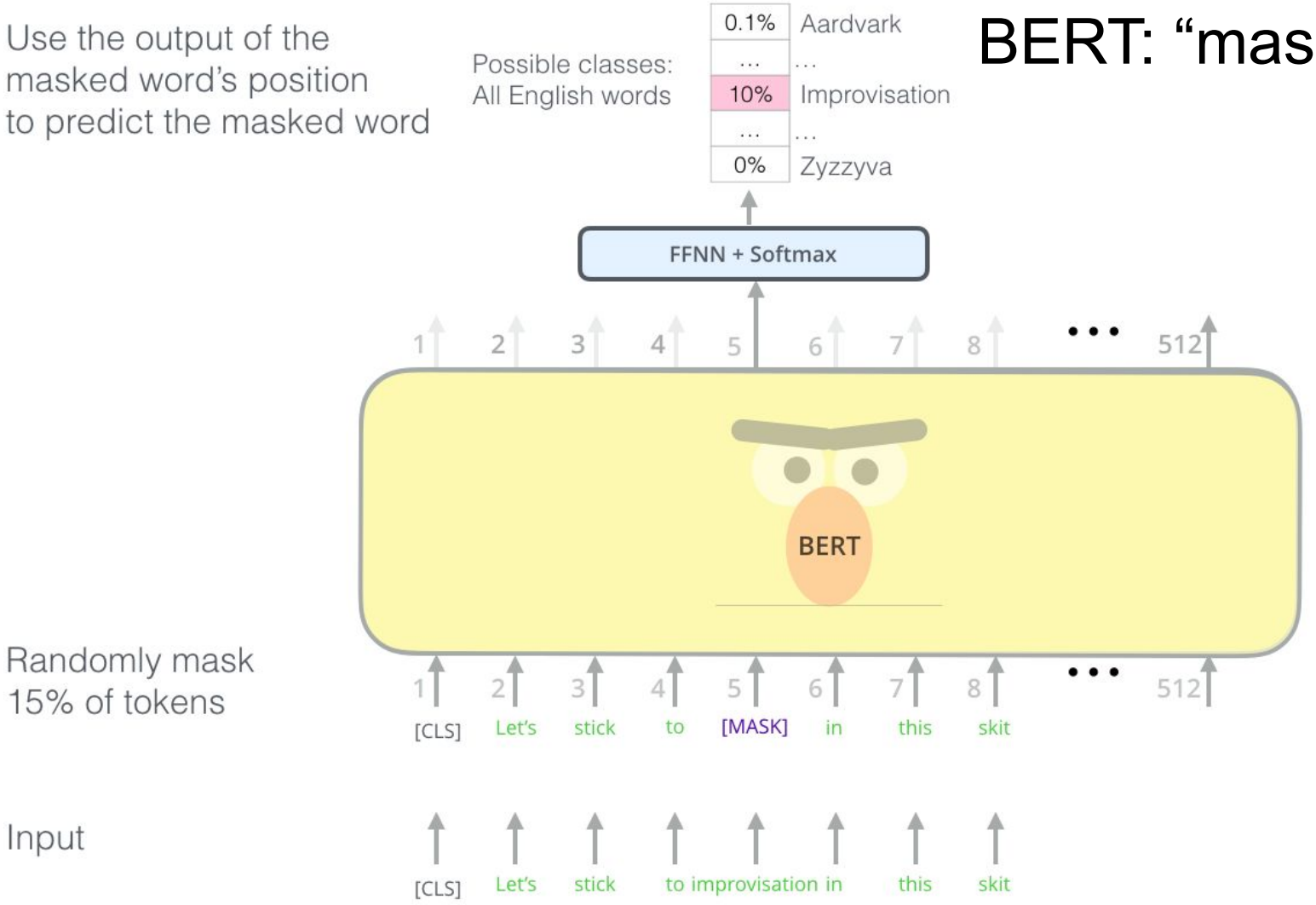


ELMo: main pipeline

ELMo represents a word as a linear combination of corresponding hidden layers:



BERT: “masked” LM



BERT: tokenization using WordPieces model

- Single model for 104 languages with a large shared vocabulary
- 119,547 tokens
- Algorithms is called WordPiece model
- Tokenization works in the following way:

Example: Unaffable -> un, ##aff, ##able

Main highlights

Transfer Learning in NLP:

- use a pre-trained Language Model to capture information about a language
- then fine-tune this model towards a goal task

Popular models:

- OpenAI Transformer
- ELMo: deep contextualized word representations
- BERT: Bidirectional Encoder Representations from Transformers
  - “Masked Language Model” approach
  - One shared dictionary for 104 languages, WordPiece tokenization

Future of NLP

The following slides copied from Stanford CS224n Lecture 20 slides

Deep Learning for NLP 5 years ago

- No Seq2Seq
- No Attention
- No large-scale QA/reading comprehension datasets
- No TensorFlow or Pytorch
- …

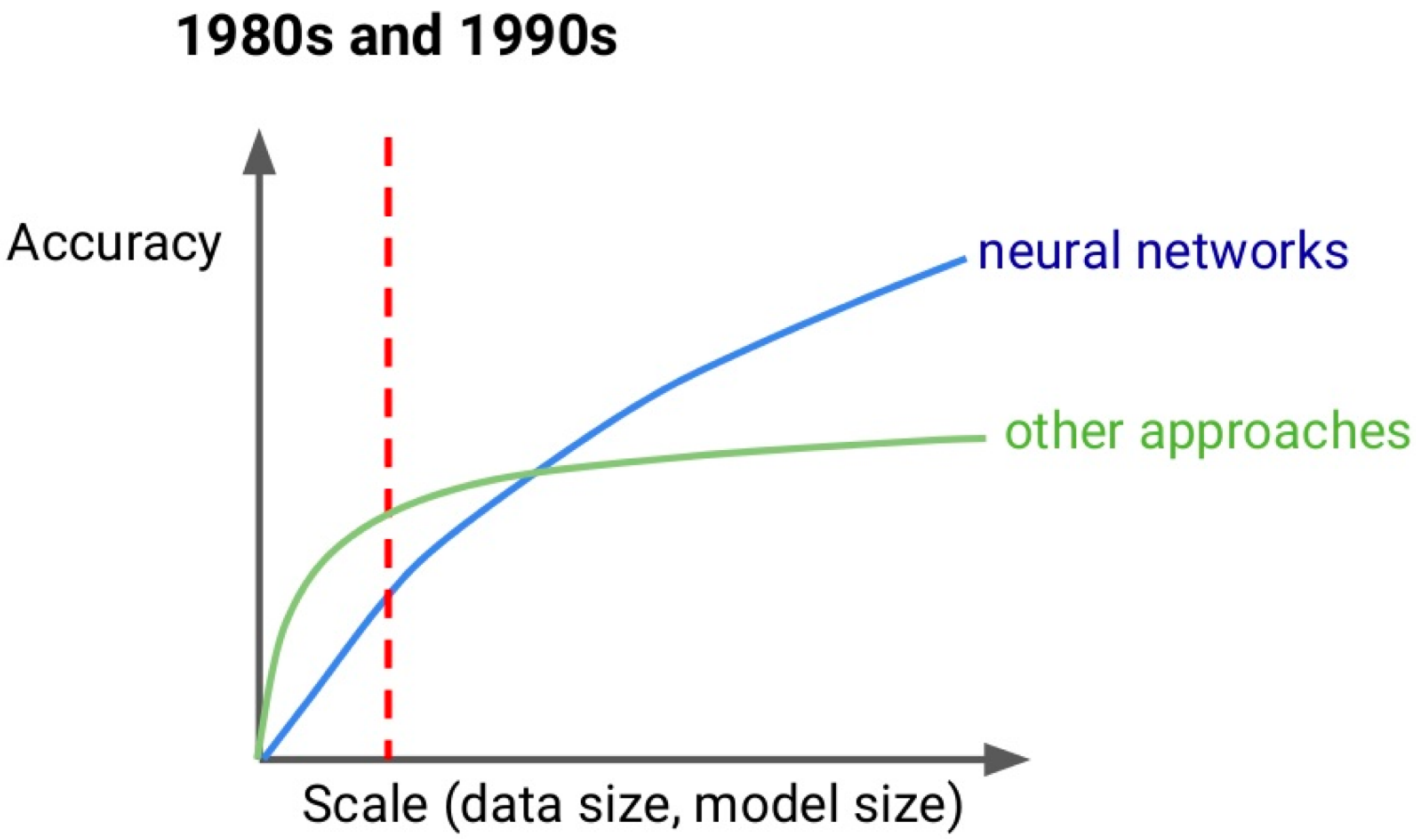
Future of Deep Learning + NLP

- **Harnessing Unlabeled Data**
  - Back-translation and unsupervised machine translation
  - Scaling up pre-training and GPT-2

- **What's next?**
  - Risks and social impact of NLP technology
  - Future directions of research

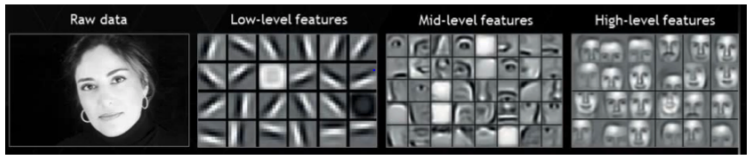
Why has deep learning been so successful recently?

Why has deep learning been so successful recently?

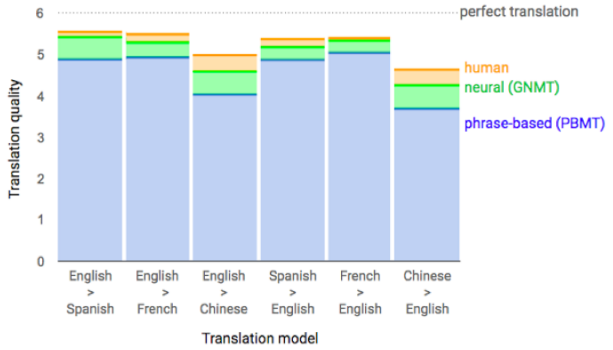


Big deep learning successes

- Image Recognition:
  Widely used by Google, Facebook, etc.



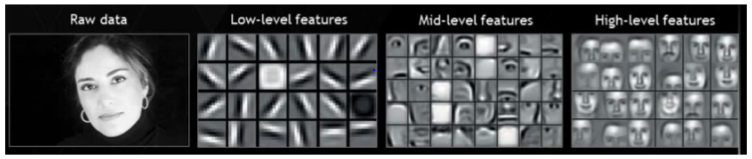
- Machine Translation:
  Google translate, etc.



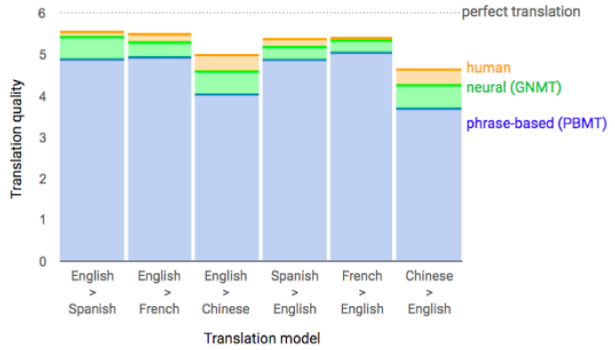
- Game Playing:
  Atari Games, AlphaGo, and more

Big deep learning successes

- Image Recognition:
  ImageNet: 14 million examples



- Machine Translation:
  WMT: Millions of sentence pairs



- Game Playing:
  10s of millions of frames for Atari AI
  10s of millions of self-play games for AlphaZero

NLP Datasets

- Even for English, most tasks have 100K or less labeled examples.

- And there is even less data available for other languages.
  - There are thousands of languages, hundreds with > 1 million native speakers
  - <10% of people speak English as their first language

- Increasingly popular solution: use **unlabeled** data.

Using Unlabeled Data for Translation

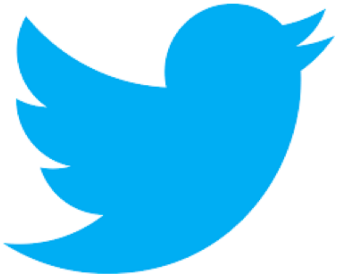
Machine Translation Data

- Acquiring translations required human expertise
  - Limits the size and domain of data

TED

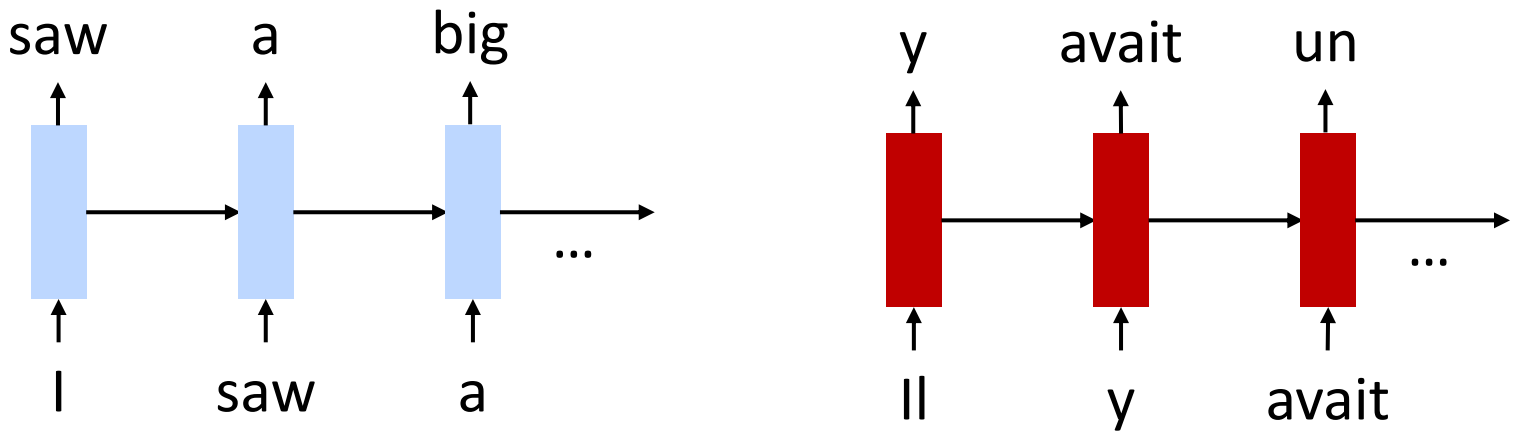
PS

- Monolingual text is easier to acquire!

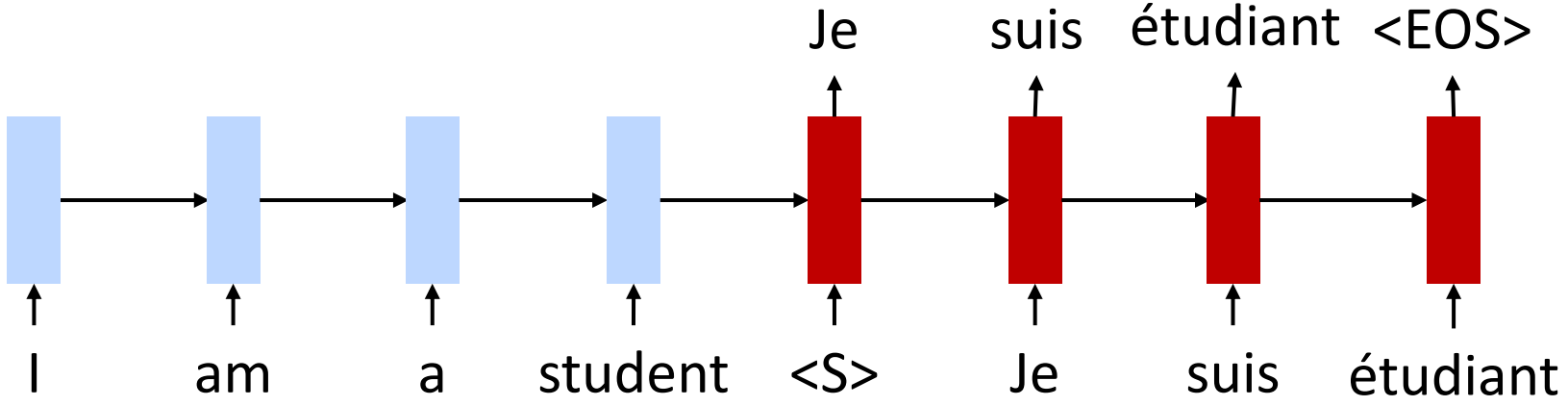


Pre-Training

1. Separately Train Encoder and Decoder as Language Models

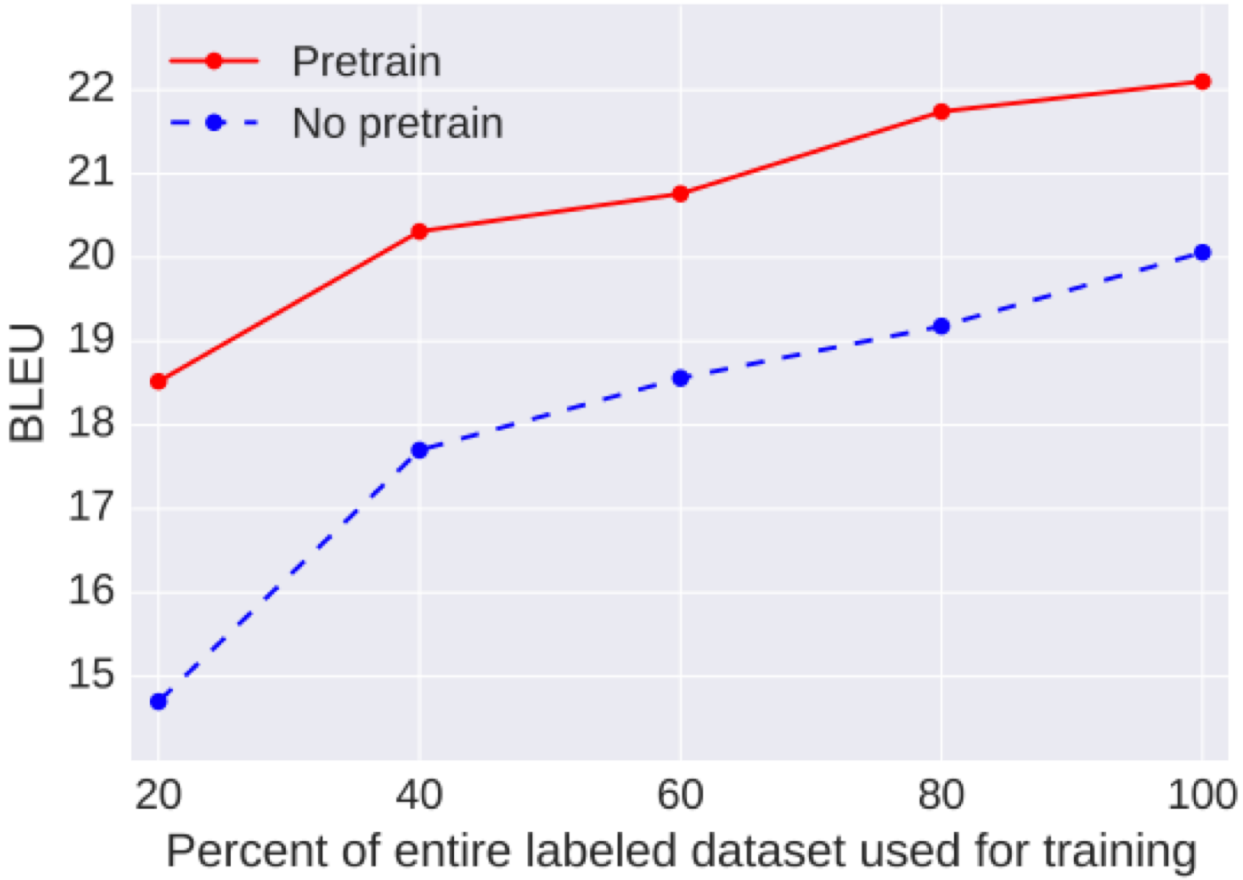


2. Then Train Jointly on Bilingual Data



Pre-Training

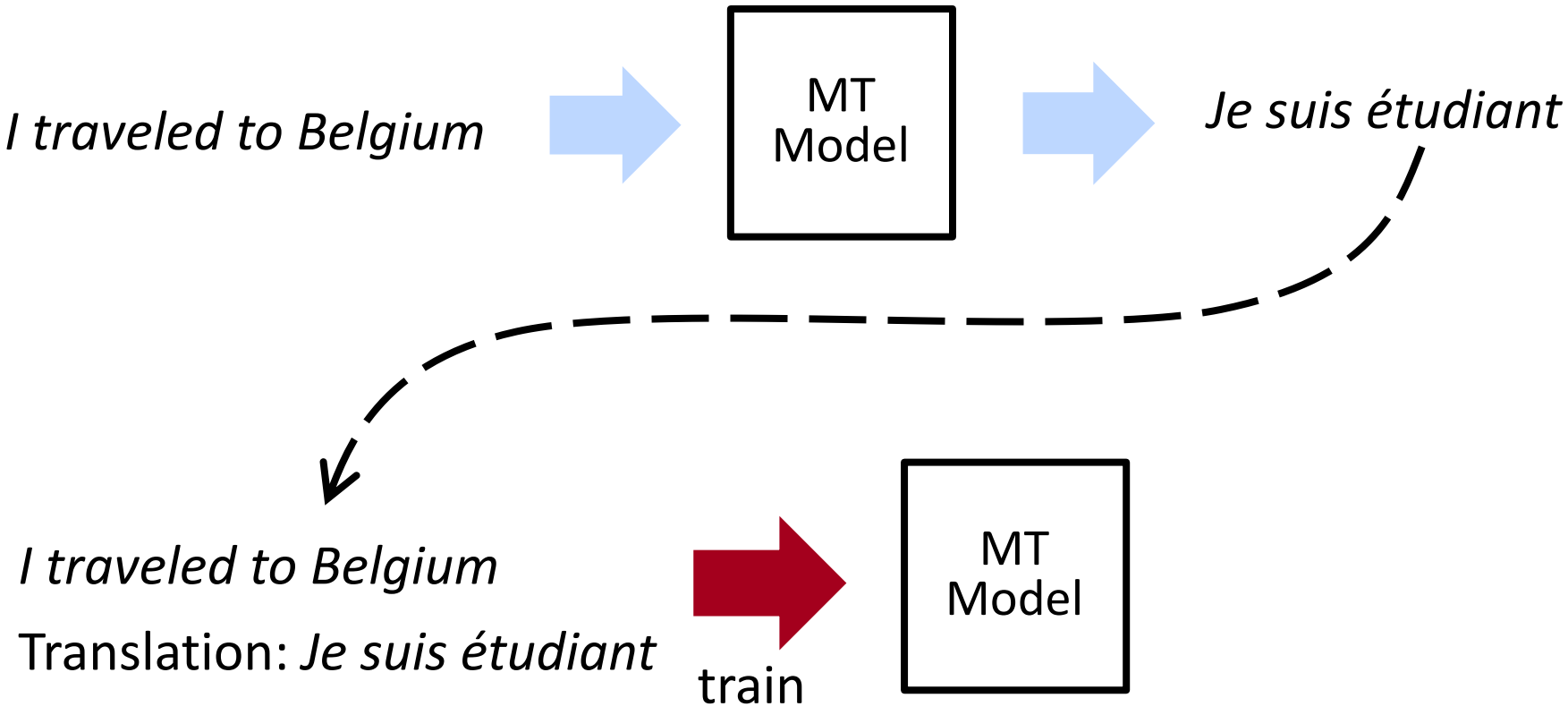
- English -> German Results: 2+ BLEU point improvement



Ramachandran et al., 2017

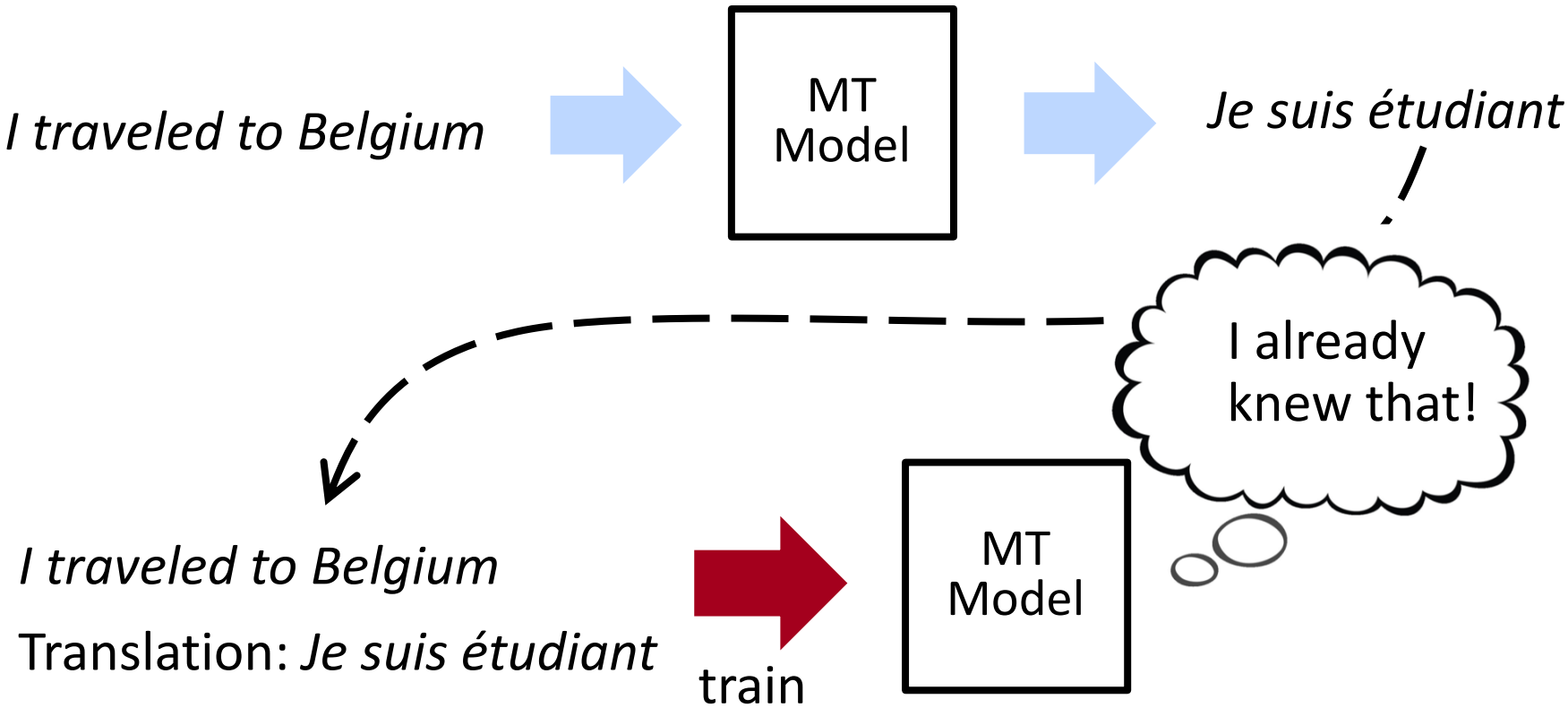
Self-Training

- Problem with pre-training: no “interaction” between the two languages during pre-training
- Self-training: label unlabeled data to get noisy training examples



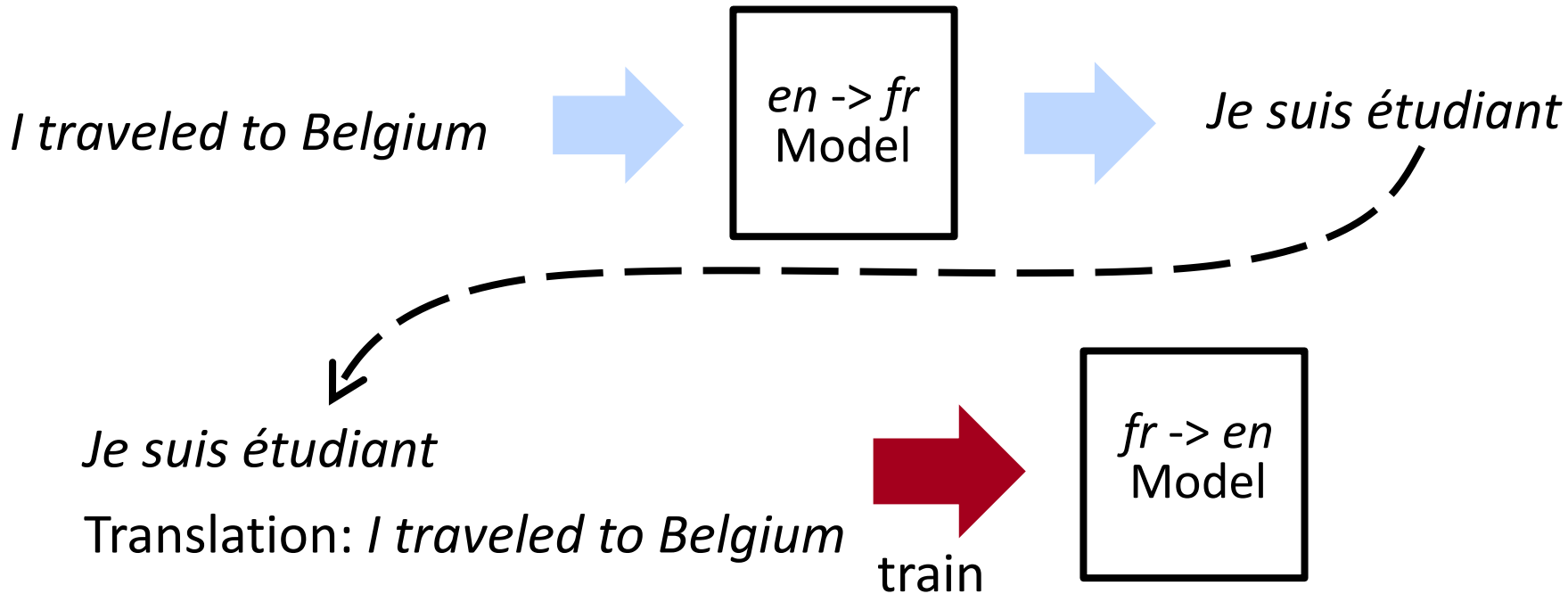
Self-Training

- Circular?



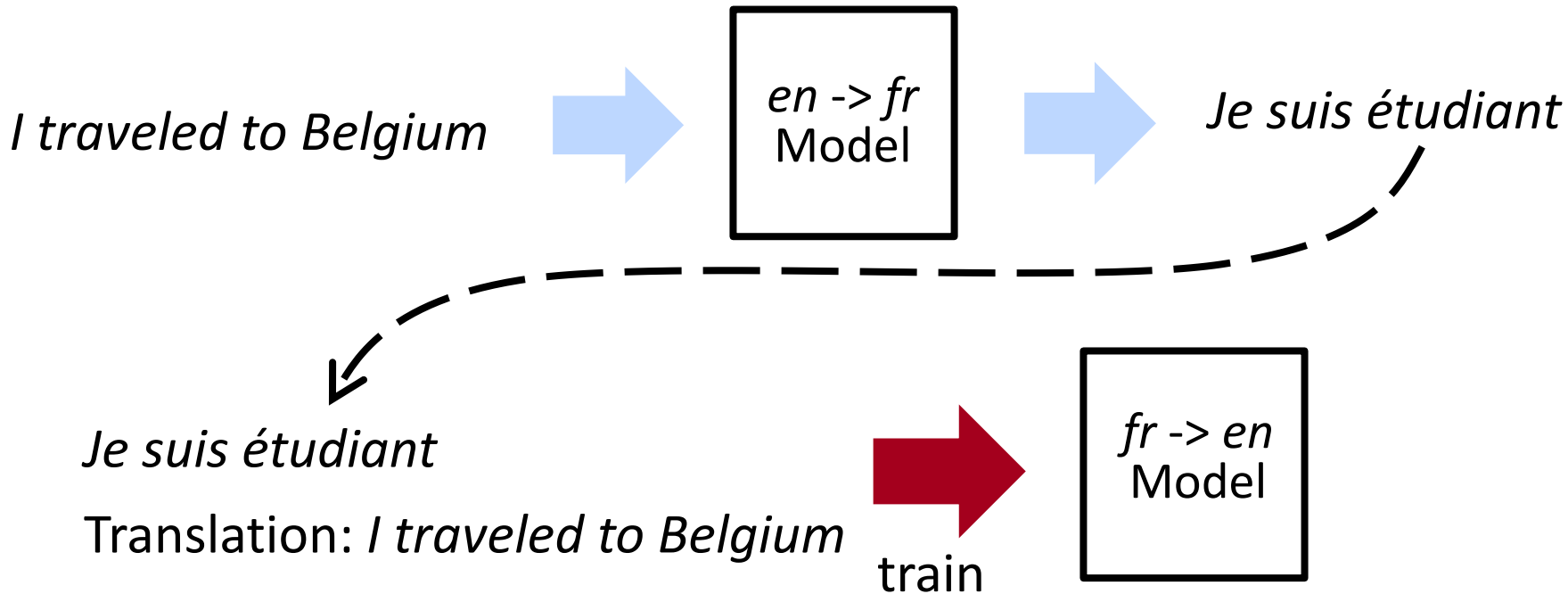
Back-Translation

- Have two machine translation models going in opposite directions (*en -> fr*) and (*fr -> en)*



Back-Translation

- Have two machine translation models going in opposite directions (*en -> fr*) and (*fr -> en)*



- No longer circular
- Models never see “bad” translations, only bad inputs

Large-Scale Back-Translation

- 4.5M English-German sentence pairs and 226M monolingual sentences

Citation	Model	BLEU
Shazeer et al., 2017	Best Pre-Transformer Result	26.0
Vaswani et al., 2017	Transformer	28.4
Shaw et al, 2018	Transformer + Improved Positional Embeddings	29.1
Edunov et al., 2018	**Transformer + Back-Translation**	**35.0**

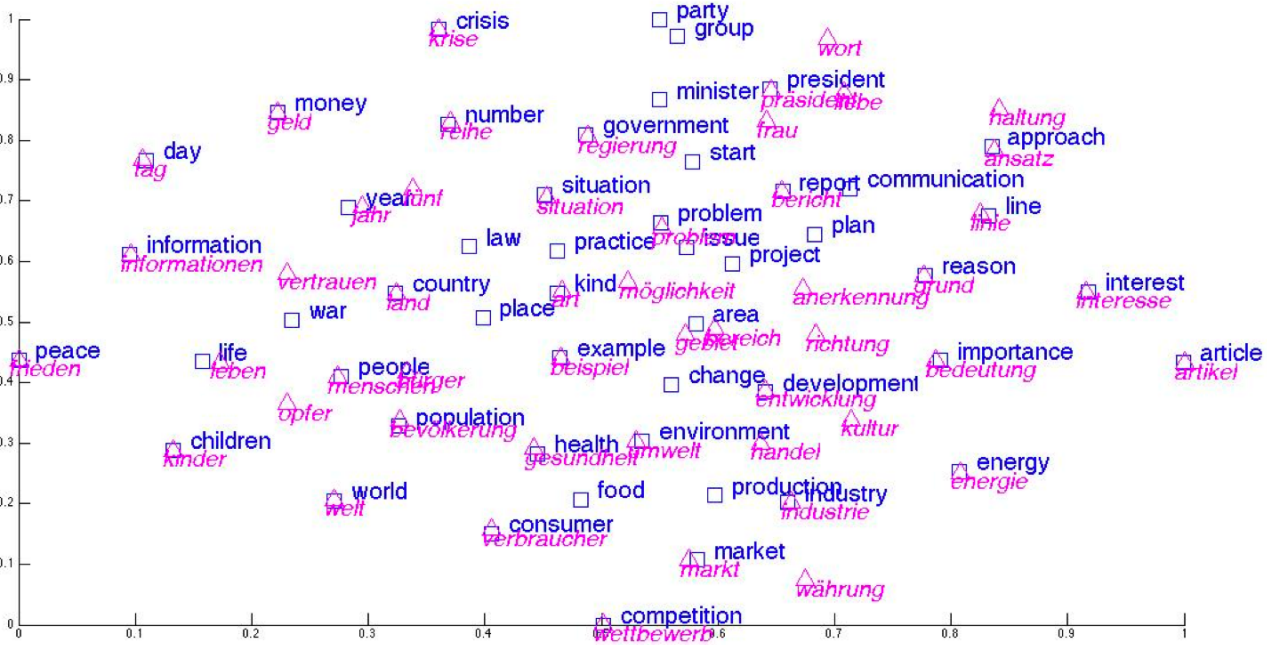
What if there is no Bilingual Data?

What if there is no Bilingual Data?

Unsupervised Word Translation

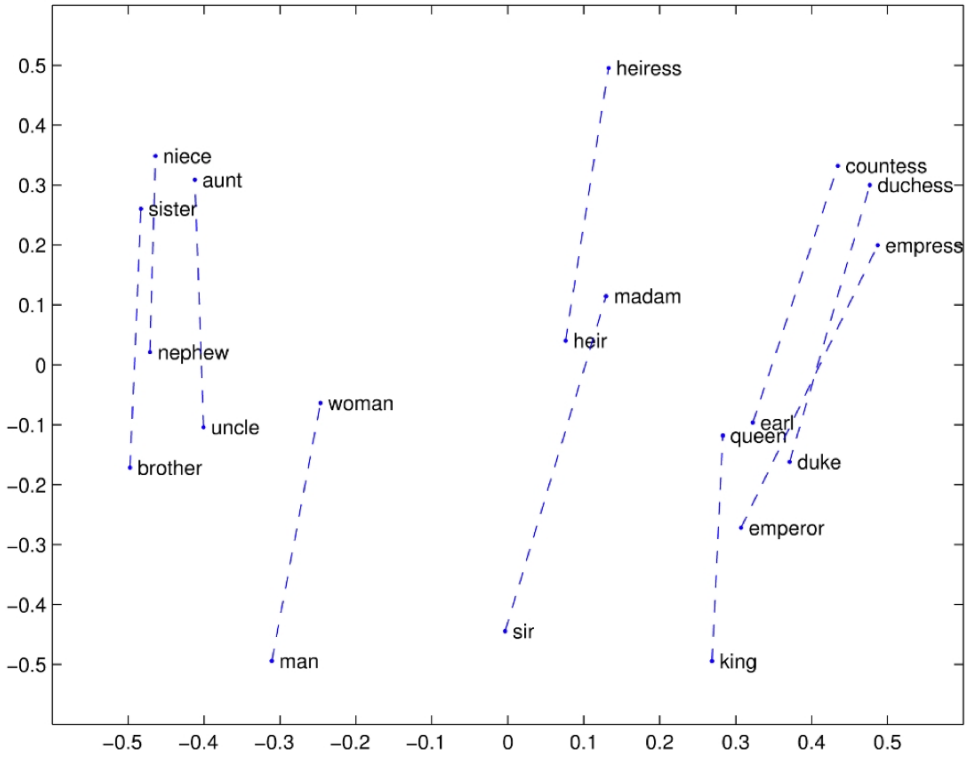
Unsupervised Word Translation

- *Cross-lingual word embeddings*
  - Shared embedding space for both languages
  - Keep the normal nice properties of word embeddings
  - But also want words close to their translations
- Want to learn from monolingual corpora



Unsupervised Word Translation

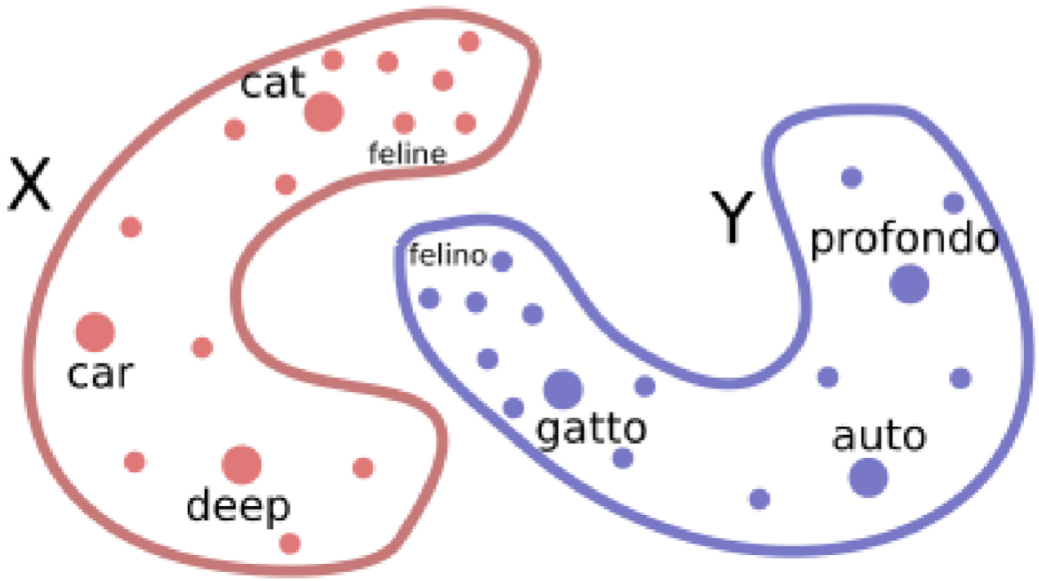
- Word embeddings have a lot of structure
- Assumption: that structure should be similar across languages



Unsupervised Word Translation

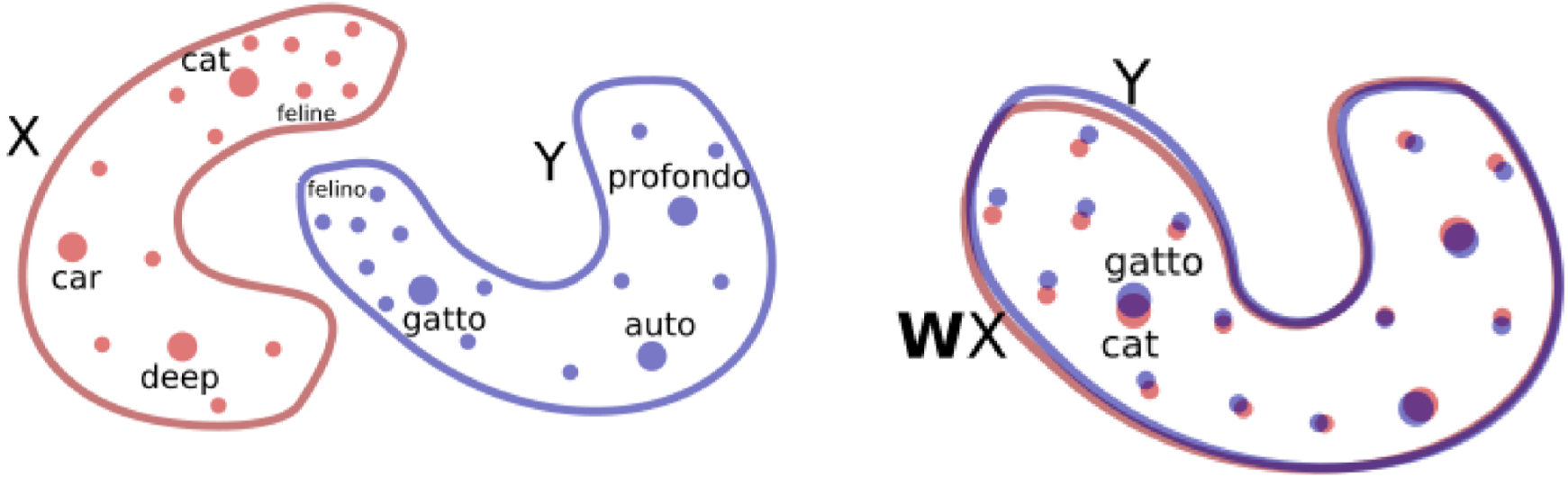
- Word embeddings have a lot of structure

- Assumption: that structure should be similar across languages



Unsupervised Word Translation

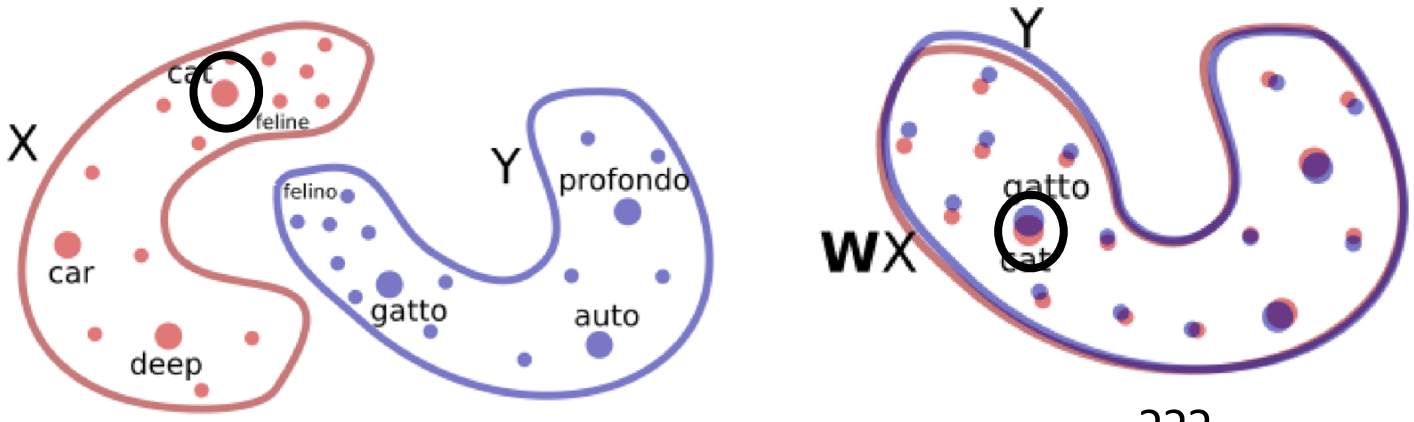
- First run word2vec on monolingual corpora, getting words embeddings *X* and *Y*
- Learn an (orthogonal) matrix *W* such that *WX* ~ *Y*



Unsupervised Word Translation

- Learn *W* with *adversarial training.*
- Discriminator**:** predict if an embedding is from *Y* or it is a transformed embedding *Wx* originally from *X.*
- Train *W* so the Discriminator gets “confused”

Discriminator predicts: is the circled point red or blue?

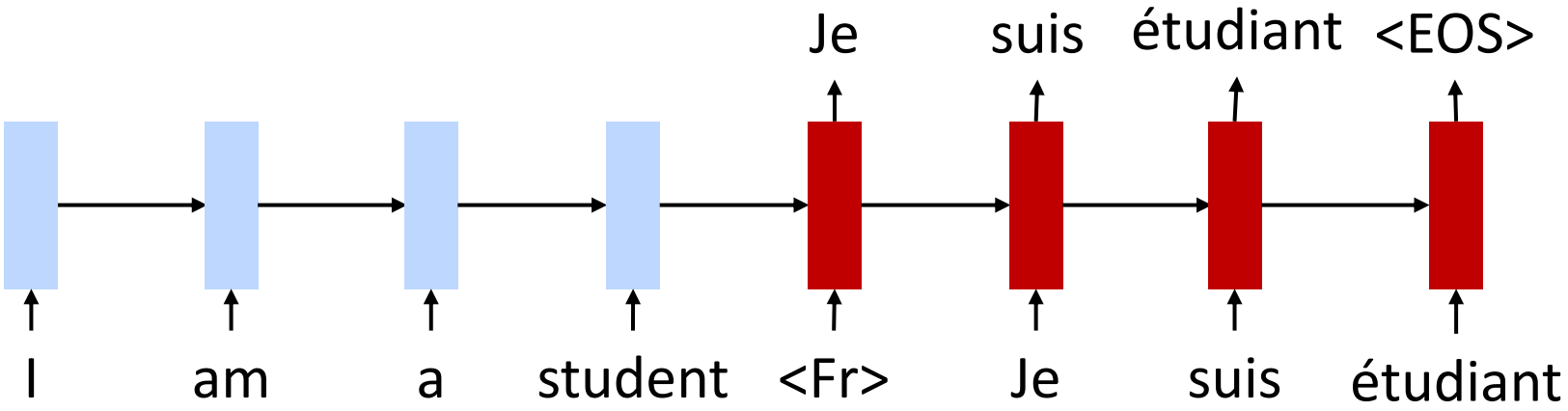


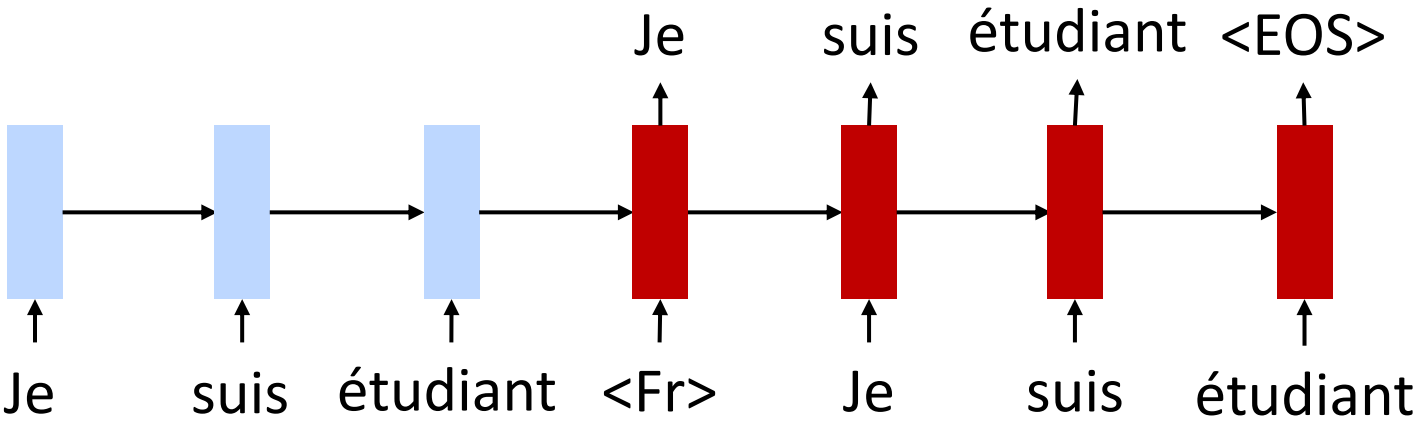
- Other tricks can be used to further improve performance, see [Word Translation without Parallel Data]

Unsupervised Machine Translation

Unsupervised Machine Translation

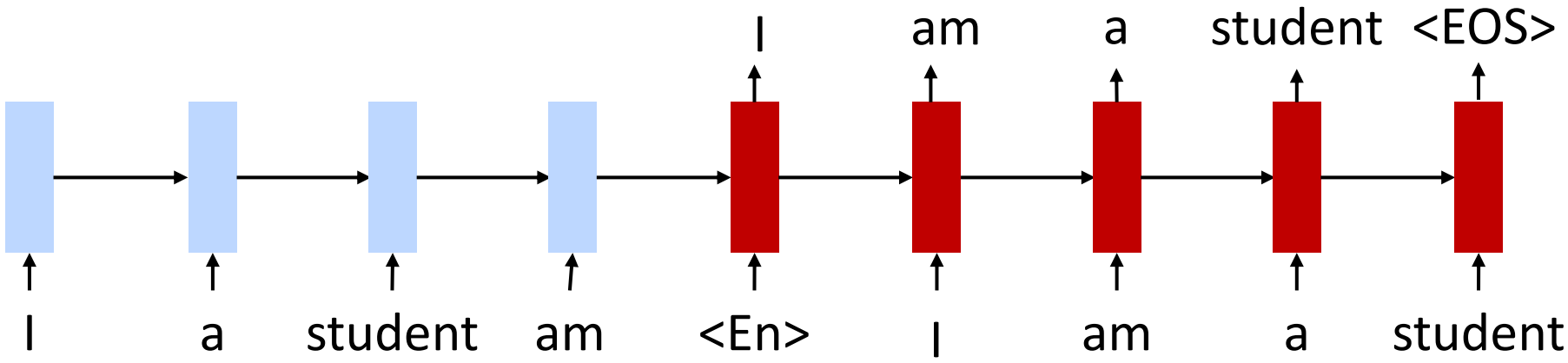
- Model: **same** encoder-decoder used for both languages
  - Initialize with cross-lingual word embeddings





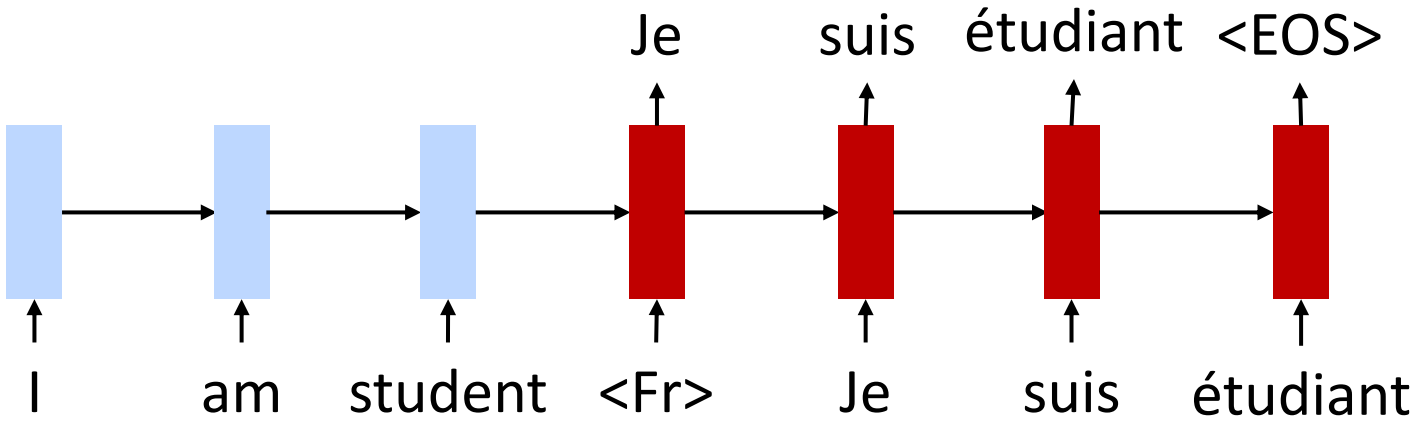
Unsupervised Neural Machine Translation

- Training objective 1: de-noising autoencoder



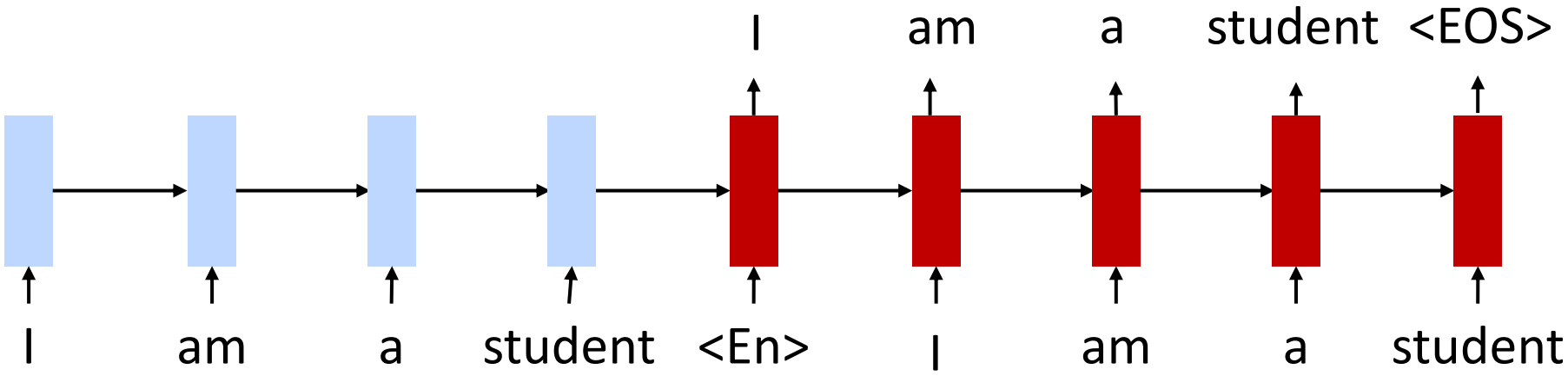
Unsupervised Neural Machine Translation

- Training objective 2: back translation
  - First translate *fr* -> *en*
  - Then use as a “supervised” example to train *en* -> *fr*



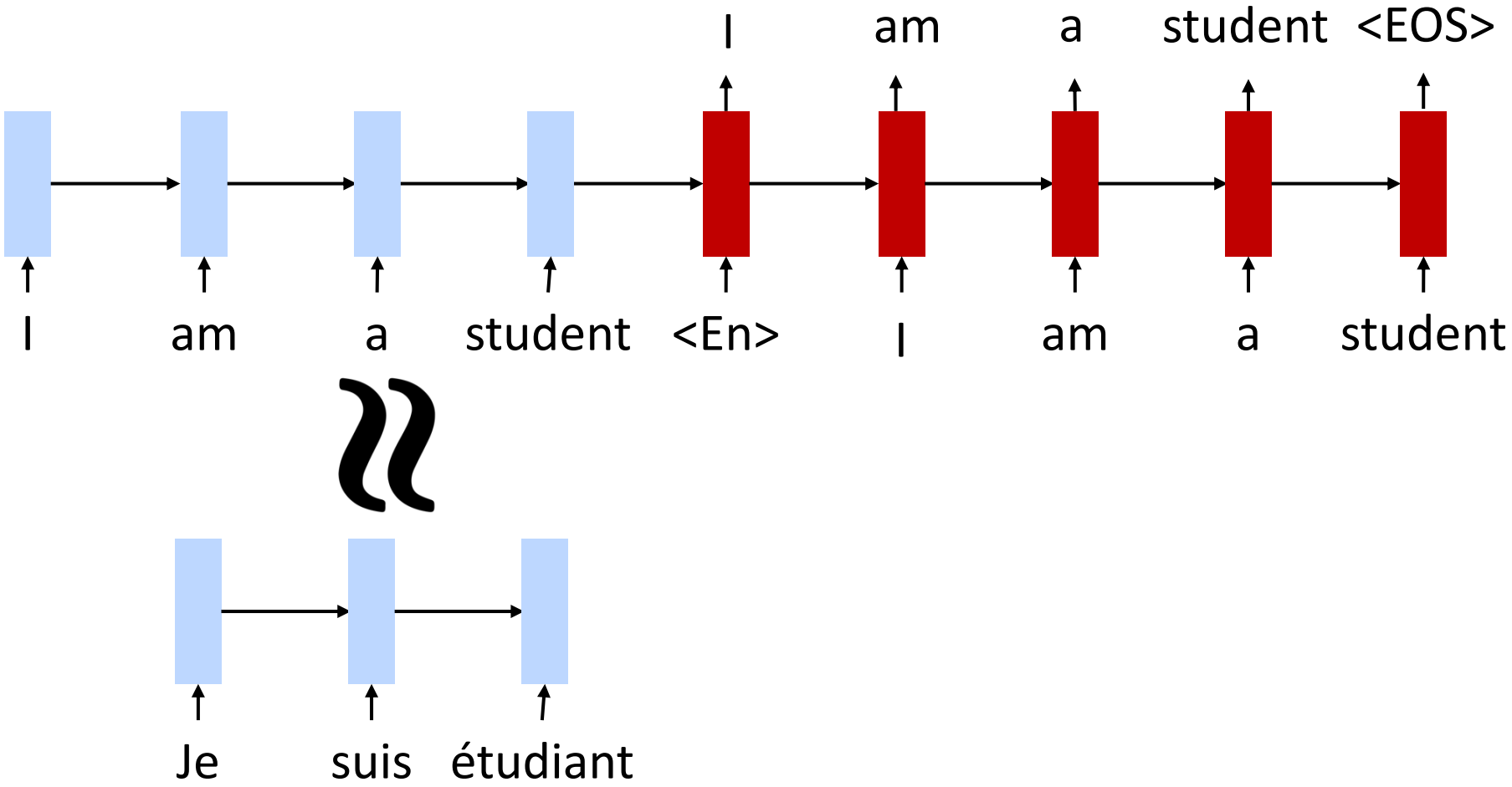
Why Does This Work?

- Cross lingual embeddings and shared encoder gives the model a starting point



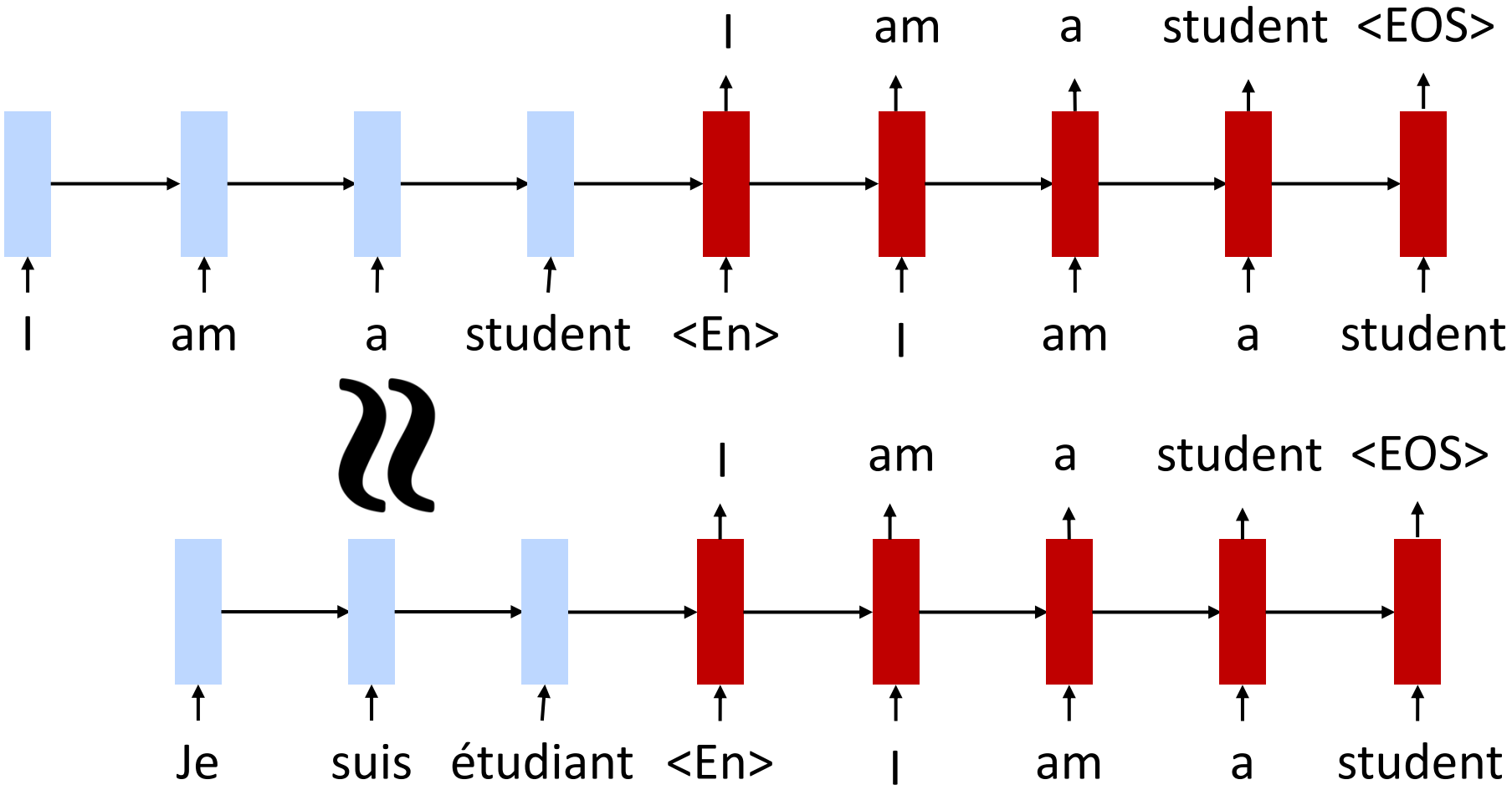
Why Does This Work?

- Cross lingual embeddings and shared encoder gives the model a starting point



Why Does This Work?

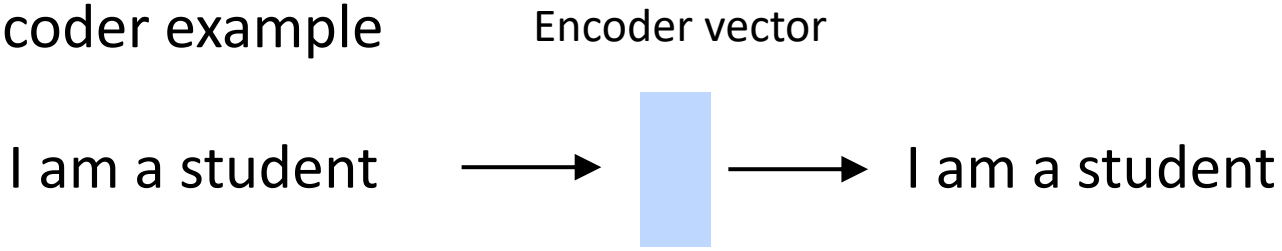
- Cross lingual embeddings and shared encoder gives the model a starting point



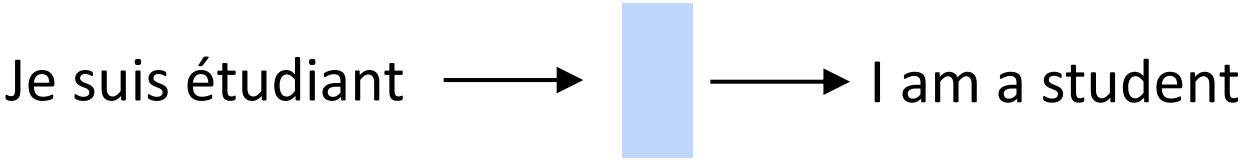
Why Does This Work?

- Objectives encourage language-agnostic representation

Auto-encoder example

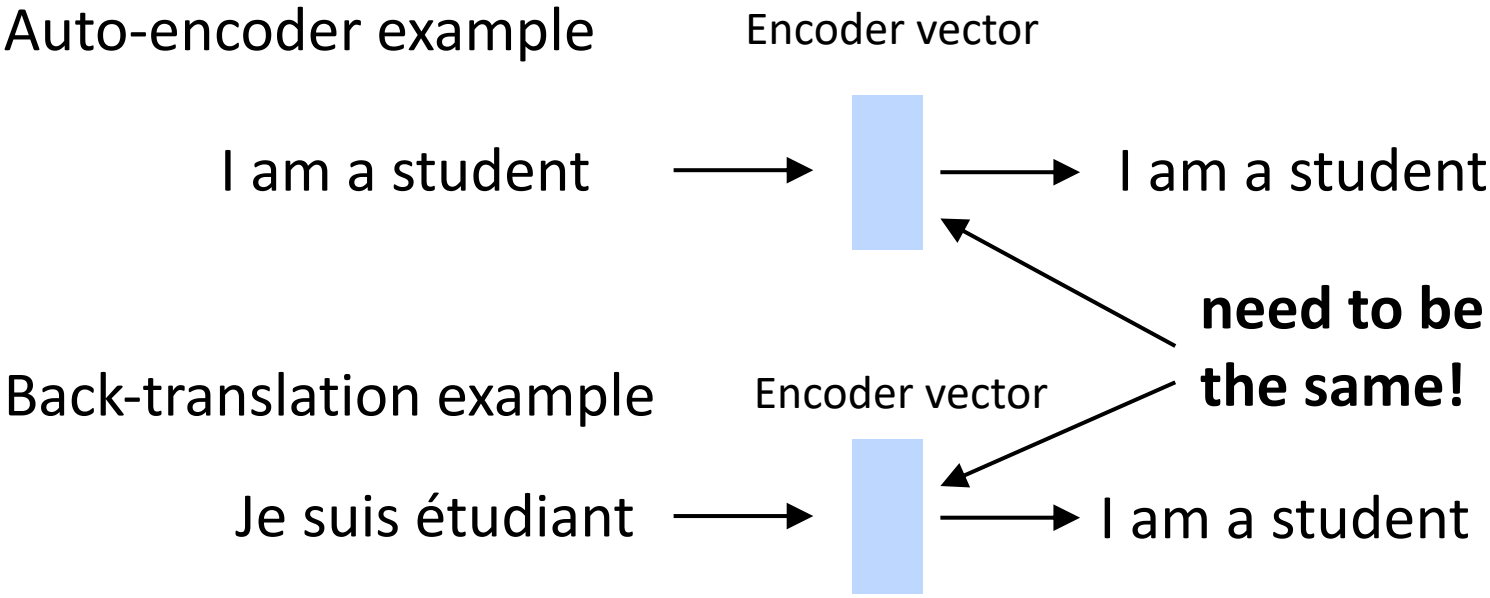


Back-translation example



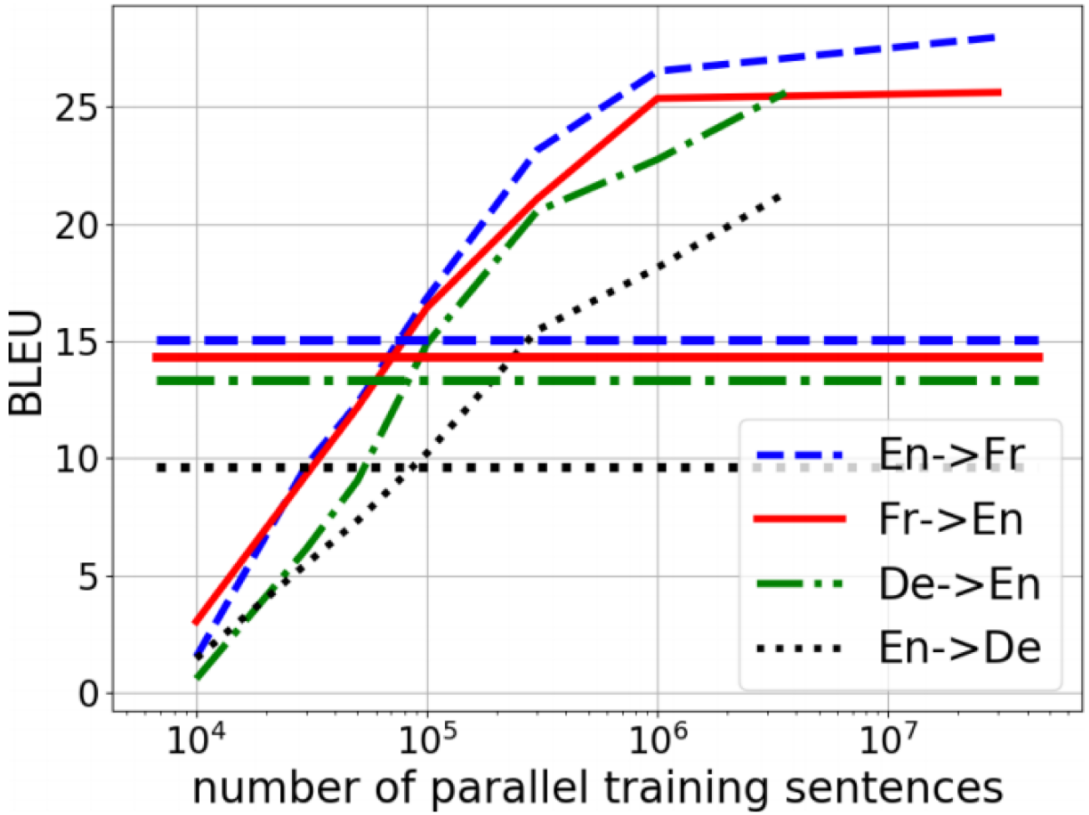
Why Does This Work?

- Objectives encourage language-agnostic representation



Unsupervised Machine Translation

- Horizontal lines are unsupervised models, the rest are supervised



Lample et al., 2018

Huge Models and GPT-2

Training Huge Models

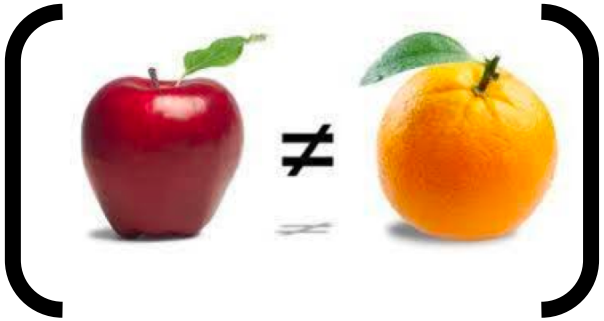
Model	# Parameters
Medium-sized LSTM	10M
ELMo	90M
GPT	110M
BERT-Large	320M
GPT-2	1.5B

Training Huge Models

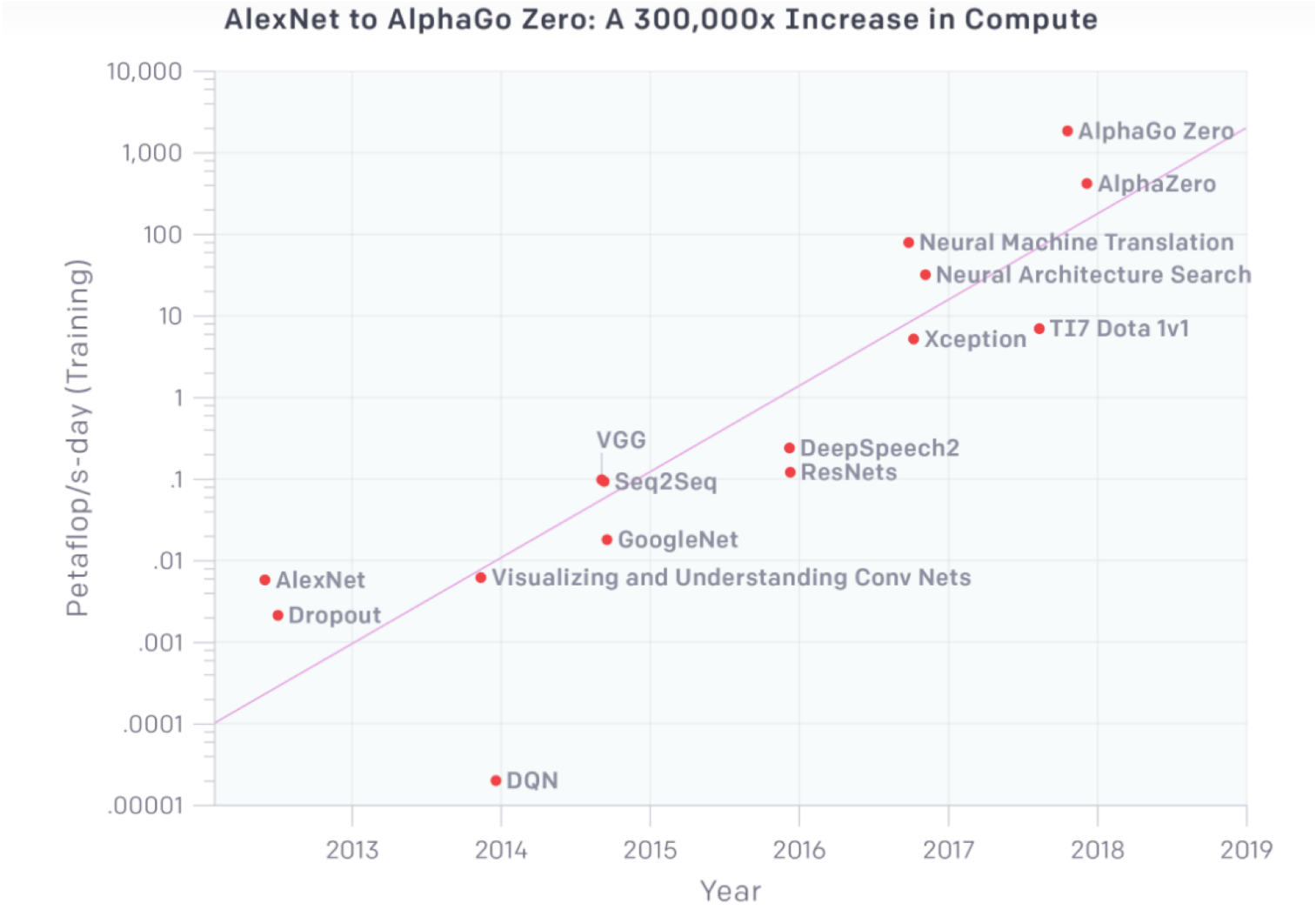
Model	# Parameters
Medium-sized LSTM	10M
ELMo	90M
GPT	110M
BERT-Large	320M
GPT-2	1.5B
Honey Bee Brain	~1B synapses

Training Huge Models

Model	# Parameters
Medium-sized LSTM	10M
ELMo	90M
GPT	110M
BERT-Large	320M
GPT-2	1.5B
Honey Bee Brain	~1B synapses



This is a General Trend in ML



Huge Models in Computer Vision

LARGE SCALE GAN TRAINING FOR HIGH FIDELITY NATURAL IMAGE SYNTHESIS

Andrew Brock[*†] Heriot-Watt University ajb5@hw.ac.uk

Jeff Donahue[†] DeepMind jeffdonahue@google.com

Karen Simonyan[†] DeepMind simonyan@google.com

- 150M parameters




See also: thispersondoesnotexist.com

Huge Models in Computer Vision

GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism

Yanping Huang
Google Brain
huangyp@google.com

Youlong Cheng
Google Brain
ylc@google.com

Dehao Chen
Google Brain
dehao@google.com

HyoukJoong Lee
Google Brain
hyouklee@google.com

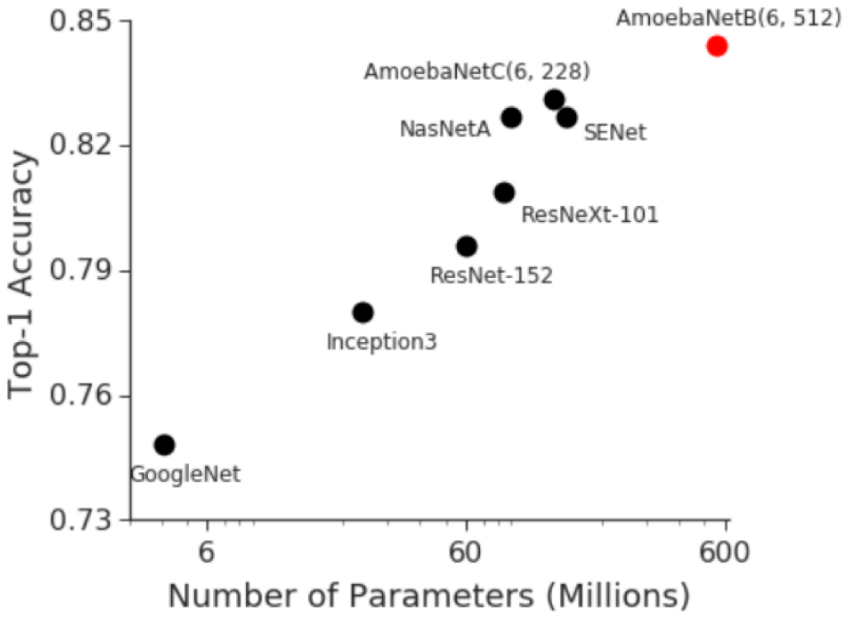
Jiquan Ngiam
Google Brain
jngiam@google.com

Quoc V. Le
Google Brain
qvl@google.com

Zhifeng Chen
Google Brain
zhifengc@google.com

- 550M parameters

ImageNet Results

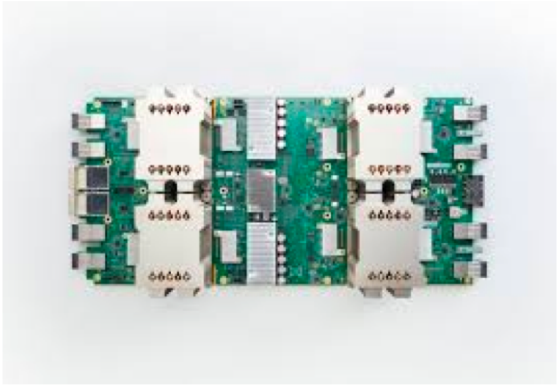


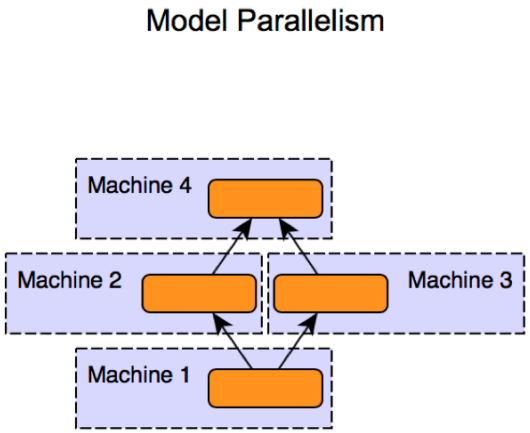
Training Huge Models

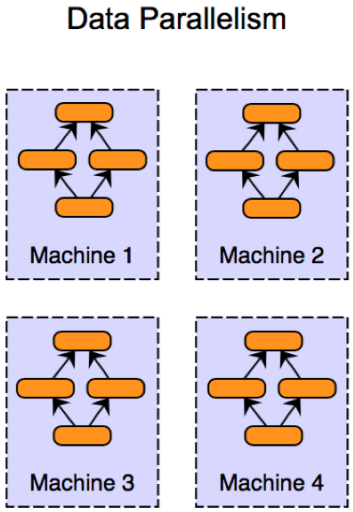
- Better hardware
- Data and Model parallelism

Mesh-TensorFlow:
Deep Learning for Supercomputers

Noam Shazeer, Youlong Cheng, Niki Parmar,
Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee
Mingsheng Hong, Cliff Young, Ryan Sepassi, Blake Hechtman
Google Brain
{noam, ylc, nikip, trandustin, avaswani, penporn, phawkins,
hyouklee, hongm, cliffy, rsepassi, blakehechtman}@google.com







GPT-2

- Just a really big Transformer LM

- Trained on 40GB of text
  - Quite a bit of effort going into making sure the dataset is good quality
  - Take webpages from reddit links with high karma

So What Can GPT-2 Do?

- Obviously, language modeling (but very well)!
- Gets state-of-the-art perplexities on datasets it's not even trained on!

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Radford et al., 2019

So What Can GPT-2 Do?

- **Zero-Shot Learning**: no supervised training data!
  - Ask LM to generate from a prompt

- **Reading Comprehension:** <context> <question> A:
- **Summarization:** <article> TL;DR:
- **Translation:**

  <English sentence1> = <French sentence1>

  <English sentence 2> = <French sentence 2>

  …..

  <Source sentence> =
- **Question Answering:** <question> A:

How can GPT-2 be doing translation?

- It's just given a big corpus of text that's almost all English

How can GPT-2 be doing translation?

- It's just given a big corpus of text that's almost all English

---

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

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

"I hate the word '**perfume**,"' Burr says. 'It's somewhat better in French: '**parfum.**'

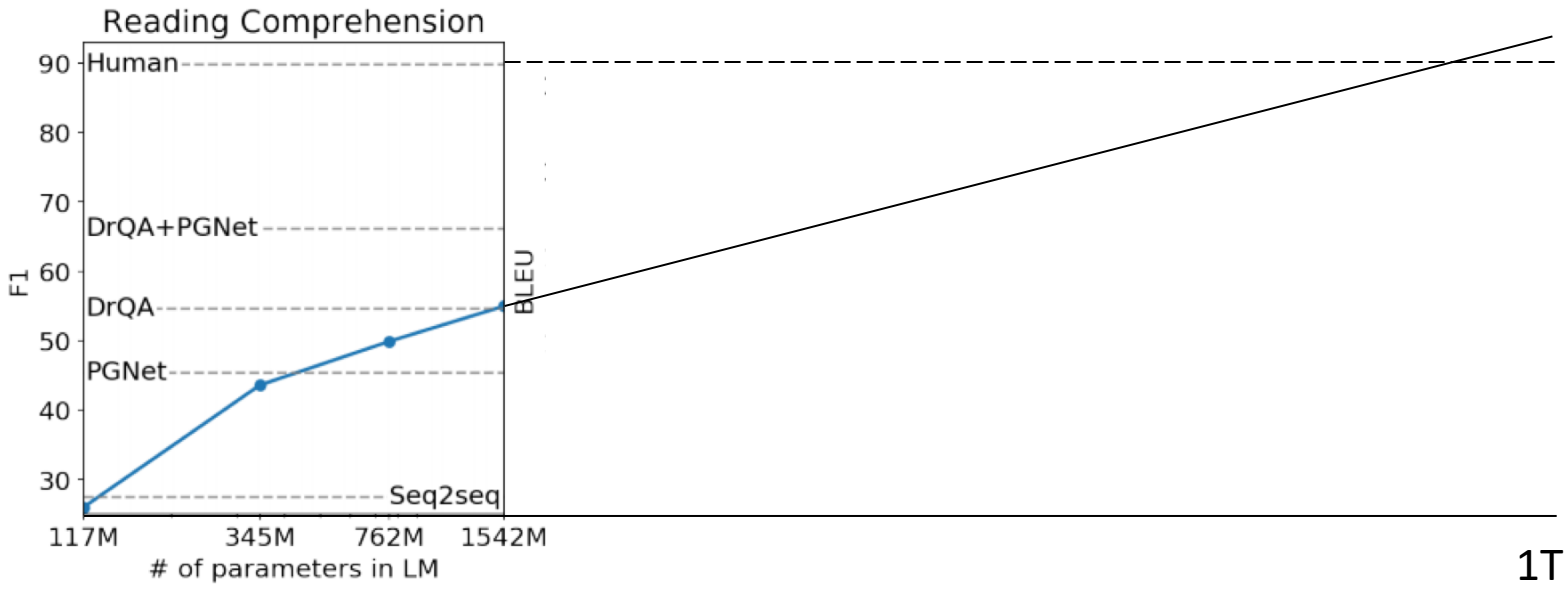
If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre coté? -Quel autre coté?**", which means "**- How do you get to the other side? - What side?**".

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

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

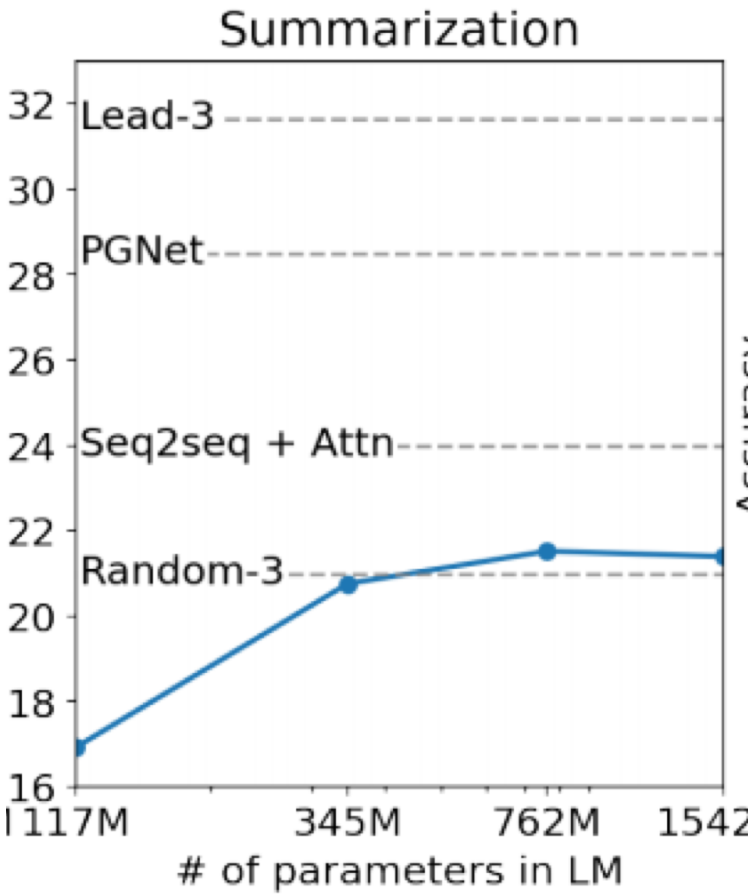
What happens as models get even bigger?

- For several tasks performance seems to increase with log(model size)



What happens as models get even bigger?

- But trend isn't clear



GPT-2 Reaction

GPT-2 Reaction

Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter:

GPT-2 Reaction

NEWS SPORT ENTERTAINMENT SOAPS MORE TRENDING

UK WORLD WEIRD TECH

Elon Musk-founded OpenAI builds artificial intelligence so powerful it must be kept locked up for the good of humanity

Jasper Hamill Friday 15 Feb 2019 10:06 am

Machine-generated text is about to break the internet

Mark Rickerby | Guest writer

OpenAI built a text generator so good, it's considered too dangerous to release

Zack Whittaker @zackwhittaker / 3 weeks ago

Comment

GPT-2 Reaction

Just wanted to give you all a heads up, our lab found an amazing breakthrough in language understanding. but we also worry it may fall into the wrong hands. so we decided to scrap it and only publish the regular *ACL stuff instead. Big respect for the team for their great work.

10:08 AM - 15 Feb 2019

118 Retweets **782** Likes

29 118 782

625

Posted by u/astonished_crofty 25 days ago 2

[Discussion] Should I release my MNIST model or keep it closed source fearing malicious use?

Discussion

Today I trained a 23064 layer ResNet and it got 99.6% accuracy on MNIST. I would love to share the model but I fear it being used maliciously. What if it is used to read documents by the Russians? What are your thoughts?

GPT-2 Reaction

OpenAI: Please Open Source Your Language Model

19.FEB.2019

Hugh Zhang

Stanford University

OpenAI Shouldn't Release Their Full Language Model

03.MAR.2019

Eric Zelikman

GPT-2 Reaction

Some arguments for release:

Some arguments against:

GPT-2 Reaction

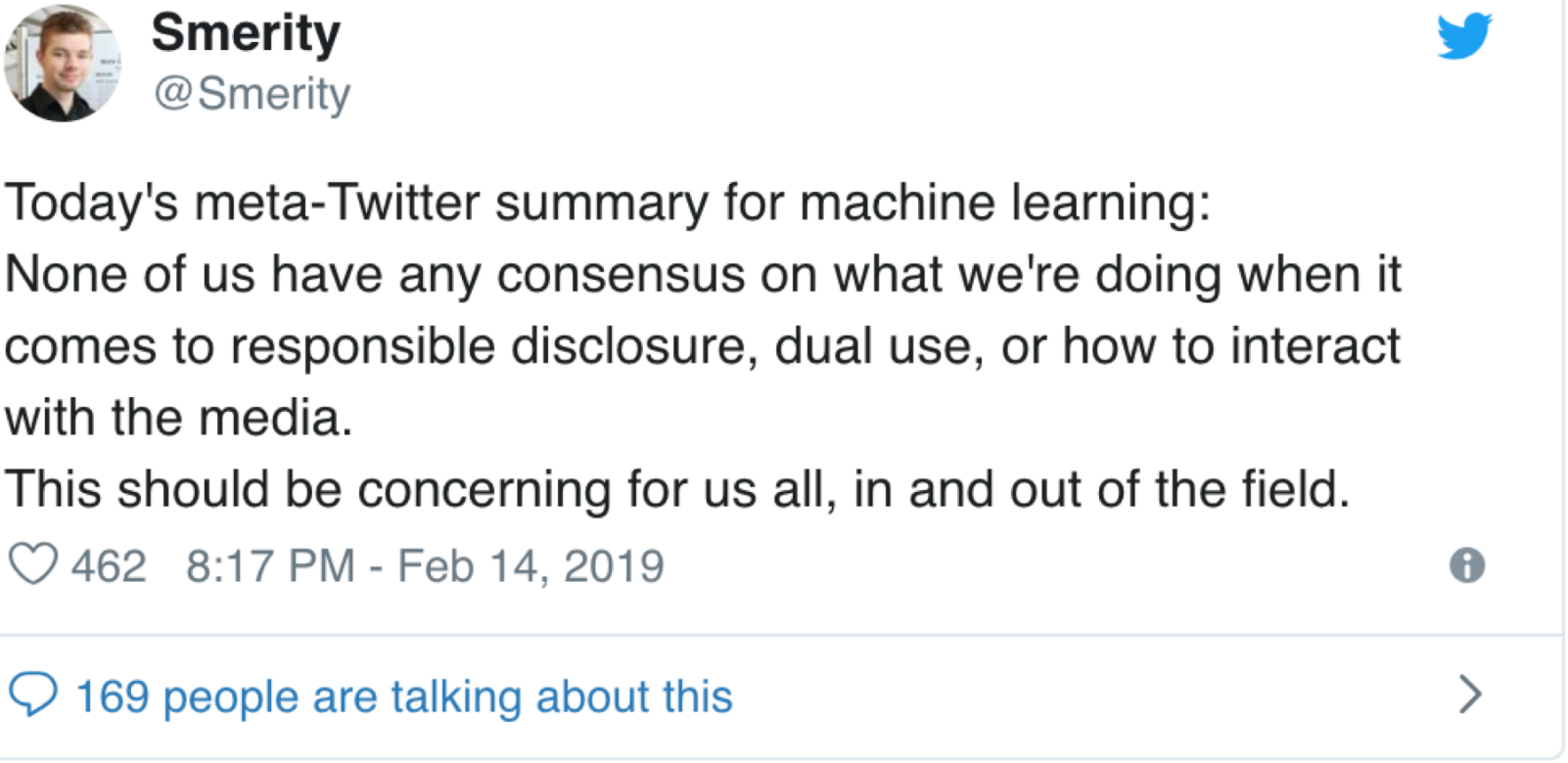
Some arguments for release:

- This model isn't much different from existing work
- Not long until these models are easy to train
  - And we're already at this point for images/speech
- Photoshop
- Researchers should study this model to learn defenses
- Dangerous PR Hype
- Reproducibility is crucial for science
- …

Some arguments against:

- Danger of fake reviews, news comments, etc.
  - Already done by companies and governments
- Precedent
  - Event if this model isn't dangerous, later ones will be even better
- Smaller model is being released
- ….

GPT-2 Reaction



GPT-2 Reaction

- Should NLP experts be the ones making these decisions?
  - Experts on computer security?
  - Experts on technology and society?
  - Experts on ethics?

- Need for more interdisciplinary science

- Many other examples of NLP with big social ramifications, especially with regards to bias/fairness

High-Impact Decisions

- Growing interest in using NLP to help with high-impact decision making
  - Judicial decisions
  - Hiring
  - Grading tests

- Plus side: can quickly evaluate a machine learning system for some kinds of bias

- However, machine learning reflects or even amplifies bias in training data
  - …which could lead to the creation of even more biased data

High-Impact Decisions

BUSINESS NEWS OCTOBER 9, 2018 / 8:12 PM / 5 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

Intelligent Machines

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

High-Impact Decisions

Ben Zimmer @bgzimmer · 2 Jul 2018

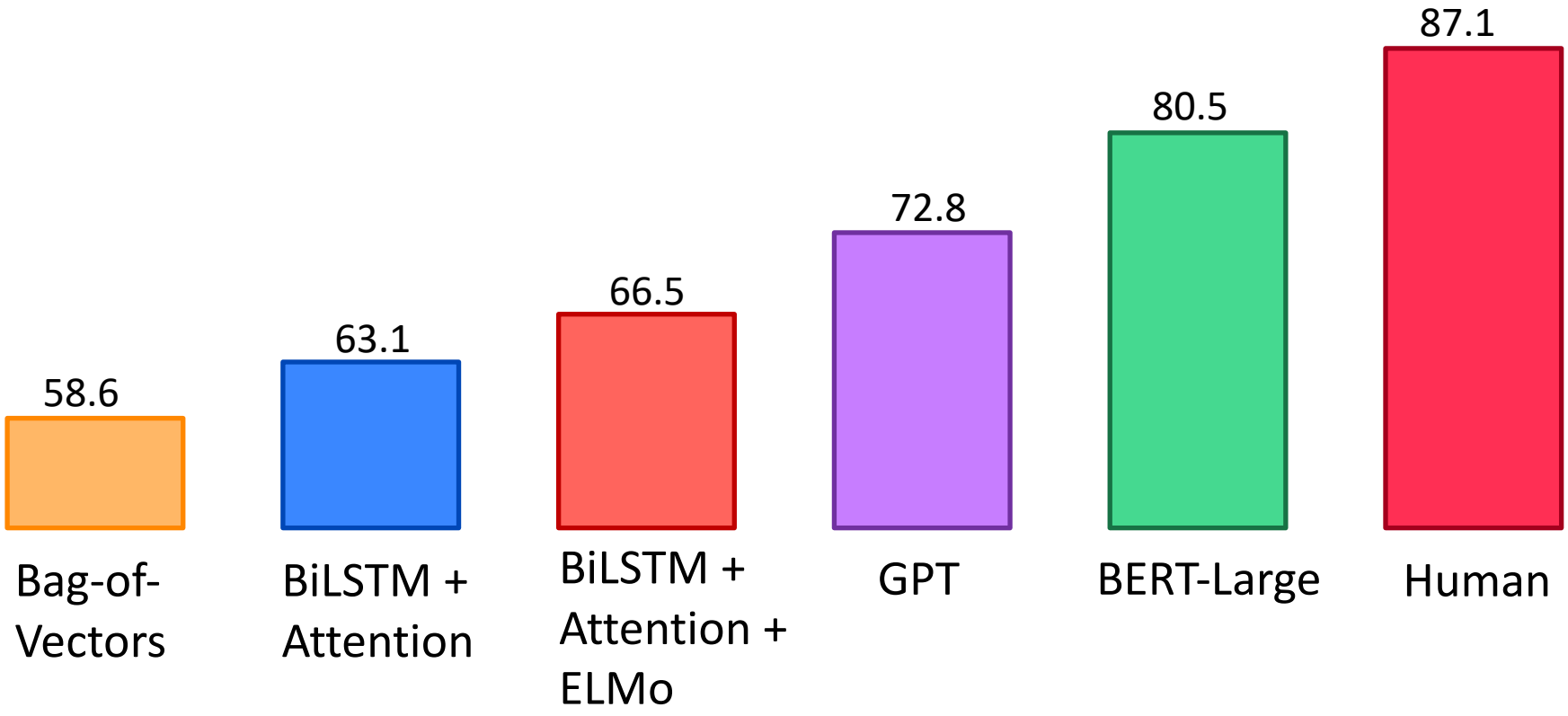
This gobbledygook earns a perfect grade from the GRE's automated essay scoring system. Algorithms writing for algorithms. npr.org/2018/06/30/624…

> *"History by mimic has not, and presumably never will be precipitously but blithely ensconced. Society will always encompass imaginativeness; many of scrutinizations but a few for an amanuensis. The perjured imaginativeness lies in the area of theory of knowledge but also the field of literature. Instead of enthralling the analysis, grounds constitutes both a disparaging quip and a diligent explanation."*

51 636 1.1K

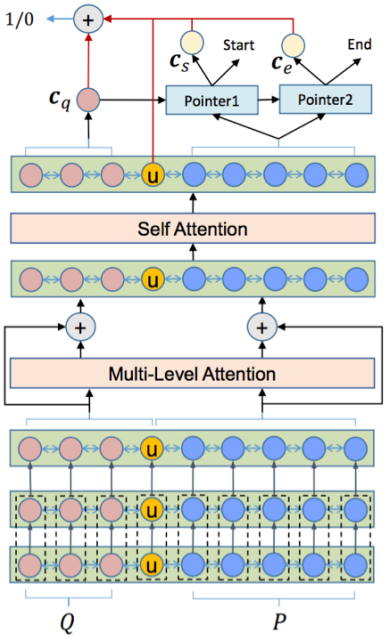
What did BERT “solve” and what do we work on next?

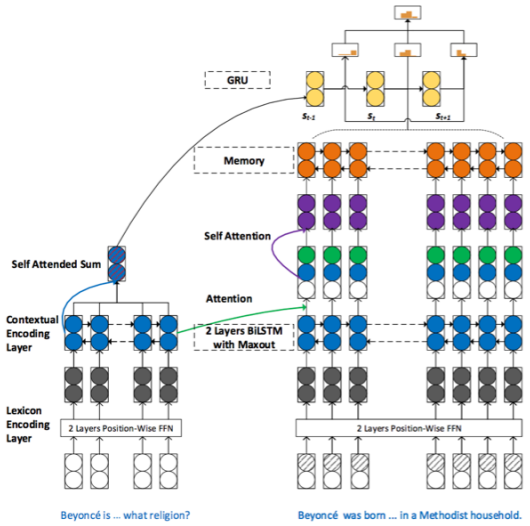
GLUE Benchmark Results

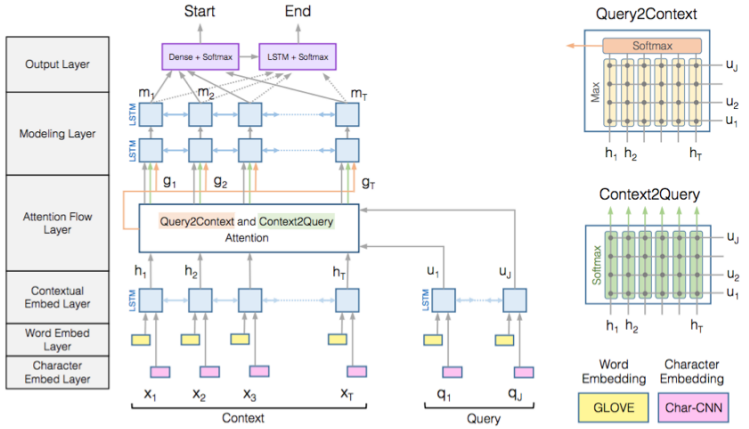


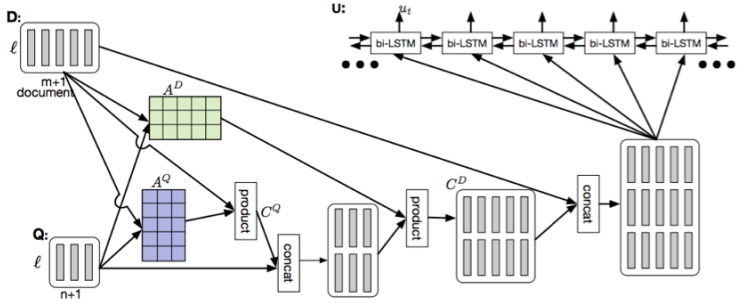
The Death of Architecture Engineering?

Some SQuAD NN Architectures



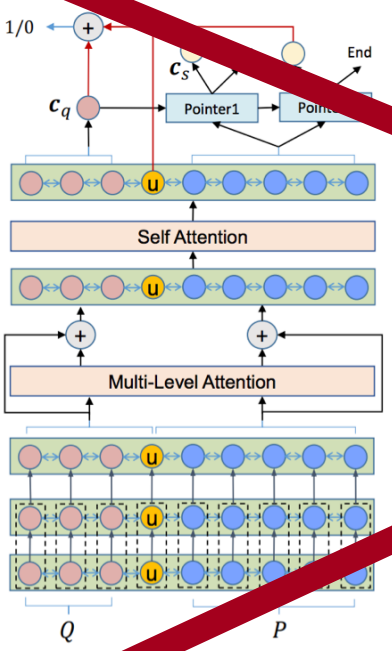


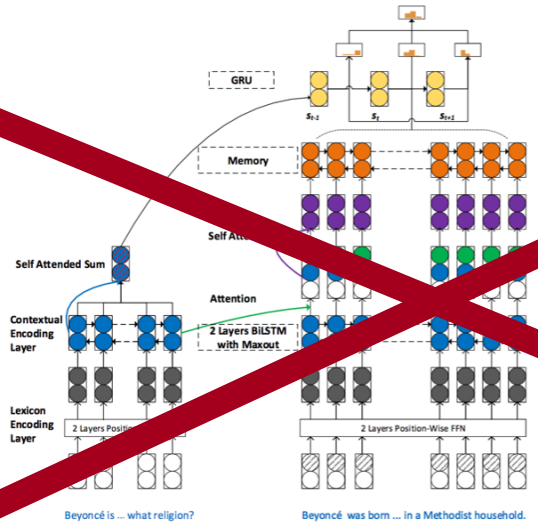


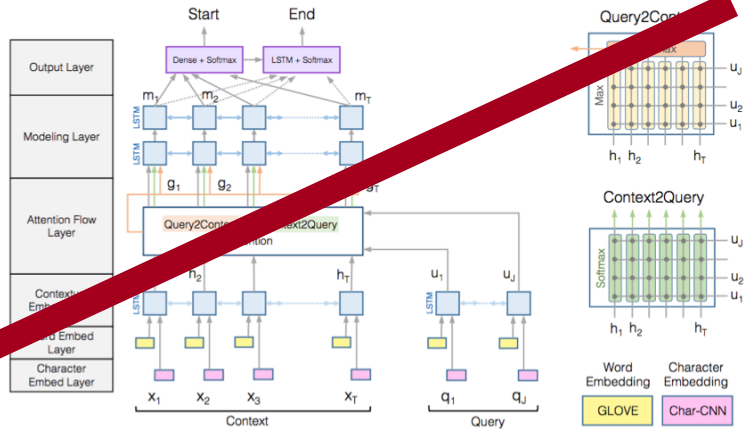


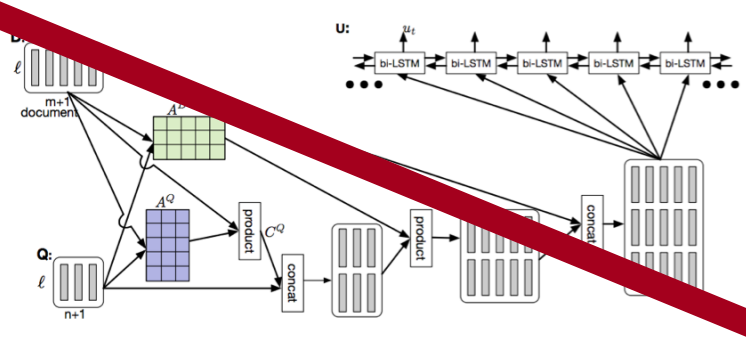
The Death of Architecture Engineering?

Some SQuAD NN Architectures









Attention Is All You Need

The Death of Architecture Engineering?

- 6 months of research on architecture design, get 1 F1 point improvement
- … Or just make BERT 3x bigger, get 5 F1 points
- Top 20 entrants on the SQuAD leaderboard all use BERT

	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	**86.673**	**89.147**
2 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715
3 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615
4 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
5 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
5 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
6 Mar 04, 2019	SemBERT (ensemble model) *Shanghai Jiao Tong University*	83.243	85.821
6 Jan 14, 2019	BERT + MMFT + ADA (single model) *Microsoft Research Asia*	83.040	85.892
7 Jan 10, 2019	BERT + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	82.972	85.810

Harder Natural Language Understanding

- Reading comprehension…
  - On longer documents or multiple documents
  - That requires multi-hop reasoning
  - Situated in a dialogue

- Key problem with many existing reading comprehension datasets: *People writing the questions see the context*
  - Not realistic
  - Encourages easy questions

QuAC: Question Answering in Context

- Dialogue between a student who asks questions and a teacher who answers
  - Teacher sees Wikipedia article on the subject, student doesn't

Section: Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↪ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
. . .

Choi et al., 2018

QuAC: Question Answering in Context

- Still a big gap to human performance

Rank	Model	F1	HEQQ	HEQD
	Human Performance (Choi et al. EMNLP '18)	81.1	100	100
 March 7, 2019	BERT w/ 2-context (single model) *NTT Media Intelligence Labs*	**64.9**	60.2	**6.1**
2 Feb 21, 2019	GraphFlow (single model) *Anonymous*	**64.9**	**60.3**	5.1
3 Sep 26, 2018	FlowQA (single model) *Allen Institute of AI* https://arxiv.org/abs/1810.06683	64.1	59.6	5.8
4 Jan 30, 2019	BERT + History Answer Embedding (single model) *Anonymous*	62.4	57.8	5.1
5 Aug 20, 2018	BiDAF++ w/ 2-Context (single model) *baseline*	60.1	54.8	4.0
6 Aug 20, 2018	BiDAF++ (single model) *baseline*	50.2	43.3	2.2

Multi-Task Learning

- Another frontier of NLP is getting one model to perform many tasks. GLUE and DecaNLP are recent examples.

- Multi-task learning yields improvements on top of BERT

BERT + Multi-task

Rank	Name	Model	URL	Score
1	GLUE Human Baselines	GLUE Human Baselines		87.1
2	王玮	ALICE large (Alibaba DAMO NLP)		83.0
3	Microsoft D365 AI & MSR AI	MT-DNNv2 (BigBird)		83.0
4	Jason Phang	BERT on STILTs		82.0
5	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hid		80.5

Low-Resource Settings

- Models that don't require lots of compute power (can't use BERT)!
  - Especially important for mobile devices
- Low-resource languages
- Low-data settings (few shot learning)
  - Meta-learning is becoming popular in ML.

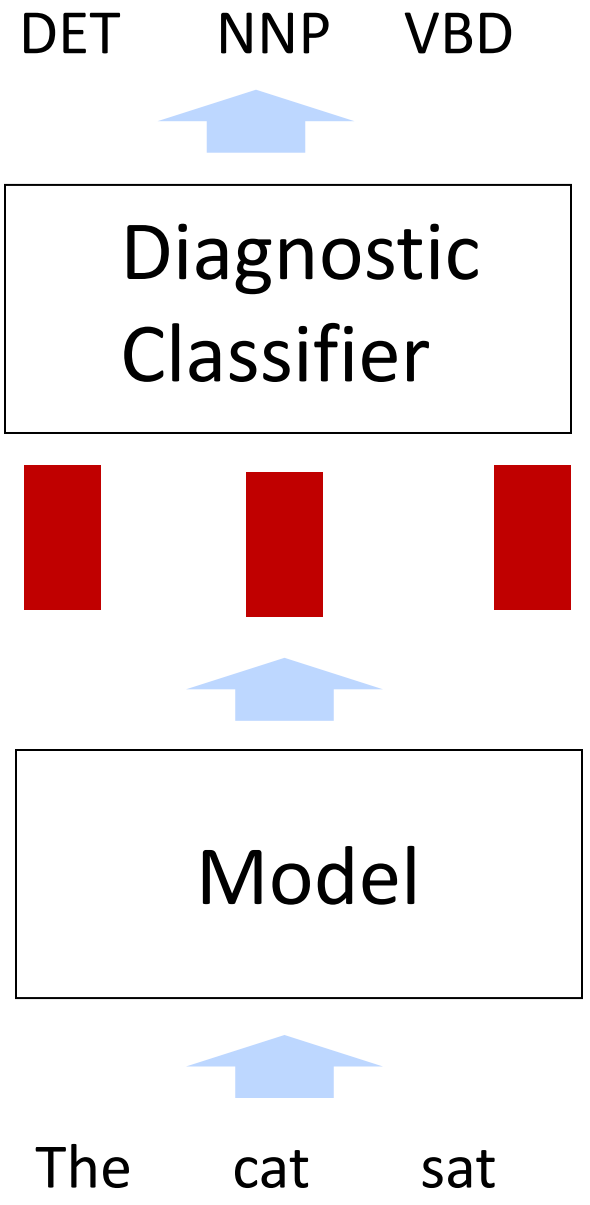
Interpreting/Understanding Models

- Can we get explanations for model predictions?
- Can we understand what models like BERT know and why they work so well?

- Rapidly growing area in NLP
- Very important for some applications (e.g., healthcare)

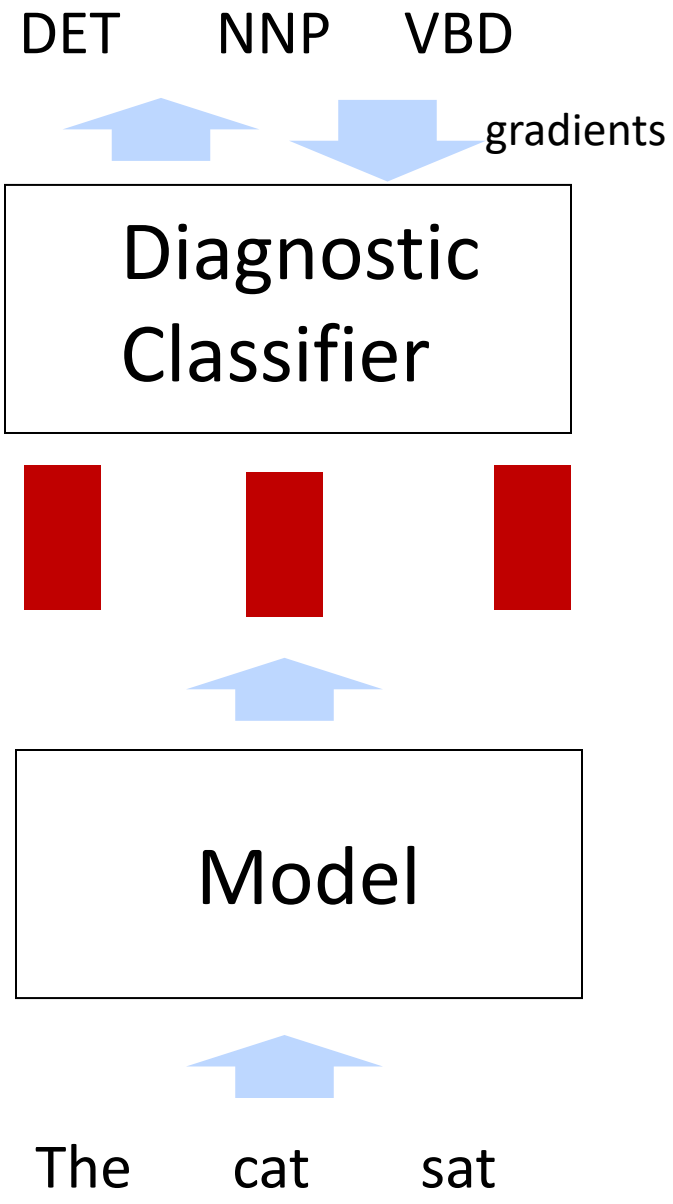
Diagnostic/Probing Classifiers

- Popular technique to see what linguistic information models "know"
- Diagnostic classifier takes representations produced by a model (e.g., BERT) as input and do some task



Diagnostic/Probing Classifiers

- Popular technique to see what linguistic information models "know"
- Diagnostic classifier takes representations produced by a model (e.g., BERT) as input and do some task
- Only the diagnositic classifier is trained



Diagnostic/Probing Classifiers

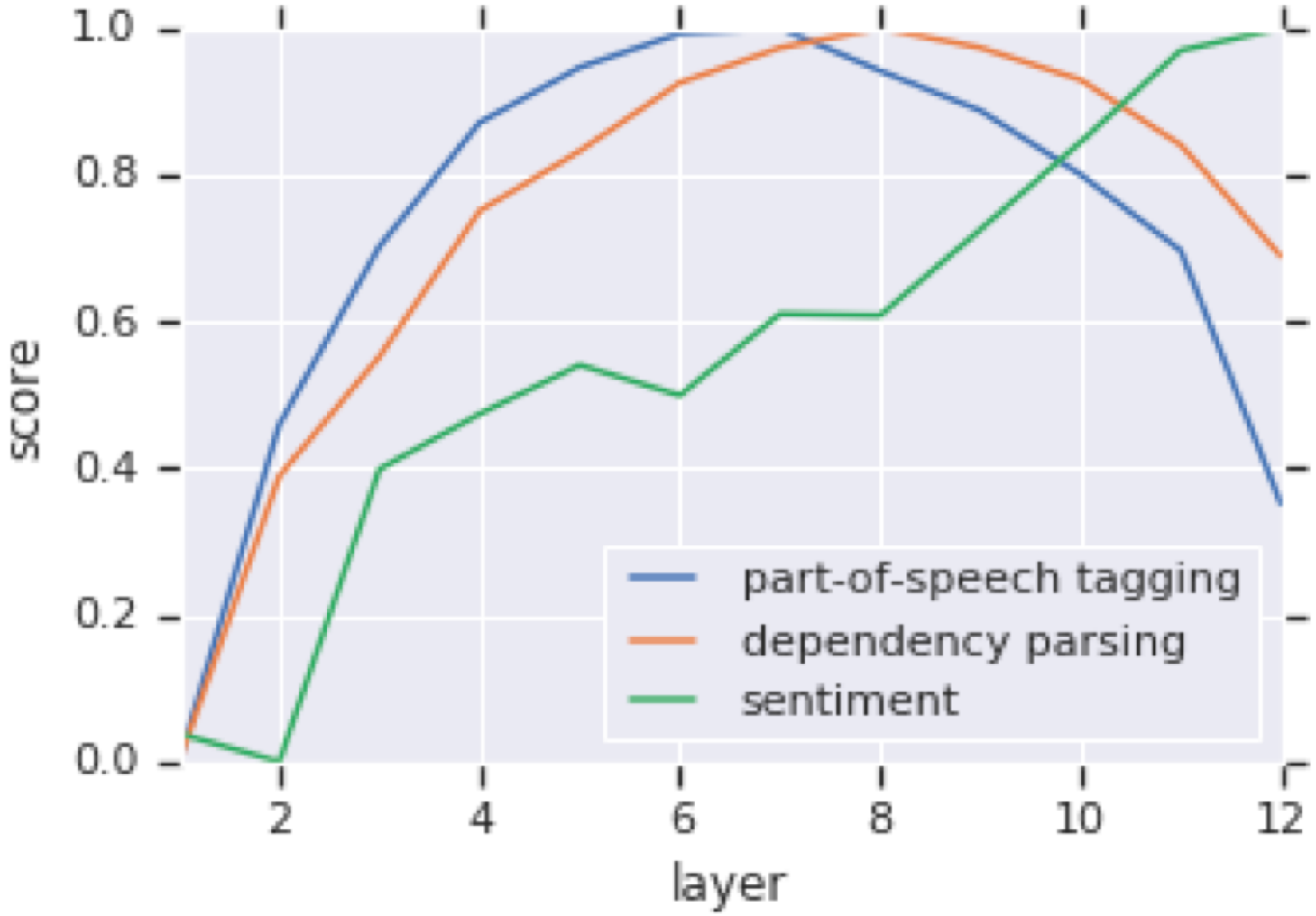
- Diagnostic classifiers are usually very simple (e.g., a single softmax). Otherwise they could learn to do the tasks without looking at the model representations

- Some diagnostic tasks

POS	The important thing about Disney is that it is a global $[\text{brand}]_1$. $\rightarrow$ NN (Noun)
Constit.	The important thing about Disney is that it $[\text{is a global brand}]_1$. $\rightarrow$ VP (Verb Phrase)
Depend.	$[\text{Atmosphere}]_1$ is always $[\text{fun}]_2$ $\rightarrow$ nsubj (nominal subject)
Entities	The important thing about $[\text{Disney}]_1$ is that it is a global brand. $\rightarrow$ Organization
SRL	$[\text{The important thing about Disney}]_2$ $[\text{is}]_1$ that it is a global brand. $\rightarrow$ Arg1 (Agent)
SPR	$[\text{It}]_1$ $[\text{endorsed}]_2$ the White House strategy... $\rightarrow$ {awareness, existed_after, ...}
Coref.$^{\text{O}}$	The important thing about $[\text{Disney}]_1$ is that $[\text{it}]_2$ is a global brand. $\rightarrow$ True
Coref.$^{\text{W}}$	$[\text{Characters}]_2$ entertain audiences because $[\text{they}]_1$ want people to be happy. $\rightarrow$ True Characters entertain $[\text{audiences}]_2$ because $[\text{they}]_1$ want people to be happy. $\rightarrow$ False
Rel.	The $[\text{burst}]_1$ has been caused by water hammer $[\text{pressure}]_2$. $\rightarrow$ Cause-Effect(e_2, e_1)

Diagnostic/ Probing Classifiers: Results

- Lower layers of BERT are better at lower-level tasks



NLP in Industry

- NLP is rapidly growing in industry as well. Two particularly big areas:

- Dialogue
  - Chatbots
  - Customer service

- Healthcare
  - Understanding health records
  - Understanding biomedical literature

Conclusion

- Rapid progress in the last 5 years due to deep learning.
- Even more rapid progress in the last year due to larger models, better usage of unlabeled data
  - Exciting time to be working on NLP!
- NLP is reaching the point of having big social impact, making issues like bias and security increasingly important.

Good luck with your projects!