Transfer Learning: Unsupervised Pretraining

Katharina Kann — CSCI/LING5832

NLI with Machine Learning

Natural Language Inference Data in 2015

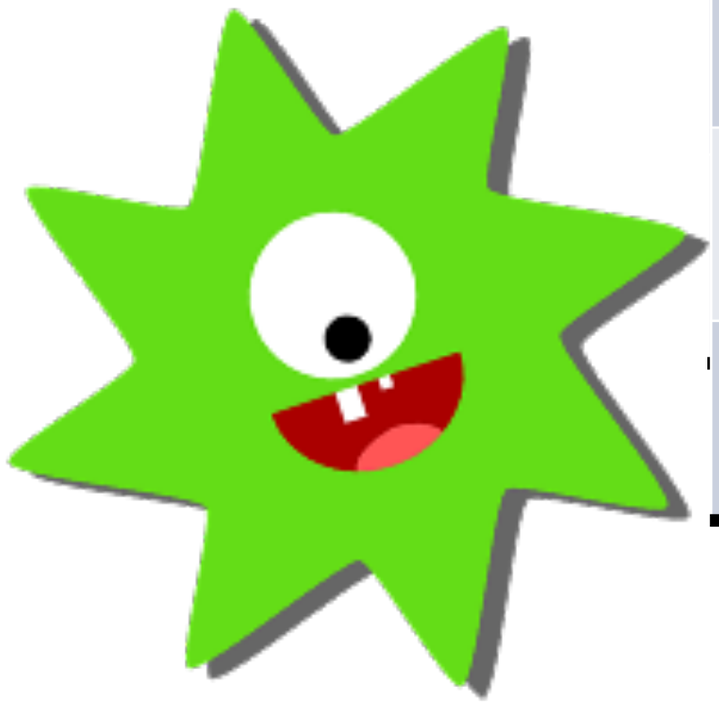
Corpus	Size
FraCaS	.3k
RTE	7k
SICK	10k

Natural Language Inference Data in 2015

- That data was not sufficient to train neural networks for NLI:
  - No successful prior applications of neural network models to NLI at that time

Natural Language Inference Data in 2015

Corpus	Size
FraCaS	.3k
RTE	7k
SICK	10k
SNLI	570k



In-Class Exercise

- Assume that we want to build a neural network for the SNLI corpus, i.e., we want to classify pairs of sentences as *entailment*, *neutral*, or *contradiction*.
- Design a neural network that can learn this task!
- Write down all dimensions explicitly. What is your output dimension?

Data Collection

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption **An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption.

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be.

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Problems (optional) *If something is wrong, have a look at the FAQ, do your best above, and let us know here.*

Source captions from Flickr30k: Young, et al. '14

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption **An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption. **Entailment**

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be. **Neutral**

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption. **Contradiction**

Problems (optional) *If something is wrong, have a look at the FAQ, do your best above, and let us know here.*

Source captions from Flickr30k: Young, et al. '14

Some Sample Results

- **Premise**: Two women are embracing while holding to go packages.
- **Hypothesis**: Two woman are holding packages.

Some Sample Results

- **Premise**: Two women are embracing while holding to go packages.
- **Hypothesis**: Two woman are holding packages.
- **Label**: Entailment

Some Sample Results

- **Premise**: A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.
- **Hypothesis**: A man is repainting a garage

Some Sample Results

- **Premise**: A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.
- **Hypothesis**: A man is repainting a garage
- **Label**: Neutral

Some Results on SNLI

Model	Test accuracy
Most frequent class	34.2%
Big lexicalized classifier	78.2%
300D CBOW	80.6%
300D BiLSTM	81.5%
BERT-based classifier	**91.1%**

Some Results on SNLI

Model	Test accuracy
Most frequent class	34.2%
Big lexicalized classifier	78.2%
300D CBOW	80.6%
300D BiLSTM	81.5%
BERT-based classifier	**91.1%**

Now basically solved!

The Multi-Genre NLI Corpus

Limitations of SNLI

- Little headroom left:
  - SotA: 91.1%
  - Human performance: ~95%

Limitations of SNLI

- Little headroom left:
  - SotA: 91.1%
  - Human performance: ~95%
- Many linguistic phenomena under-represented or ignored

Limitations of SNLI

- Gururangan et al. (2018):
  - Some cues in SNLI hypotheses give clues to the label:
    - Negation is most common with contradiction
    - Some content words more common in contradiction (‘sleeping’)
    - Very short sentences tend to be entailment

Limitations of SNLI

- Gururangan et al. (2018):
  - Some cues in SNLI hypotheses give clues to the label:
    - Negation is most common with contradiction
    - Some content words more common in contradiction (‘sleeping’)
    - Very short sentences tend to be entailment
- A trained NN classifier can reach 67% without access to the premise.

The MultiGenre NLI Corpus

Genre	Train	Dev	Test
Captions (SNLI	(550,152)	(10,000)	(10,000)
Fiction	77,348	2,000	2,000
Government	77,350	2,000	2,000
Slate	77,306	2,000	2,000
Switchboard	83,348	2,000	2,000
Travel Guides	77,350	2,000	2,000
9/11 Report	0	2,000	2,000
Face-to-Face Speech	0	2,000	2,000
Letters	0	2,000	2,000
OUP (Nonfiction	0	2,000	2,000
Verbatim (Magazine)	0	2,000	2,000
Total	**392,702**	**20,000**	**20,000**

The MultiGenre NLI Corpus

Genre	Train	Dev	Test
Captions (SNLI	(550,152)	(10,000)	(10,000)
Fiction	77,348	2,000	2,000
Government	77,350	2,000	2,000
Slate	77,306	2,000	2,000
Switchboard	83,348	2,000	2,000
Travel Guides	77,350	2,000	2,000
9/11 Report	0	2,000	2,000
Face-to-Face Speech	0	2,000	2,000
Letters	0	2,000	2,000
OUP (Nonfiction	0	2,000	2,000
Verbatim (Magazine)	0	2,000	2,000
Total	**392,702**	**20,000**	**20,000**

The MultiGenre NLI Corpus

Genre	Train	Dev	Test
Captions (SNLI	(550,152)	(10,000)	(10,000)
Fiction	77,348	2,000	2,000
Government	77,350	2,000	2,000
Slate	77,306	2,000	2,000
Switchboard	83,348	2,000	2,000
Travel Guides	77,350	2,000	2,000
9/11 Report	0	2,000	2,000
Face-to-Face Speech	0	2,000	2,000
Letters	0	2,000	2,000
OUP (Nonfiction	0	2,000	2,000
Verbatim (Magazine)	0	2,000	2,000
Total	**392,702**	**20,000**	**20,000**

Typical Dev Set Example

- **Premise**: In contrast, suppliers that have continued to innovate and expand their use of the four practices, as well as other activities described in previous chapters, keep outperforming the industry as a whole.
- **Hypothesis**: The suppliers that continued to innovate in their use of the four practices consistently underperformed in the industry.

Typical Dev Set Example

- **Premise**: In contrast, suppliers that have continued to innovate and expand their use of the four practices, as well as other activities described in previous chapters, keep outperforming the industry as a whole.
- **Hypothesis**: The suppliers that continued to innovate in their use of the four practices consistently underperformed in the industry.
- **Label**: Contradiction
- **Genre**: Oxford University Press (Nonfiction books)

Typical Dev Set Example

- **Premise**: someone else noticed it and i said well i guess that’s true and it was somewhat melodious in other words it wasn’t just you know it was really funny
- **Hypothesis**: No one noticed and it wasn’t funny at all.

Typical Dev Set Example

- **Premise**: someone else noticed it and i said well i guess that’s true and it was somewhat melodious in other words it wasn’t just you know it was really funny
- **Hypothesis**: No one noticed and it wasn’t funny at all.
- **Label**: Contradiction
- **Genre**: Switchboard (Telephone Speech)

Key Figures

Tag	SNLI	MultiNLI
Entire Corpus	100	100
Pronouns (PTB)	34	68
Quantifiers	33	63
Modals (PTB)	<1	28
Negation (PTB)	5	31
WH terms (PTB)	5	30
Belief Verbs	<1	19
Time Terms	19	36
Discourse Mark.	<1	14
Presup. Triggers	8	22
Compr./Supr.(PTB)	3	17
Conditionals	4	15
Tense Match (PTB)	62	69
Interjections (PTB)	<1	5
>20 words	<1	5

From Williams et al. (2018)

Some Results on MultiNLI

Model	Matched Test Acc.	Mismatched Test Acc.
Most frequent class	36.5%	35.6%
CBOW	65.2%	64.6%
Multitask learning with BERT (Xiaodong Liu et al., 2019)	87.9%	87.4%

Some Results on MultiNLI

Model	Matched Test Acc.	Mismatched Test Acc.
Most frequent class	36.5%	35.6%
CBOW	65.2%	64.6%
Multitask learning with BERT (Xiaodong Liu et al., 2019)	87.9%	87.4%

Now basically solved!

Still Some Limitations

- Models that do well on SNLI and MultiNLI are still easy to trick with certain kinds of examples

Premise/Hypothesis	Label
The man is holding a saxophone The man is holding an electric guitar	contradiction[1]
A little girl is very sad. A little girl is very unhappy.	entailment
A couple drinking wine A couple drinking champagne	neutral

(Glockner et al. (2018))

Still Some Limitations

- Models that do well on SNLI and MultiNLI are still easy to trick with certain kinds of examples

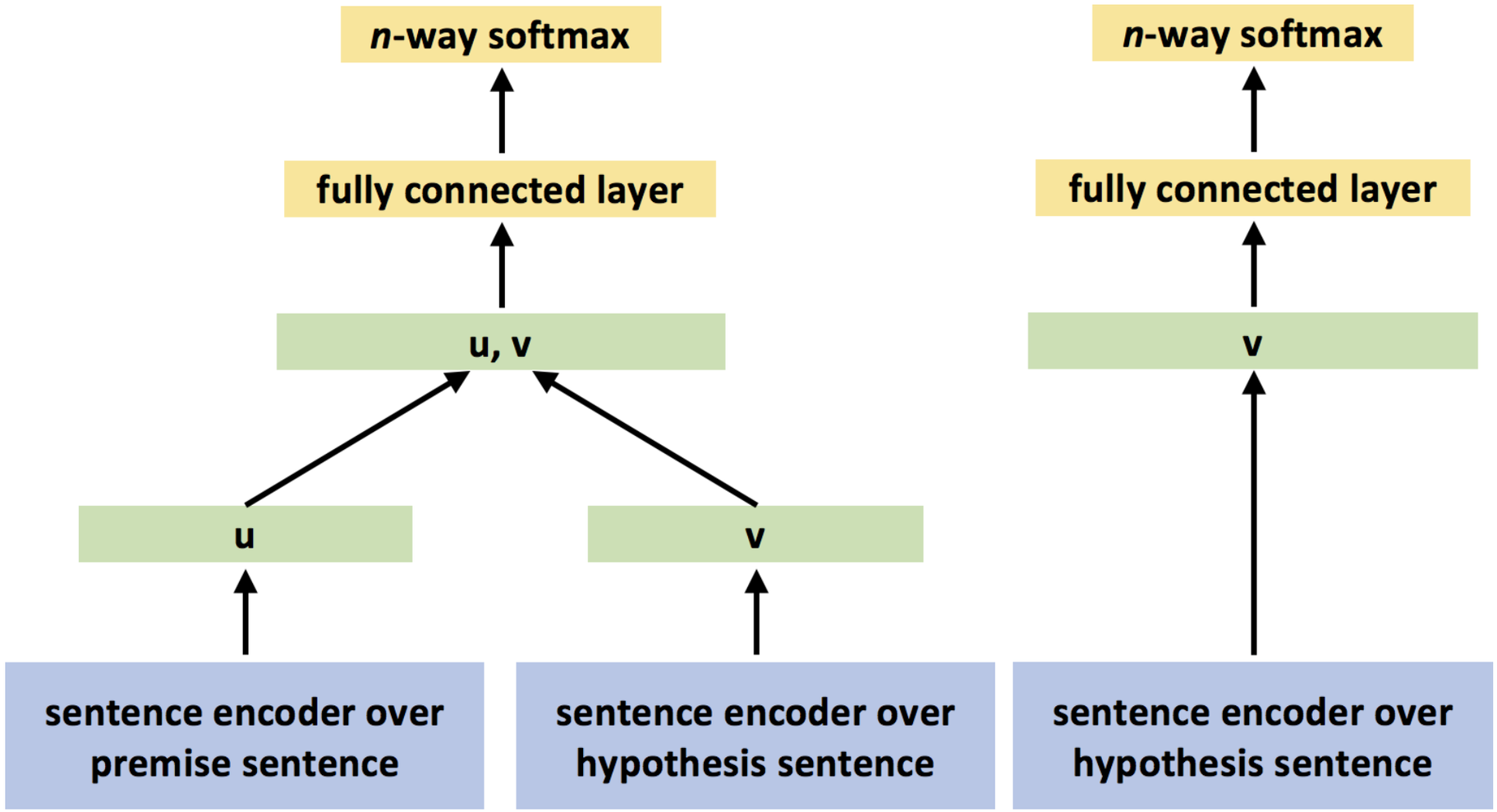
Alice believes Mary is lying. $\not\mapsto$
Alice believes Mary.

The book on the table is blue. $\not\mapsto$
The table is blue.

(McCoy and Linzen (2019))

Hypotheses Only!

- Gururangan et al. (2018) and Poliak et al. (2018):



From Poliak et al. (2018)

Hypotheses Only!

- Gururangan et al. (2018) and Poliak et al. (2018):
  - Many clues are in the hypothesis sentences.
  - NN classifier performance without access to premise:
    - SNLI: 67% (~69%) (vs. SotA 89%)
    - MultiNLI: 54% (~55%) (vs. SotA 80%)

Hypotheses Only!

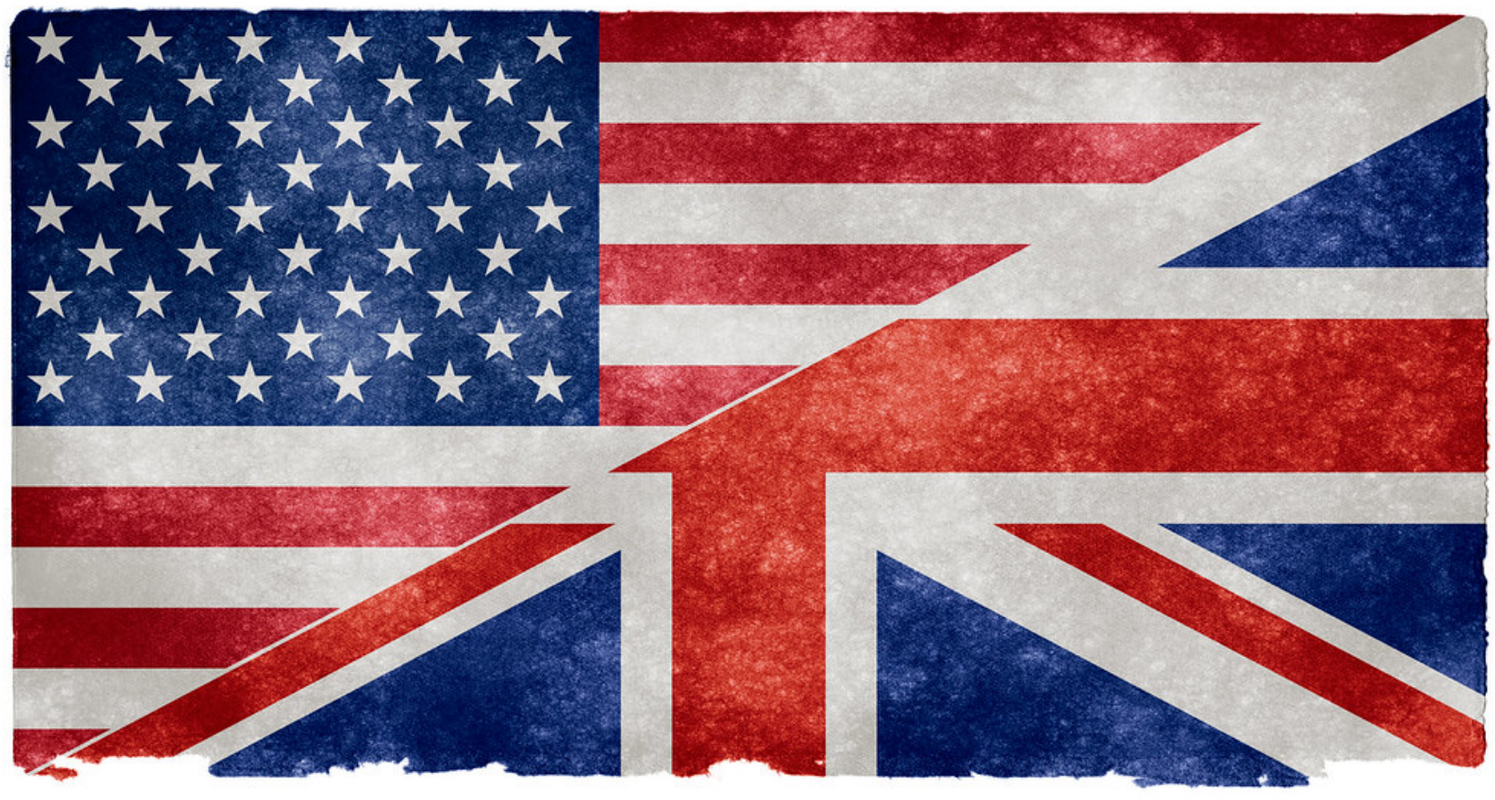
- Gururangan et al. (2018) and Poliak et al. (2018):
  - Many clues are in the hypothesis sentences.
  - NN classifier performance without access to premise:
    - SNLI: 67% (~69%) (vs. SotA 89%)
    - MultiNLI: 54% (~55%) (vs. SotA 80%)
- Why is MultiNLI better? No deliberate intervention, but…
  - More diverse content (fewer content cues)
  - More diverse hypothesis structure (fewer structural cues)
  - More communication with annotators

Other Recent NLI Datasets

- Adversarial NLI (Nie et al., 2020)
- XNLI (Conneau et al., 2018)
- OCNLI: Original Chinese Natural Language Inference (Hu et al., 2020)

Transfer Learning: Unsupervised Pretraining

Current NLP



The Problem: Limited Resources

- Deep learning models need a lot of training data

The Problem: Limited Resources

- Deep learning models need a lot of training data
- Data can be limited for:
  - Languages

The Problem: Limited Resources

- Deep learning models need a lot of training data
- Data can be limited for:
  - Languages
  - Tasks

The Problem: Limited Resources

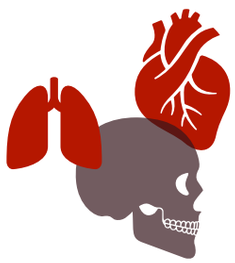
- Deep learning models need a lot of training data

- Data can be limited for:

  - Languages

  - Tasks

  - Domains



Wikipedia

- English: ~49M pages
- Chinese: ~5M pages
- Spanish: ~5M pages
- German: ~5M pages
- Norwegian: ~1M pages
- Afrikaans: ~200k pages

Wikipedia

- English: ~49M pages
- Chinese: ~5M pages
- Spanish: ~5M pages
- German: ~5M pages
- Norwegian: ~1M pages
- Afrikaans: ~200k pages

NLP systems cannot just be trained for Afrikaans as they can for English!

What is Transfer Learning?

- Sharing knowledge across tasks, languages, domains

What is Transfer Learning?

- Sharing knowledge across tasks, languages, domains
- Different sources complement and inform each other

What is Transfer Learning?

- Sharing knowledge across tasks, languages, domains
- Different sources complement and inform each other
- Reduces the amount of data required for certain…
  - …languages
  - …tasks
  - …domains

What is Transfer Learning?

- Sharing knowledge across tasks, languages, domains
- Different sources complement and inform each other
- Reduces the amount of data required for certain…
  - …languages
  - …tasks
  - …domains

Has lead to important improvements in NLP in the last years!

Why Transfer Learning?

SuperGLUE GLUE Paper Code Tasks Leaderboard FAQ Diagnostics Submit Login

Leaderboard Version: **2.0**

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
2	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
3	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	38.0	99.4/51.4
		BERT		69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	23.0	97.8/51.7
		Most Frequent Class		47.1	62.3	21.7/48.4	50.0	61.1/0.3	33.4/32.5	50.3	50.0	65.1	0.0	100.0/50.0
		CBoW		44.5	62.2	49.0/71.2	51.6	0.0/0.5	14.0/13.6	49.7	53.1	65.1	-0.4	100.0/50.0
		Outside Best		-	80.4	-	84.4	70.4/24.5	74.8/73.0	82.7	-	-	-	-
4	IBM Research AI	BERT-mtl		71.3	84.8	89.6/94.0	72.2	73.2/30.5	74.6/74.0	84.1	50.0	61.0	29.6	97.8/57.3
-	Stanford Hazy Research	Snorkel [SuperGLUE v1.9]		-	-	88.6/93.2	76.2	76.4/36.3	-	78.9	72.1	72.6	47.6	-

Click on a submission to see more information

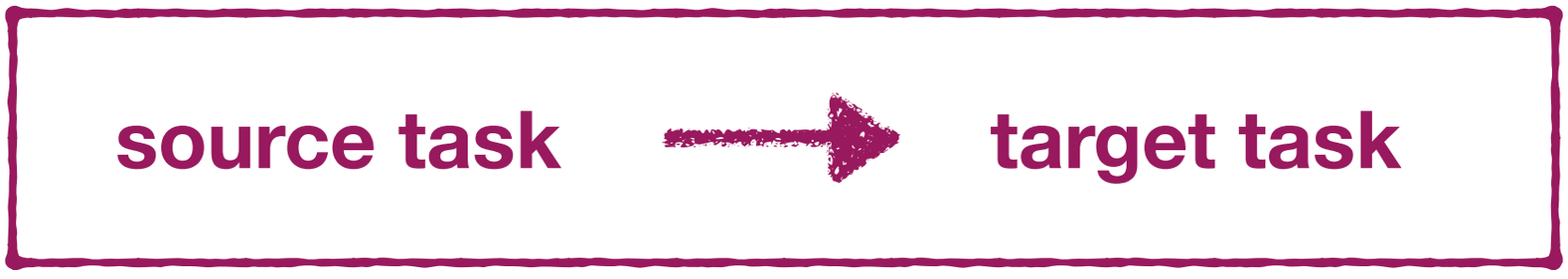
Pretraining

Pretraining

- Train a model on large amounts of data from a source task which is different from the target task
- Learned knowledge will (hopefully) be useful for target task
- General properties of language can even be learned from raw text

Pretraining

- Train a model on large amounts of data from a source task which is different from the target task
- Learned knowledge will (hopefully) be useful for target task
- General properties of language can even be learned from raw text

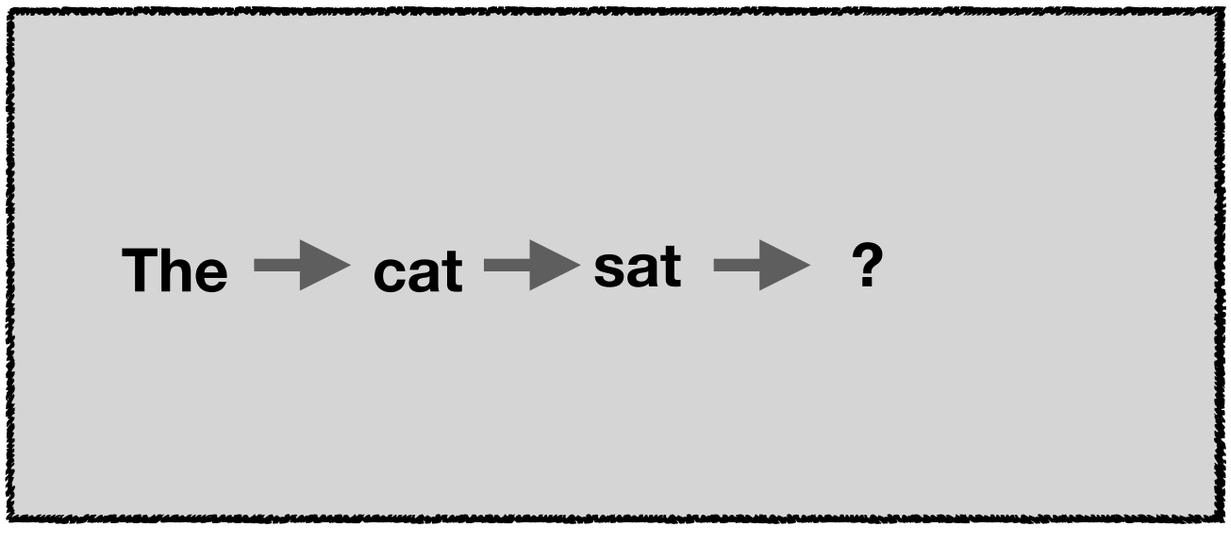


Pretraining

- Any task could be used in theory
- Popular examples:
  - Language modeling
  - Copy task
  - Natural language inference

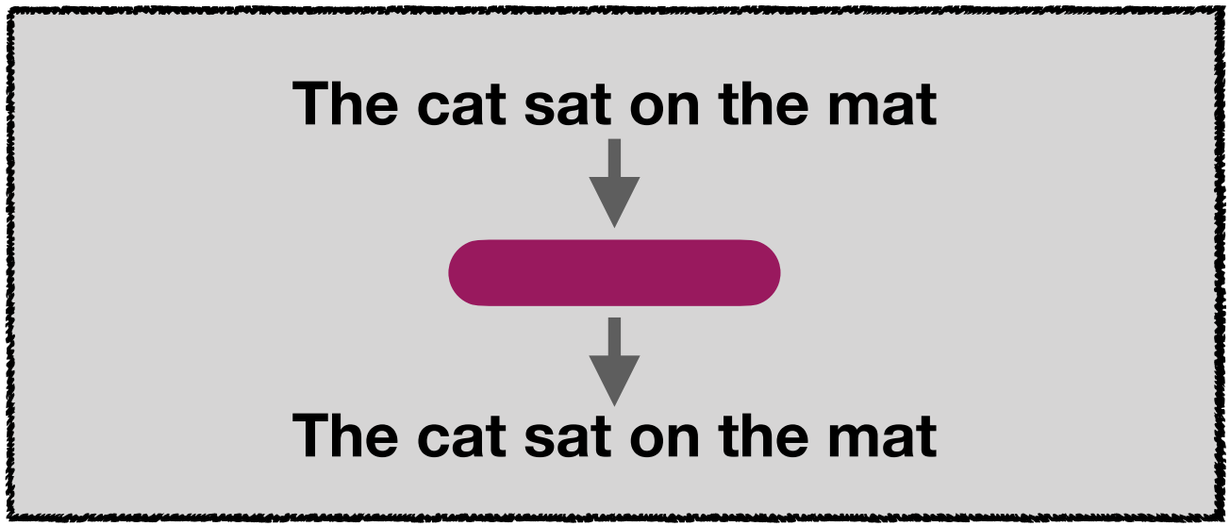
Pretraining

- Any task could be used in theory
- Popular examples:
  - Language modeling
  - Copy task
  - Natural language inference



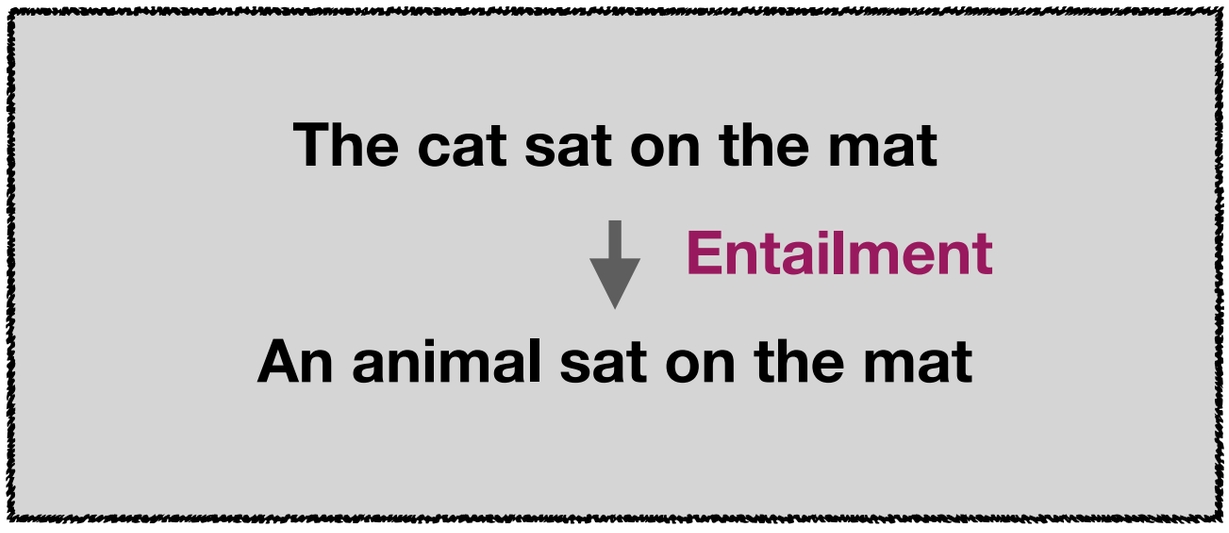
Pretraining

- Any task could be used in theory
- Popular examples:
  - Language modeling
  - Copy task
  - Natural language inference



Pretraining

- Any task could be used in theory
- Popular examples:
  - Language modeling
  - Copy task
  - Natural language inference



Pretraining

- Any task could be used in theory
- Popular examples:
  - Language modeling
  - Copy task
  - Natural language inference

Many source tasks are chosen because a lot of data are available

Pretraining

- Any task could be used in theory
- Popular examples:
  - Language modeling
  - Copy task
  - Natural language inference

Many source tasks are chosen because a lot of data are available

Example: Word Embeddings

- **Firth** (1957): “You shall know a word by the company it keeps.”

- **Harris** (1954): “distributional statements can cover all of the material of a language without requiring support from other types of information.”

Example (this slide is a repetition!)

“<word> are performed around the world.”

Example (this slide is a repetition!)

“<word> are performed around the world.”

“Several shorter <word> on Broadway and in the West End have been presented in one act in recent decades.”

Example (this slide is a repetition!)

“<word> are performed around the world.”

“Several shorter <word> on Broadway and in the West End have been presented in one act in recent decades.”

“Some of the most famous <word> through the decades that followed include West Side Story (1957), [...] and Hamilton (2015).”

Example (this slide is a repetition!)

“<word> are performed around the world.”

“Several shorter <word> on Broadway and in the West End have been presented in one act in recent decades.”

“Some of the most famous <word> through the decades that followed include West Side Story (1957), [...] and Hamilton (2015).”

What is <word>?

Word2vec (Mikolov et al., 2013)

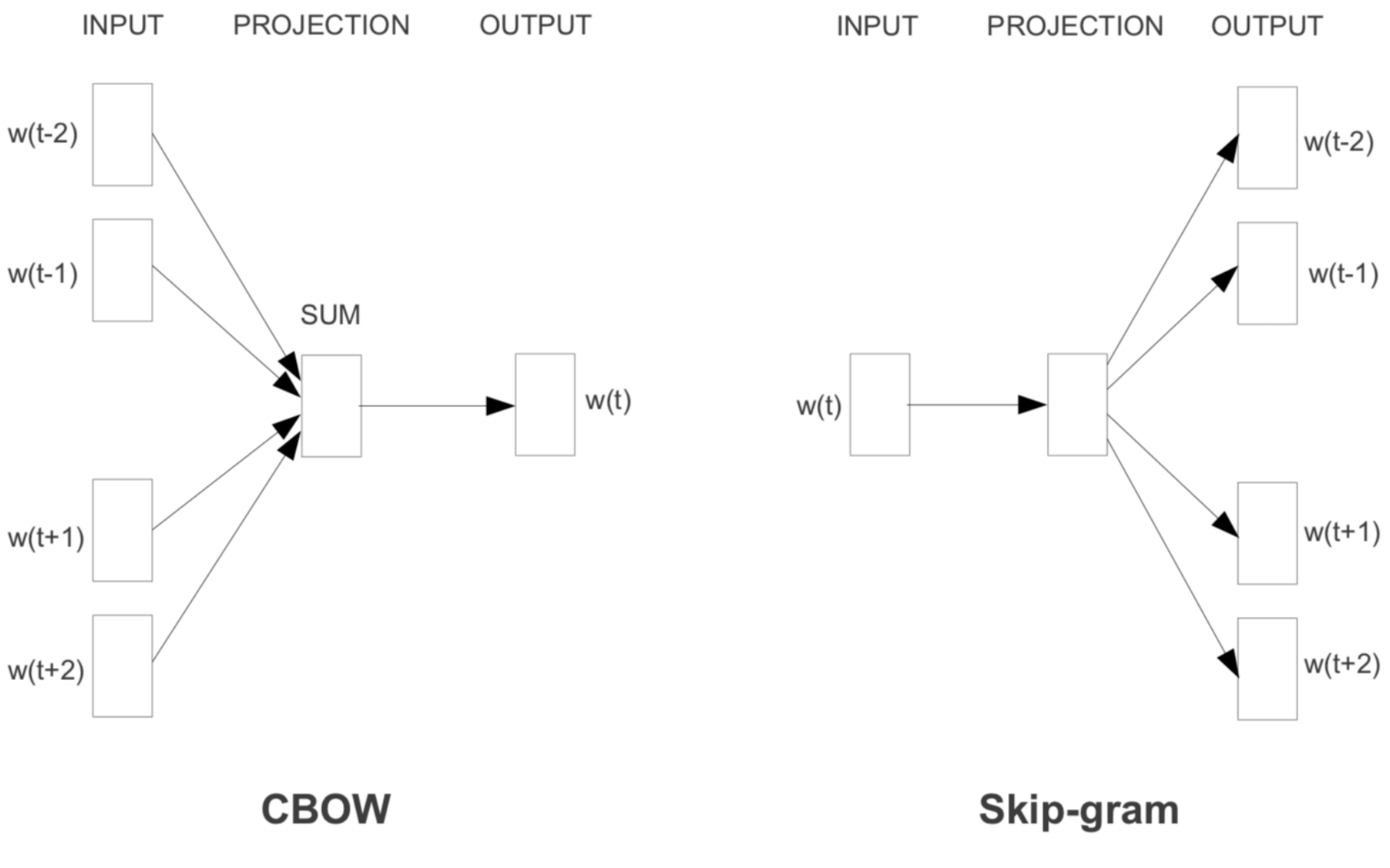


Figure from Mikolov et al. (2013)

Shortcomings of Word Embeddings

- **Apple** will release a new iPhone in September.
- This **apple** tastes better than it looks.

Shortcomings of Word Embeddings

- **Apple** will release a new iPhone in September.
- This **apple** tastes better than it looks.

Contextualized word embeddings!

Contextualized Word Embeddings

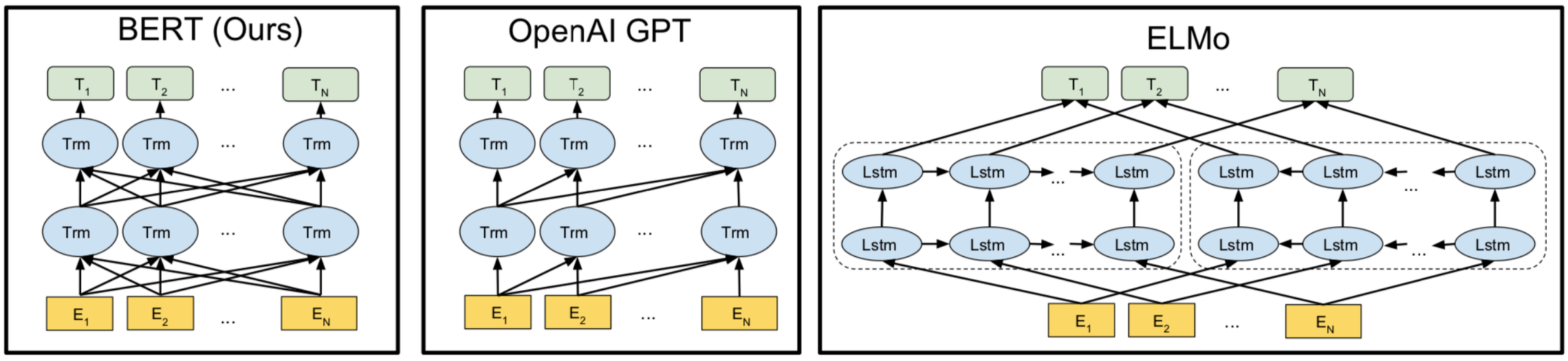


Figure from Devlin et al. (2018)

ELMo (Peters et al., 2018), GPT (Radford et al., 2018)

ELMo

How to Read Papers

- Jason Eisner (2009):
  - Skim the paper first
    - Give yourself a limited time budget!
  - If the paper seems worthwhile, read it in more detail afterwards

How to Read Papers

- Jason Eisner (2009):
  - Skim the paper first
    - Give yourself a limited time budget!
  - If the paper seems worthwhile, read it in more detail afterwards
  - Keep low-level notes

How to Read Papers

- Jason Eisner (2009):
  - Skim the paper first
    - Give yourself a limited time budget!
  - If the paper seems worthwhile, read it in more detail afterwards
  - Keep low-level notes
  - Keep high-level notes:
    - "*They don't say this, but equation (2) is basically the same as the method of Pookie (2001), except that they add a reconfabulation step after the data purée. I was surprised at their reconfabulator, which doesn't match what I would have expected from Kachu (2004), but it does cure the exponential growth problem in this domain. To see the difference, I found it useful to think about this example: ...*"

How to Read Papers

- Jason Eisner (2009):
  - Do you need to read the whole paper in the second pass?

How to Read Papers

- Jason Eisner (2009):
  - Do you need to read the whole paper in the second pass?
    - Depends on the paper!

How to Read Papers

- Jason Eisner (2009):
  - Do you need to read the whole paper in the second pass?
    - Depends on the paper!
  - Motivation needs to be understood

How to Read Papers

- Jason Eisner (2009):
  - Do you need to read the whole paper in the second pass?
    - Depends on the paper!
  - Motivation needs to be understood
  - Technical parts can sometimes be skipped

How to Read Papers

- Jason Eisner (2009):
  - Do you need to read the whole paper in the second pass?
    - Depends on the paper!
  - Motivation needs to be understood
  - Technical parts can sometimes be skipped
  - You might want to care about the experiments as long as you learn from them

In-Class Exercise

- Read the ELMo paper (Peters et al, 2019)!
  - Skim the paper first, read then in more detail.
  - Make notes for all sections (not necessarily all subsections)!
- State in one sentence:
  - **What is the paper about and what contributions does it make?**

GLUE: General Language Understanding Evaluation

GLUE

The General Language Understanding Evaluation (GLUE):

An open-ended competition and evaluation platform for general-purpose sentence encoders.

Wang, Singh, Michael, Hill, Levy & Bowman '19, ICLR

GLUE, in short

- Nine English-language sentence understanding tasks based on existing data, varying in:
  - Task difficulty
  - Training data volume and degree of training set–test set similarity
  - Language style/genre

GLUE, in short

- Nine English-language sentence understanding tasks based on existing data, varying in:
  - Task difficulty
  - Training data volume and degree of training set–test set similarity
  - Language style/genre
- Kaggle-style evaluation platform with private test data.

GLUE, in short

- Nine English-language sentence understanding tasks based on existing data, varying in:
  - Task difficulty
  - Training data volume and degree of training set–test set similarity
  - Language style/genre
- Kaggle-style evaluation platform with private test data.
- Built completely on open source/open data.

GLUE: The Main Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks			
CoLA	8.5k	1k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks			
MRPC	3.7k	408	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	40k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks			
MNLI	393k	20k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	108k	5.7k	5.7k	QA/NLI	acc.	Wikipedia
RTE	2.5k	276	3k	NLI	acc.	misc.
WNLI	634	71	**146**	coreference/NLI	acc.	fiction books

GLUE: The Main Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks			
CoLA	8.5k	1k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks			
MRPC	3.7k	408	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	40k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks			
MNLI	393k	20k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	108k	5.7k	5.7k	QA/NLI	acc.	Wikipedia
RTE	2.5k	276	3k	NLI	acc.	misc.
WNLI	634	71	**146**	coreference/NLI	acc.	fiction books

GLUE: The Main Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks			
CoLA	8.5k	1k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks			
MRPC	3.7k	408	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	40k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks			
MNLI	393k	20k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	108k	5.7k	5.7k	QA/NLI	acc.	Wikipedia
RTE	2.5k	276	3k	NLI	acc.	misc.
WNLI	634	71	**146**	coreference/NLI	acc.	fiction books

The Corpus of Linguistic Acceptability (Warstadt et al. '18)

- **Binary classification: Is some string of words a possible English sentence.**
- **Data of this form is a major source of evidence in linguistic theory. Sentences derived from books and articles on morphology, syntax, and semantics.**

* *Who do you think that will question Seamus first?*

✓ *The gardener planted roses in the garden.*

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
Single-Sentence Tasks						
CoLA	8.5k	1k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks						
MRPC	3.7k	408	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	40k	**391k**	paraphrase	acc./F1	social QA questions
Inference Tasks						

The Quora Question Pairs

- **Binary classification for pairs of user generated questions. Positive pairs can be answered with the same answer.**

What are the best tips for outlining/planning a novel?
How do I best outline my novel?
positive

Corpus						
CoLA						
SST-2						
MR						news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	40k	**391k**	paraphrase	acc./F1	social QA questions
Inference Tasks						
MNLI	393k	20k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	108k	5.7k	5.7k	QA/NLI	acc.	Wikipedia
RTE	2.5k	276	3k	NLI	acc.	misc.
WNLI	634	71	**146**	coreference/NLI	acc.	fiction books

The Multi-Genre Natural Language Inference Corpus (Williams et al., 2018)

- **Balanced classification for pairs of sentences into *entailment*, *contradiction*, and *neutral***
- **Training set sentences drawn from five written and spoken genres. Dev/test sets divided into a matched set and a *mismatched* set with five more.**

The Old One always comforted Ca'daan, except today.

Ca'daan knew the Old One very well.

neutral

Corpus						
CoLA						
SST-2						
MRPC						
STS-B						
QQ						
Inference Tasks						
MNLI	393k	20k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	108k	5.7k	5.7k	QA/NLI	acc.	Wikipedia
RTE	2.5k	276	3k	NLI	acc.	misc.
WNLI	634	71	**146**	coreference/NLI	acc.	fiction books

The Winograd Schema Challenge (based on Levesque et al., 2011)

- **Binary classification for expert-constructed pairs of sentences: What does the pronoun refer to?**
- **Manually constructed to foil superficial statistical cues.**

Jane gave Joan candy because she was hungry.
Jane was hungry.
not-entailment

Jane gave Joan candy because she was hungry.
Joan was hungry.
entailment

Corpus						
CoLA						
SST-2						
MRPC						
STS-B						
QQP						
MN						
QN						Wikipedia
RTE	2.5k	276	3k	NLI	acc.	misc.
WNLI	634	71	**146**	coreference/NLI	acc.	fiction books

Limitations of GLUE

- GLUE is built only on English data.
    - Sentence representation learning may look quite different in lower-resource languages!

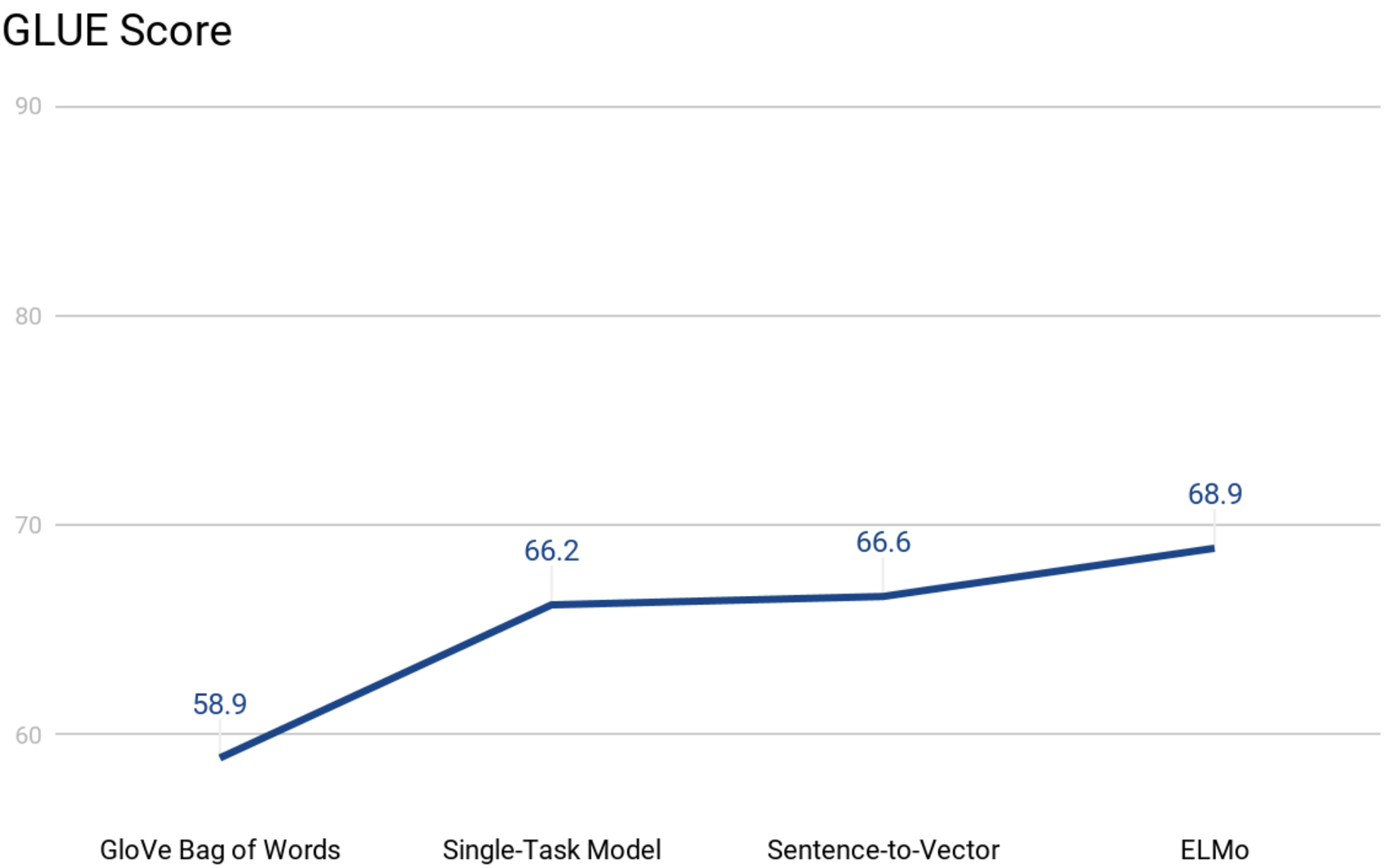
Limitations of GLUE

- GLUE is built only on English data.
    - Sentence representation learning may look quite different in lower-resource languages!
- GLUE does not evaluate text ***generation***, and uses only small amounts of context.
    - Isolates the problem of extracting sentence meaning, but avoids other hard parts of NLP.

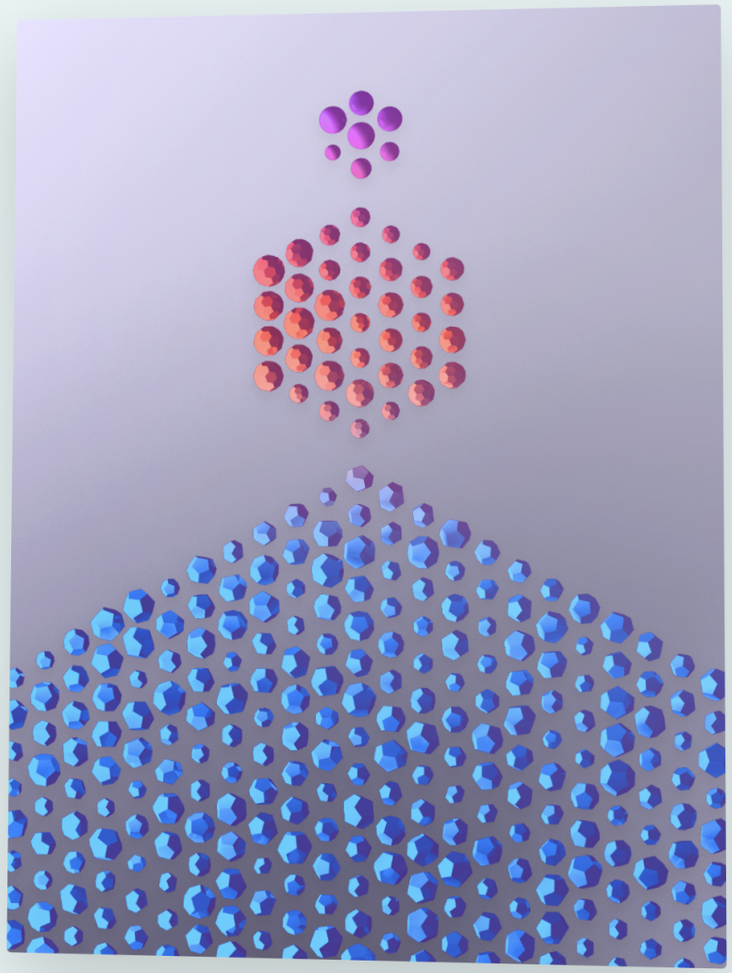
Limitations of GLUE

- GLUE is built only on English data.
    - Sentence representation learning may look quite different in lower-resource languages!
- GLUE does not evaluate text ***generation***, and uses only small amounts of context.
    - Isolates the problem of extracting sentence meaning, but avoids other hard parts of NLP.
- GLUE uses naturally occurring and crowdsourced data.
    - Like any such data, the GLUE evaluation sets contain evidence of implicit bias (gender, race, etc.).
    - Models that reflect these biases will ***tend to do better on GLUE***.

What Methods Work?



OpenAI's Submission



JUNE 11, 2018

Improving Language Understanding with Unsupervised Learning

We've obtained state-of-the-art results on a suite of diverse language tasks with a scalable, task-agnostic system, which we're also releasing. Our approach is a combination of two existing ideas: transformers and unsupervised pre-training. These results provide a convincing example that pairing supervised learning methods with

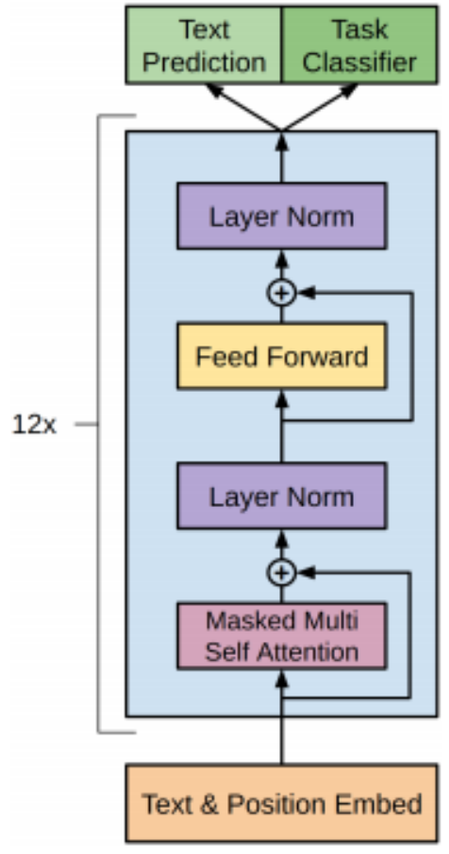
Radford et al. '18

OpenAI's Transformer Language Model

- Same basic idea as ELMo, but many small differences (and many open questions!), including:

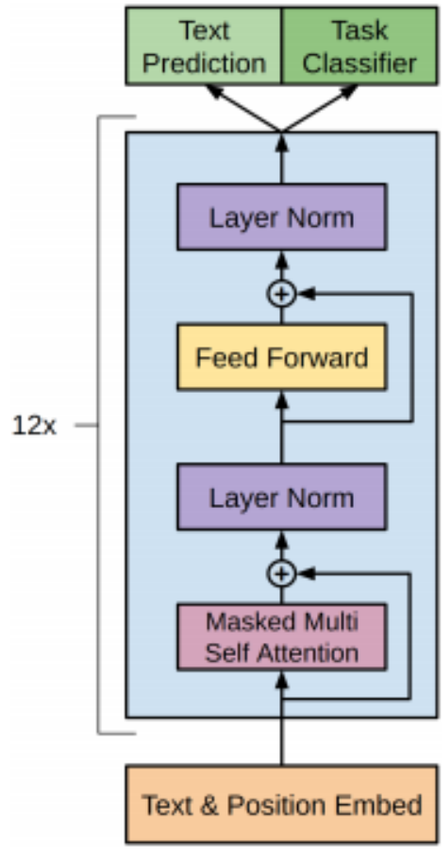
OpenAI's Transformer Language Model

- Same basic idea as ELMo, but many small differences (and many open questions!), including:
- *Transformer* encoder architecture.



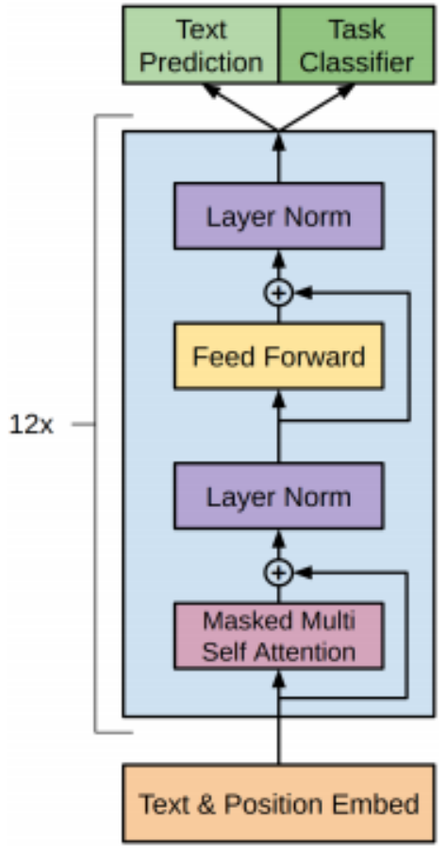
OpenAI's Transformer Language Model

- Same basic idea as ELMo, but many small differences (and many open questions!), including:
  - *Transformer* encoder architecture.
  - Entire network is *fine-tuned* for each task; few new parameters are added.

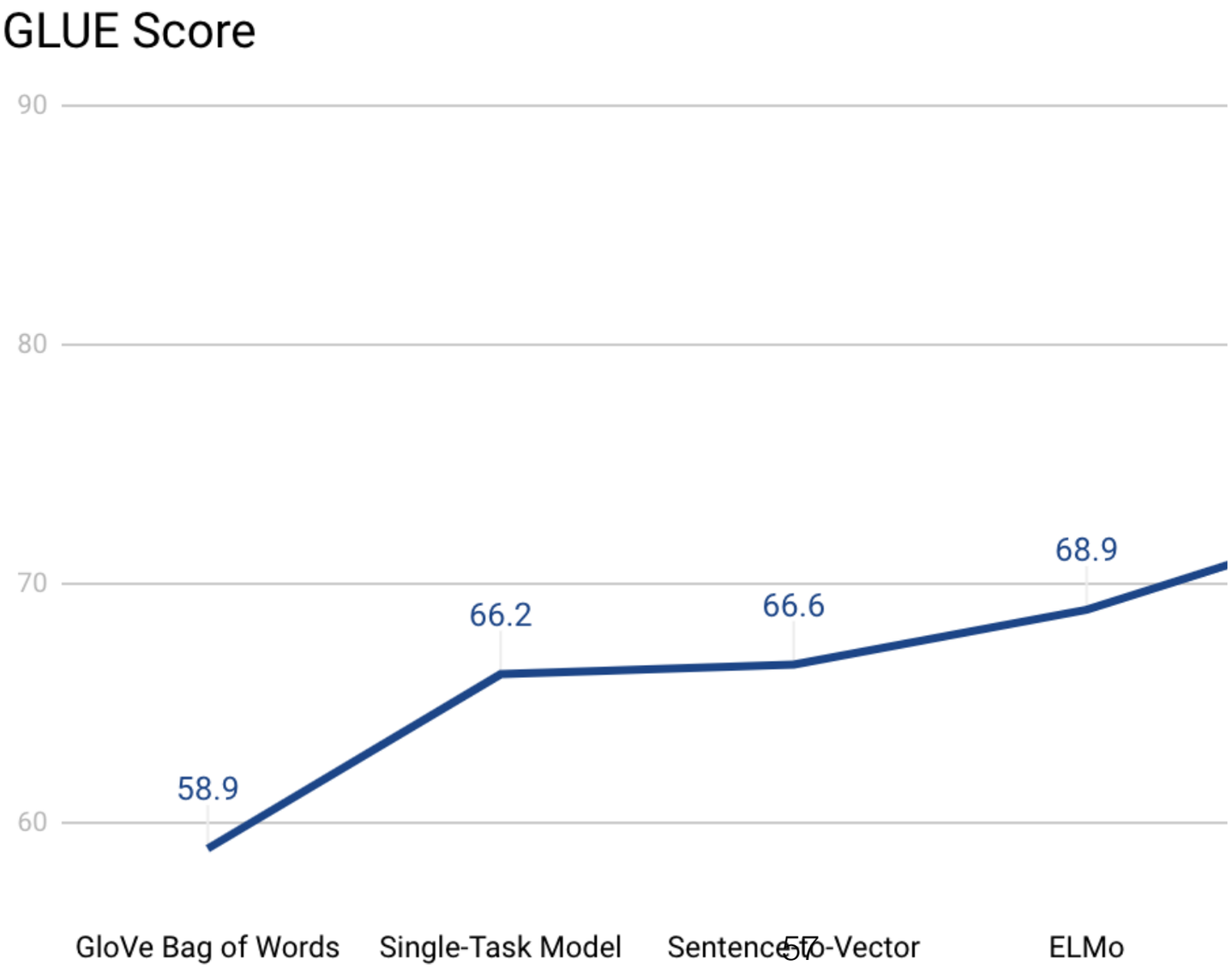


OpenAI's Transformer Language Model

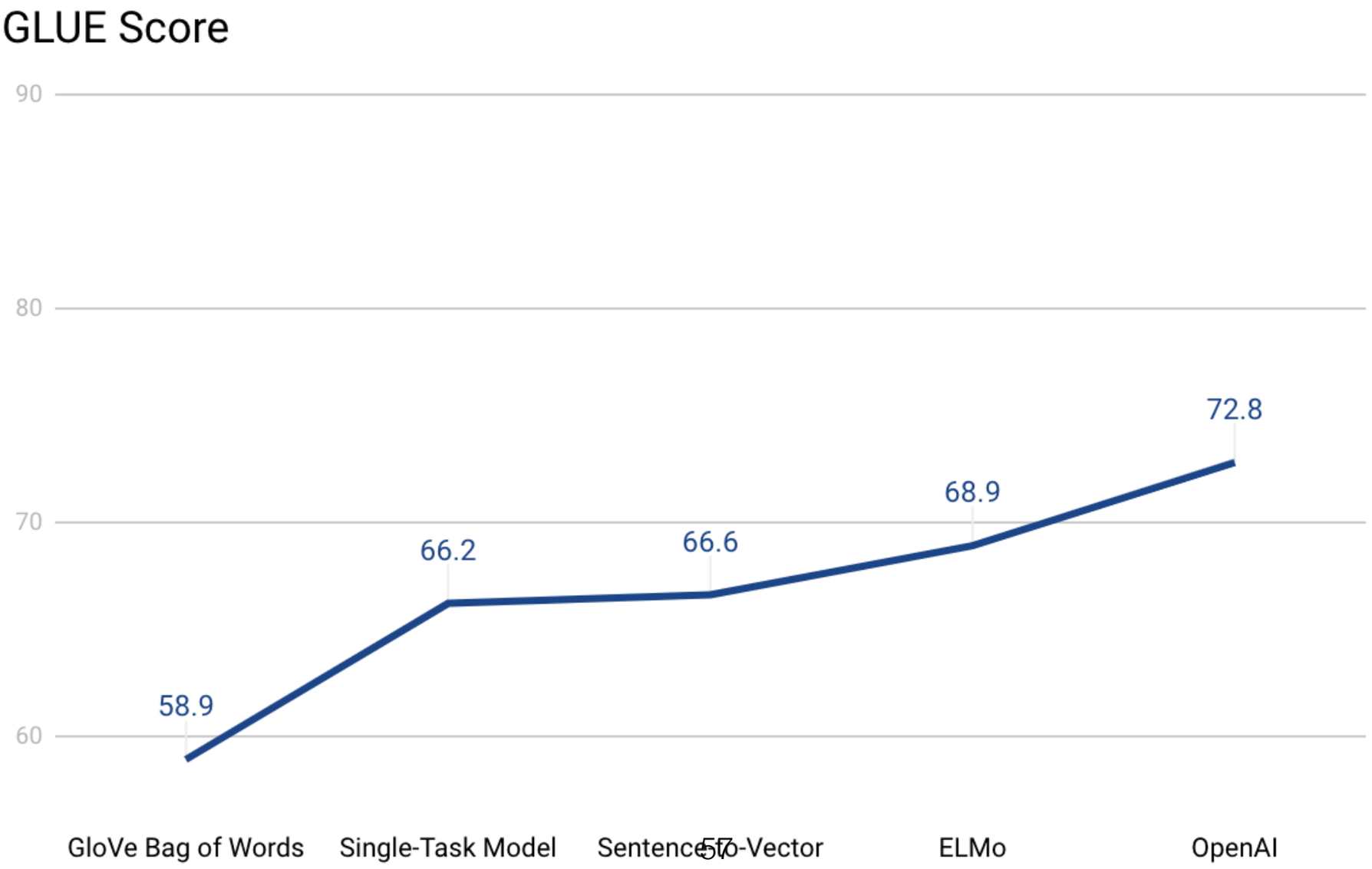
- Same basic idea as ELMo, but many small differences (and many open questions!), including:
  - *Transformer* encoder architecture.
  - Entire network is *fine-tuned* for each task; few new parameters are added.
  - Pretraining is on long spans of running text, not just isolated sentences.



What Methods Work?



What Methods Work?



The Transformer Architecture

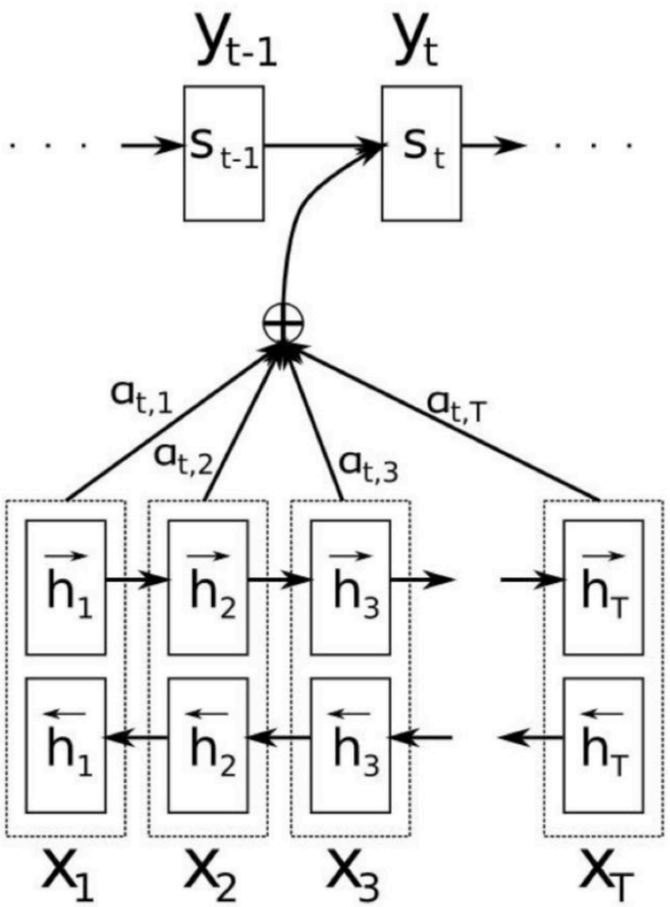
Attention (Bahdanau et al., 2015)

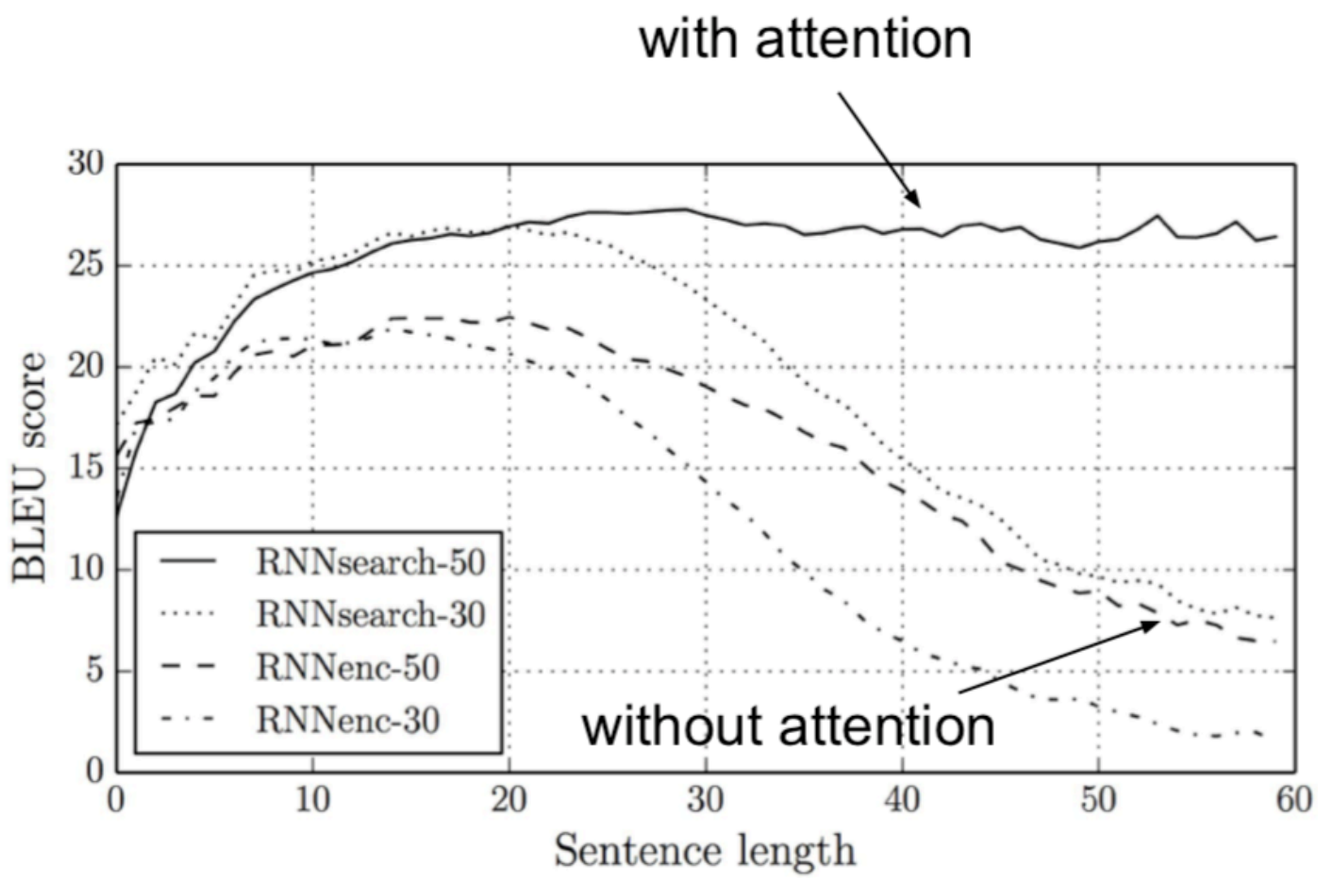
- We want to compute how important each input is for a given output

Attention (Bahdanau et al., 2015)

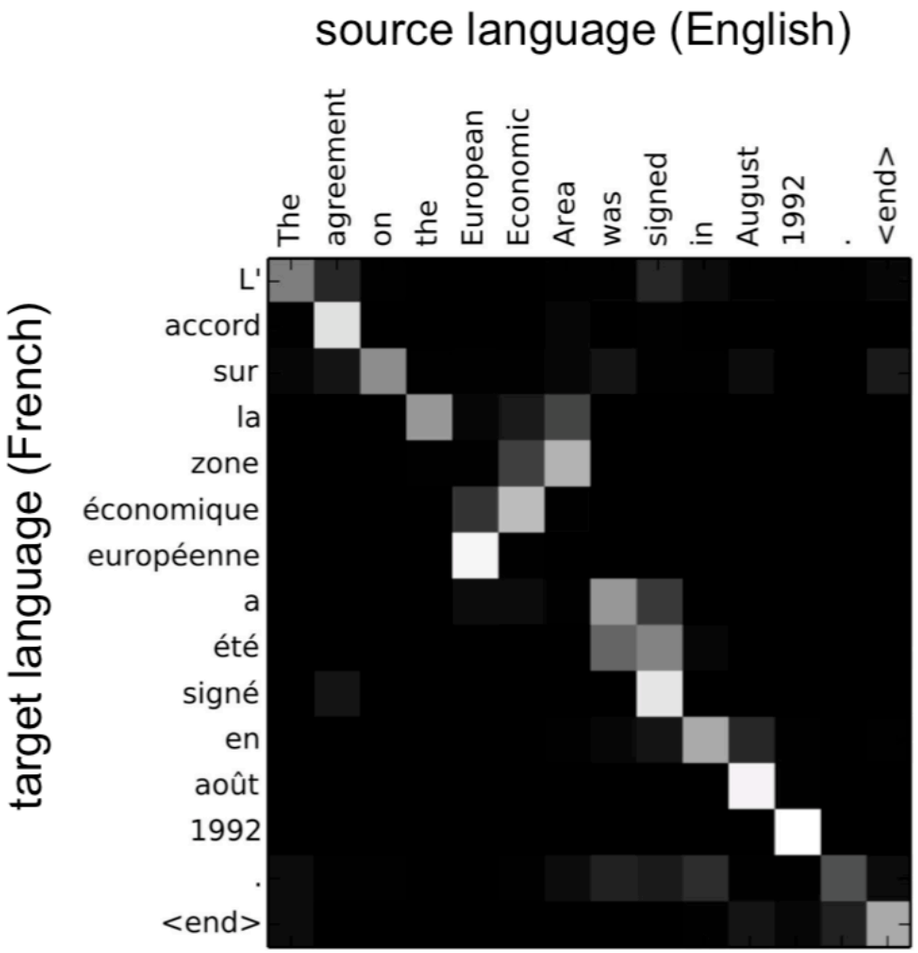
- We want to compute how important each input is for a given output
- “Attention” corresponds to an automatically computed weight for each input (at each time step in sequence-to-sequence models)
  - (We will discuss sequence-to-sequence models in detail when discussion machine translation!)

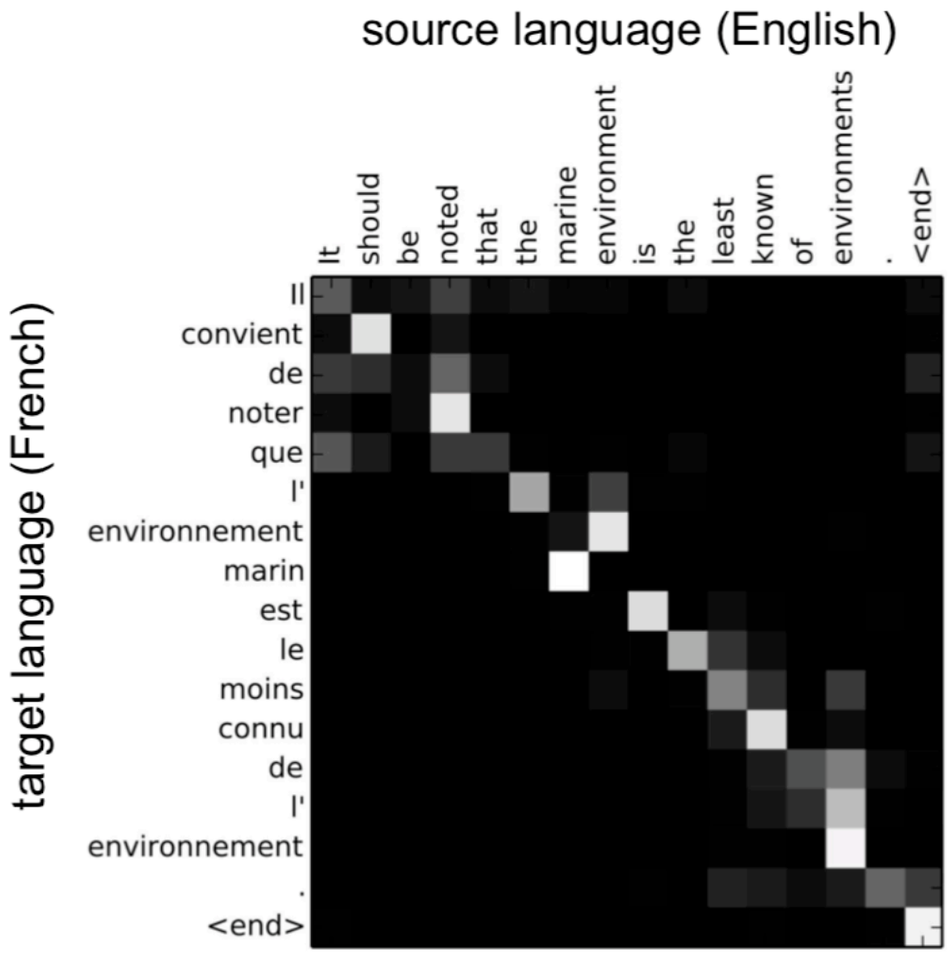
Attention (Bahdanau et al., 2015)





Attention (Bahdanau et al., 2015)





“Attention is all you need” (Vaswani et al., 2017)

Problem with RNN seq2seq architecture:

- Not parallelizable! (Why?)

“Attention is all you need” (Vaswani et al., 2017)

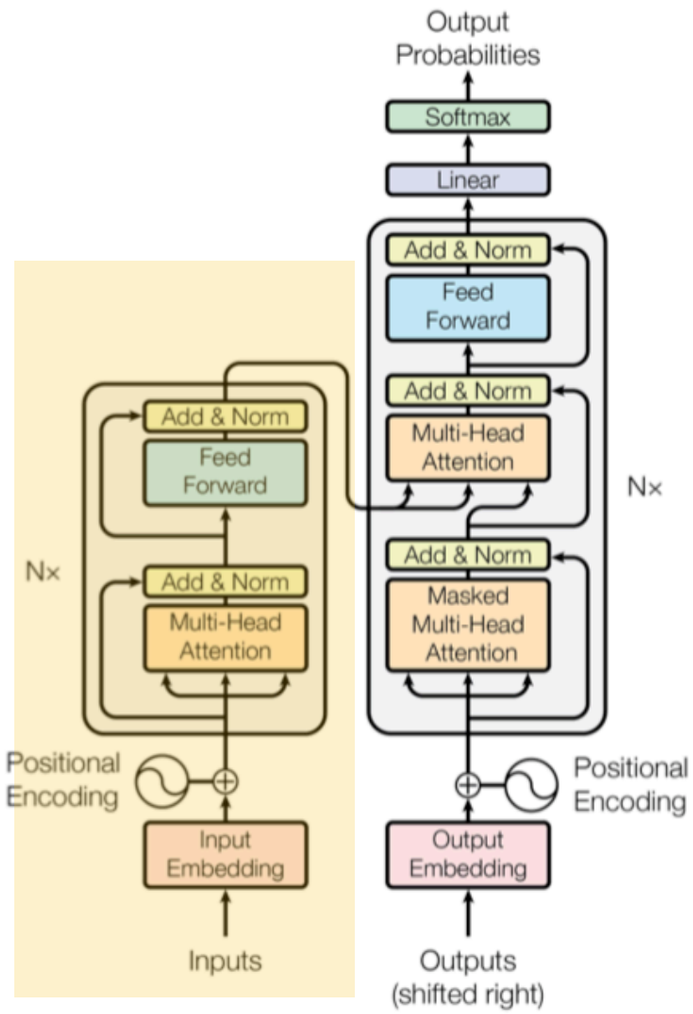
Problem with RNN seq2seq architecture:

- Not parallelizable! (Why?)

Suggested solution:

- Transformer architecture: parallelizable!
- Can reach state-of-the-art performance in 12h of training

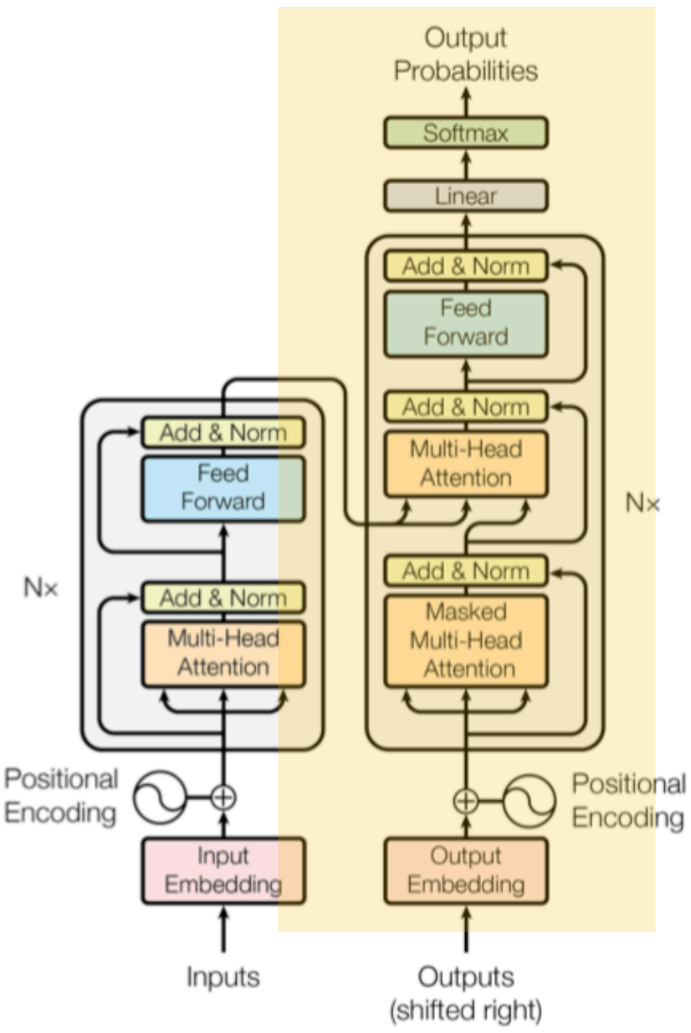
“Attention is all you need” (Vaswani et al., 2017)



Encoder:

- Consists of N=6 layers
- Each layer consists of 2 sub-layers
  - Self-attention (access to all positions in the encoder)
  - Feed-forward network
- Inputs:
  - Current embedding
  - Position embedding

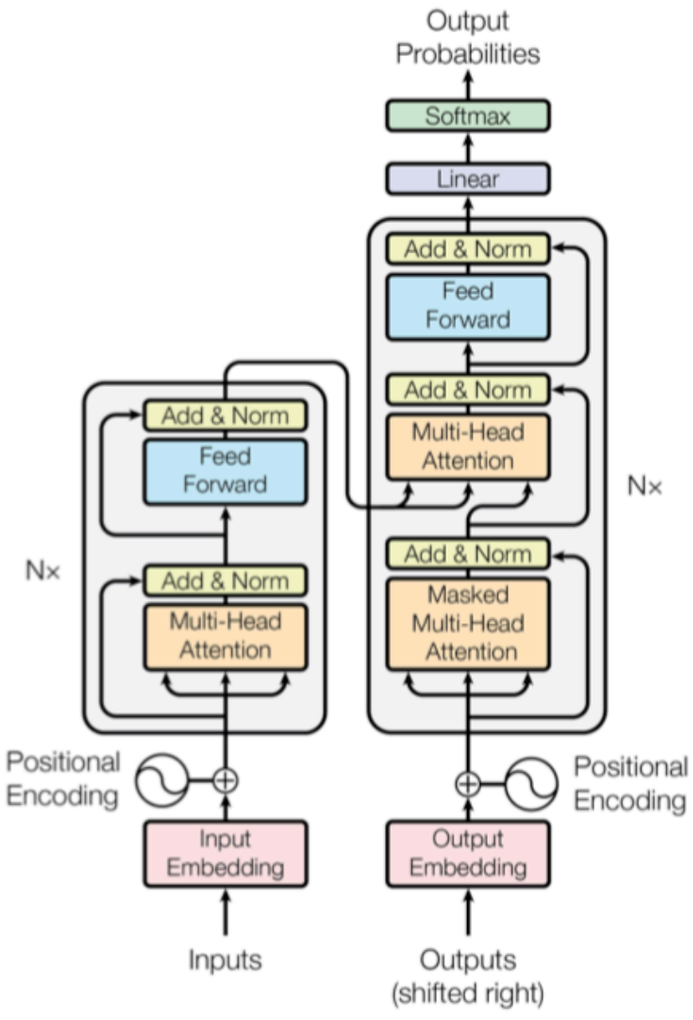
“Attention is all you need” (Vaswani et al., 2017)



Decoder:

- Also consists of N=6 layers
- Each layer consists of **3 sub-layers**
  - Self-attention (access to all previous positions in the decoder)
  - **Encoder-decoder attention**
  - Feed-forward network
- Inputs:
  - Current embedding
  - Position embedding

“Attention is all you need” (Vaswani et al., 2017)



Training:

- Can be done in parallel for all output positions

How about testing?

“Attention is all you need” (Vaswani et al., 2017)

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

“Attention is all you need” (Vaswani et al., 2017)

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

BERT and Co.

BERT (Devlin et al., 2018)

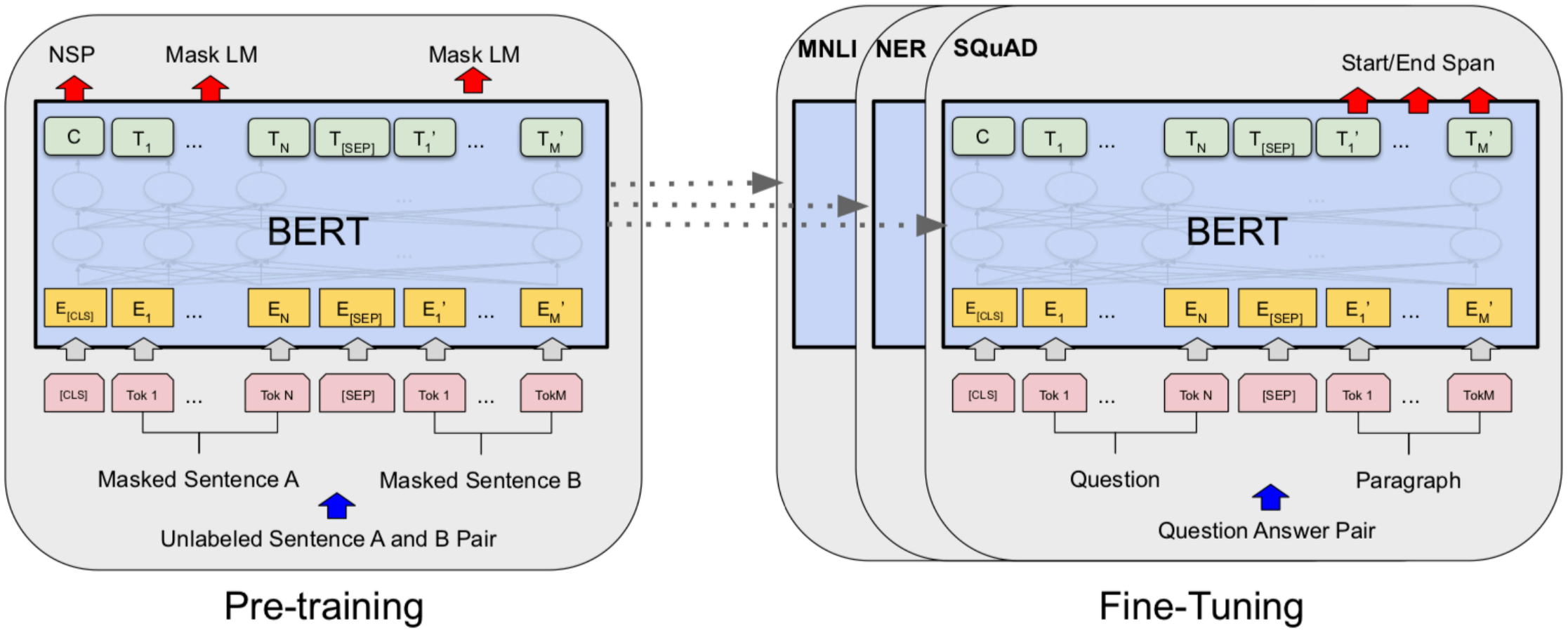


Figure from Devlin et al. (2018)

BERT (Devlin et al., 2018)

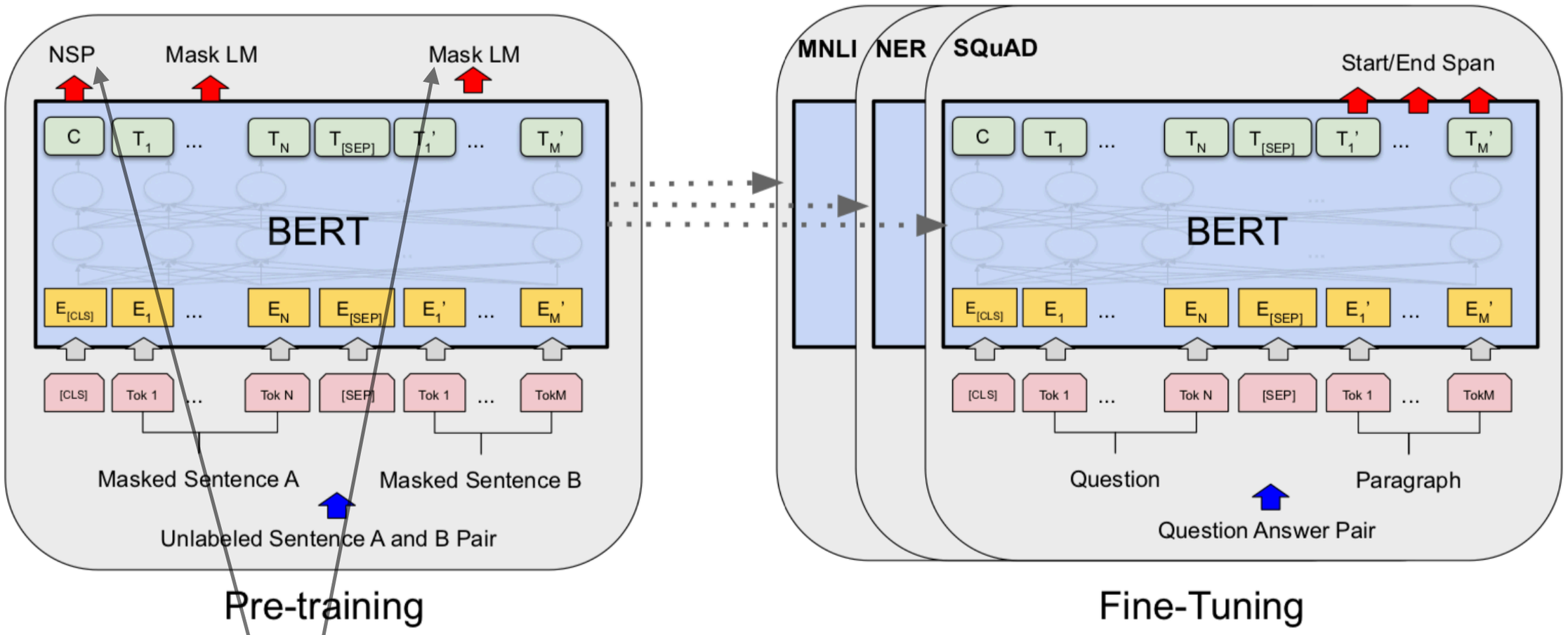


Figure from Devlin et al. (2018)

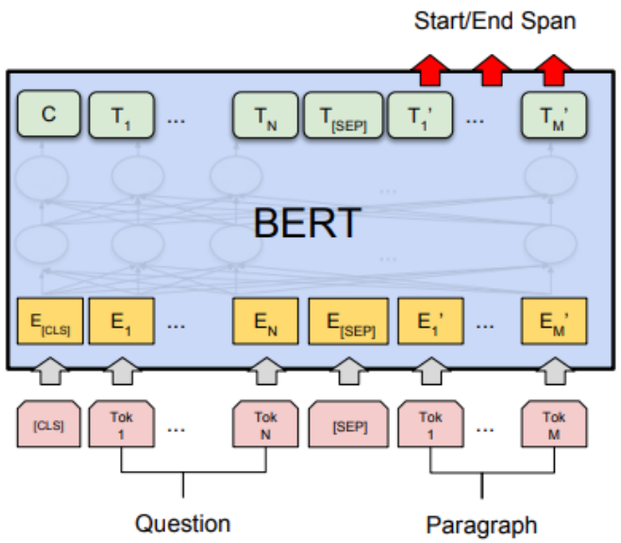
Training objectives

The BERT Model

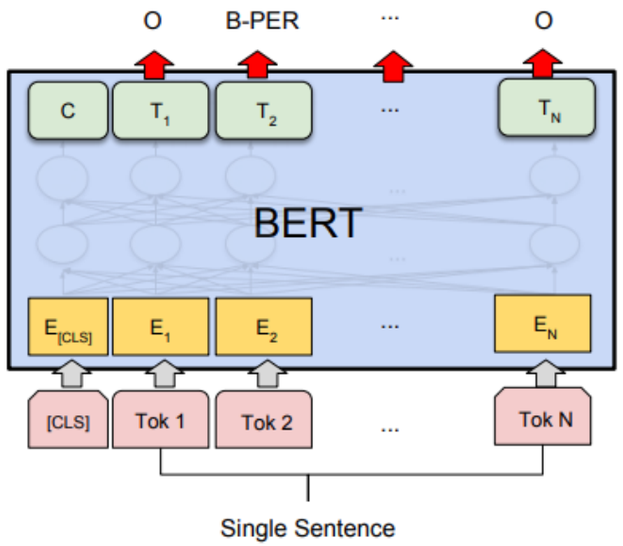
- Same basic idea as OpenAI, but many small differences, including:
  - Two different unlabeled data tasks in place of language modeling.
    - Neither requires 'predicting the future', so we can use an encoder-style Transformer rather than decoder-style.
- Very big (24 layers, >300M params).

The BERT Model

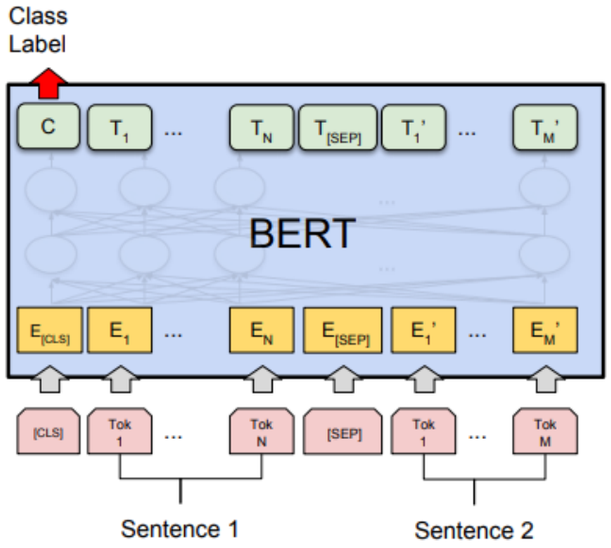
- For downstream tasks, an additional classification layer is added
  - The original output layer is discarded
- On top of what exactly we add the layer depends on the task



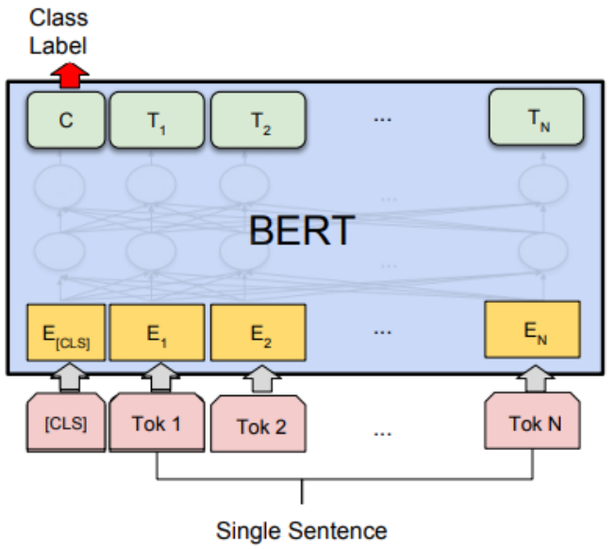
(c) Question Answering Tasks:
SQuAD v1.1



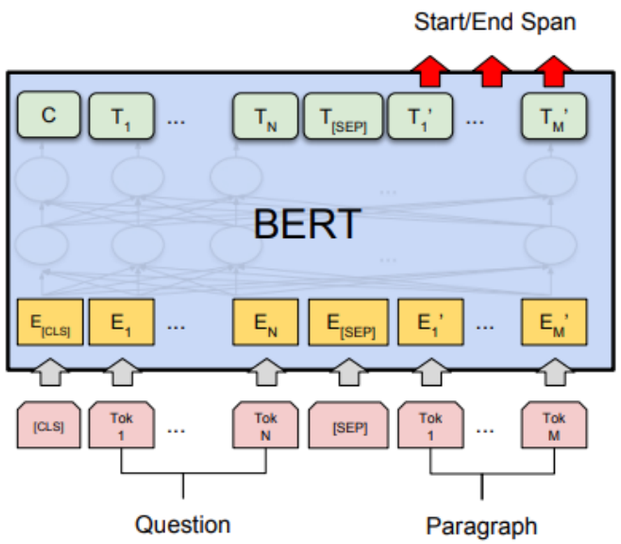
(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER



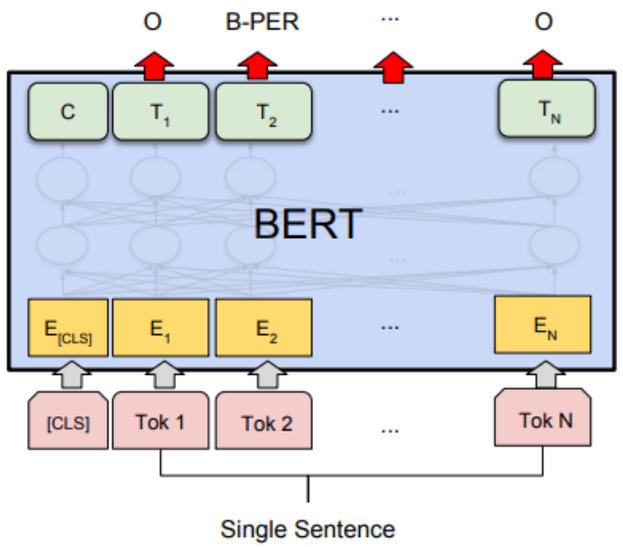
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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

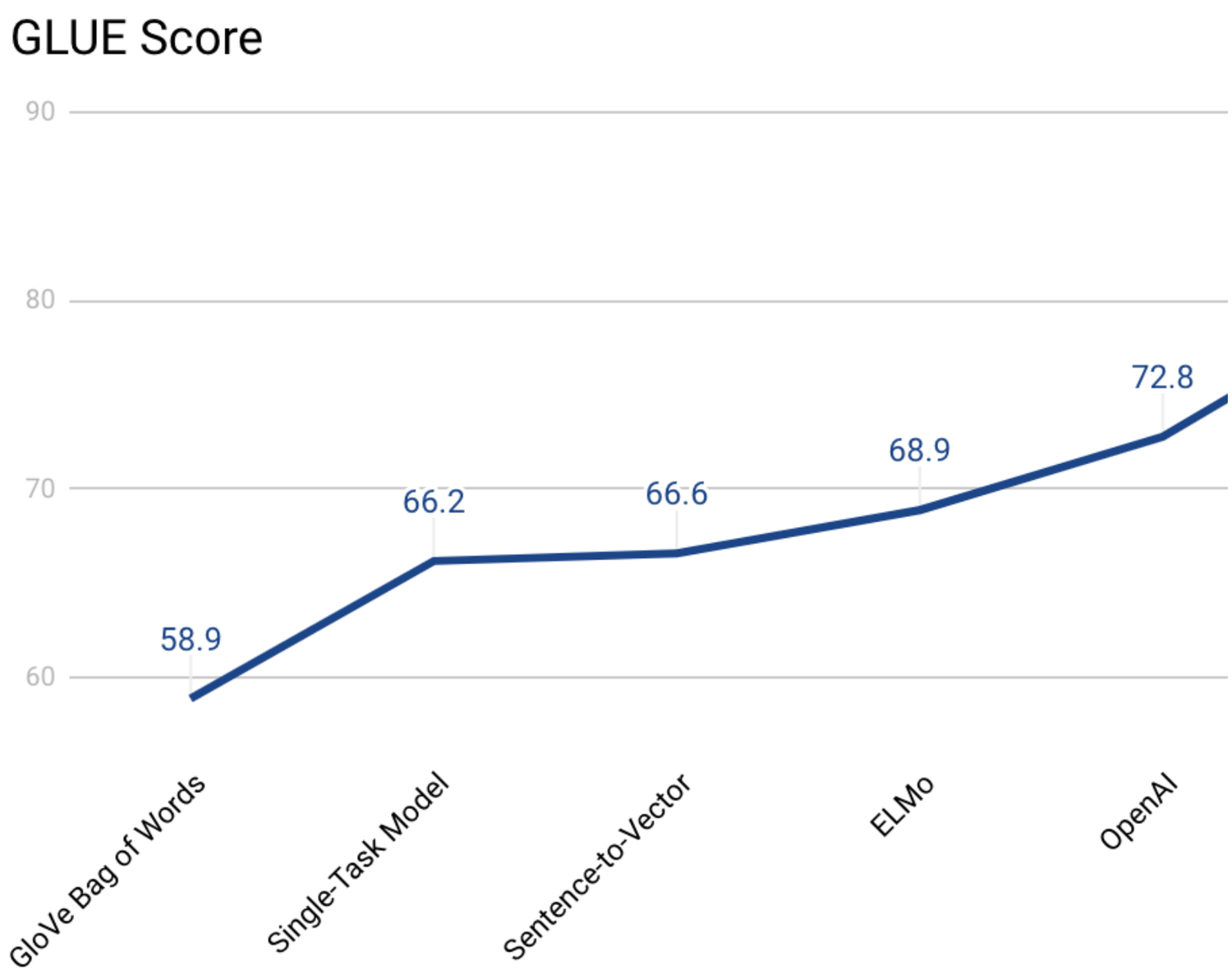


(c) Question Answering Tasks: SQuAD v1.1

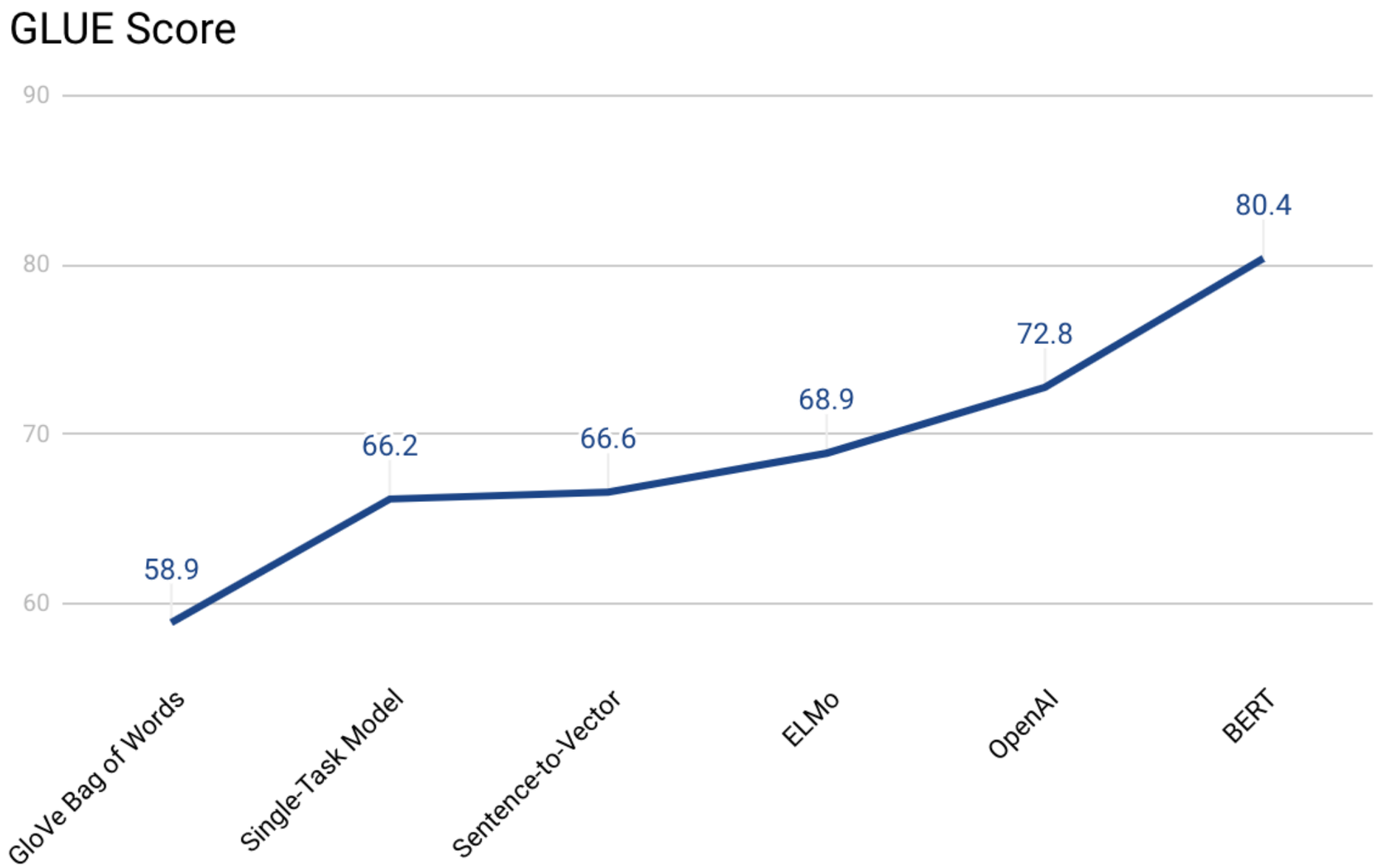


(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

What Methods Work?



What Methods Work?



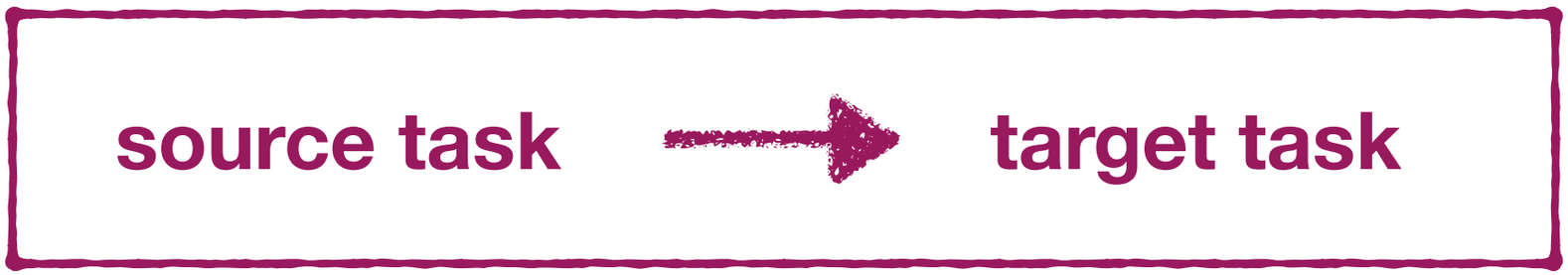
After BERT…

- Improved versions of BERT are constantly being proposed
- Keep an eye out for that
- Make sure you know the state of the art for your task

Model Adaptation

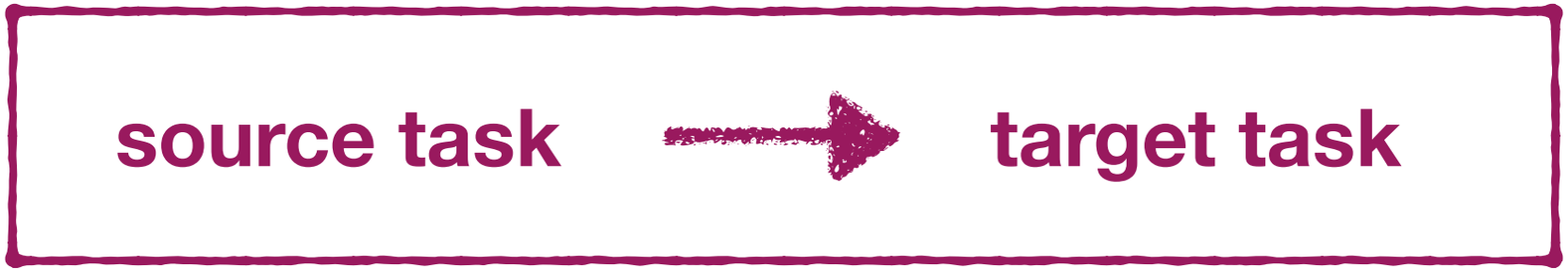
When to Use Finetuning?

- Target task is different from source task



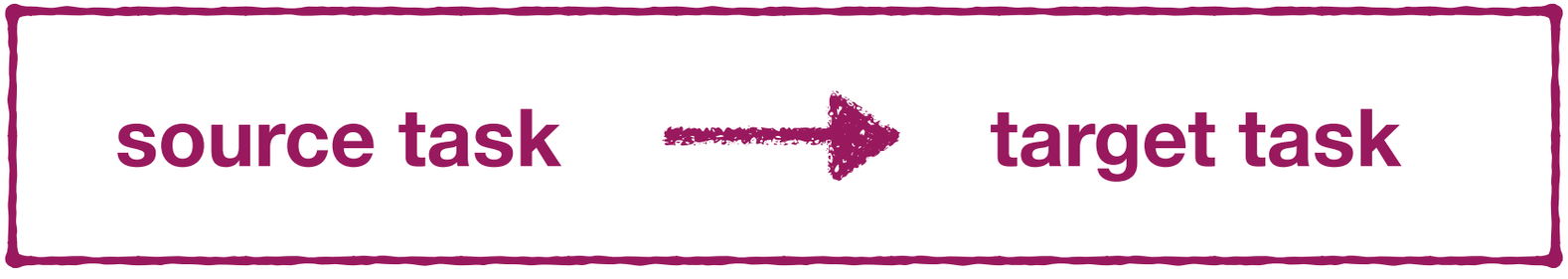
When to Use Finetuning?

- Target task is different from source task
- Target domain is different from source domain (“domain adaptation”)



When to Use Finetuning?

- Target task is different from source task
- Target domain is different from source domain (“domain adaptation”)
- Target language is different from source language

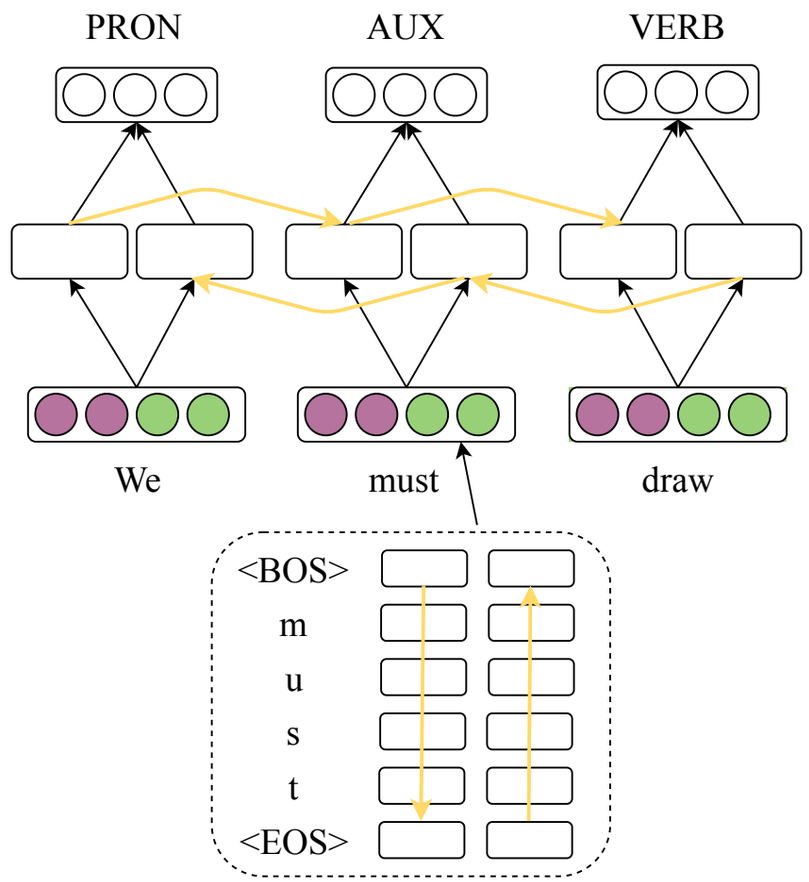


When to Use Finetuning?

- Only if we have data available
- What can we do if we don’t have task-specific training data?

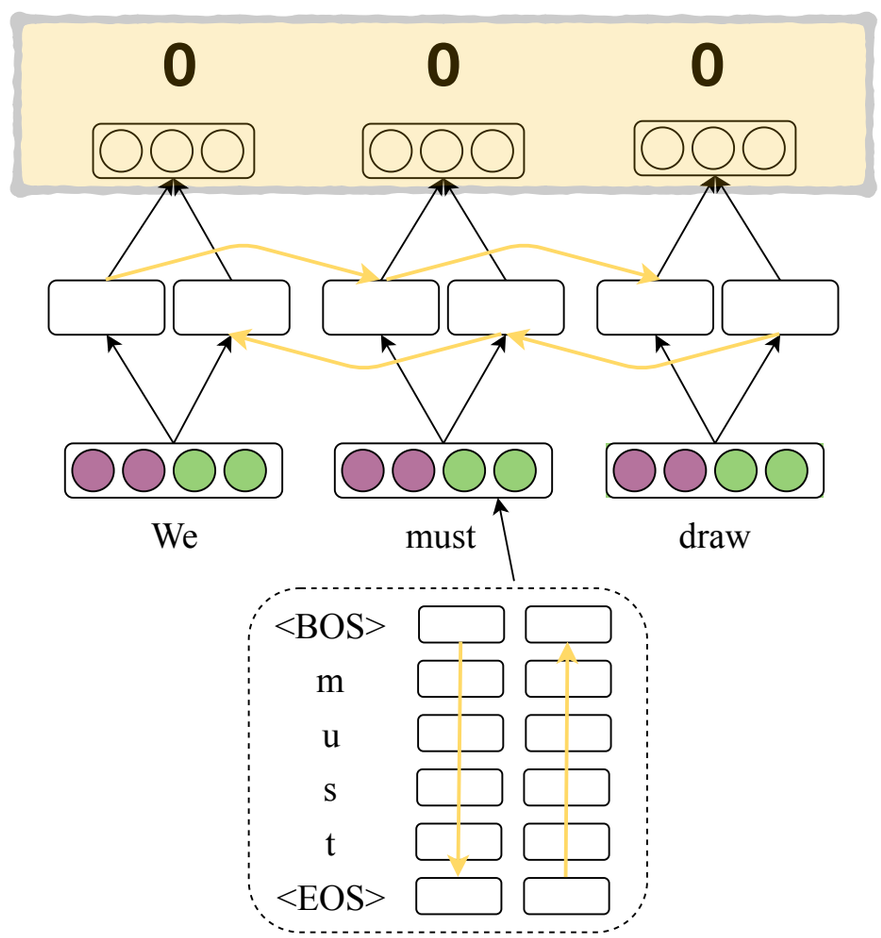
Design Decisions

- Sharing Information
  - Which parameters should be updated?
  - How about unfreezing?
- Training Regime
  - Learning rate?
  - How many epochs?



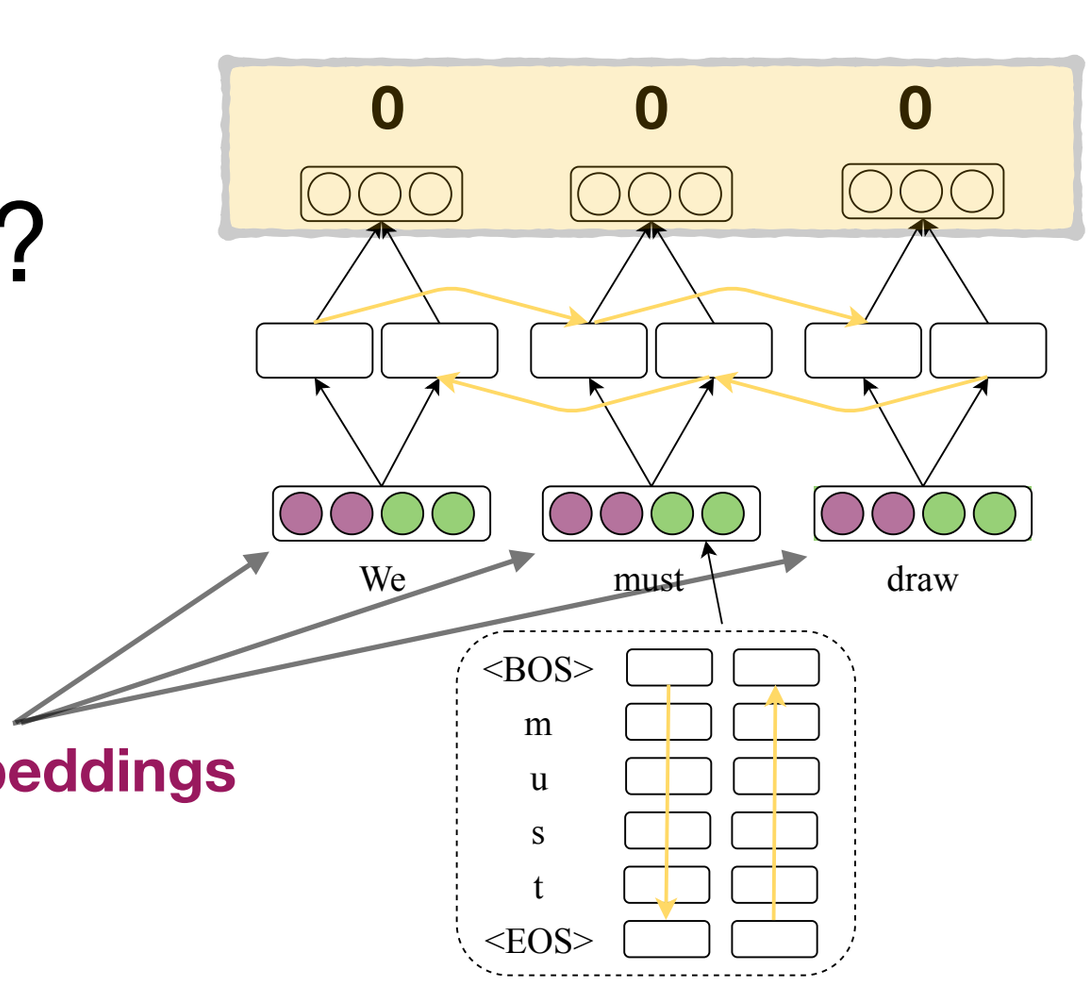
Design Decisions

- Sharing Information
  - Which parameters should be updated?
  - How about unfreezing?
- Training Regime
  - Learning rate?
  - How many epochs?



Design Decisions

- Sharing Information
  - Which parameters should be updated?
  - How about unfreezing?
- Training Regime
  - Learning rate?
  - How many epochs?



Wrapping up

- Discussed today:
  - Pretraining
  - How to read research papers and ELMo
  - GLUE
  - The transformer architecture
  - BERT and Co.
  - Model adaptation

- On Monday: Transfer Learning: Supervised Pretraining and Multi-task Training