

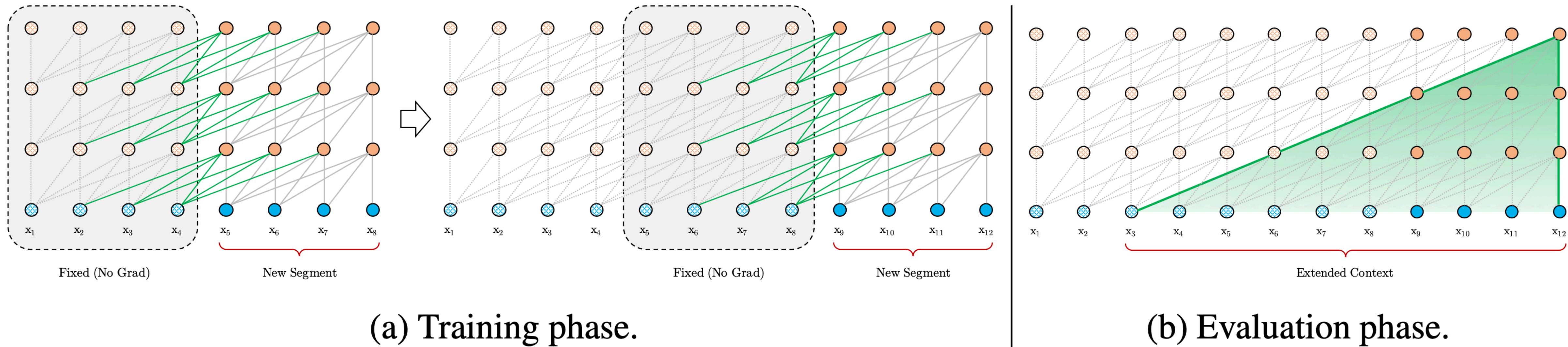
Pre-training & Transfer Learning

Tim Rocktäschel & Sebastian Riedel
COMP0087 Natural Language Processing



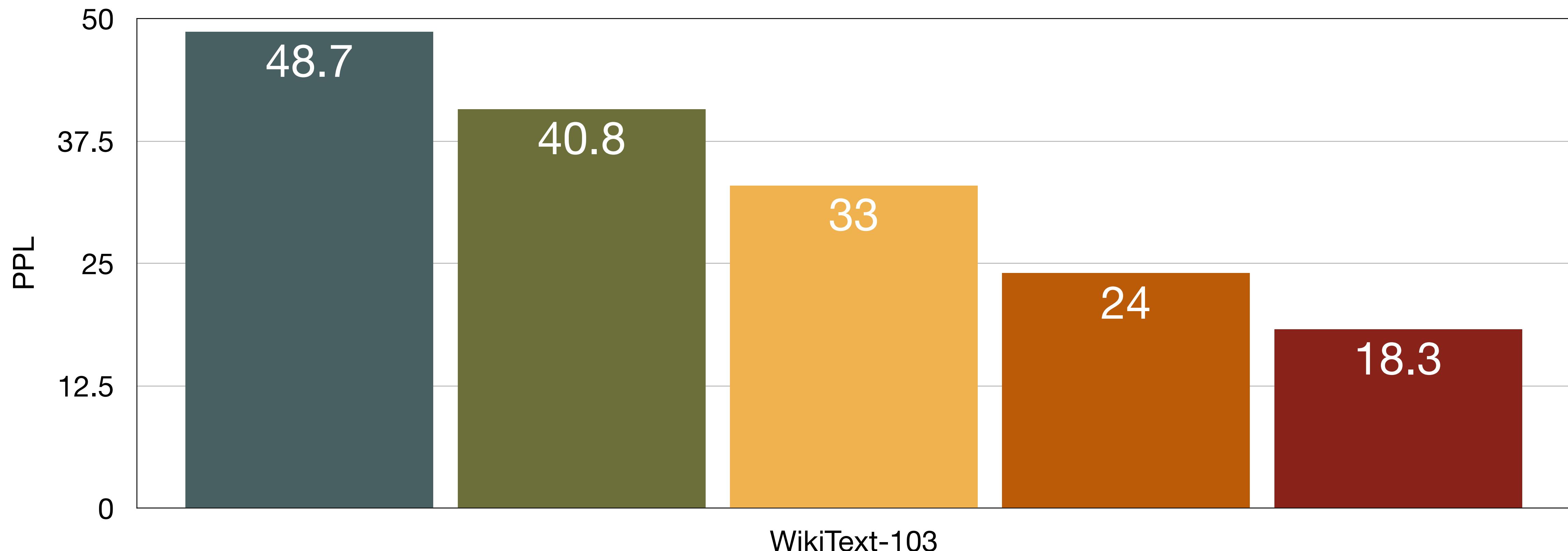
Transformer XL

- Large-scale self-attention model for language modeling
- Idea: **Cache** outputs of Transformer for previous **segments**
- Allow TransformerXL to **attend over cached outputs** without backpropagating beyond current segment



Results

LSTM LSTM+Neural cache QRNN
TransformerXL (standard) TransformerXL (large)



OpenAI GPT

- Use decoder in Transformer architecture for Language Modeling
- v2
 - Trained on 8 million documents (40GB of text)
 - Scrapped from outgoing links of Reddit articles with at least 3 karma
 - 50k vocab
 - 1024 tokens context
 - >1.5 billion parameters

"Dataset Engineering"

"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.**"

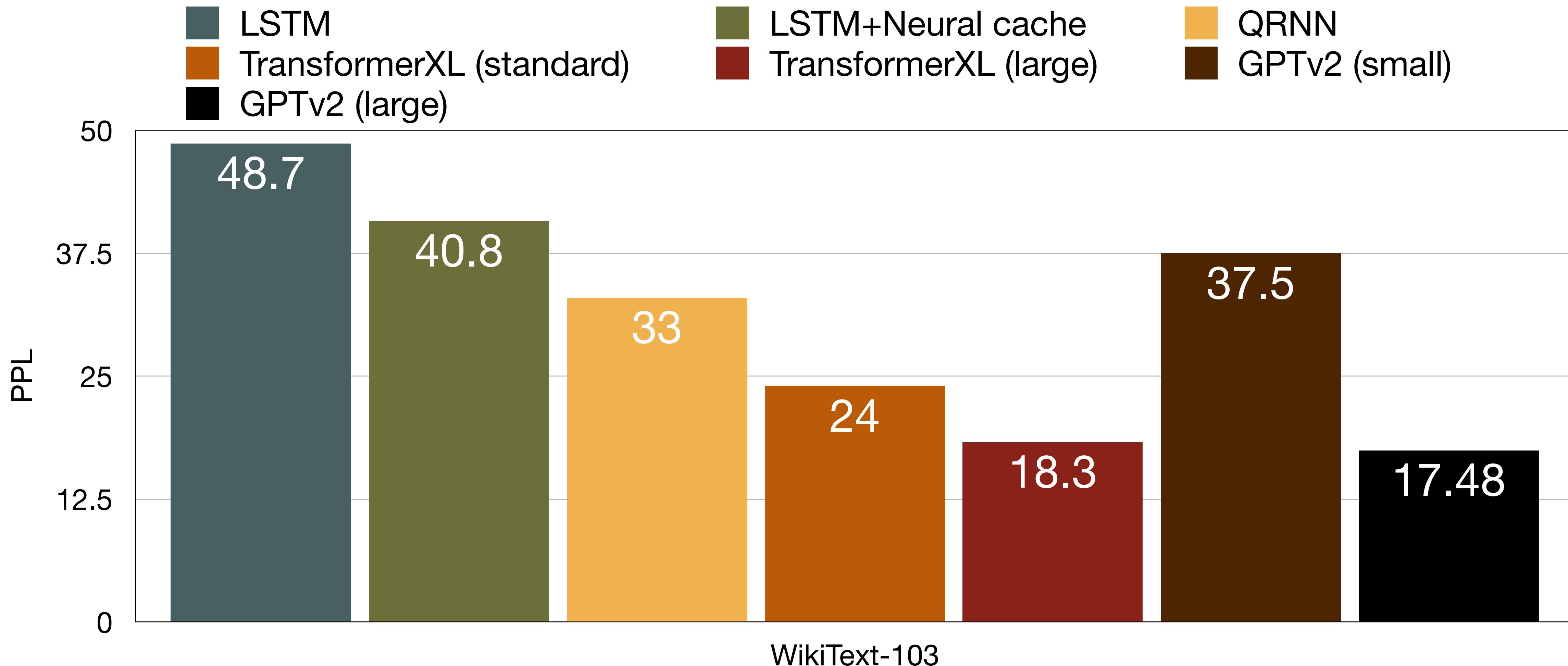
"I hate the word '**perfume**','" Burr says. 'It's somewhat better in French: '**parfum**'.'

If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre côté? -Quel autre côté?**", 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**".

Results



Generating “Wikipedia”

Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by **John Clair, An Imperial Japanese Revolt**, associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the **Protestant Immineners**, which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to **Antioch, Perth, October 25|21** to note, the Kingdom of Costa Rica, unsuccessful fashioned the **Thrales, Cynth's Dajoard**, known in western **Scotland**, near Italy to the conquest of India with the conflict. Copyright was the succession of independence in the slop of Syrian influence that was a famous German movement based on a more popular servicious, non-doctrinal and sexual power post.

Context (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.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

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.”

GPTv2

=====Model prompt >>> I believe nobody in this room knows this, but Sebastian Riedel was a famous underground rapper in Hamburg before starting an academic career. In his raps, he would frequently make references to Noam Chomsky, secretly educating the hood on linguistics.

===== SAMPLE 1 =====

His nickname was Violent. For some reason Sebastian chose to call him Calvin White.

The Worm

This act nevertheless alerted the drug Users Manual into the dangers of Lynn, and I think replaying and his signature tread appears to be one of the reason behind his initiation as electronic jungle kingpin. Deep down, he would like porn to become a Raw Language, and digital video to become Internet Porn.

At one point, at about age 45, Sebastian worked for Hermann Gothenburg Systems, a kind of revisionist Marxist investment agency selling the "Basic Income" Ben Gurion study. He sneered at Eminem and watched General Gigigi's shareholders float in panic. "The monetary system has cost more than the incomes of American workers," he writes in the reverse of his provocative drink.

Not surprisingly, both the films Psycho and Spirit Of Disputation were enormously success for a young Gothenburg software engineer who had grown tired of the complex monthly payments, which got him homesick as often night work becomes her job title drama. In his insomniac rant his wiser brother suggested, "It's like ... reading for nine hours in the most deliciously pornographic western movie ever."

Memorization

```
=====
Model prompt >>> import torch
===== SAMPLE 1 =====
glow in ( stop )
end

def spawn_vump ( errs )

if has_grenade ( develop_clip ) :

# Set end_toggled to false so we don't see holo
clip_badge = defenderish_clip_badge ()

elif not clip_badge. weapon >= 5 :

msg [ " Defending " ] = True

msg [ " Damaged and at the outskirts of the lab. " ] = True

font_file name = " NotepadEx Enemy.png "

time = player. reload_minute ( value )

if path. exists ( " file/item_drop_drop.ogg " ) and ( " xp.h_mgr.y_smenu.txt " in name ) :
log. dump ( test ( d " Use totem " ) )

if self ._listen_to_sound ( msg ) and ( - 60000 . flags & path. bloody ) is None :
start_turn = None

pcov2. touch_camera_truthfulin ( name , true )

fnv_defence_freemove = self ._detach_freemove
```

```
self ._corpse_is_killed = true

def destroy ( iter = None , verdict = 1 ):

""" Randomly extinguish a hand grenade on the ground right player turn.

including where it hit. With an addition of cleanup it, The
hand is then removed.

"""

if len ( iter )== 1 and len ( verdict ) >= verdict . len () :
knapsacks. remove ( )

for lock in lock. enumerate ( items [ len ( verdict )]):
```

return verdict

```
def destroy_weakgrenade ( iter , verdict , salt = None ):

""" Spawn individual frag grenades. Use with caution. Evac
high is high and low if high. Ctrl + key events will bring grenade
tying up. If high, all grenade tying will be changed to low.

subkill_quick ( self .recruit, jampot_time, destroy = true )

def threadingwith_weapons ( self , slot , player = False ):

if slot == slot and player == None :

nickname = server. nickname if slot in os. path .
=====
```

Concerns

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](#): “we expect that safety and security concerns will reduce our traditional publishing in the future, while increasing the importance of sharing safety, policy, and standards research,” and we see this current work as potentially representing the early beginnings of such concerns, which we expect may grow over time.

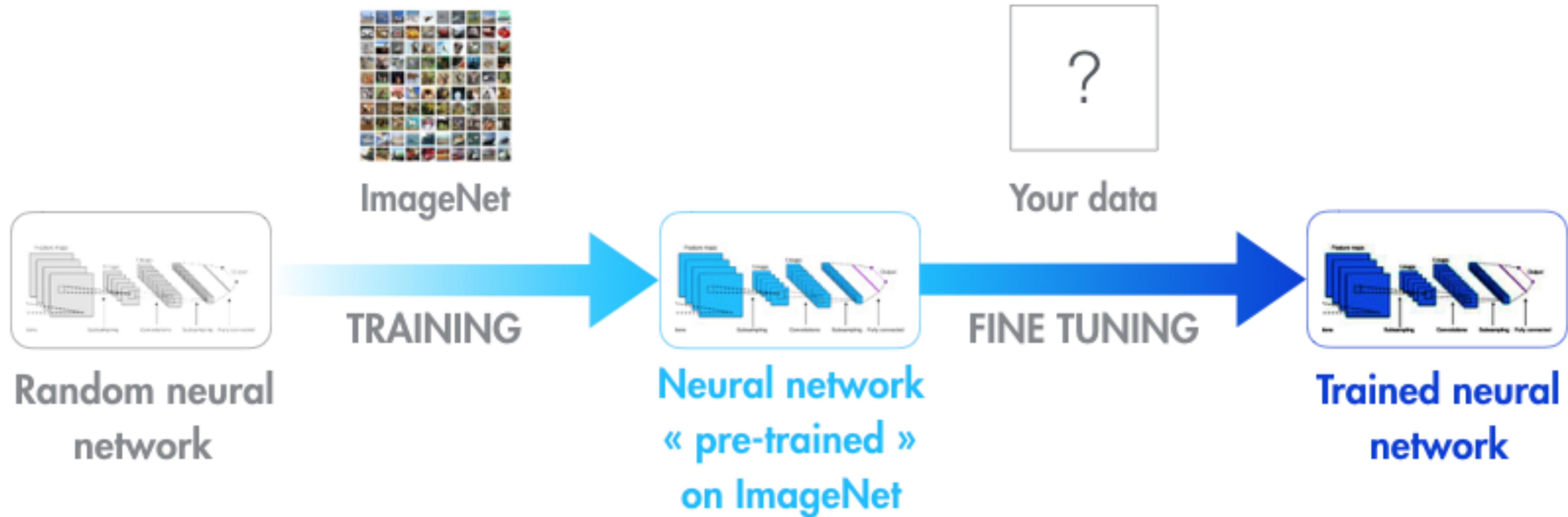
Transfer Learning

- In general, large amounts of annotated data are needed for training machine learning models in NLP
- Can we transfer representations learned on one task to other tasks?
- If done successfully, for another task we could
 - learn quicker (better sample efficiency)
 - learn better (better generalization)
 - cut down annotation cost
 - apply machine learning to tasks for which it is difficult to collect data

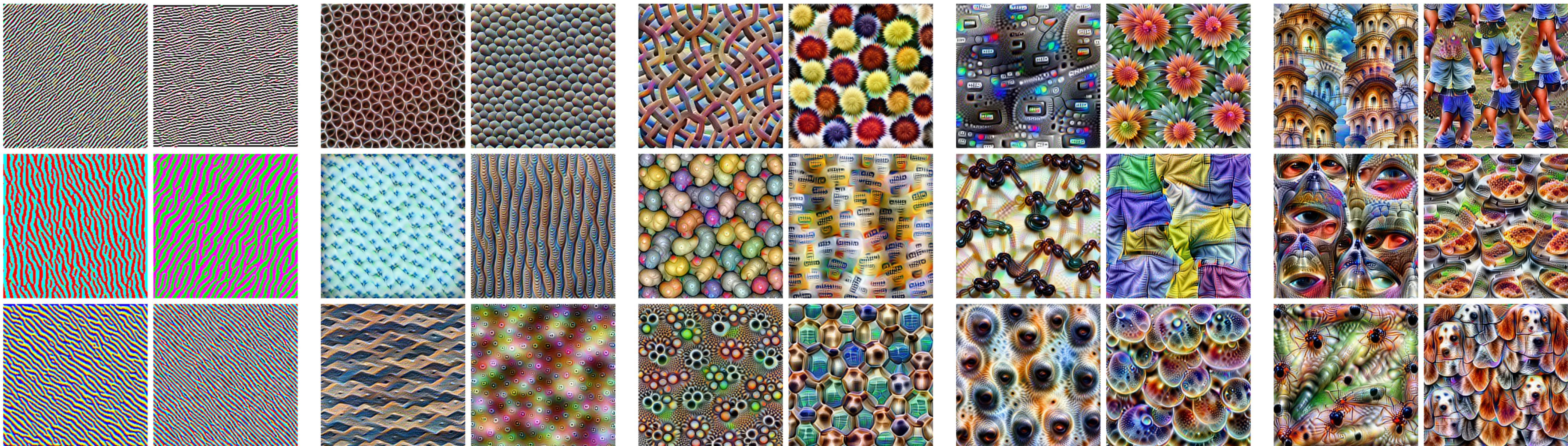
Pre-training in Computer Vision



Pre-training in Computer Vision



What does that look like?



Edges (layer conv2d0)

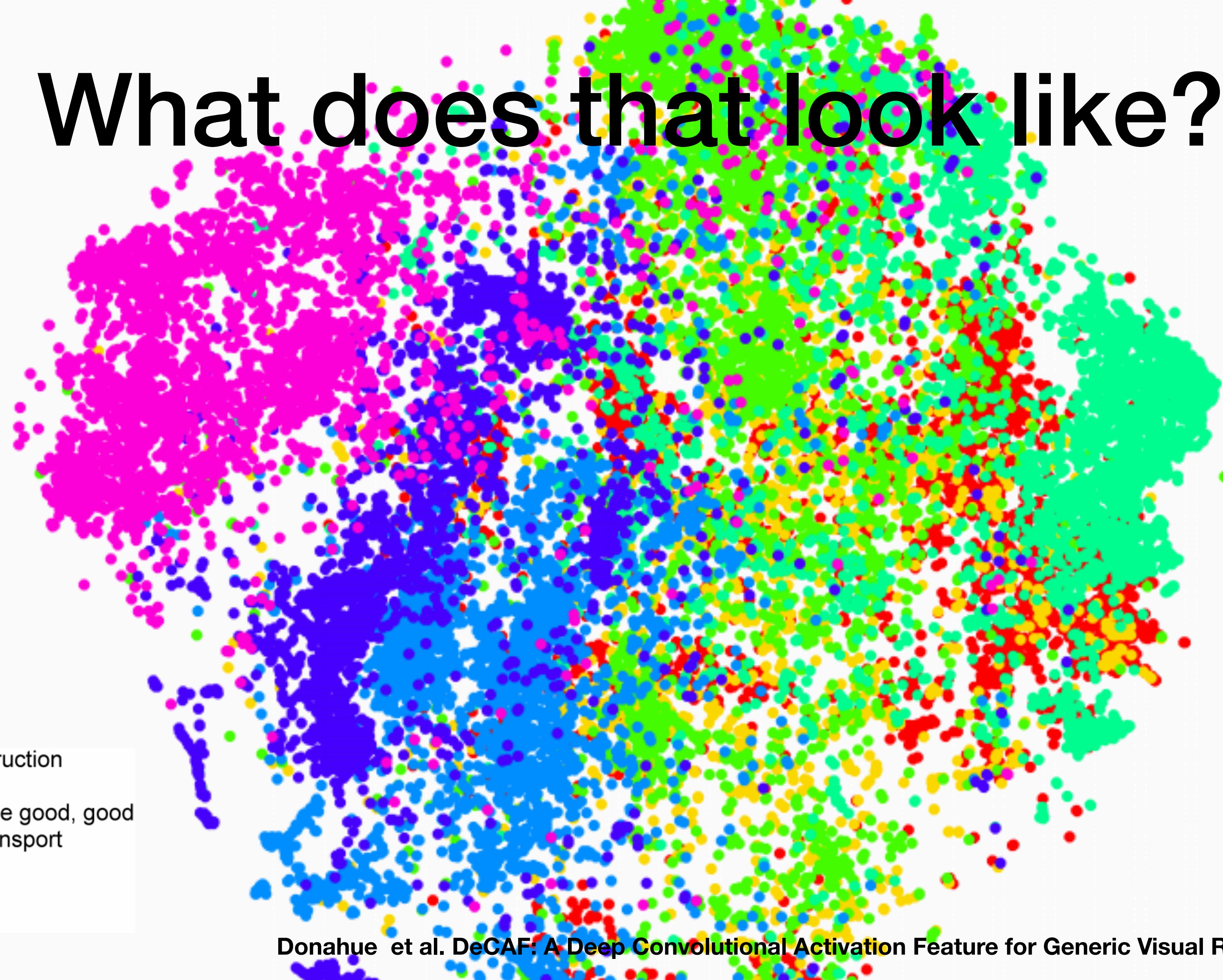
Textures (layer mixed3a)

Patterns (layer mixed4a)

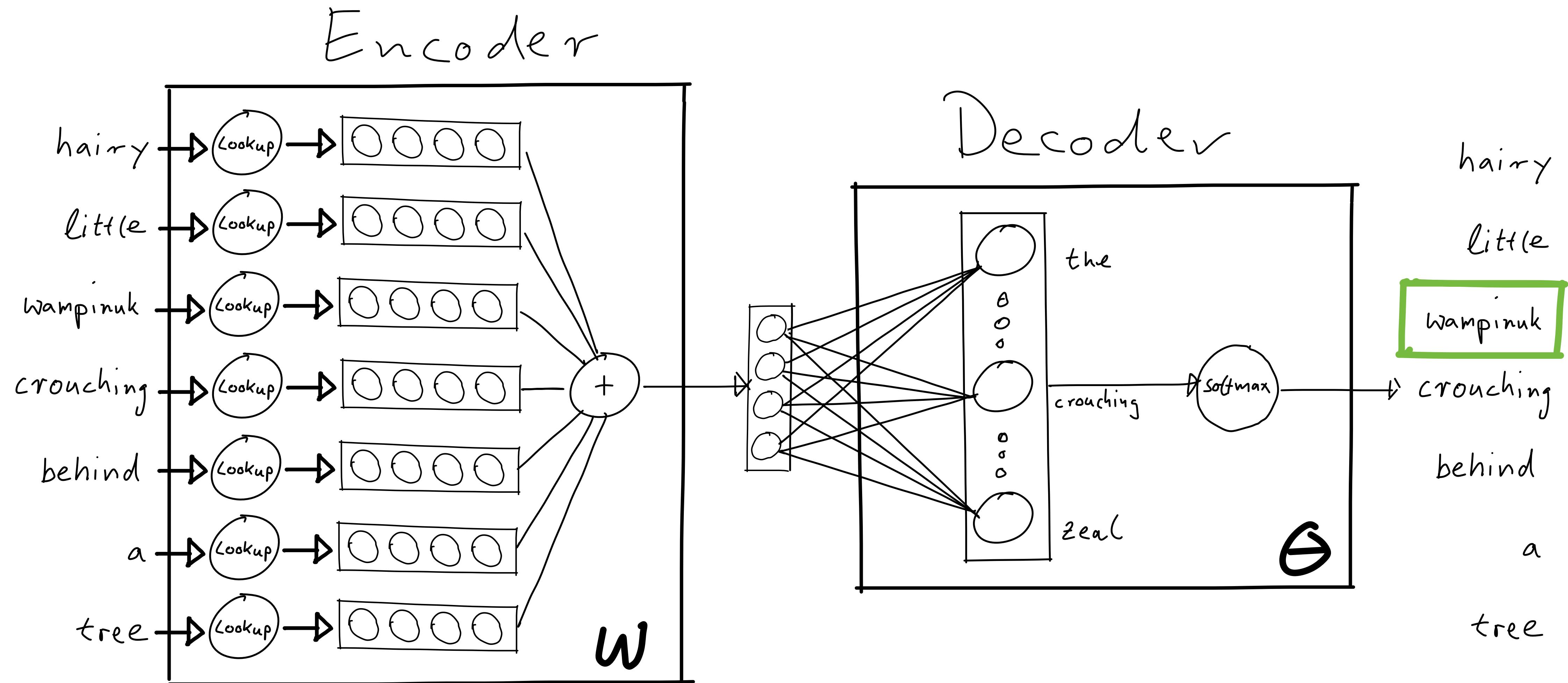
Parts (layers mixed4b & mixed4c)

Objects (layers mixed4d & mixed4e)

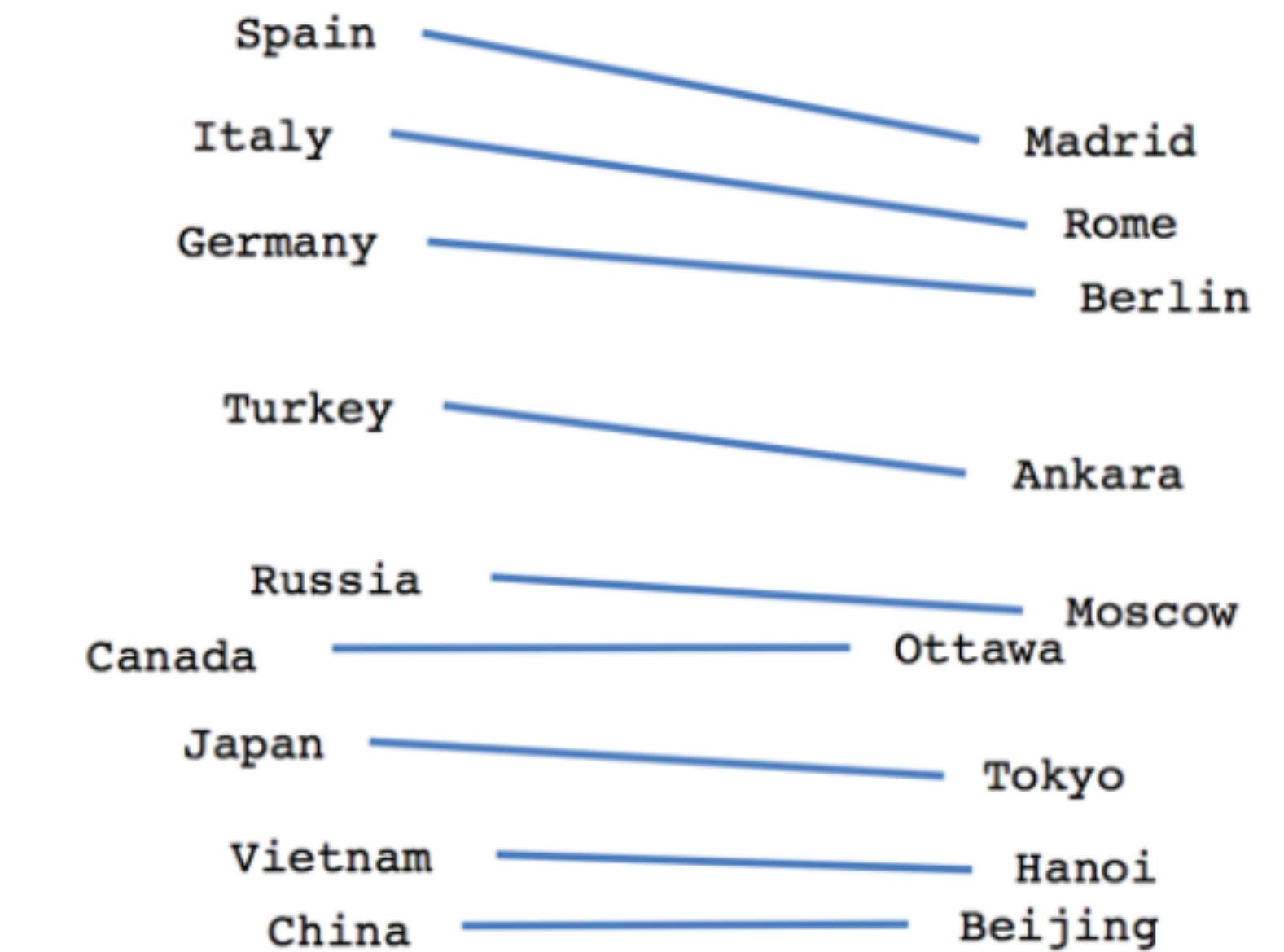
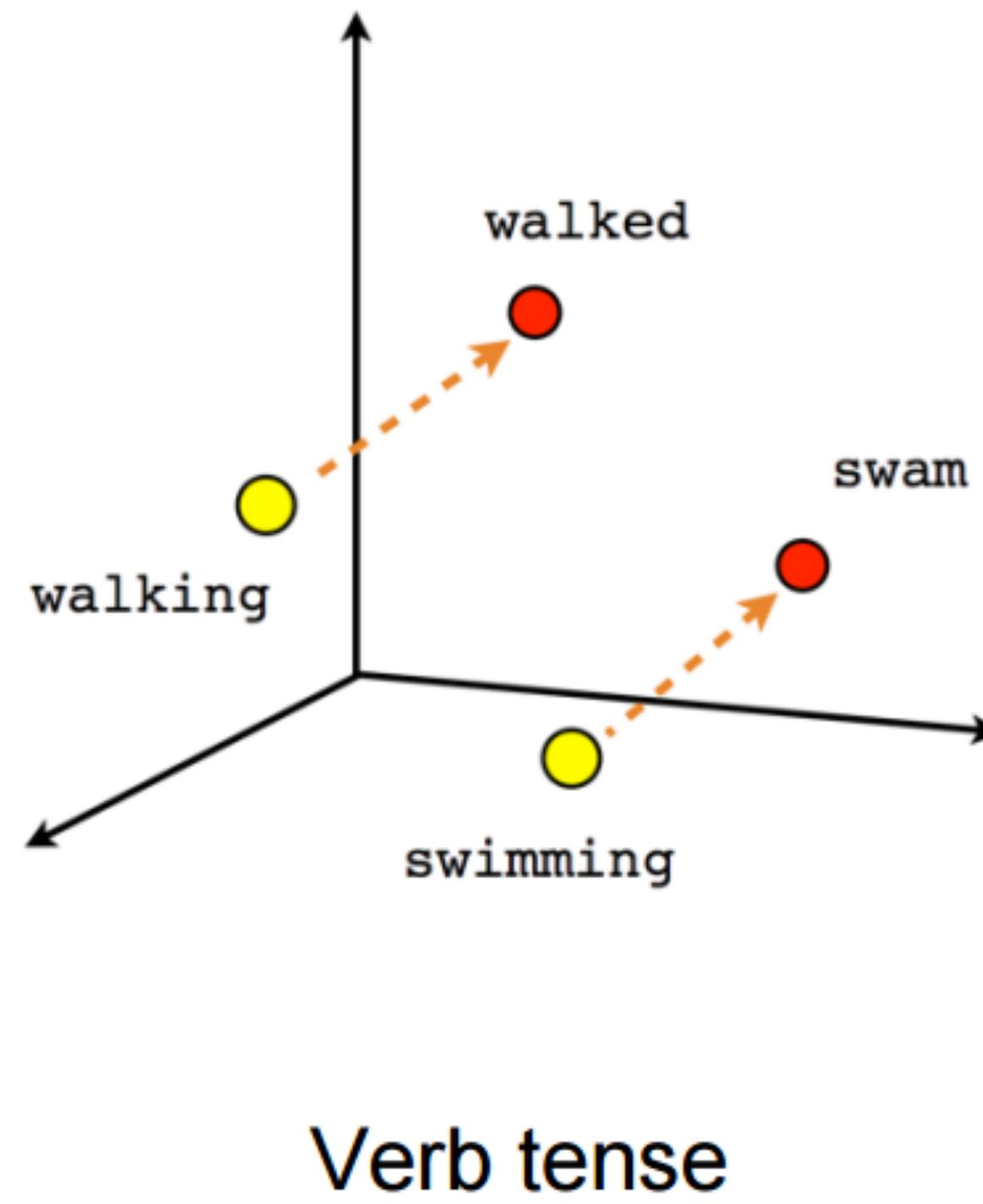
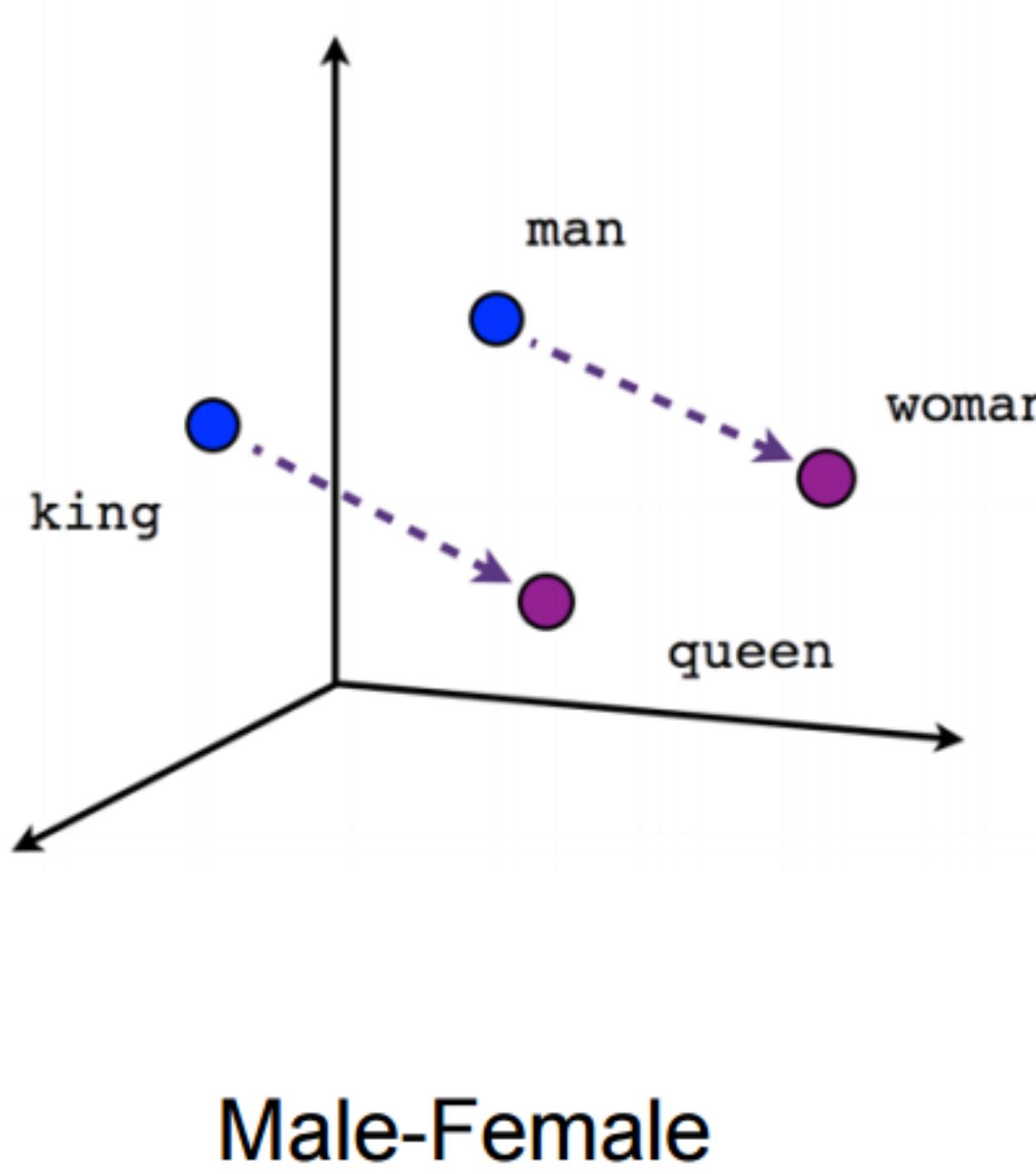
What does that look like?



Pre-training Word Representations

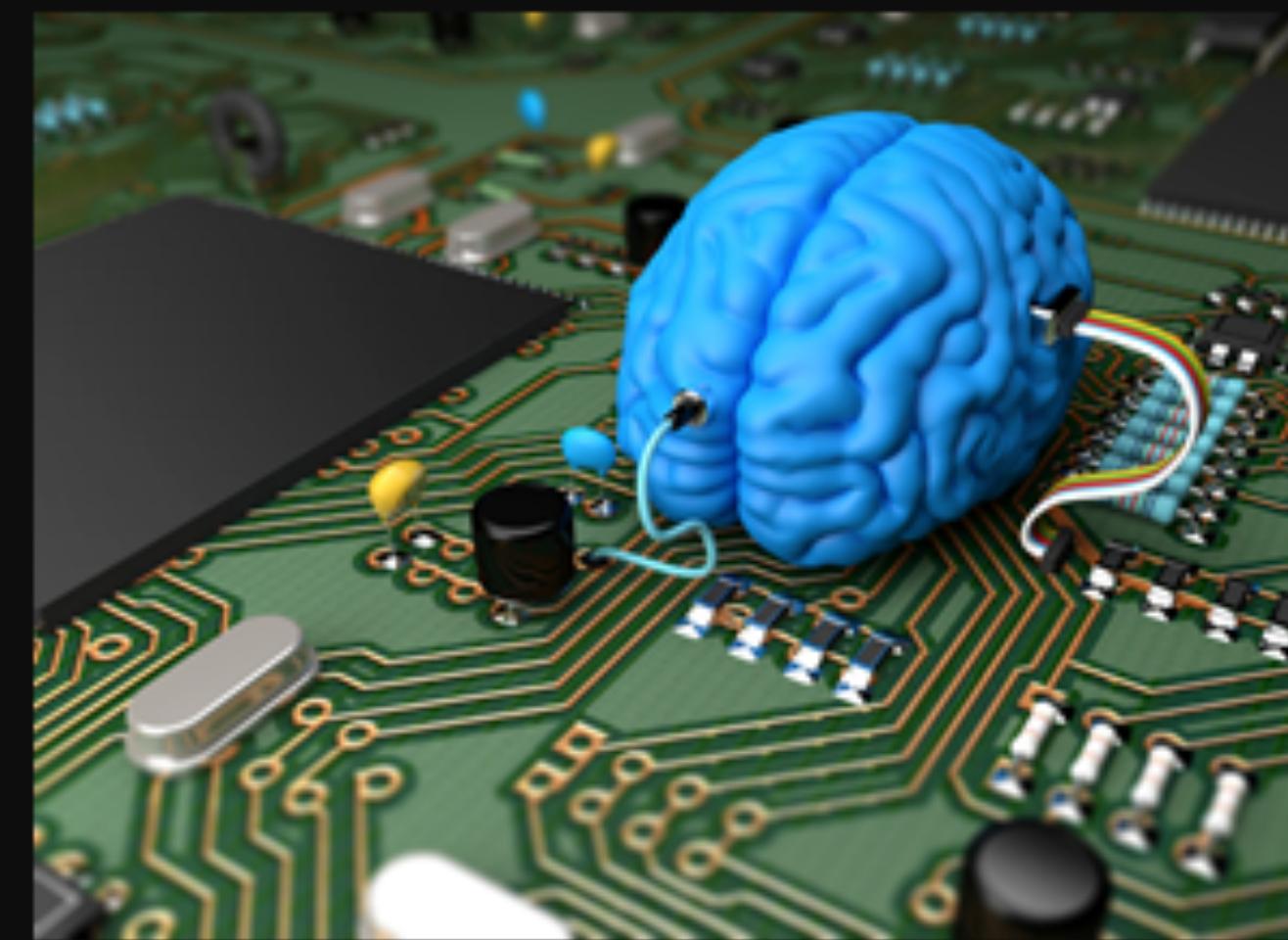


word2vec & GloVe





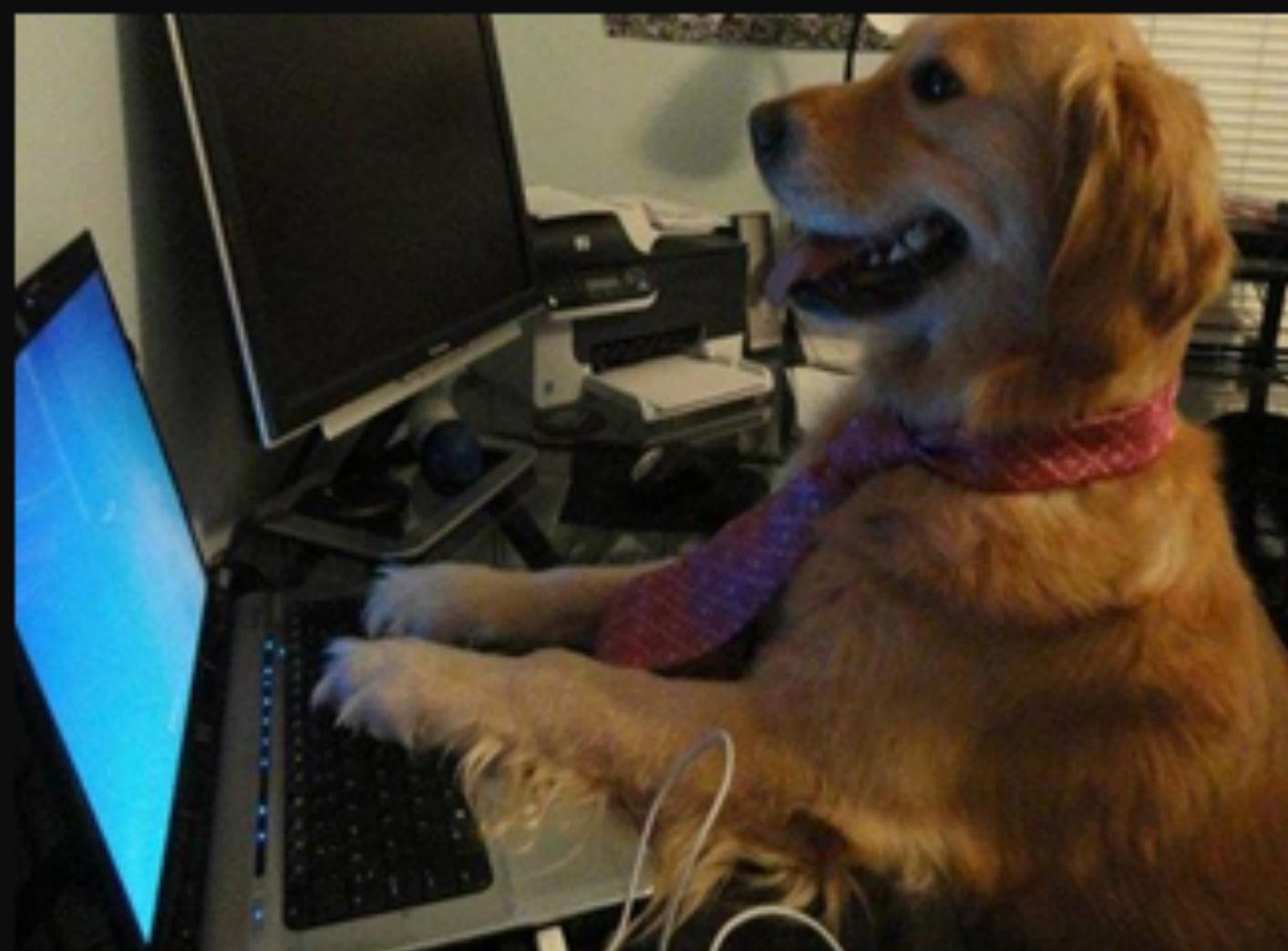
What society thinks I do



What my friends think I do



What other computer
scientists think I do



What mathematicians think I do

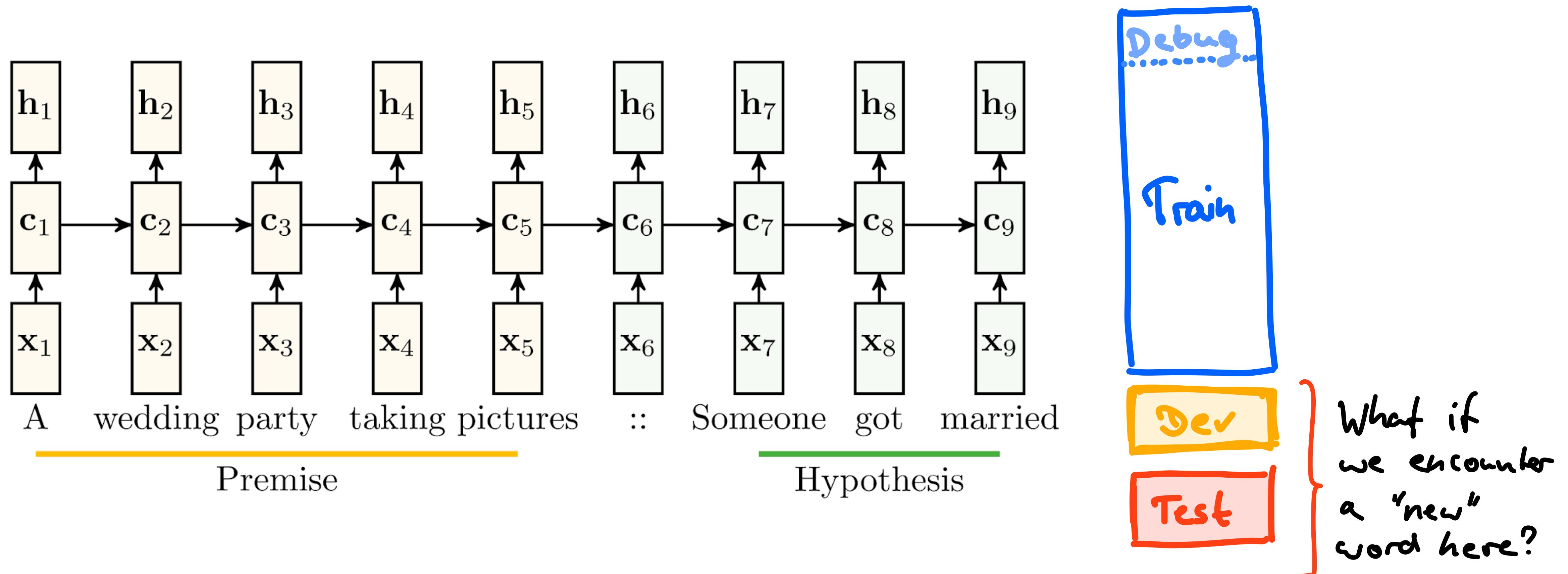


What I think I do

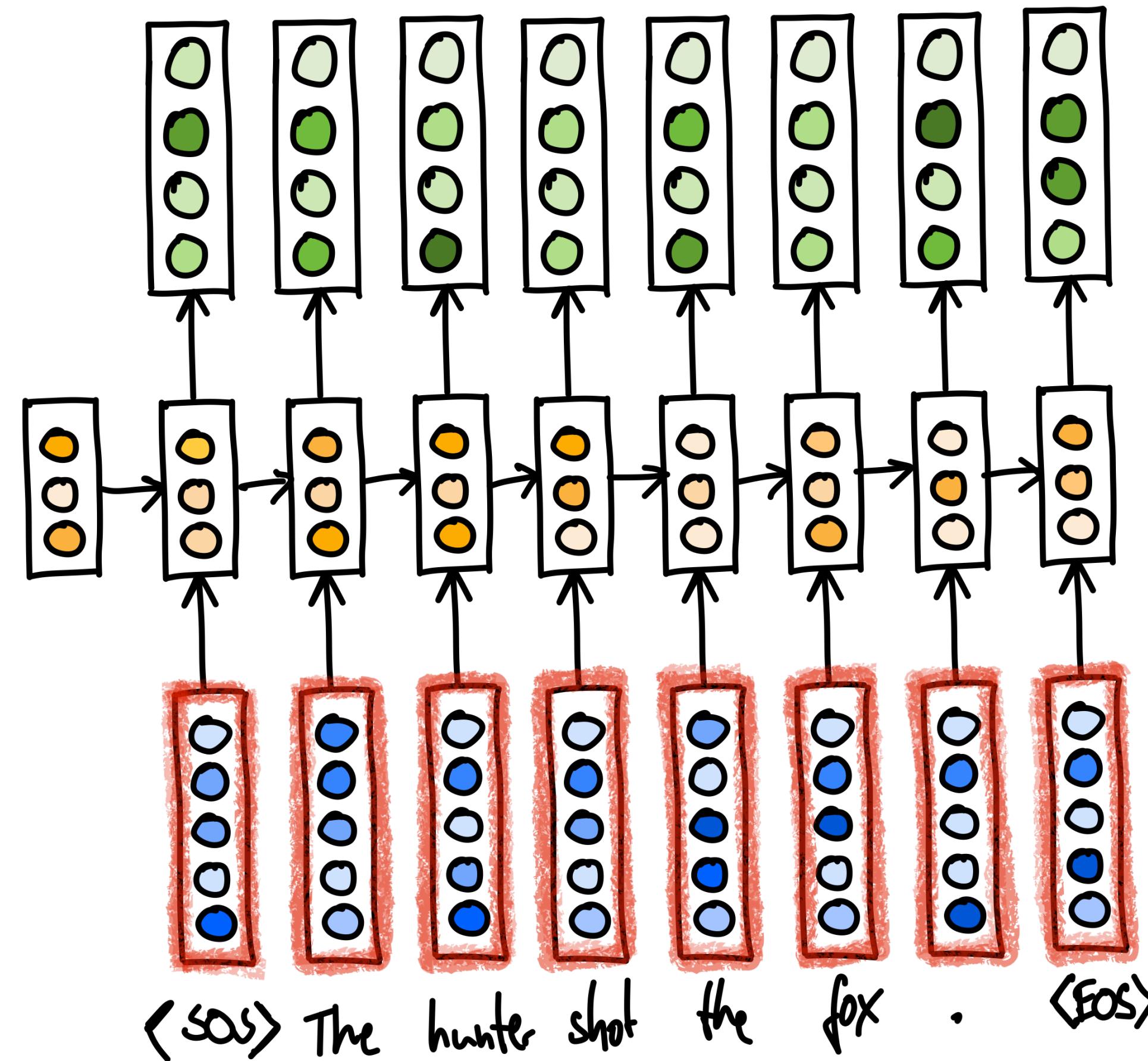
```
from gensim.models import  
Word2vec
```

What I actually do

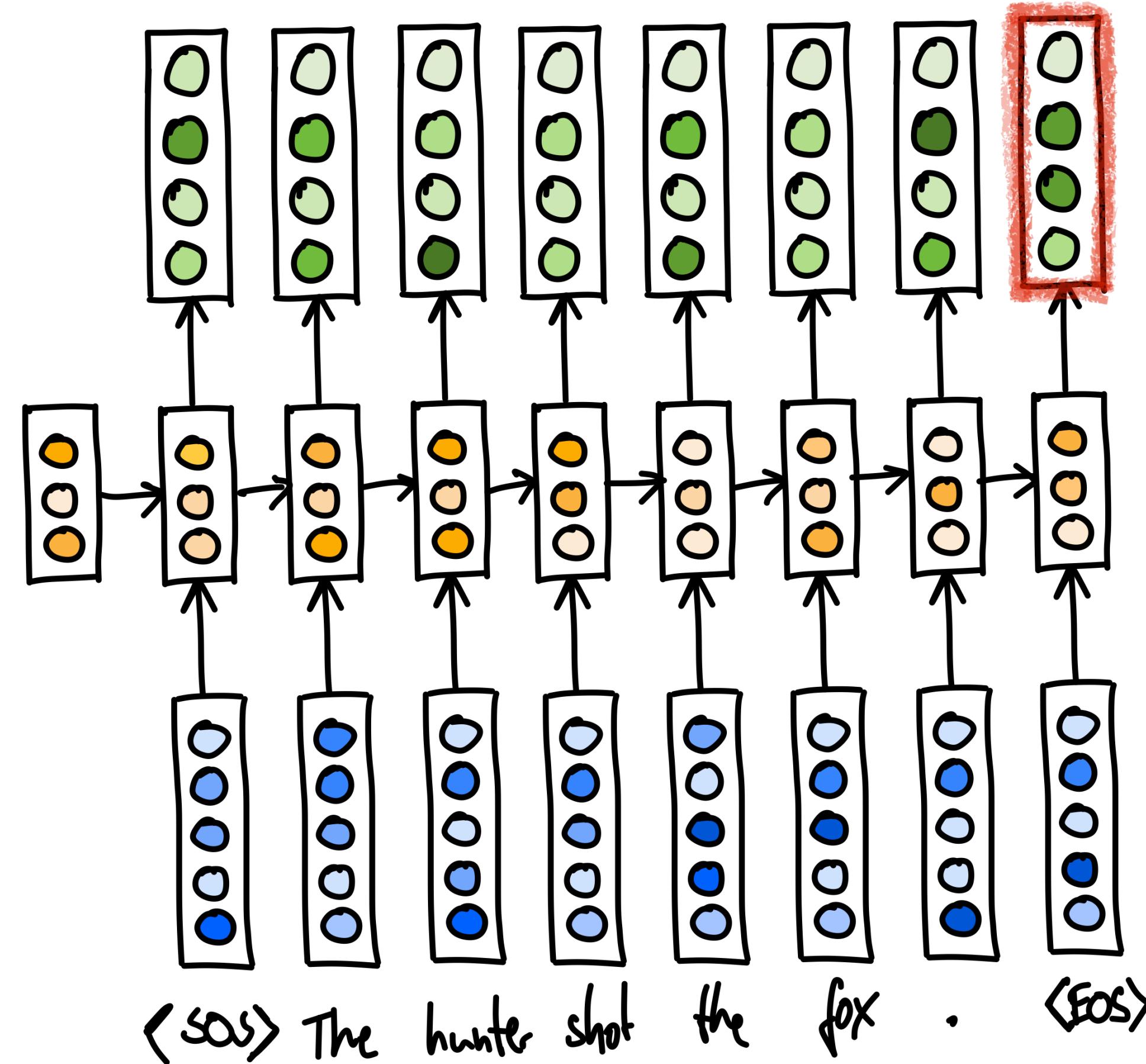
Fine-tuning on the downstream task?



Beyond Word Representations?



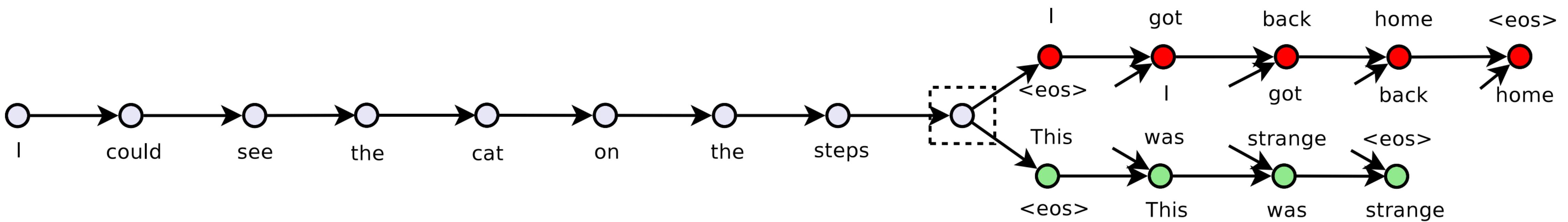
Beyond Word Representations?



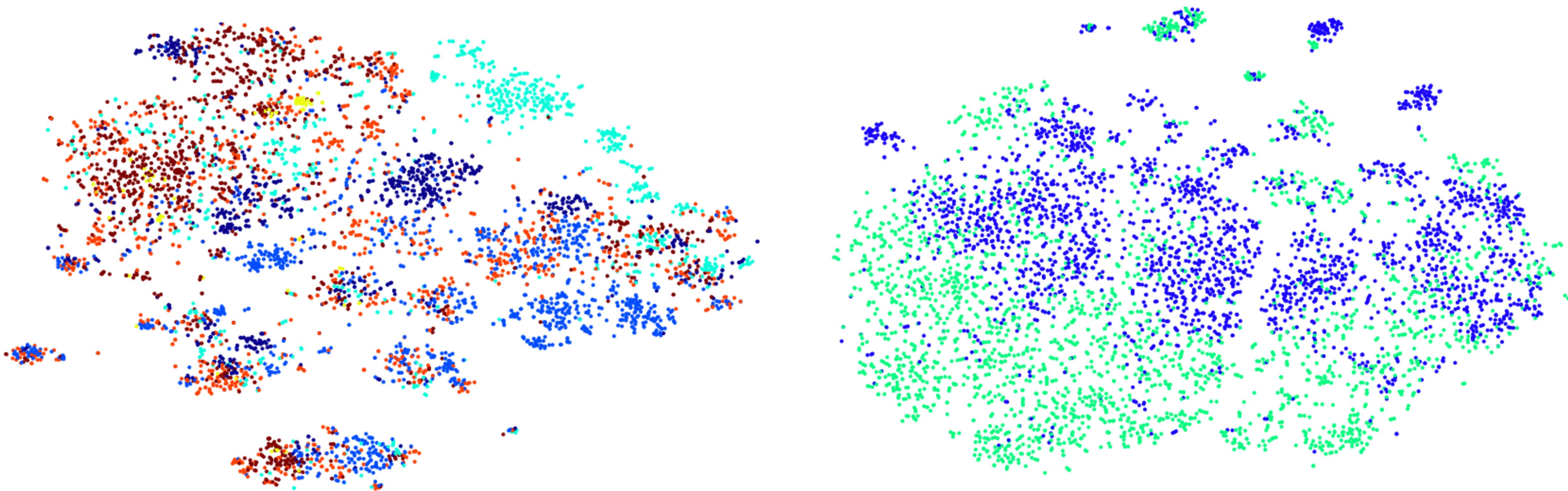
Pre-training Sentence Representations

"Skip-Thought Vectors"

- Large sentence encoder trained on the One Billion Book Corpus
 - 11k books
 - 74M sentences
 - 984M words
 - 1.3M unique words
 - Idea: encode a sentence and generate previous and next sentence



Disentanglement for Downstream Task



TREC

SUBJ

Nearest Neighbor Search

Q: he ran his hand inside his coat , double-checking that the unopened letter was still there .

N: he slipped his hand between his coat and his shirt , where the folded copies lay in a brown envelope .

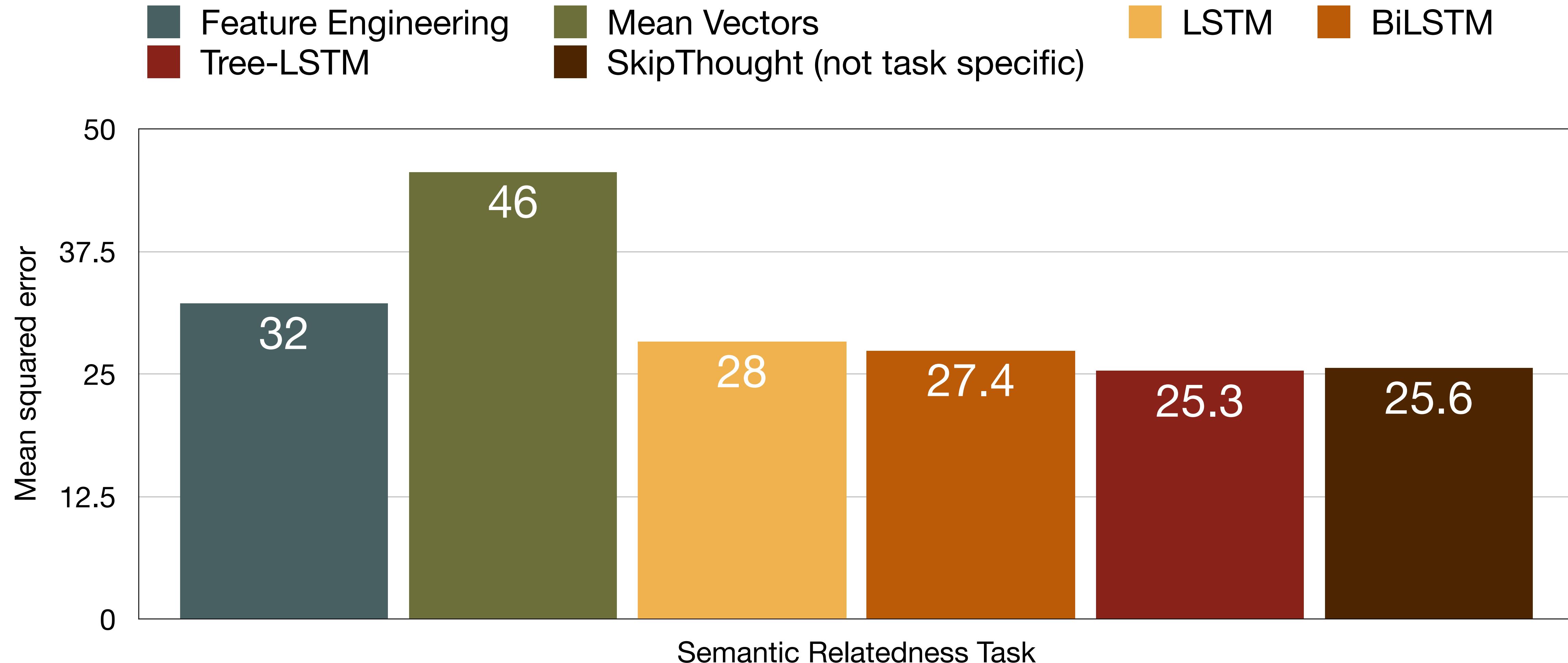
Q: im sure youll have a glamorous evening , she said , giving an exaggerated wink .

N: im really glad you came to the party tonight , he said , turning to her .

Q: although she could tell he had n't been too invested in any of their other chitchat , he seemed genuinely curious about this .

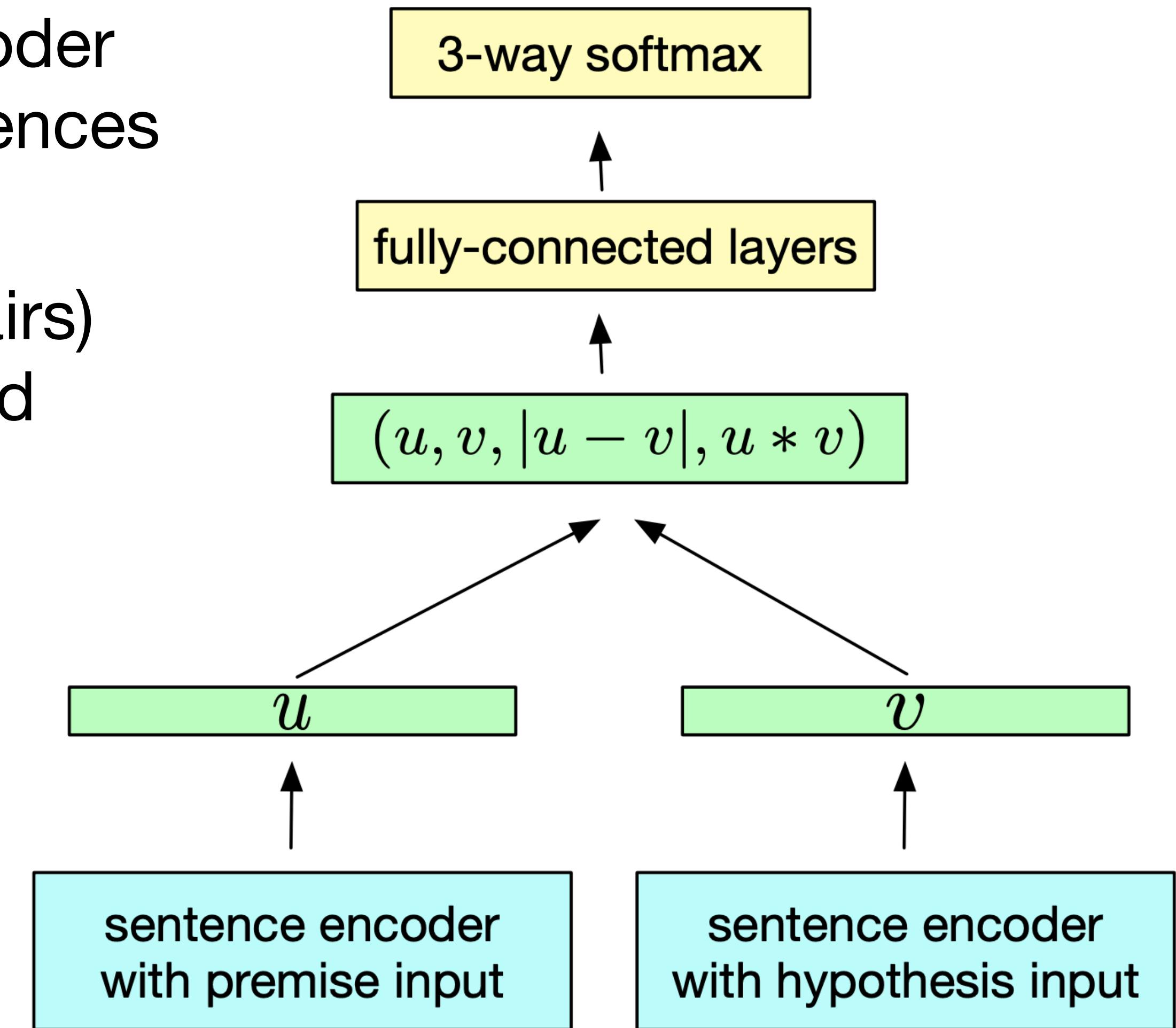
N: although he had n't been following her career with a microscope , he 'd definitely taken notice of her appearances .

Results

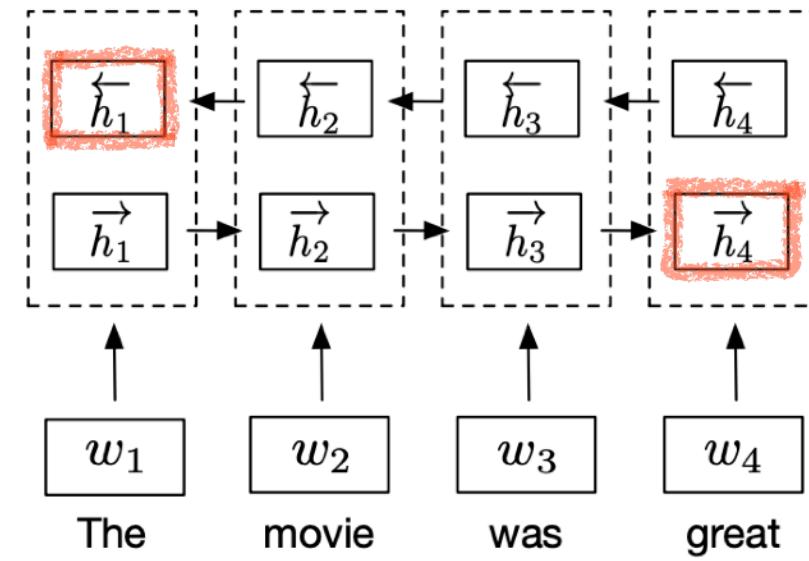


InferSent

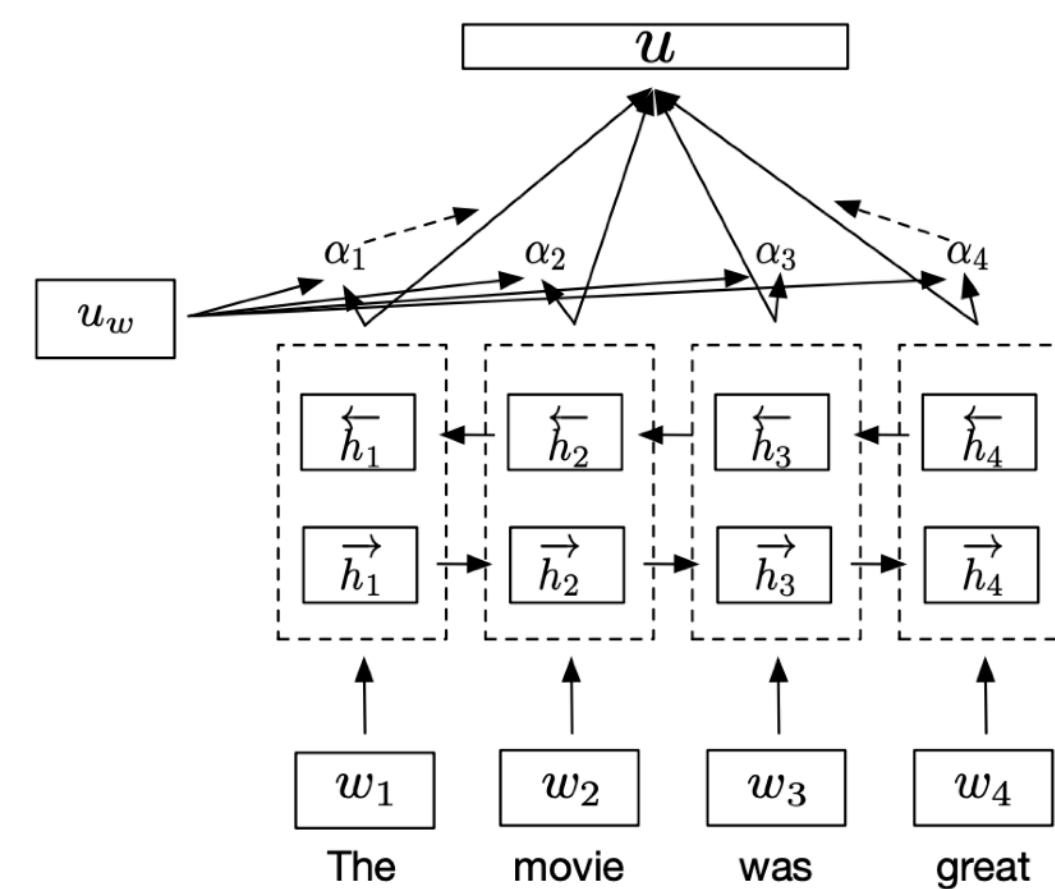
- Unsupervised training a sentence encoder and decoder on a corpus of 74M sentences is expensive
- Can we use a small (0.5M sentence pairs) supervised dataset (SNLI) to learn good sentence representations?



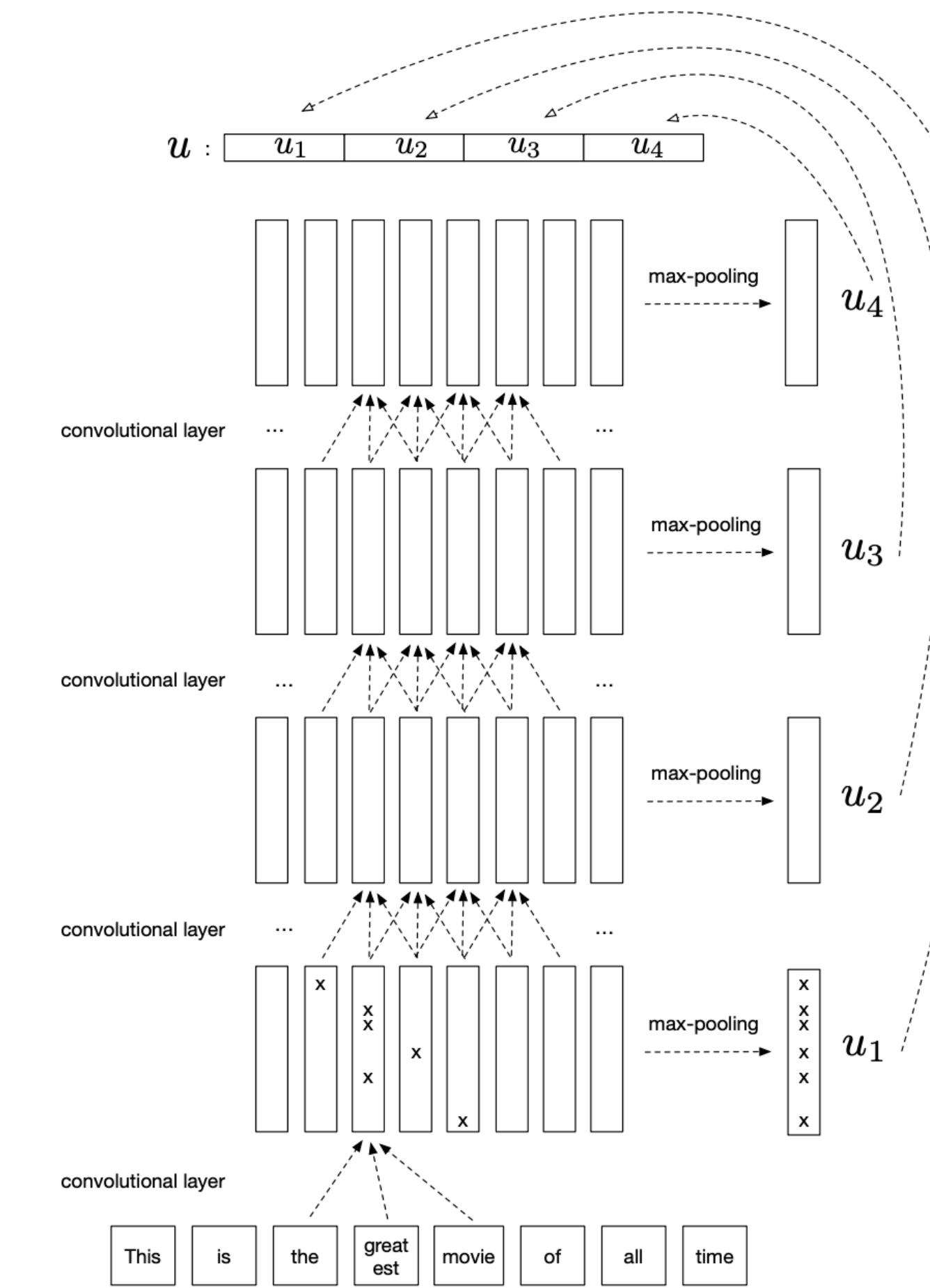
Different ways to build a Sentence Representation



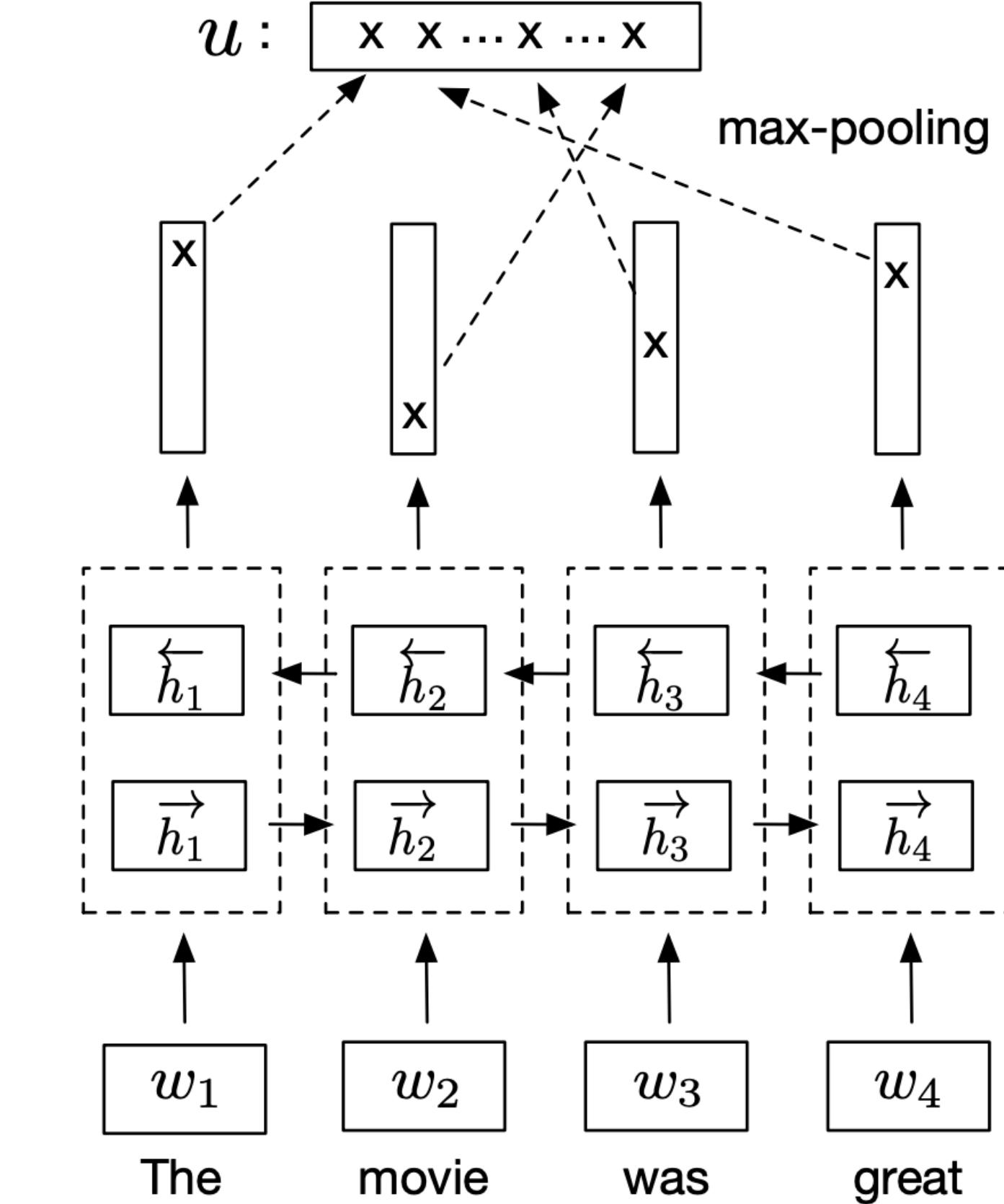
Bi GRU



Self Attention



HConvNet



BiLSTM - Max

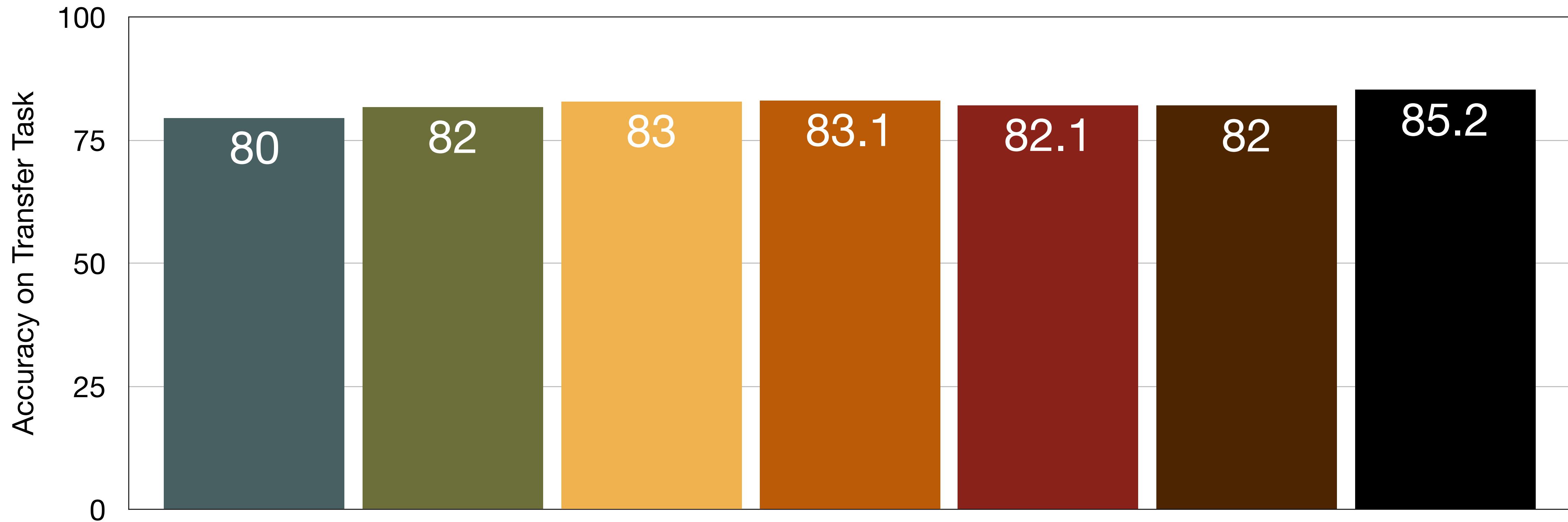
SentEval Benchmark

- Benchmark for evaluating fixed-length sentence representations
- One common evaluation pipeline with fixed hyperparameters
 - You provide a way to map a sentence to a vector
 - They learn a logistic regressor using the sentence representations

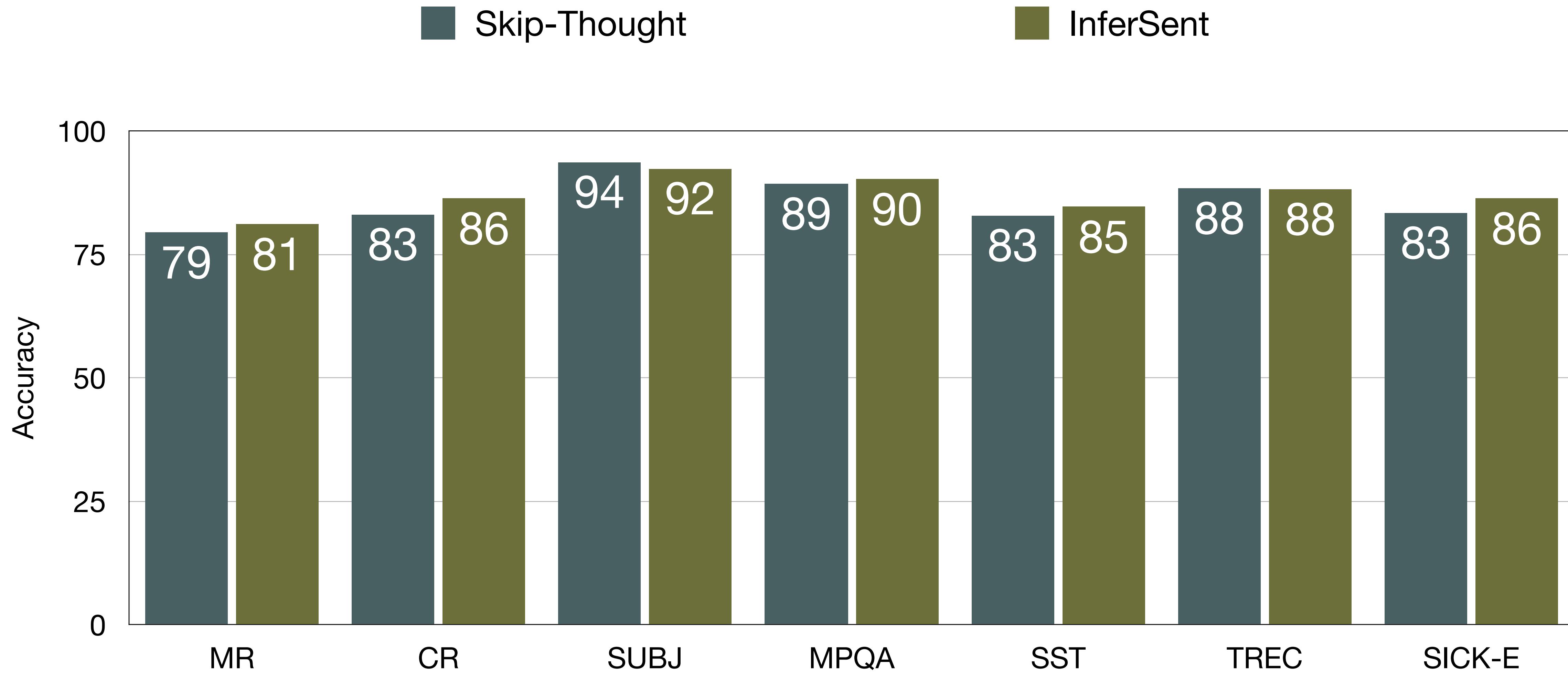
name	N	task	C	examples
MR	11k	sentiment (movies)	2	"Too slow for a younger crowd , too shallow for an older one." (neg)
CR	4k	product reviews	2	"We tried it out christmas night and it worked great ." (pos)
SUBJ	10k	subjectivity/objectivity	2	"A movie that doesn't aim too high , but doesn't need to." (subj)
MPQA	11k	opinion polarity	2	"don't want"; "would like to tell"; (neg, pos)
TREC	6k	question-type	6	"What are the twin cities ?" (LOC:city)
SST	70k	sentiment (movies)	2	"Audrey Tautou has a knack for picking roles that magnify her [..]" (pos)

Results

LSTM GRU BiGRU (last) BiLSTM-Mean Self-Attention
HConvNet BiLSTM-Max



Comparison to Skip-Thought



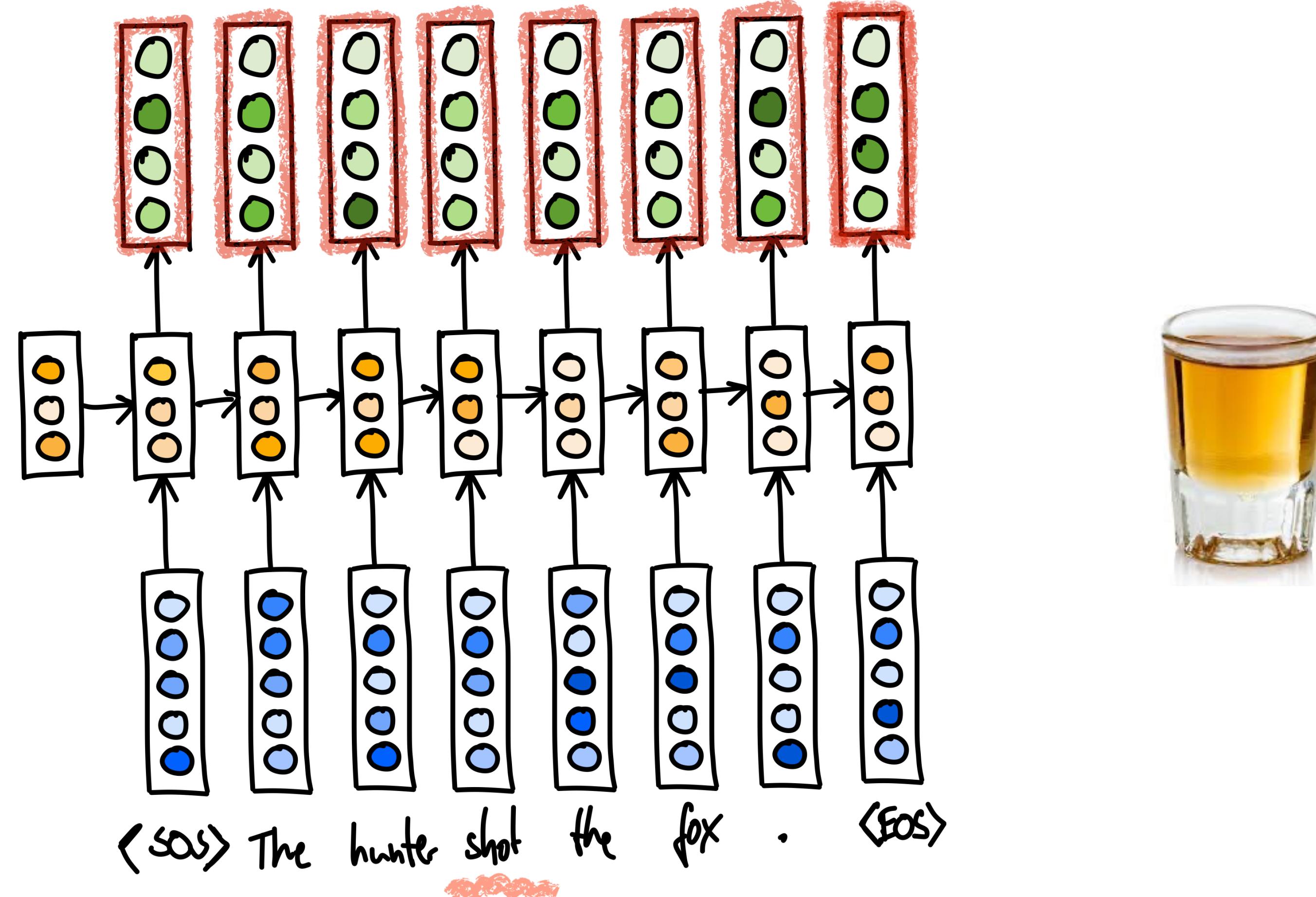
Beyond Word and Sentence Representations?

“You can’t cram the meaning of a whole %&!\$# sentence into a single \$&!#* vector!”

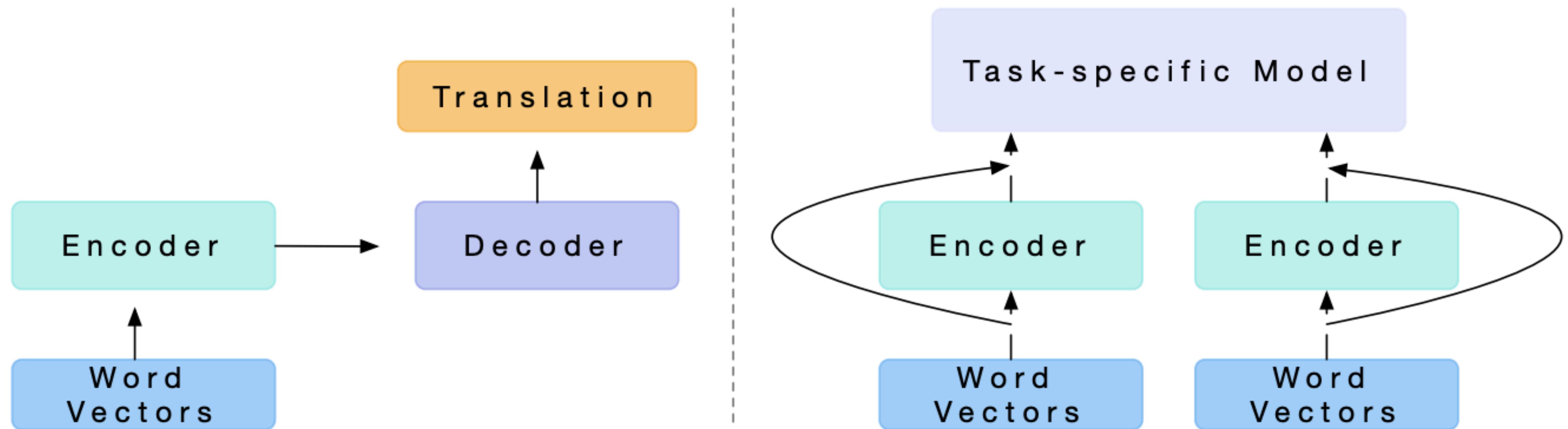
— *Raymond J. Mooney*



Pre-training Words vs Sentences vs Contexts



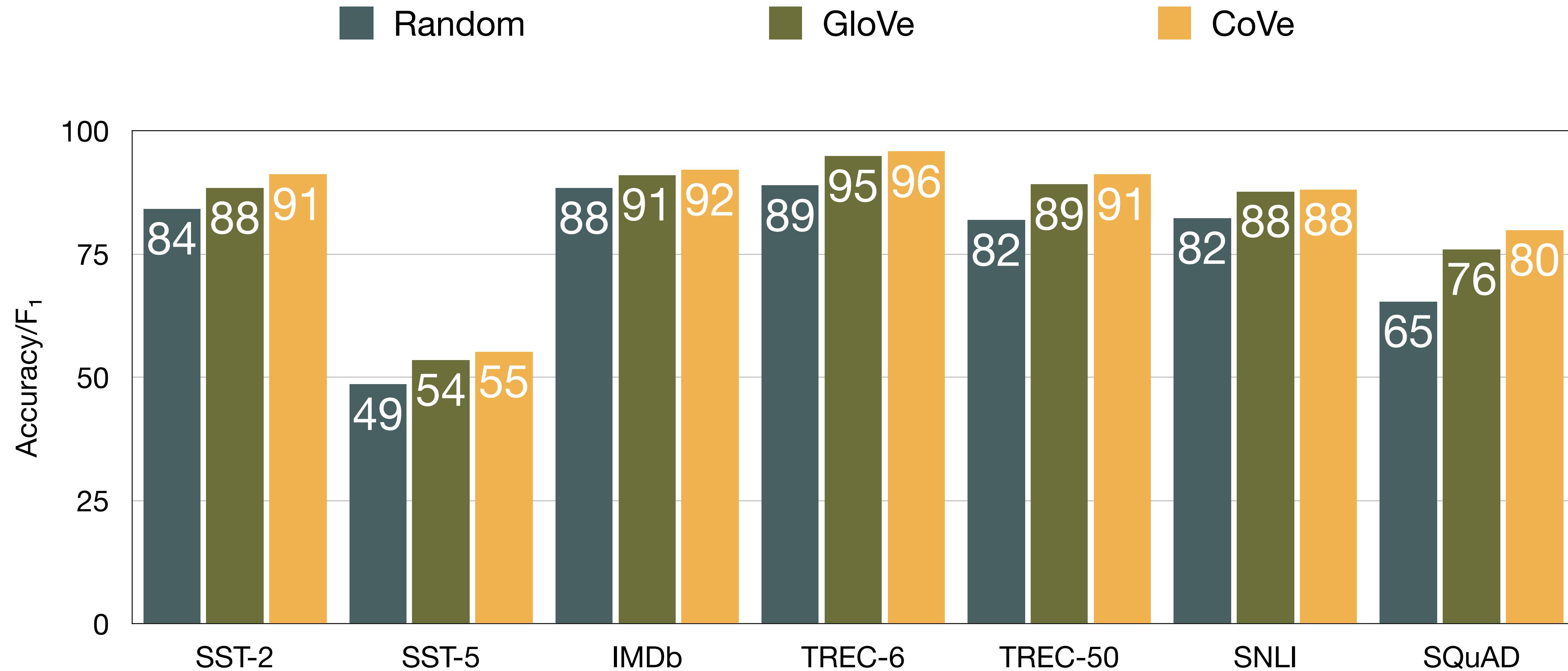
Contextualized Word Vectors (CoVe)



Downstream Tasks

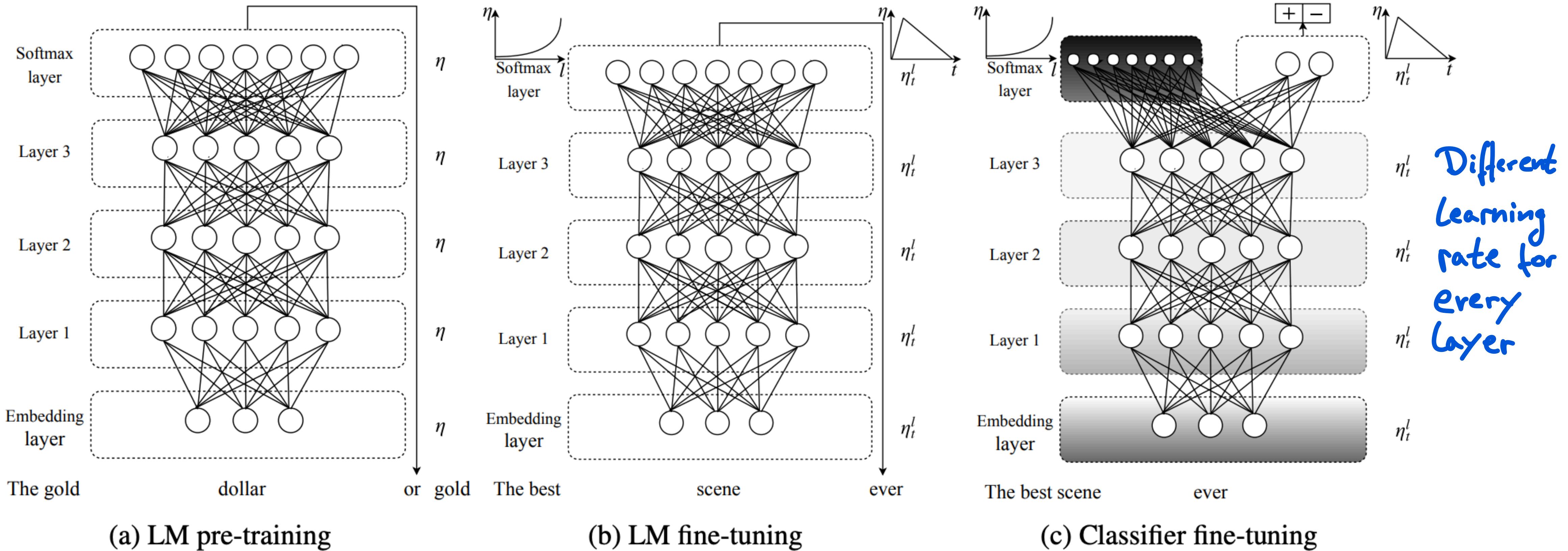
Dataset	Task	Details	Examples
SST-2	Sentiment Classification	2 classes, single sentences	56.4k
SST-5	Sentiment Classification	5 classes, single sentences	94.2k
IMDb	Sentiment Classification	2 classes, multiple sentences	22.5k
TREC-6	Question Classification	6 classes	5k
TREC-50	Question Classification	50 classes	5k
SNLI	Entailment Classification	2 classes	550k
SQuAD	Question Answering	open-ended (answer-spans)	87.6k

Results

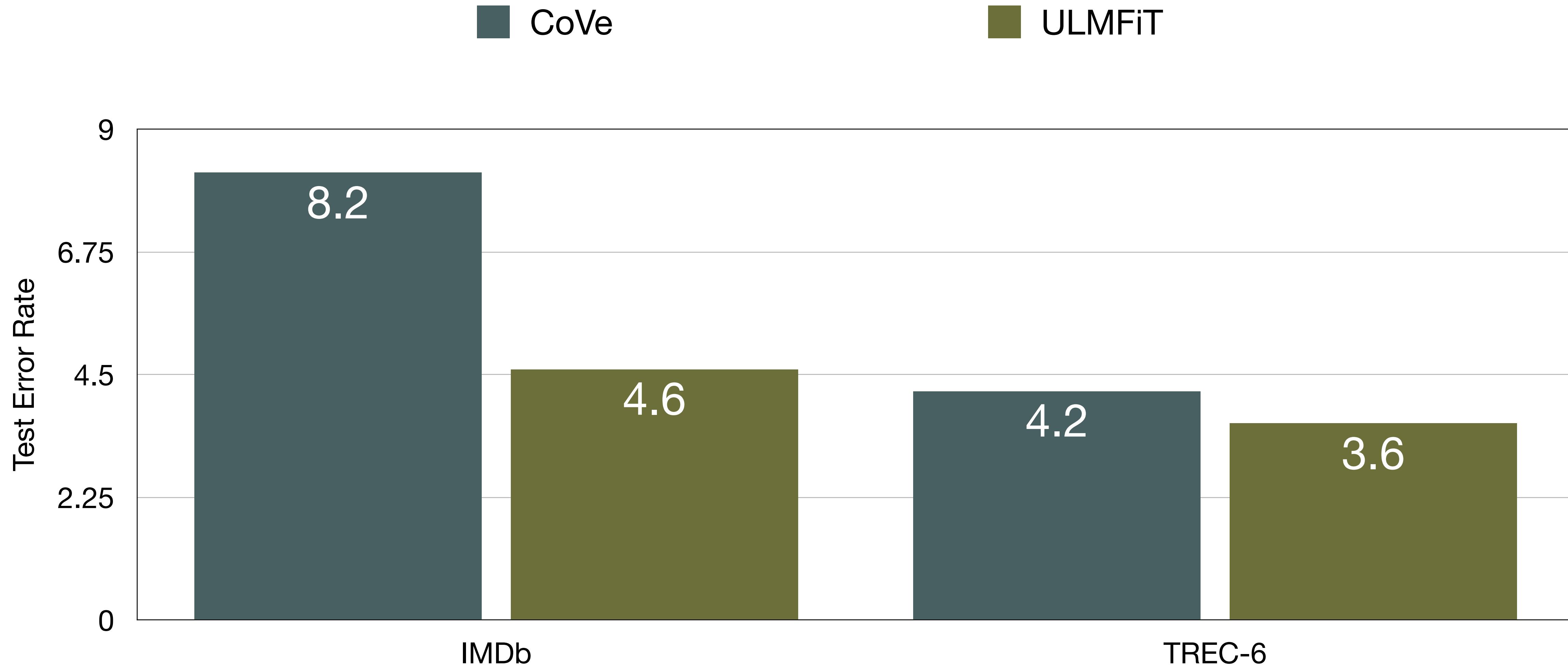


ULMFiT

- Problem with CoVe: limited by the amount of supervised (MT) data
- Can we learn contextual word representations in an unsupervised way?

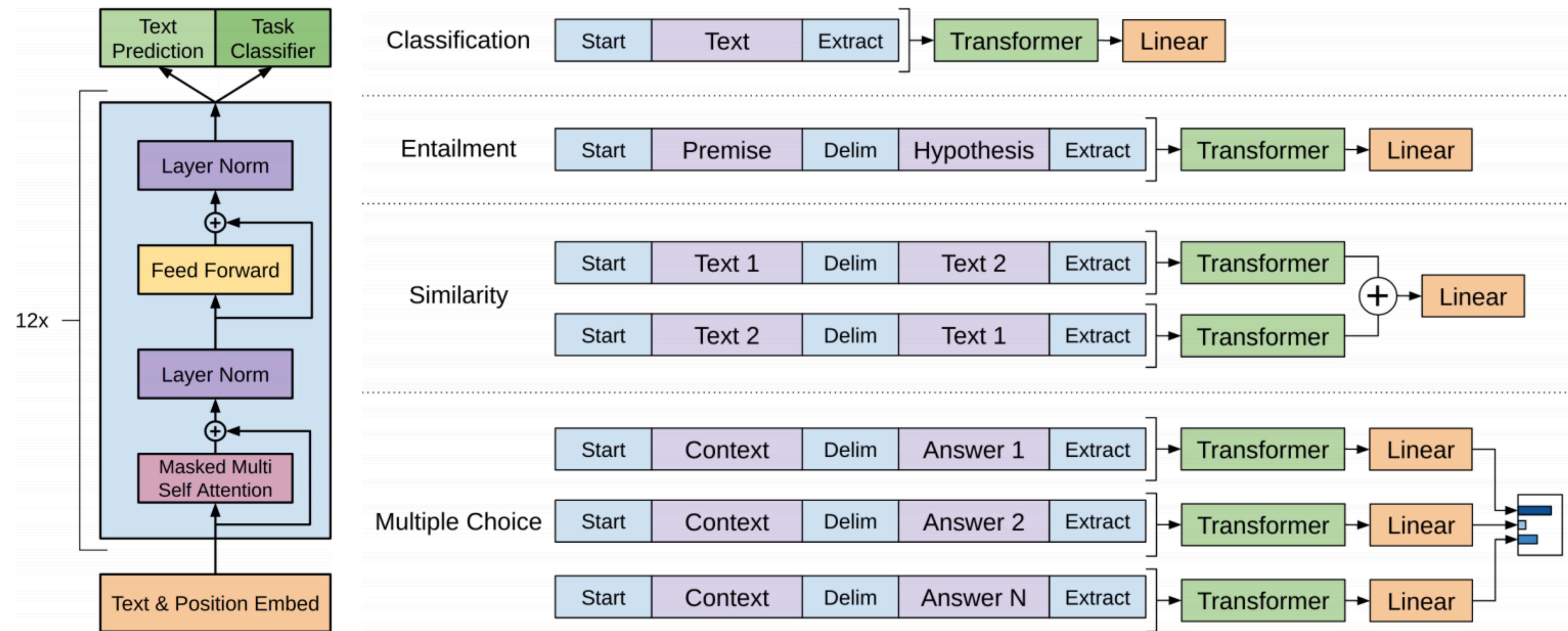


Results

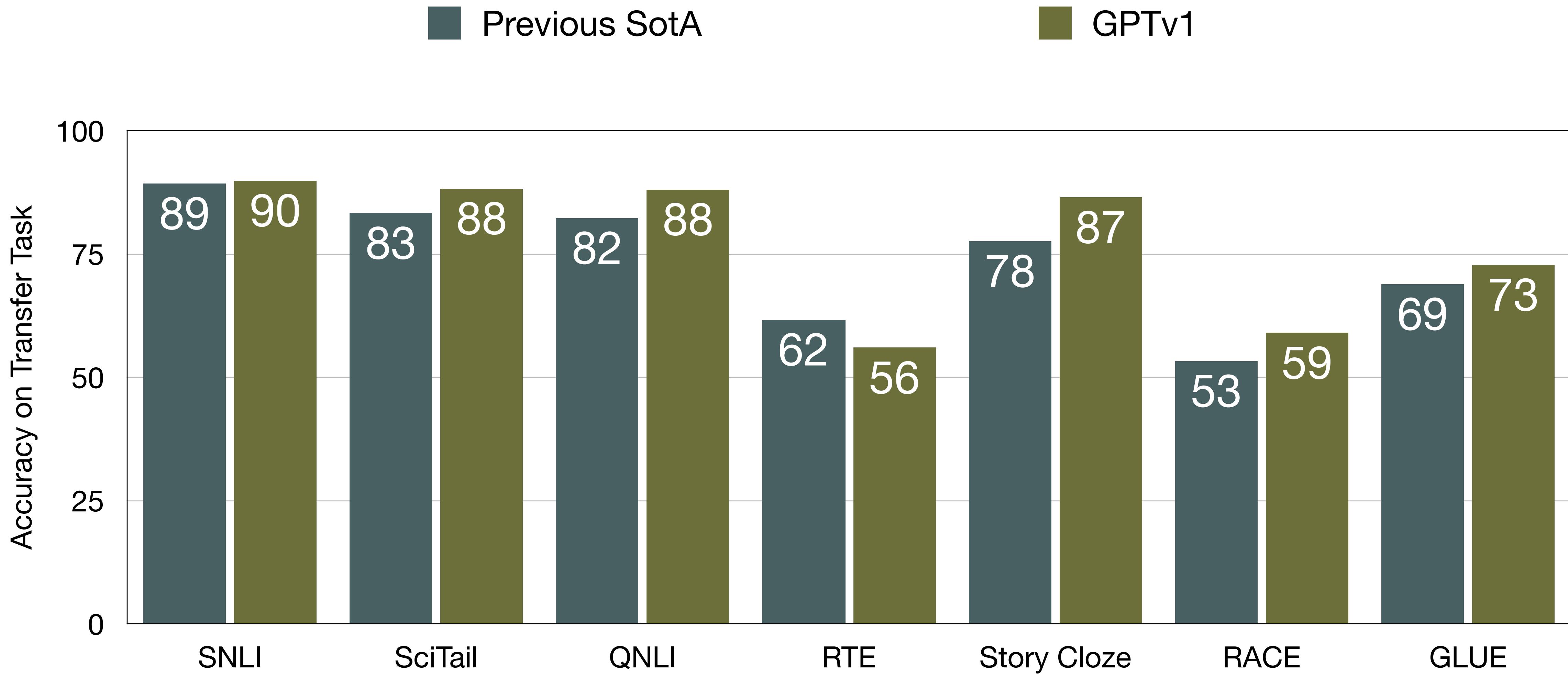


Generative Pre-training (GPTv1)

- Pre-train a large transformer language model on a lot of data
- Fine-tune on downstream task



Results



ELMo

- Train bidirectional RNN language model
- Use output of BiRNN as input to downstream models (difference to GPTv1 which fine-tunes the model)
- Bidirectional language modeling???

$$p(w_1, w_2, \dots, w_N) = \prod_{i=1}^N p_\theta(w_t | w_{t-1}, w_{t-2}, \dots, w_1)$$

$$\vec{p}_\theta(w_t | w_{t-1}, w_{t-2}, \dots, w_1) = \text{softmax}(\vec{\mathbf{V}} \vec{\mathbf{h}}_t + \mathbf{b})$$

$$\overleftarrow{p}_\theta(w_t | w_{t+1}, w_{t+2}, \dots, w_N) = \text{softmax}(\overleftarrow{\mathbf{V}} \overleftarrow{\mathbf{h}}_t + \mathbf{b})$$

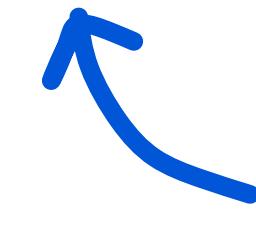
$$L(w, \hat{w}) = -\frac{1}{N} \sum_i \hat{y}_{\hat{w}}^T \log \vec{p}_\theta(w_i | w_{i-1}, \dots, w_1) + \hat{y}_{\hat{w}}^T \log \overleftarrow{p}_\theta(w_i | w_{i+1}, \dots, w_N)$$

Task-specific adaptation

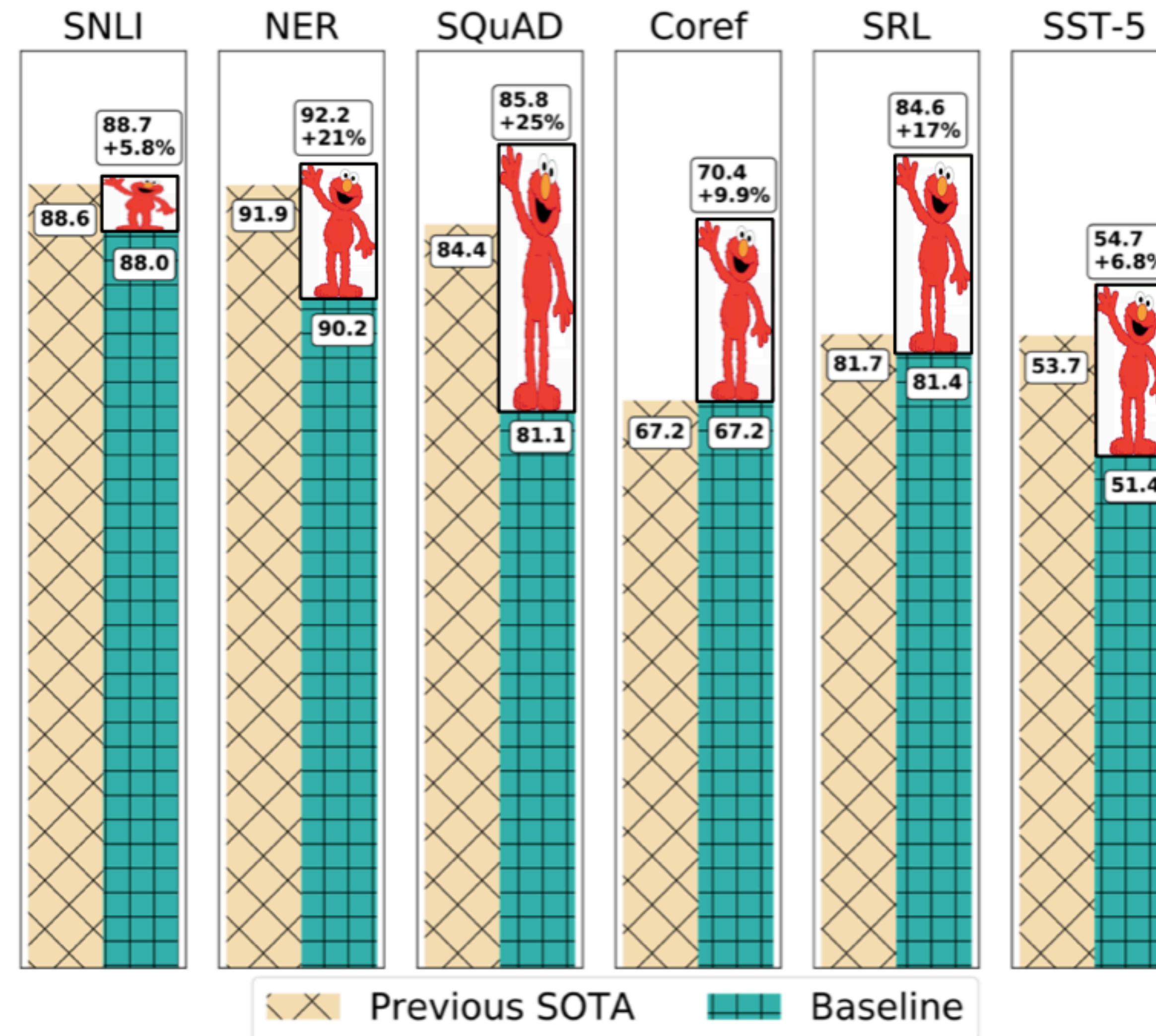
- ELMo has L layers
- At time step t, we have $2L+1$ vector representations
 - $2L$ due to BiRNN
 - $+1$ due to token representations
- Single vector at step t for downstream task T is calculated as

$$h_t^{\text{ELMo}} = \gamma_T \sum_{j=0}^L s_j^T \begin{bmatrix} \vec{h}_t^j; \vec{h}_t^j \end{bmatrix}$$

Scale whole vector 

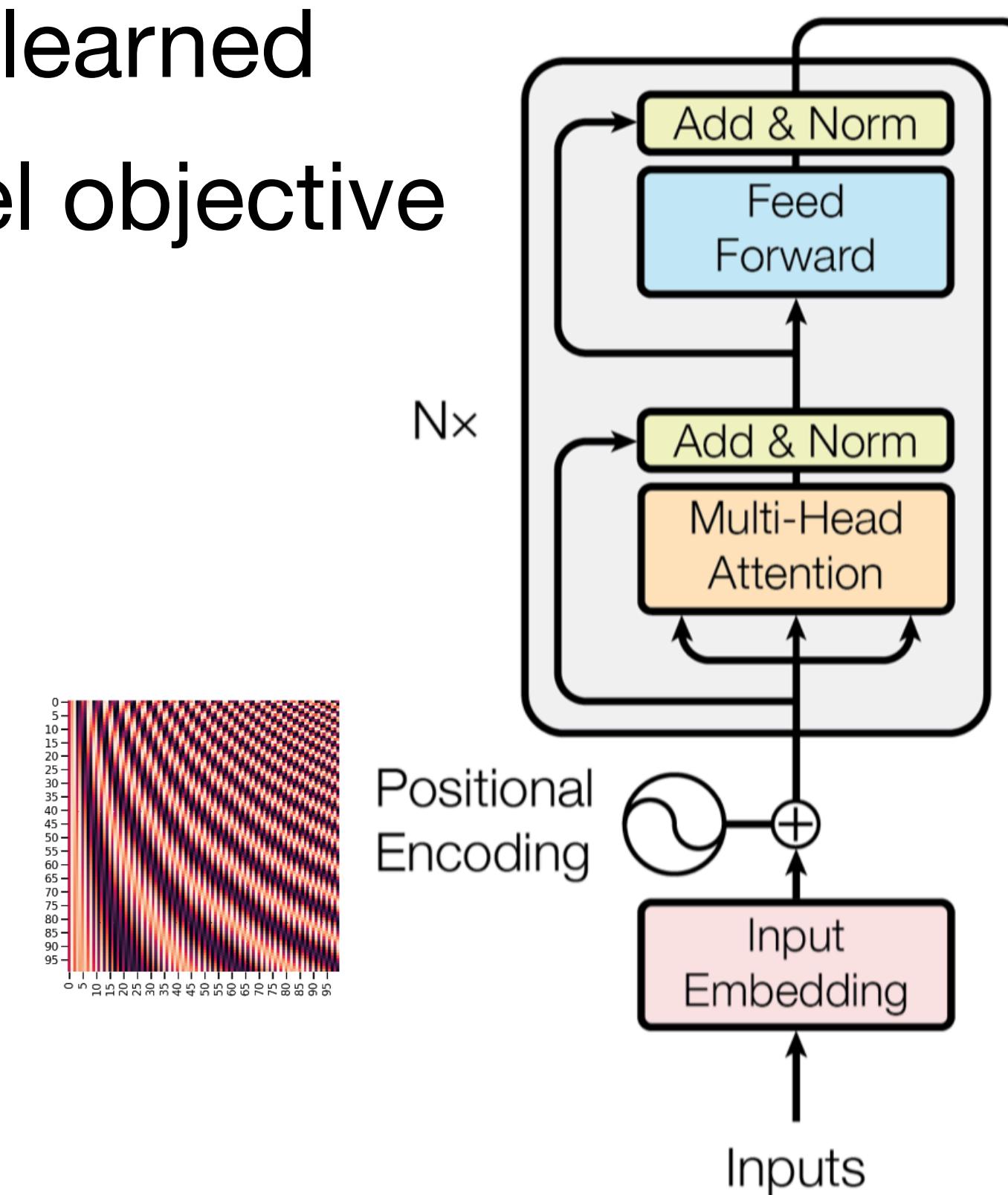
Layer-dependent scaling 

Results

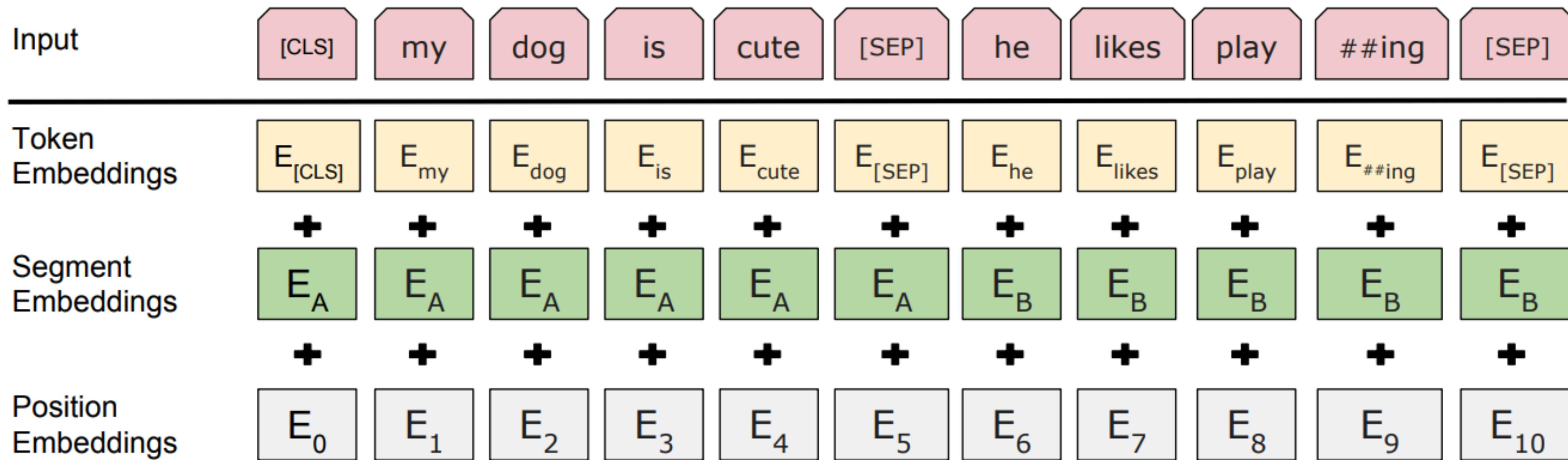


Bidirectional Encoder Representations from Transformers (BERT)

- Difference to GPTv1: Bi-directional
- Change to vanilla Transformer: positional embedding is learned
- Difference to ELMo: Not directly using a language model objective
 - Sample 15% of tokens to be replaced
 - 80% of the time, replace token with [MASK]
 - 10% of the time, replace with random word
 - 10% of the time, keep word unchanged



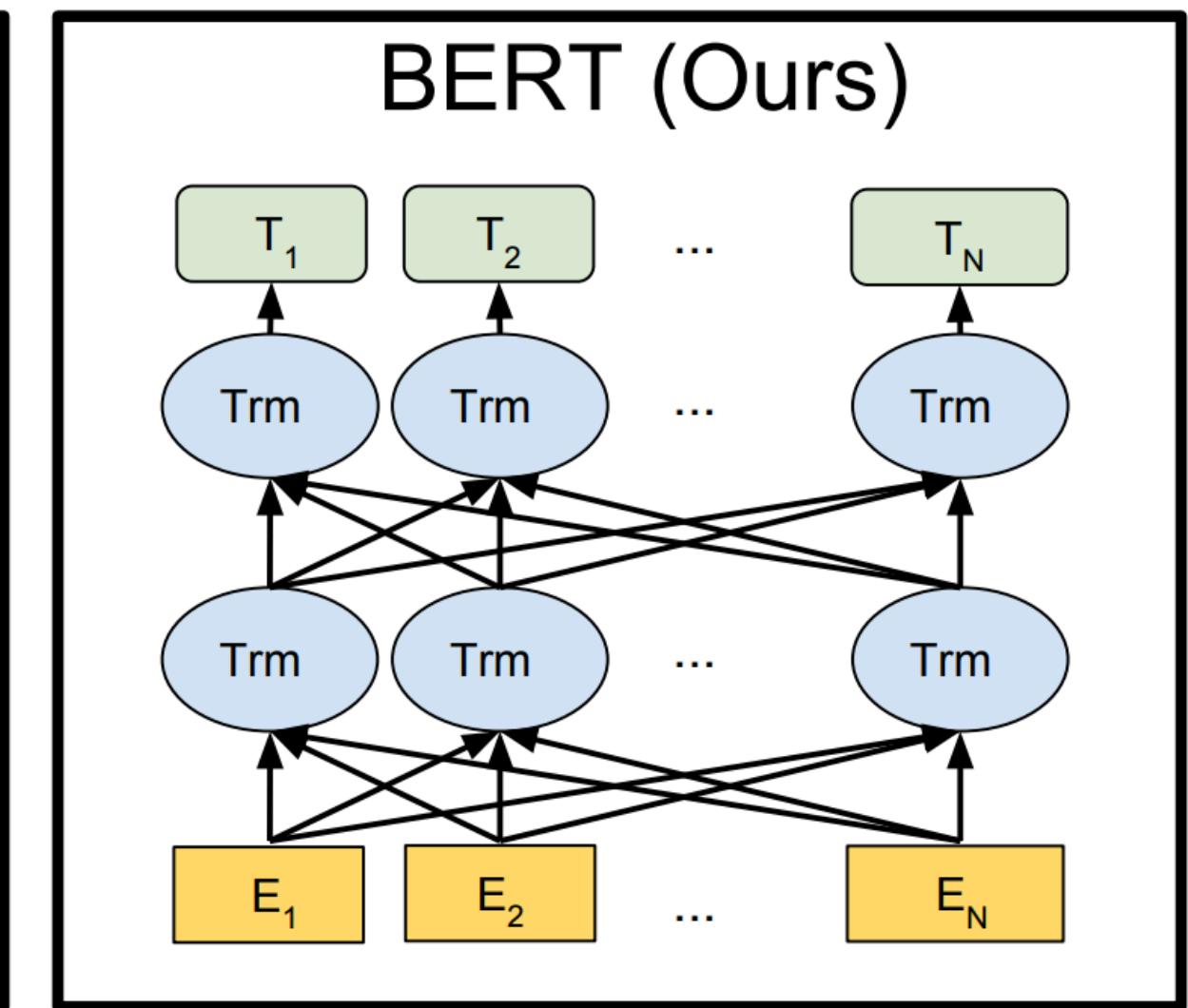
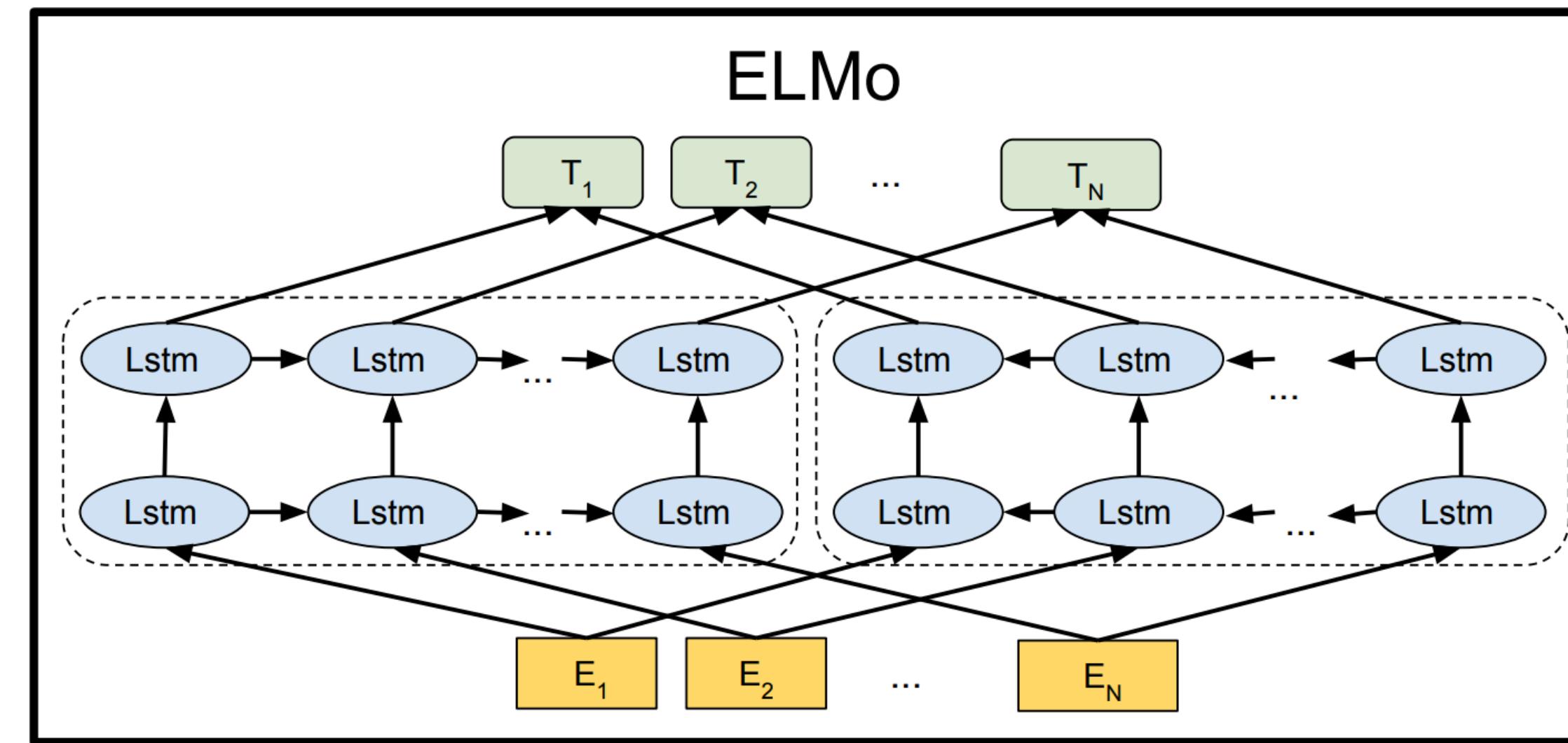
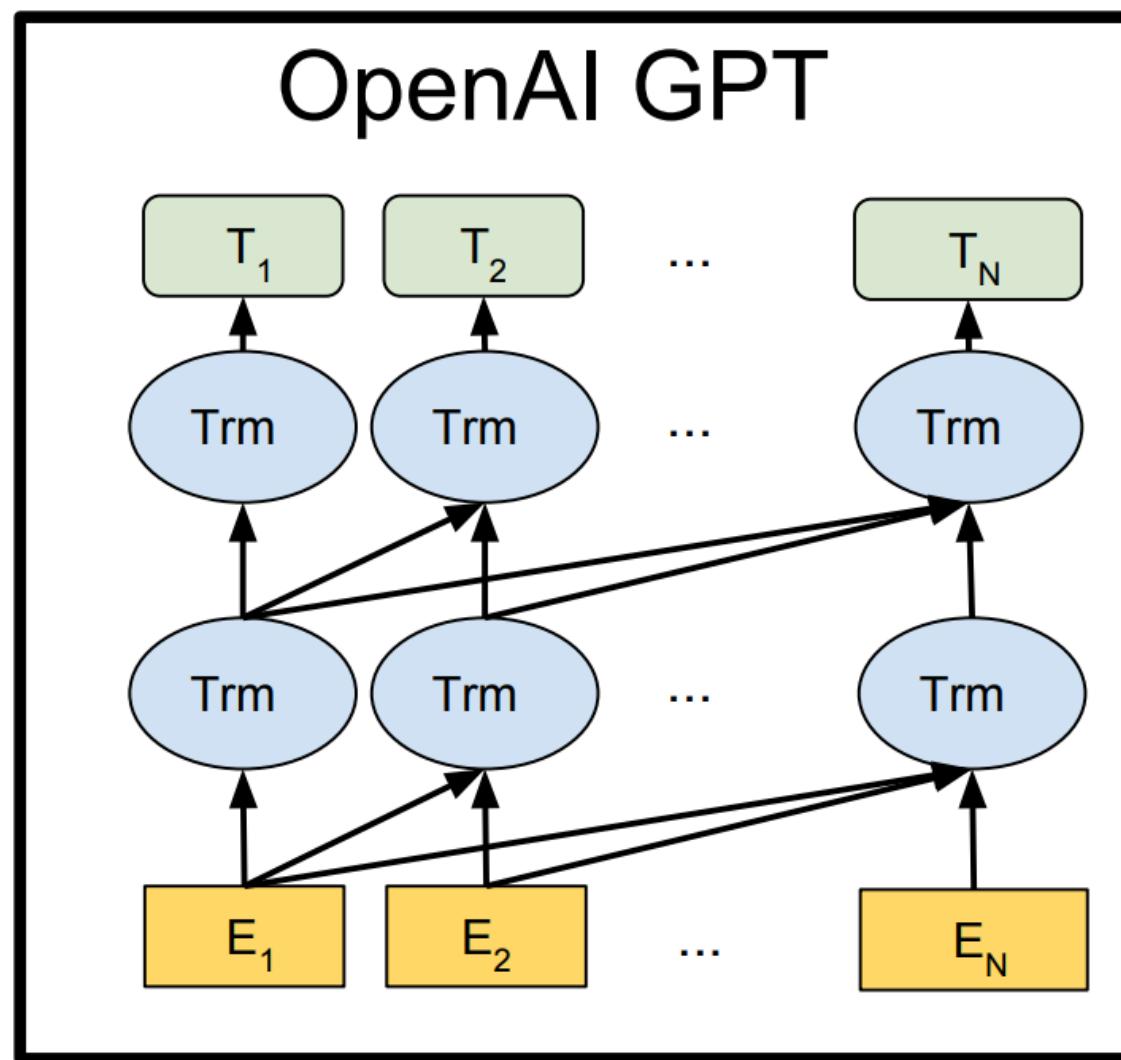
BERT Input Representation



What does BERT "know"?

template	answer	predictions
Francesco Bartolomeo Conti was born in [MASK]. Adolphe Adam died in [MASK].	Florence	Rome [-1.8] , Florence [-1.8] , Naples [-1.9] , Milan [-2.4] , Bologna [-2.5]
English bulldog is a subclass of [MASK].	Paris	Paris [-0.5] , London [-3.5] , Vienna [-3.6] , Berlin [-3.8] , Brussels [-4.0]
The official language of Mauritius is [MASK].	dog	dogs [-0.3] , breeds [-2.2] , dog [-2.4] , cattle [-4.3] , sheep [-4.5]
Patrick Oboya plays in [MASK] position.	English	English [-0.6] , French [-0.9] , Arabic [-6.2] , Tamil [-6.7] , Malayalam [-7.0]
Alfred John Shout was awarded the [MASK].	midfielder	centre [-2.0] , center [-2.2] , midfielder [-2.4] , forward [-2.4] , midfield [-2.7]
James Merry was educated at the University of [MASK].	VC	medal [-2.0] , MBE [-2.1] , OBE [-2.2] , DSO [-2.4] , VC [-2.5]
Albanian shares border with [MASK].	Glasgow	Edinburgh [-1.0] , Glasgow [-1.6] , Aberdeen [-2.5] , Cambridge [-2.6] , Oxford [-2.9]
Hamburg Airport is named after [MASK].	Kosovo	Montenegro [-2.0] , Macedonia [-2.0] , Serbia [-2.5] , Greece [-2.6] , Bulgaria [-2.7]
The original language of Mon oncle Benjamin is [MASK].	Hamburg	Hess [-7.0] , Hermann [-7.1] , Schmidt [-7.1] , Hamburg [-7.5] , Ludwig [-7.5]
Dani Alves plays with [MASK].	French	French [-0.2] , Breton [-3.3] , English [-3.8] , Dutch [-4.2] , German [-4.9]
FPF is a member of [MASK].	Barcelona	Santos [-2.4] , Porto [-2.5] , Sporting [-3.1] , Brazil [-3.3] , Portugal [-3.7]
[MASK] participated in the Mundialito.	UEFA	NATO [-1.1] , UNESCO [-2.7] , IUCN [-3.4] , UEFA [-3.6] , FIFA [-3.7]
Paul Toungui is a [MASK] by profession .	Argentina	He [-0.4] , It [-2.8] , They [-3.4] , She [-3.8] , Also [-3.9]
sodium sulfide consists of [MASK].	politician	lawyer [-1.1] , journalist [-2.4] , teacher [-2.7] , doctor [-3.0] , physician [-3.7]
Gordon Scholes is a member of the [MASK] political party.	sodium	water [-1.2] , sulfur [-1.7] , sodium [-2.5] , zinc [-2.8] , salt [-2.9]
Kenya maintains diplomatic relations with [MASK].	Labor	Labour [-1.3] , Conservative [-1.6] , Green [-2.4] , Liberal [-2.9] , Labor [-2.9]
iPod Touch is produced by [MASK].	Uganda	India [-3.0] , Uganda [-3.2] , Tanzania [-3.5] , China [-3.6] , Pakistan [-3.6]
Thomas Earle is [MASK] citizen.	Apple	Apple [-1.6] , Nokia [-1.7] , Sony [-2.0] , Samsung [-2.6] , Intel [-3.1]
	American	American [-3.6] , US [-4.0] , honorary [-5.1] , Canadian [-5.7] , British [-5.9]

GPTv1 vs ELMo vs BERT



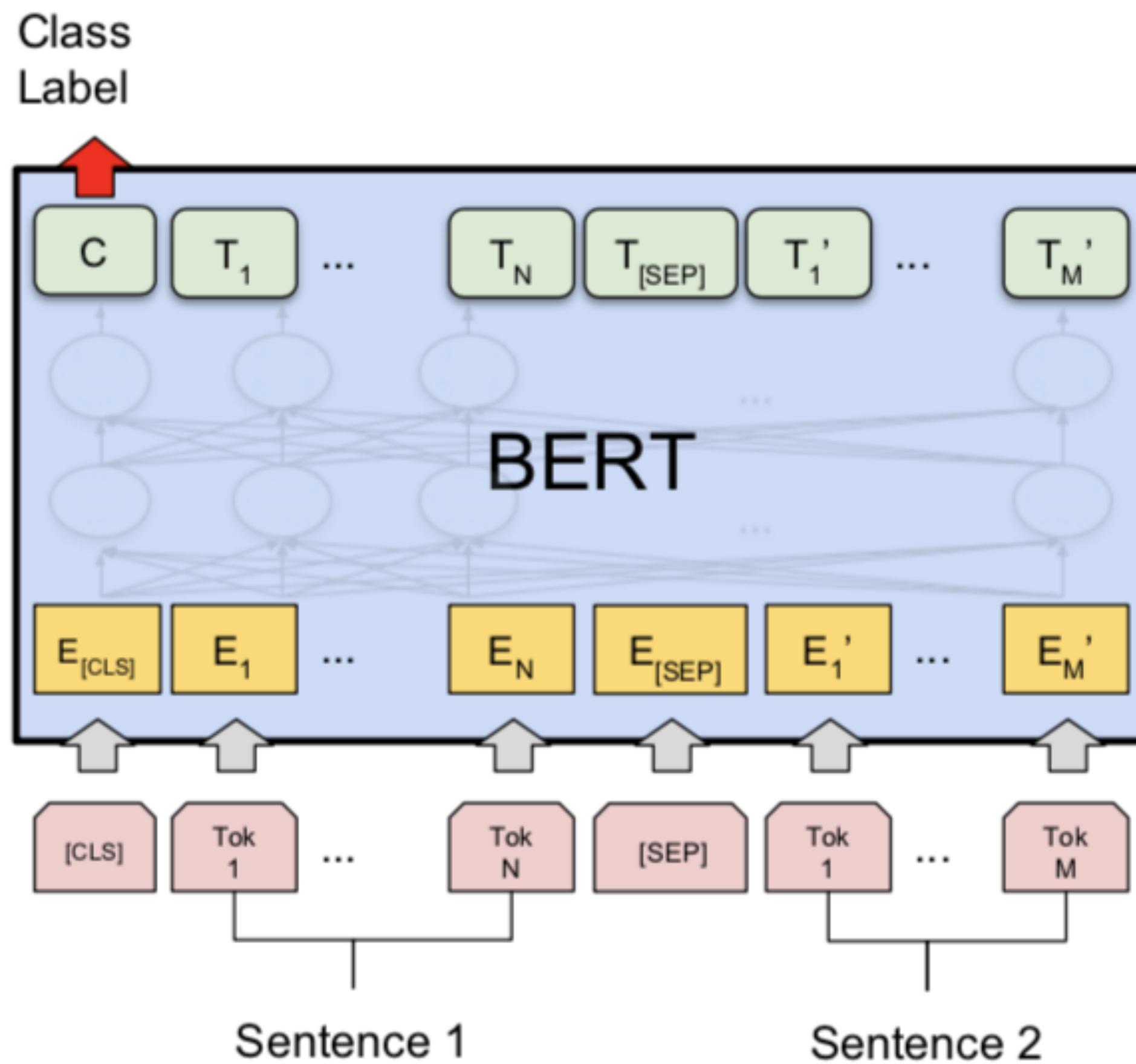
Training Data Comparison

Model	Base Model	#Parameters	Training Corpus	Corpus Size
ELMo (original) (Peters et al., 2018a)	BiLSTM	93.6M	Google Billion Word	800M Words
ELMo 5.5B (Peters et al., 2018a)	BiLSTM	93.6M	Wikipedia (en) & WMT 2008-2012	5.5B Words
BERT (base) (Devlin et al., 2018)	Transformer	110M	Wikipedia (en) & BookCorpus	3.3B Words
BERT (large) (Devlin et al., 2018)	Transformer	340M	Wikipedia (en) & BookCorpus	3.3B Words
GPTv2 (large) (Radford et al., 2019b)	Transformer	1.5B	WebText	8M Documents

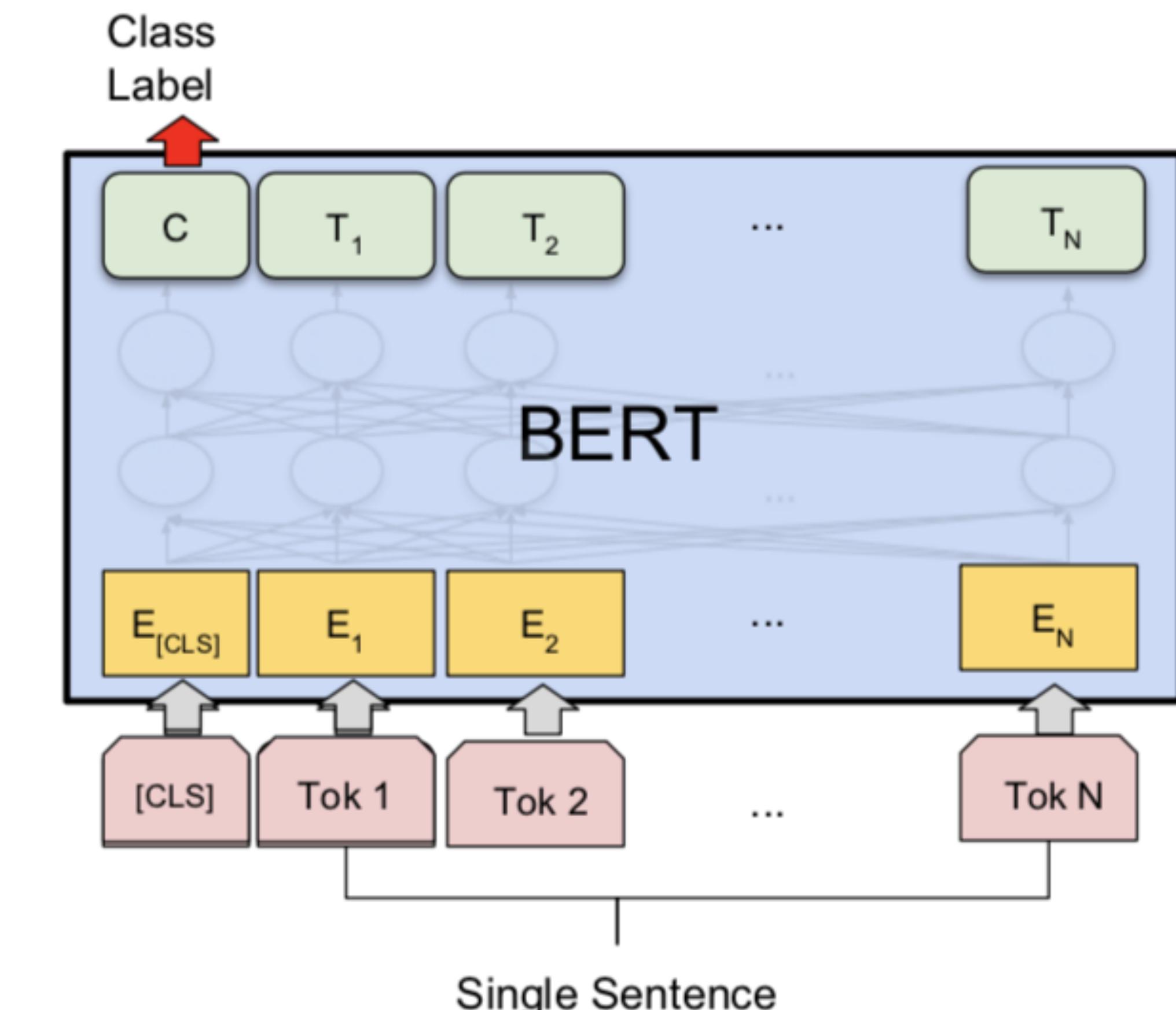
GLUE Benchmark

Corpus	Train	Dev	Test	Task	Metric	Domain
Single-Sentence Tasks						
CoLA	10k	1k	1.1k	acceptability	Matthews	linguistics literature
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks						
MRPC	4k	N/A	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman	misc.
QQP	400k	N/A	391k	paraphrase	acc./F1	social QA Questions
Inference Tasks						
MNLI	393k	20k	20k	NLI	acc. (match/mismatch)	misc.
QNLI	108k	11k	11k	QA/NLI	acc.	Wikipedia
RTE	2.7k	N/A	3k	NLI	acc.	misc.
WNLI	706	N/A	146	coreference/NLI	acc.	fiction books

Applying BERT to Downstream Tasks

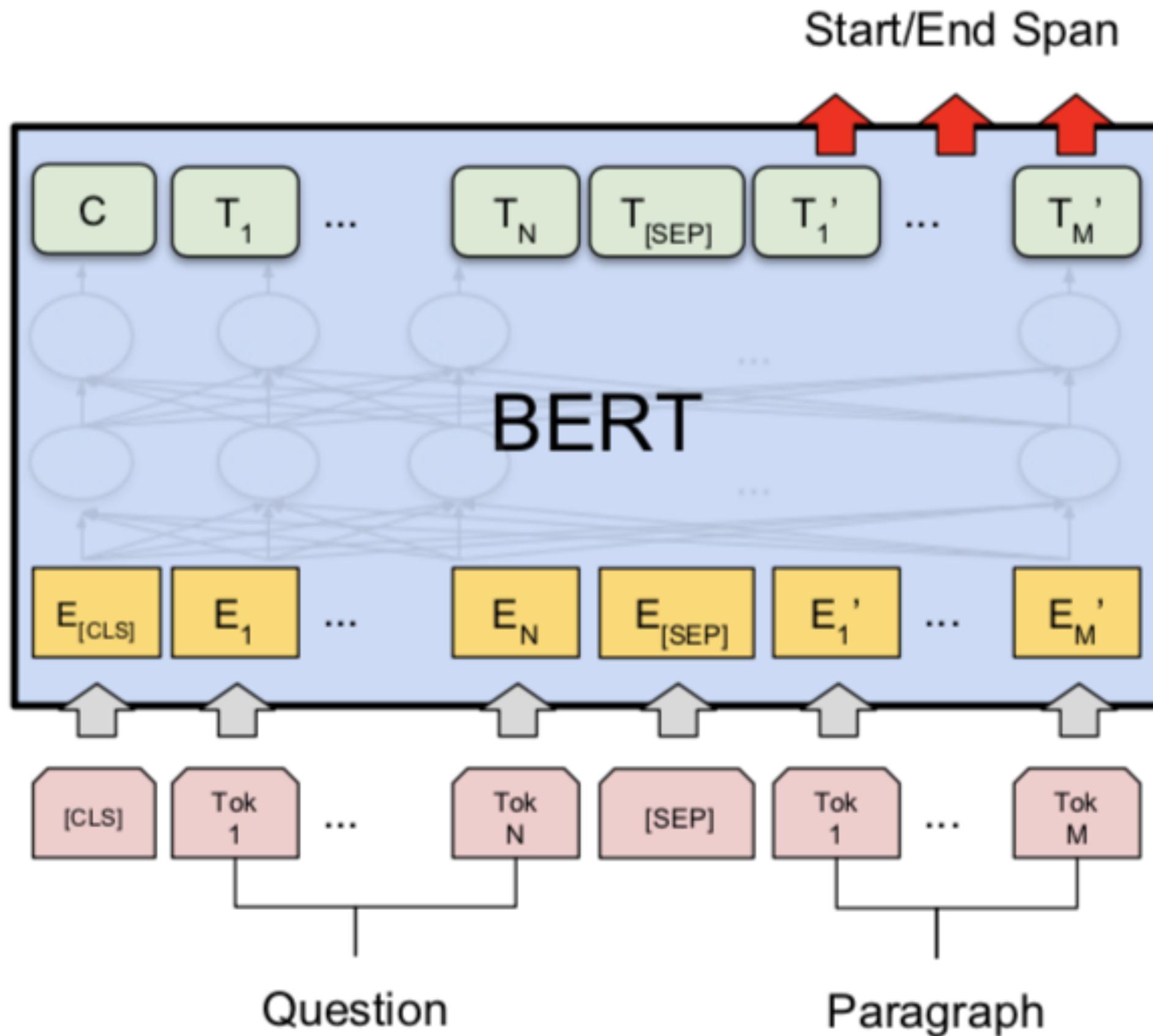


(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG

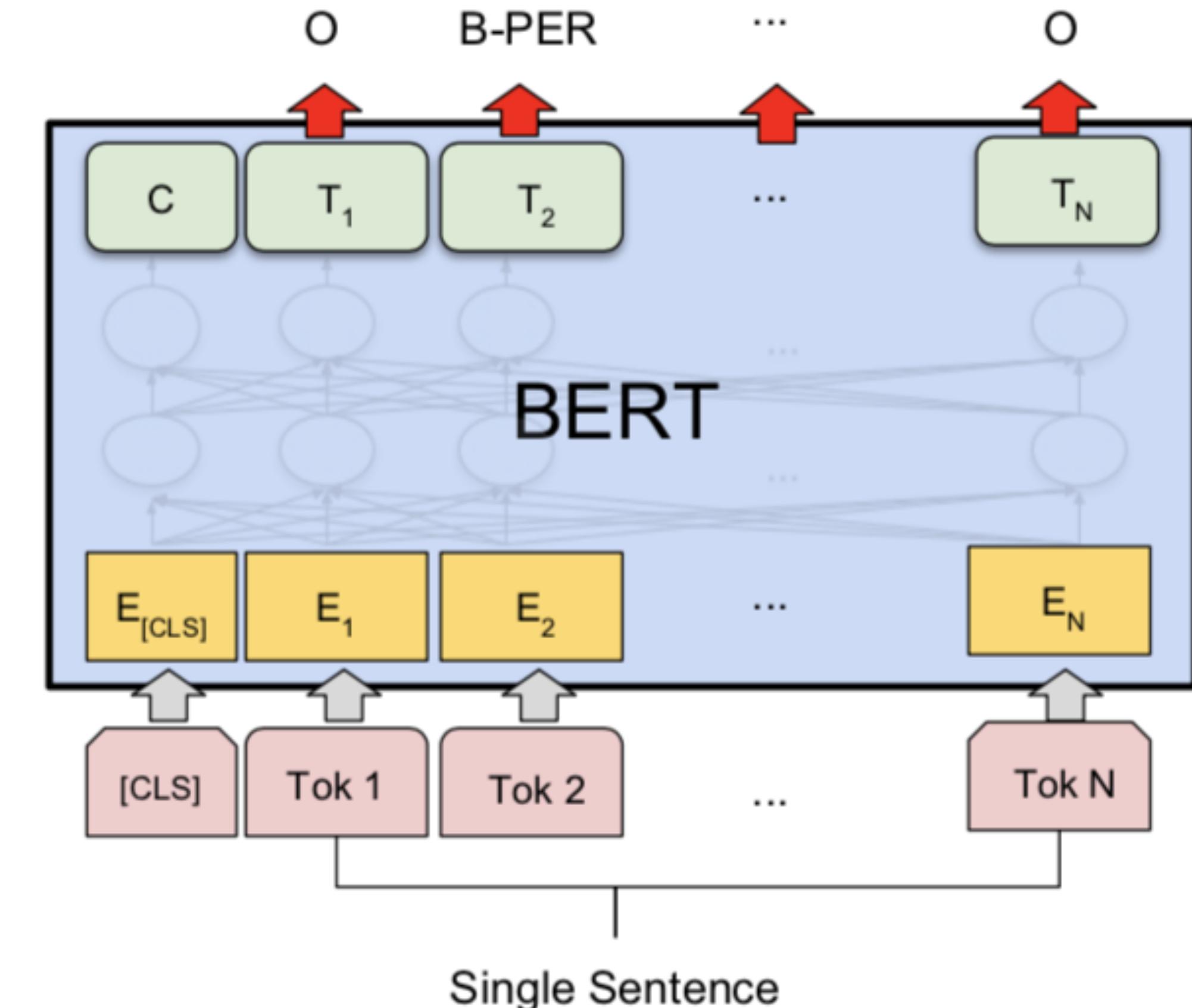


(b) Single Sentence Classification Tasks:
SST-2, CoLA

Applying BERT to Downstream Tasks

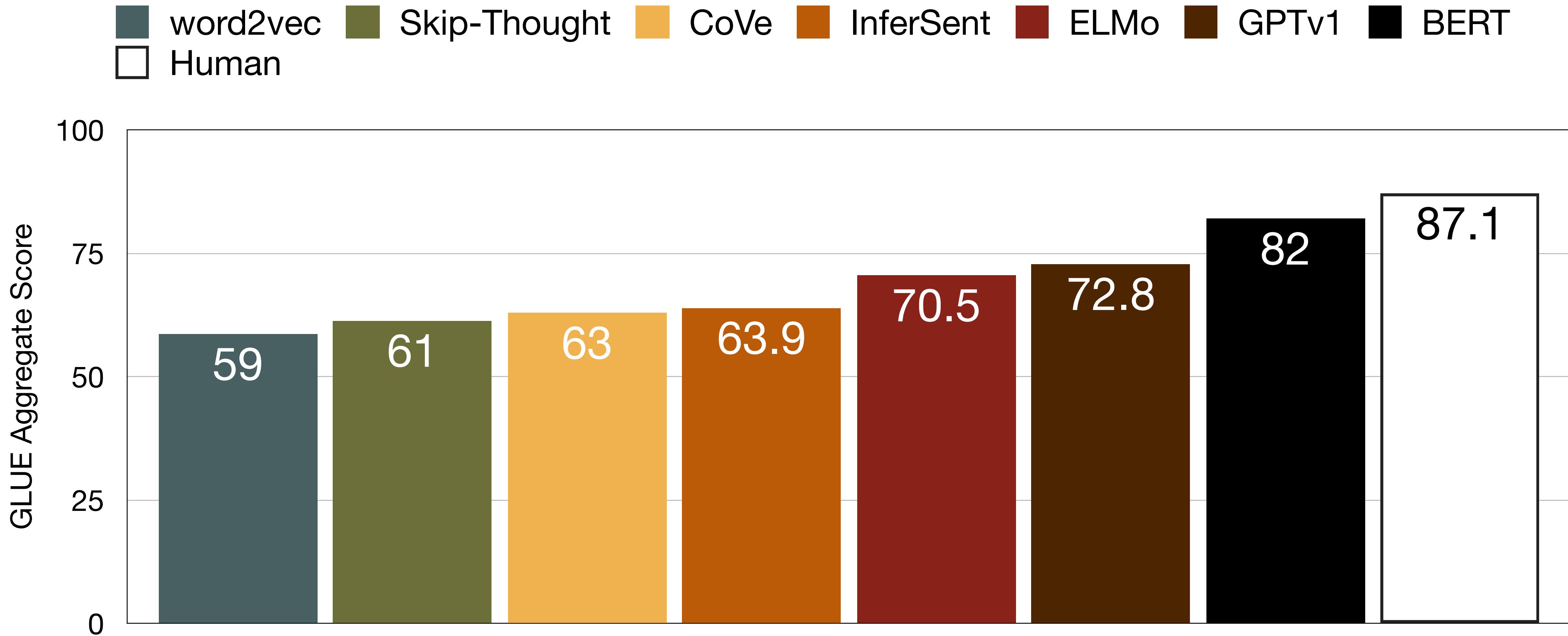


(c) Question Answering Tasks:
SQuAD v1.1

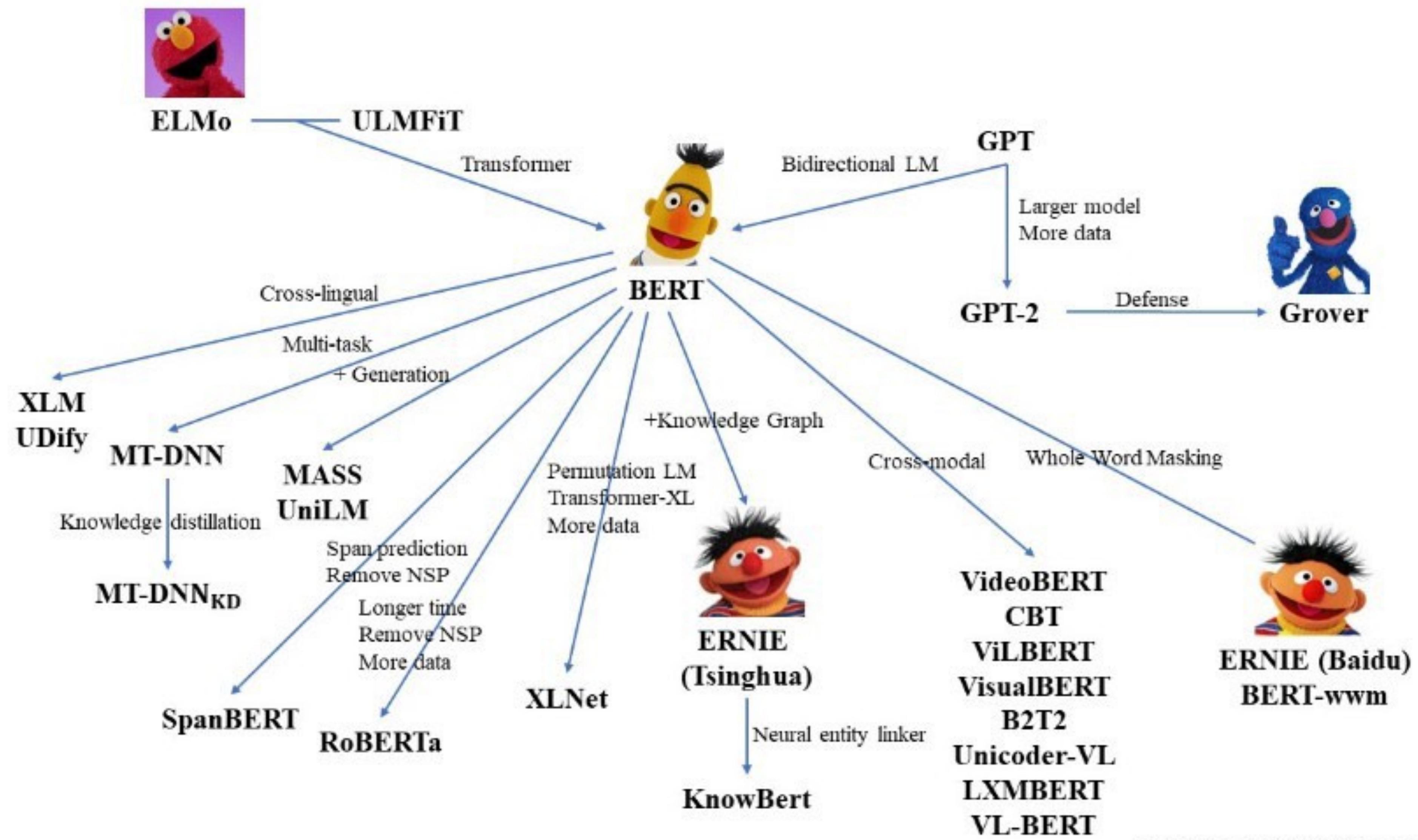


(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

Results



Many Muppets on Sesame Street



GPTv2

=====Model prompt >>> I believe nobody in this room knows this, but Sebastian Riedel was a famous underground rapper in Hamburg before starting an academic career. In his raps, he would frequently make references to Noam Chomsky, secretly educating the hood on linguistics.

===== SAMPLE 1 =====

His nickname was Violent. For some reason Sebastian chose to call him Calvin White.

The Worm

This act nevertheless alerted the drug Users Manual into the dangers of Lynn, and I think replaying and his signature tread appears to be one of the reason behind his initiation as electronic jungle kingpin. Deep down, he would like porn to become a Raw Language, and digital video to become Internet Porn.

At one point, at about age 45, Sebastian worked for Hermann Gothenburg Systems, a kind of revisionist Marxist investment agency selling the "Basic Income" Ben Gurion study. He sneered at Eminem and watched General Gigigi's shareholders float in panic. "The monetary system has cost more than the incomes of American workers," he writes in the reverse of his provocative drink.

Not surprisingly, both the films Psycho and Spirit Of Disputation were enormously success for a young Gothenburg software engineer who had grown tired of the complex monthly payments, which got him homesick as often night work becomes her job title drama. In his insomniac rant his wiser brother suggested, "It's like ... reading for nine hours in the most deliciously pornographic western movie ever."

Recap: GPTv2

”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 côté? -Quel autre côté?”**, which means **“- How do you get to the other side? - What side?”**.

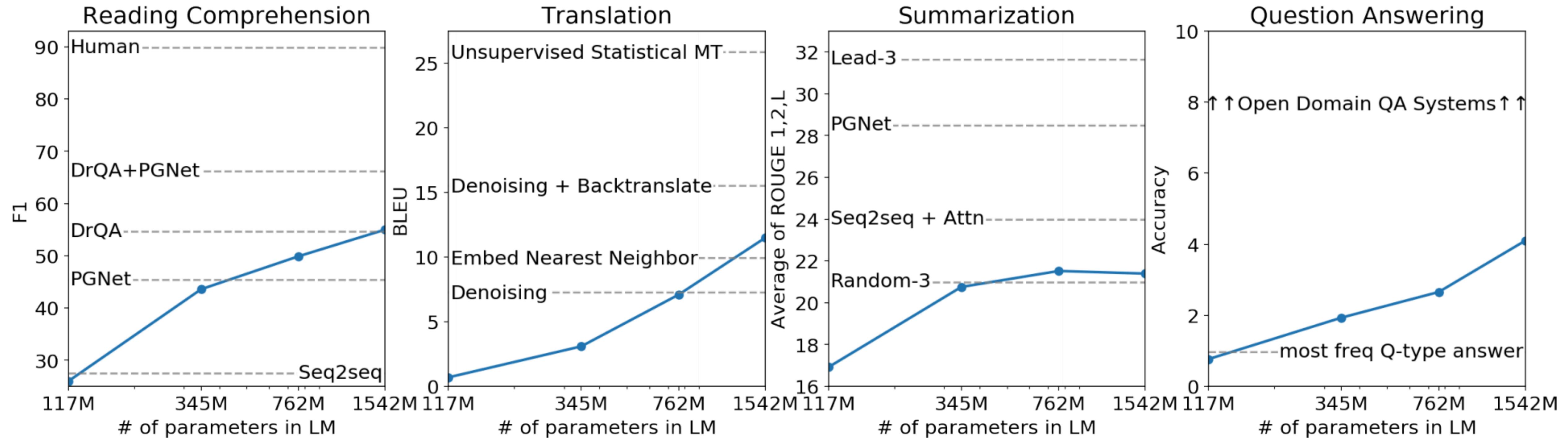
If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

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

Zero-shot Transfer

- Zero-shot Transfer: Transfer model without any fine-tuning/learning on downstream the task
- GPTv2 has been trained on so much text, seeing so many different contexts that the following (kind of) works:
 - Text generation by using Language Model (straight forward)
 - Machine Translation by conditioning model on translation pairs
 - "The hunter shot the fox. = Der Jäger erschoss den Fuchs.
 - This model is dangerous. = "
 - Q&A by conditioning on "Context + Question"
 - Sentence Summarization by conditioning on "Article + TL;DR"

Size Matters

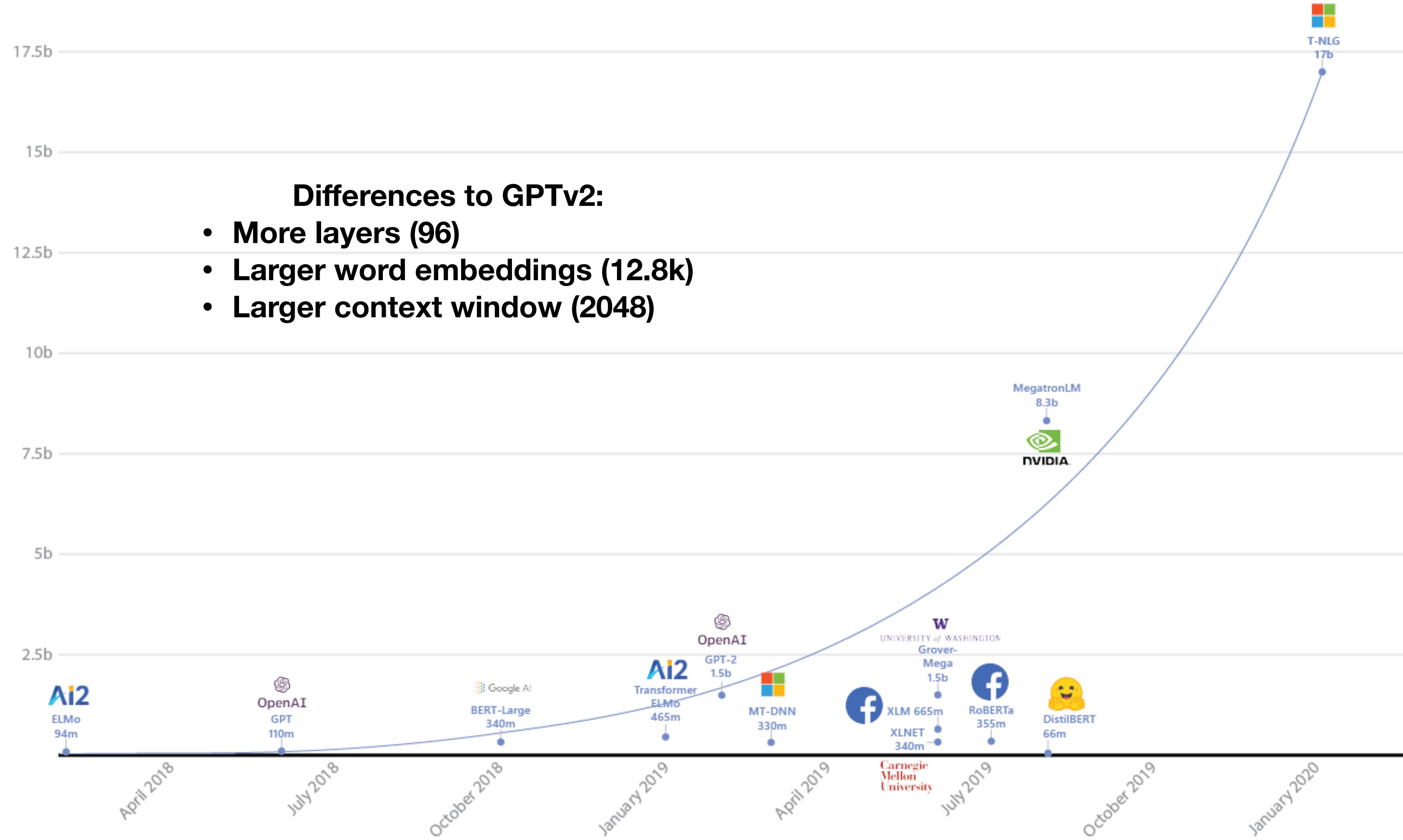


What does GPTv2 "know"?

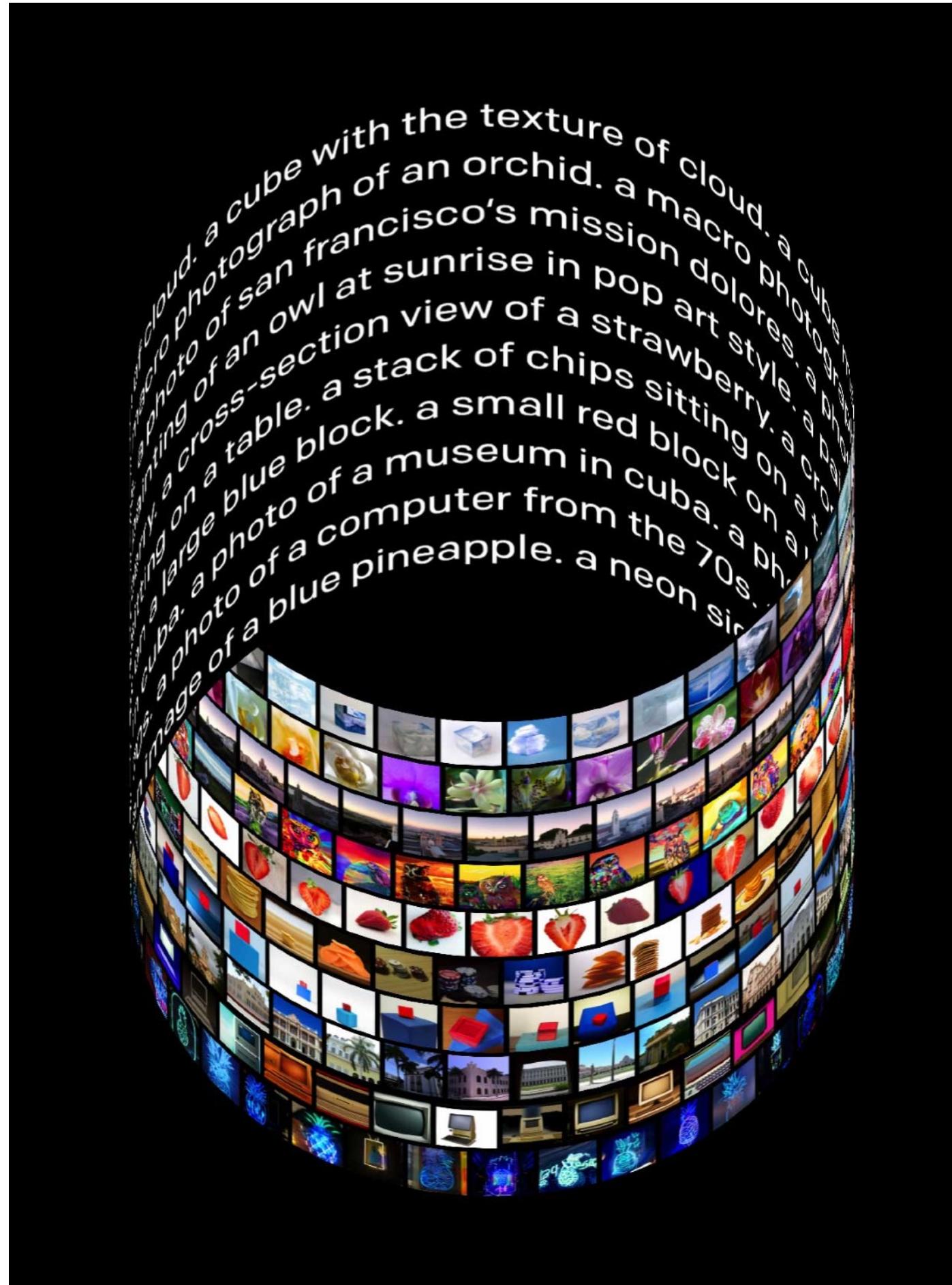
Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%

GPTv3

175B ↑ GPTv3



Multi-modal Pre-training



TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES



[Edit prompt or view more images ↓](#)

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES



[Edit prompt or view more images ↓](#)

Example

Highlights

- Sarah Palin from Alaska meets with McCain
- Fareed Zakaria says John McCain did not put country first with his choice
- Zakaria: This is “hell of a time” for Palin to start thinking about national, global issues

Questions

The questions can refer directly to the highlights, for example:

- Where is Palin from?
- What did Fareed say about John McCain’s choice?
- Who is thinking about global issues?

Questions must always be related to the highlights but their answers don’t have to be in the highlights. You can assume that the highlights summarize a document which can answer other questions for example:

- What was the meeting about?
- What was McCain’s choice?
- What issues is Palin thinking about?

Other Rules

- Do not re-use the same or very similar questions.
- Questions should be written to have short answers.
- Do not write “how” nor “why” type questions since their answers are not short. “How far/long/many/much” are okay.

GPTvN?

Here are the highlights:

- Math geeks and others celebrate Pi Day every March 14
- Pi, or roughly 3.14, is the ratio of circumference to diameter of a circle
- The Pi Day holiday idea started at the Exploratorium museum in San Francisco
- Albert Einstein was also born on March 14

Write questions about them:

Here are questions about the highlights:

1. When is Pi Day celebrated?
 2. What is the value of Pi up to the second decimal digit?
- Another thing is that Pi is important in mathematics.

Model-agnostic Meta-Learning

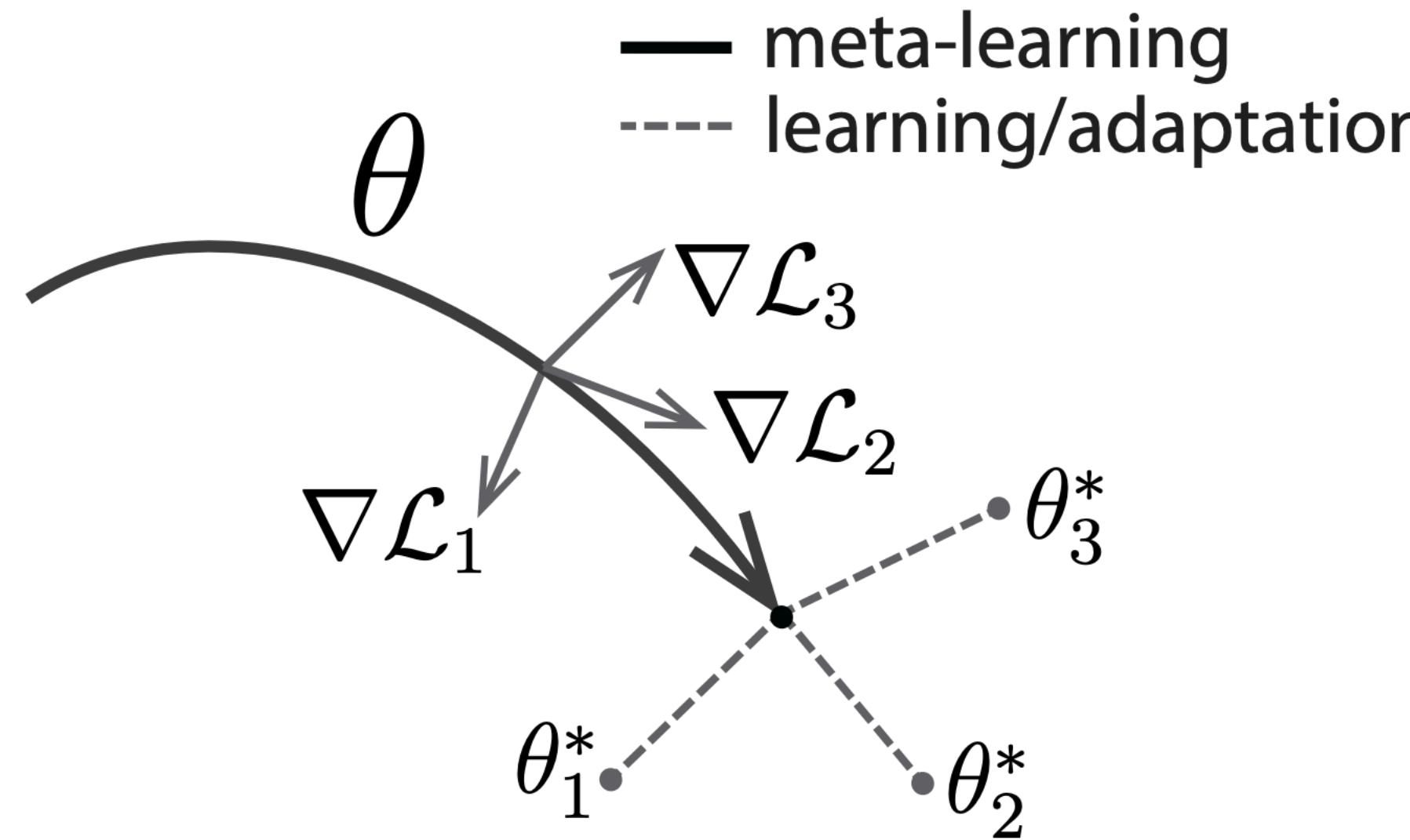


Figure 1. Diagram of our model-agnostic meta-learning algorithm (MAML), which optimizes for a representation θ that can quickly adapt to new tasks.

Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β : step size hyperparameters

- 1: randomly initialize θ
- 2: **while** not done **do**
- 3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
- 4: **for all** \mathcal{T}_i **do**
- 5: Evaluate $\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})$ with respect to K examples
- 6: Compute adapted parameters with gradient descent: $\theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})$
- 7: **end for**
- 8: Update $\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_i})$
- 9: **end while**
