Sentence Representation Learning: Evaluation and the State of the Art

Sam Bowman
Asst. Prof. of Data Science and Linguistics, NYU
General Purpose Sentence Representation Learning Team, JSALT

Includes work with Alex Wang (NYU CS/JSALT), Amanpreet Singh (NYU CS), Julian Michael (UW), Felix Hill (DeepMind) & Omer Levy (UW)

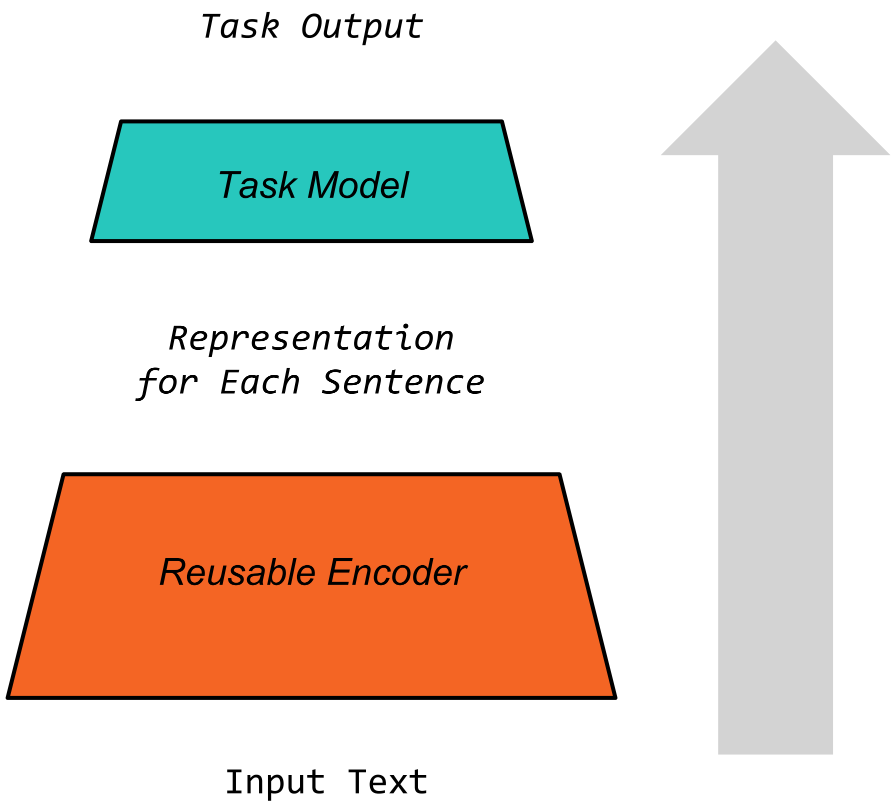
JHU HLT Summer School

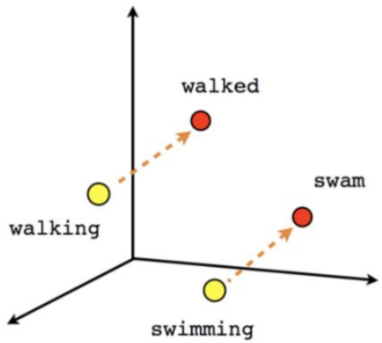
Sentence Representation Learning

The Long-Term Goal

To develop a general-purpose neural network sentence encoder which produces substantial gains in performance and data efficiency across diverse NLU tasks.

A general-purpose sentence encoder





General purpose representation learning

Words:

- Distributional word vectors:
  SENNA, word2vec, GloVe, fastText, etc.

Images:

- ImageNet-trained deep CNNs

Sentences:

- Promising results just emerging this Spring!

Where might this be valuable?

Scenario 1: *An engineer wants to solve some English sentence understanding task for which no data exists.*

Examples:

- Intent detection for a new Alexa skill
- Customer service ticket classification for a new business

  ...

Where might this be valuable?

Scenario 1: *An engineer wants to solve some English sentence understanding task for which no data exists.*

Now:

- Pay to annotate 10k–1m examples at $0.05–0.50 each
- Train a BiLSTM-based classification/regression model over word embeddings

With effective sentence representations:

- Train a model over the outputs of an existing encoder.

➔ Comparable performance with ~1–10% the parameters.

Where might this be valuable?

Scenario 2: *An engineer wants to solve some English sentence understanding task for which ample labeled data exists, but performance is still inadequate.*

Examples:

- Major language machine translation
- Question answering over short texts

  ...

Where might this be valuable?

Scenario 2: *An engineer wants to solve some English sentence understanding task for which ample labeled data exists, but performance is still inadequate.*

Now:

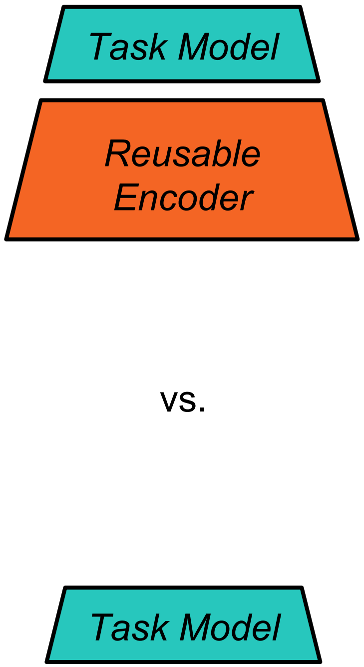
- Train BiLSTM/Seq.-to-seq./etc. over word embeddings

With effective sentence representations:

- Use a general-purpose encoder as the input layer(s) of the model

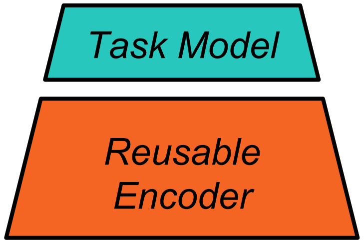
➔ Prior knowledge of English makes learning more effective

A general-purpose sentence encoder



Is This Possible?

- Yes. (It's easy at least *harmless*.)
- If you have an oracle to give you the optimal task model (in teal), then the identity function will be at least as good as any other encoder…
    - ... but we have no such oracle. Since we must search for the task model using supervised learning with as few as 100 training examples, having a pre-trained encoder extract informative features will improve the odds that we can identify an adequate function.



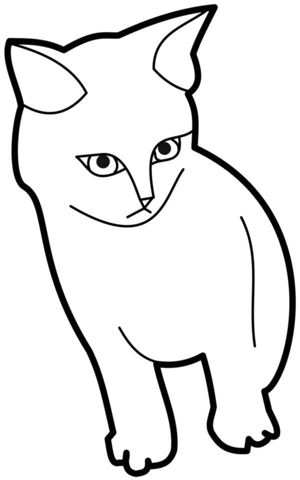
A general-purpose sentence encoder

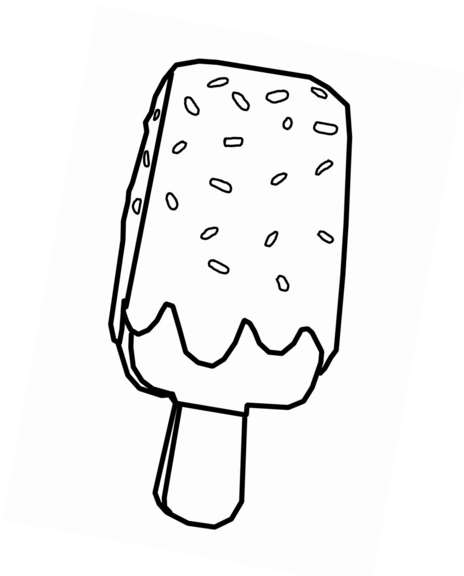
Roughly, we might expect effective encodings to capture:

- Lexical contents and word order.
- (Rough) syntactic structure.
- Cues to idiomatic/non-compositional phrase meanings.
- Cues to connotation and social meaning.
- Disambiguated semantic information of the kind expressed in a semantic parse (or formal semantic analysis).

$$\forall x[\text{patient}'(x) \rightarrow \exists y[\text{doctor}'(y) \wedge \text{treat}'(y, x)]]$$

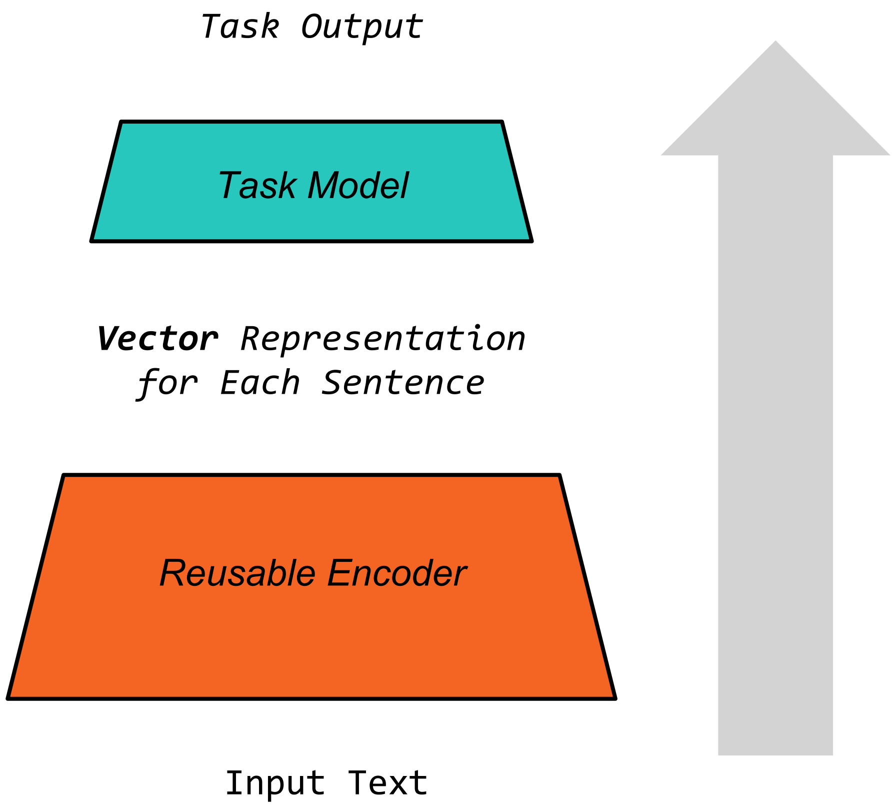
Outline

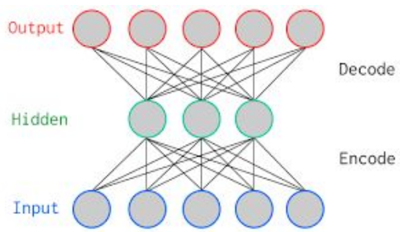




- **Background: Sentence-to-vector Encoders**
- Recent progress: Newer Encoders
- Evaluation: GLUE
- *Very* recent progress: OpenAI
- The JSALT Project

A general-purpose sentence encoder





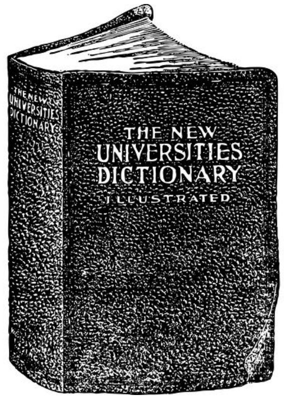
Progress to date

Unsupervised training on single sentences:

- Sequence autoencoders (Dai and Le ‘15)
- Paragraph vector (Le and Mikolov ‘15)
- Variational Autoencoder LM (Bowman et al. ‘16)
- Denoising autoencoders (Hill et al. ‘16)

Unsupervised training on running text:

- Skip Thought (Kiros et al. ‘15)
- FastSent (Hill et al. ‘16)
- DiscSent/DisSent (Jernite et al. ‘17/Nie et al. ‘17)



Progress to date

Supervised training on large corpora:

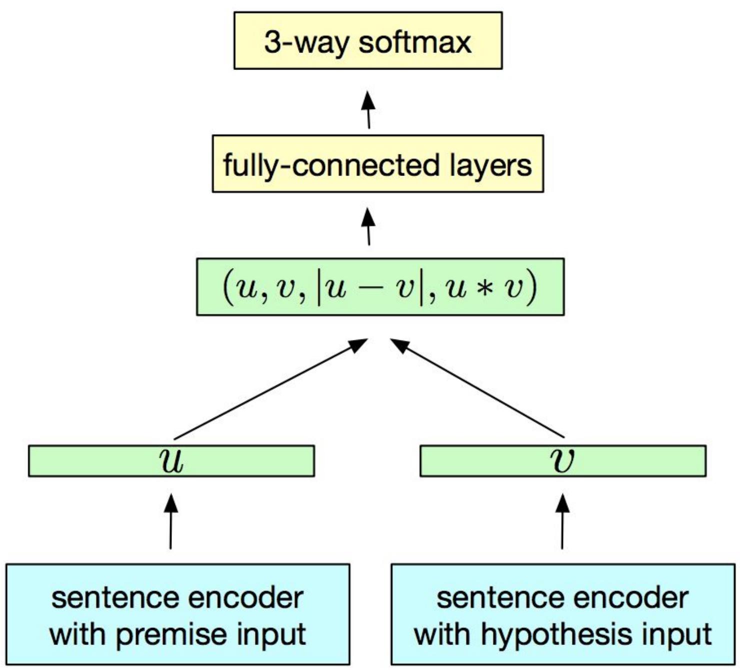
- Dictionaries (Hill et al. ‘15)
- Image captions (Hill et al. ‘16)
- Natural language inference data (Conneau et al. ‘17)
- Translated parallel corpora (McCann et al. ‘17)

The Standard Evaluation: SentEval

- Informal evaluation standard formalized by Conneau and Kiela (2018).
- Suite of ten tasks:
  - MR, CR, SUBJ, MPQA, SST, TREC, MRPC, SICK-R, SICK-E, STS-B
- Software package automatically trains and evaluates per-task linear classifiers using supplied representations.

Case Study: InferSent

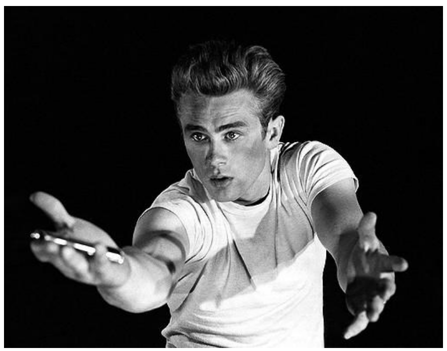
Sentence encoder pretrained for *natural language inference*.



Conneau et al. '17

Natural Language Inference (NLI)

also known as recognizing textual entailment (RTE)



James Byron Dean refused to move without blue jeans

{**entails**, contradicts, neither}

James Dean didn't dance without pants

Example: MacCartney thesis '09

Judging Understanding with NLI

To reliably perform well at NLI, your representations of meaning must handle with the full complexity of compositional semantics…

- Lexical entailment (*cat* vs. *animal*, *cat* vs. *dog*)
- Quantification (*all*, *most*, *fewer than eight*)
- Lexical ambiguity and scope ambiguity (*bank*, ...)
- Modality (*might*, *should*, ...)
- Common sense background knowledge

…while *avoiding* most of the other hard problems in NLP: grounding, text generation, knowledge base access, and structured prediction.

Background: SNLI and MultiNLI

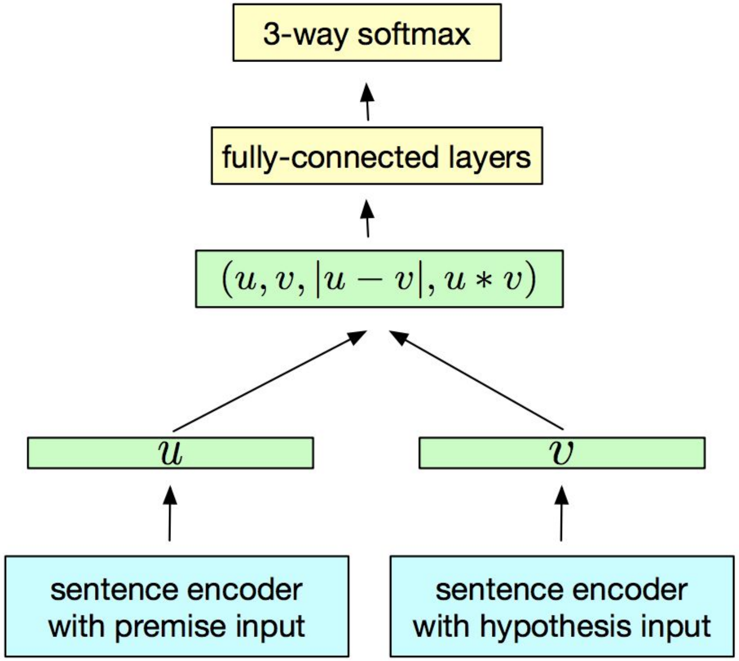
- ~1m sentence pairs created and labeled by crowd workers.
- Balanced classification task: *Entailment*, *contradiction*, *neutral*.
- Split across several genres of written and spoken language.

At 8:34, the Boston Center controller received a third transmission from American 11	9/11 **entailment** E E E E	The Boston Center controller got a third transmission from American 11.
I am a lacto-vegetarian.	SLATE **neutral** N N E N	I enjoy eating cheese too much to abstain from dairy.
someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny	TELEPHONE **contradiction** C C C C	No one noticed and it wasn't funny at all.

Bowman et al. '15; Williams et al. '18

Case Study: InferSent

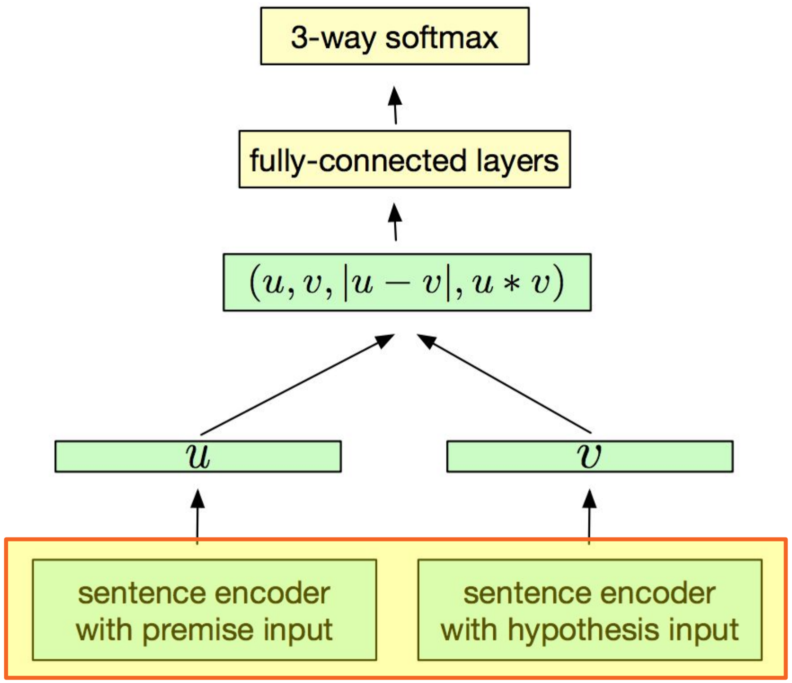
Sentence encoder pretrained for *natural language inference*.



Conneau et al. '17

Case Study: InferSent

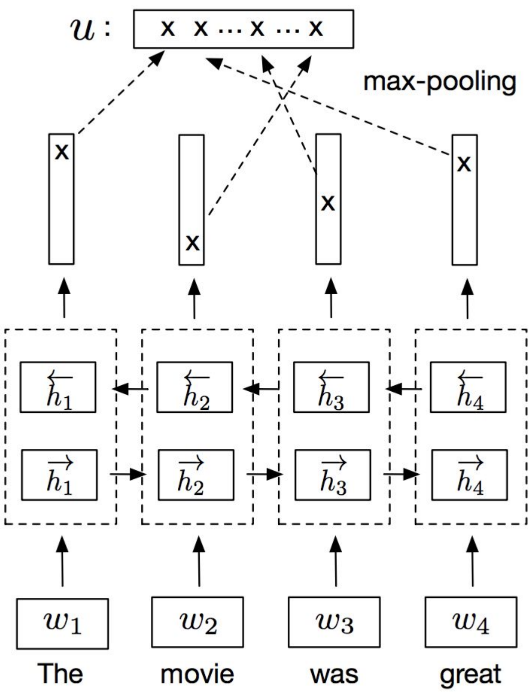
Sentence encoder pretrained for *natural language inference*.



Conneau et al. '17

Case Study: InferSent

Encoder: Bidirectional LSTM RNN with max pooling



Conneau et al. ‘17

Subramanian et al. '18

Results on SentEval

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-R	SICK-E	STSB
Transfer approaches										
FastSent	70.8	78.4	88.7	80.6	-	76.8	72.2/80.3	-	-	-
FastSent+AE	71.8	76.7	88.8	81.5	-	80.4	71.2/79.1	-	-	-
NMT En-Fr	64.7	70.1	84.9	81.5	-	82.8	-	-	-	-
CNN-LSTM	77.8	82.1	93.6	89.4	-	92.6	76.5/83.8	0.862	-	-
Skipthought	76.5	80.1	93.6	87.1	82.0	92.2	73.0/82.0	0.858	82.3	-
Skipthought + LN	79.4	83.1	93.7	89.3	82.9	88.4	-	0.858	79.5	72.1/70.2
Word Embedding Average	-	-	-	-	82.2	-	-	0.860	84.6	-
DiscSent + BiGRU	-	-	88.6	-	-	81.0	71.6/-	-	-	-
DiscSent + unigram	-	-	92.7	-	-	87.9	72.5/-	-	-	-
DiscSent + embed	-	-	93.0	-	-	87.2	75.0/-	-	-	-
Byte mLSTM	**86.9**	**91.4**	**94.6**	88.5	-	-	75.0/82.8	0.792	-	-
Infersent (SST)	(*)	83.7	90.2	89.5	(*)	86.0	72.7/80.9	0.863	83.1	-
Infersent (SNLI)	79.9	84.6	92.1	89.8	83.3	88.7	75.1/82.3	0.885	86.3	-
Infersent (AllNLI)	81.1	86.3	92.4	90.2	**84.6**	88.2	76.2/83.1	0.884	86.3	75.8/75.5

Case Study: GenSen

Same model as InferSent, but trained on five different tasks at once:

- NLI
- Four sequence to sequence tasks:
  - English–French translation
  - English–German translation
  - Predicting the next sentence in a book (language modeling, aka Skip-Thought)
  - Sequence-to-sequence parsing

Subramanian et al. ‘18

Subramanian et al. '18

Results on SentEval

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-R	SICK-E	STSB	Δ
Transfer approaches											
FastSent	70.8	78.4	88.7	80.6	-	76.8	72.2/80.3	-	-	-	-
FastSent+AE	71.8	76.7	88.8	81.5	-	80.4	71.2/79.1	-	-	-	-
NMT En-Fr	64.7	70.1	84.9	81.5	-	82.8	-	-	-	-	-
CNN-LSTM	77.8	82.1	93.6	89.4	-	92.6	76.5/83.8	0.862	-	-	-
Skipthought	76.5	80.1	93.6	87.1	82.0	92.2	73.0/82.0	0.858	82.3	-	-
Skipthought + LN	79.4	83.1	93.7	89.3	82.9	88.4	-	0.858	79.5	72.1/70.2	-
Word Embedding Average	-	-	-	-	82.2	-	-	0.860	84.6	-	-
DiscSent + BiGRU	-	-	88.6	-	-	81.0	71.6/-	-	-	-	-
DiscSent + unigram	-	-	92.7	-	-	87.9	72.5/-	-	-	-	-
DiscSent + embed	-	-	93.0	-	-	87.2	75.0/-	-	-	-	-
Byte mLSTM	**86.9**	**91.4**	**94.6**	88.5	-	-	75.0/82.8	0.792	-	-	-
Infersent (SST)	(*)	83.7	90.2	89.5	(*)	86.0	72.7/80.9	0.863	83.1	-	-
Infersent (SNLI)	79.9	84.6	92.1	89.8	83.3	88.7	75.1/82.3	0.885	86.3	-	-
Infersent (AllNLI)	81.1	86.3	92.4	90.2	**84.6**	88.2	76.2/83.1	0.884	86.3	75.8/75.5	0.0
Our Models											
+STN	78.9	85.8	93.7	87.2	80.4	84.2	72.4/81.6	0.840	82.1	72.9/72.4	-2.56
+STN +Fr +De	80.3	85.1	93.5	90.1	83.3	92.6	77.1/83.3	0.864	84.8	77.1/77.1	0.01
+STN +Fr +De +NLI	81.2	86.4	93.4	90.8	84.0	93.2	76.6/82.7	0.884	87.0	79.2/79.1	0.99
+STN +Fr +De +NLI +L	81.7	87.3	94.2	90.8	84.0	**94.2**	77.1/83.0	0.887	87.1	78.7/78.2	1.33
+STN +Fr +De +NLI +L +STP	82.7	88.0	94.1	91.2	84.5	92.4	77.8/83.9	0.885	86.8	78.7/78.4	1.44
+STN +Fr +De +NLI +2L +STP	82.8	88.3	94.0	**91.3**	83.6	92.6	77.4/83.3	0.884	87.6	**79.2/79.1**	1.47
+STN +Fr +De +NLI +L +STP +Par	82.5	87.7	94.0	90.9	83.2	93.0	**78.6/84.4**	**0.888**	**87.8**	78.9/78.6	**1.48**

Caveat: Results on SentEval

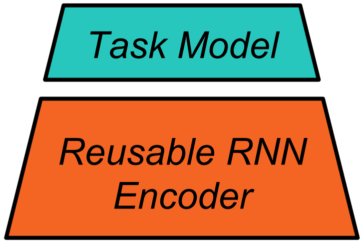
Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-R	SICK-E	STSB	Δ
Transfer approaches											
FastSent	70.8	78.4	88.7	80.6	-	76.8	72.2/80.3	-	-	-	-
FastSent+AE	71.8	76.7	88.8	81.5	-	80.4	71.2/79.1	-	-	-	-
NMT En-Fr	64.7	70.1	84.9	81.5	-	82.8	-	-	-	-	-
CNN-LSTM	77.8	82.1	93.6	89.4	-	92.6	76.5/83.8	0.862	-	-	-
Skipthought	76.5	80.1	93.6	87.1	82.0	92.2	73.0/82.0	0.858	82.3	-	-
Skipthought + LN	79.4	83.1	93.7	89.3	82.9	88.4	-	0.858	79.5	72.1/70.2	-
Word Embedding Average	-	-	-	-	82.2	-	-	0.860	84.6	-	-
DiscSent + BiGRU	-	-	88.6	-	-	81.0	71.6/-	-	-	-	-
DiscSent + unigram	-	-	92.7	-	-	87.9	72.5/-	-	-	-	-
DiscSent + embed	-	-	93.0	-	-	87.2	75.0/-	-	-	-	-
Byte mLSTM	**86.9**	**91.4**	**94.6**	88.5	-	-	75.0/82.8	0.792	-	-	-
Infersent (SST)	(*)	83.7	90.2	89.5	(*)	86.0	72.7/80.9	0.863	83.1	-	-
Infersent (SNLI)	79.9	84.6	92.1	89.8	83.3	88.7	75.1/82.3	0.885	86.3	-	-
Infersent (AllNLI)	81.1	86.3	92.4	90.2	**84.6**	88.2	76.2/83.1	0.884	86.3	75.8/75.5	0.0
Our Models											
+STN	78.9	85.8	93.7	87.2	80.4	84.2	72.4/81.6	0.840	82.1	72.9/72.4	-2.56
+STN +Fr +De	80.3	85.1	93.5	90.1	83.3	92.6	77.1/83.3	0.864	84.8	77.1/77.1	0.01
+STN +Fr +De +NLI	81.2	86.4	93.4	90.8	84.0	93.2	76.6/82.7	0.884	87.0	79.2/79.1	0.99
+STN +Fr +De +NLI +L	81.7	87.3	94.2	90.8	84.0	**94.2**	77.1/83.0	0.887	87.1	78.7/78.2	1.33
+STN +Fr +De +NLI +L +STP	82.7	88.0	94.1	91.2	84.5	92.4	77.8/83.9	0.885	86.8	78.7/78.4	1.44
+STN +Fr +De +NLI +2L +STP	82.8	88.3	94.0	**91.3**	83.6	92.6	77.4/83.3	0.884	87.6	**79.2/79.1**	1.47
+STN +Fr +De +NLI +L +STP +Par	82.5	87.7	94.0	90.9	83.2	93.0	**78.6/84.4**	**0.888**	**87.8**	78.9/78.6	**1.48**
Approaches trained from scratch on these tasks											
Naive Bayes SVM	79.4	81.8	93.2	86.3	83.1	-	-	-	-	-	
AdaSent	83.1	86.3	95.5	93.3	-	92.4	-	-	-	-	
TF-KLD	-	-	-	-	-	-	80.4/85.9	-	-	-	
Illinois LH	-	-	-	-	-	-	-	-	84.5	-	
Dependency tree LSTM	-	-	-	-	-	-	-	0.868	-	-	
Neural Semantic Encoder	-	-	-	-	89.7	-	-	-	-	-	
BLSTM-2DCNN	82.3	-	94.0	-	89.5	96.1	-	-	-	-	

The Standard Evaluation: SentEval

- Informal evaluation standard formalized by Conneau and Kiela (2018).
- Suite of ten tasks:
    - MR, CR, SUBJ, MPQA, SST, TREC, MRPC, SICK-R, SICK-E, STS-B
- Software package automatically trains and evaluates per-task linear classifiers using supplied representations.

The Standard Evaluation: SentEval

- Informal evaluation standard formalized by Conneau and Kiela (2018).
- Suite of ten tasks:
  - MR, CR, SUBJ, MPQA, SST, TREC, MRPC, SICK-R, SICK-E, STS-B
- Software package automatically trains and evaluates per-task linear classifiers using supplied representations.
- Limited to sentence-to-vector models.



A general-purpose sentence encoder

General-purpose sentence representations probably won't be fixed length vectors.

- For most tasks, a sequence of vectors is preferable.
- For others, you can pool the sequence into one vector.

"You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!"

—Ray Mooney (UT Austin)

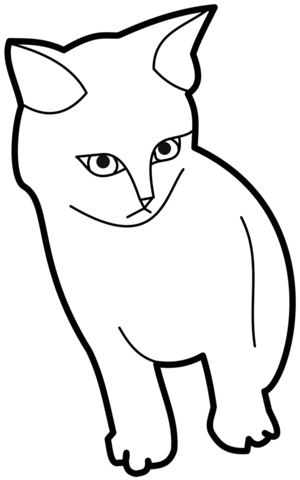
The Standard Evaluation: SentEval

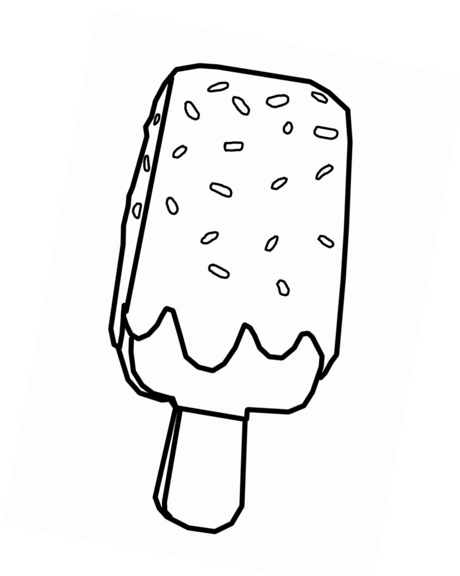
- Informal evaluation standard formalized by Conneau and Kiela (2018).
- Suite of ten tasks:
  - MR, CR, SUBJ, MPQA, SST, TREC, MRPC, SICK-R, SICK-E, STS-B
- Software package automatically trains and evaluates per-task linear classifiers using supplied representations.
- Limited to sentence-to-vector models.

The Standard Evaluation: SentEval

- Informal evaluation standard formalized by Conneau and Kiela (2018).
- Suite of ten tasks:
    - **MR**, **CR**, **SUBJ**, **MPQA**, **SST**, TREC, MRPC, SICK-R, SICK-E, STS-B
- Software package automatically trains and evaluates per-task linear classifiers using supplied representations.
- Limited to sentence-to-vector models.
- Heavy skew toward **sentiment-related** tasks.

Outline





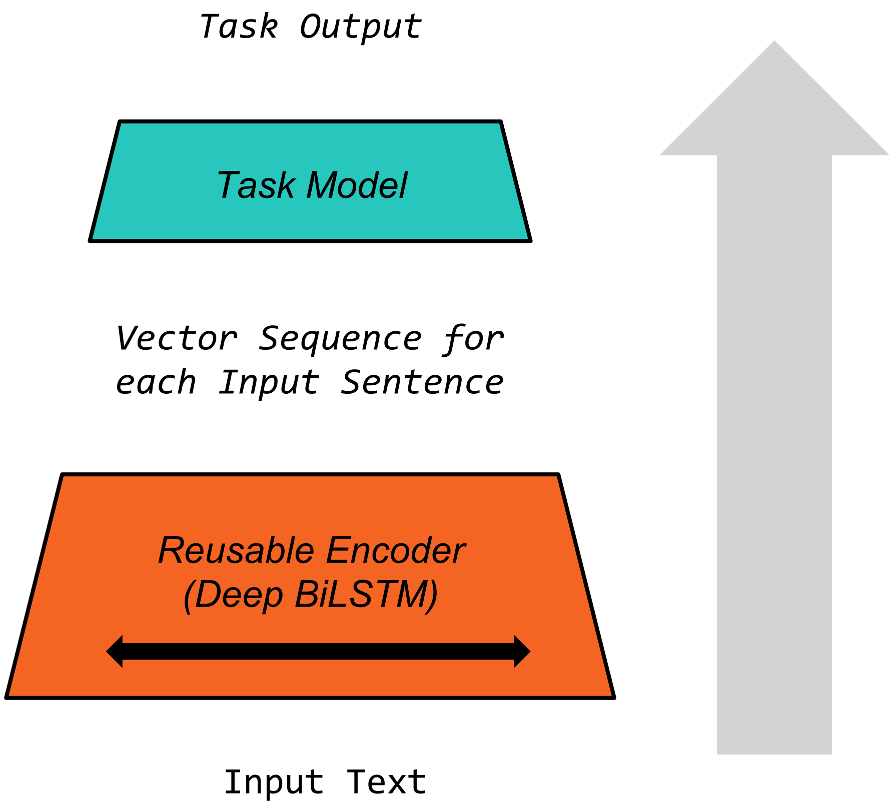
- Background: Sentence-to-vector Encoders
- **Recent progress: Newer Encoders**
- Evaluation: GLUE
- *Very* recent progress: OpenAI
- The JSALT Project

Progress to date:
Beyond $&!#* Vectors

Training objectives:

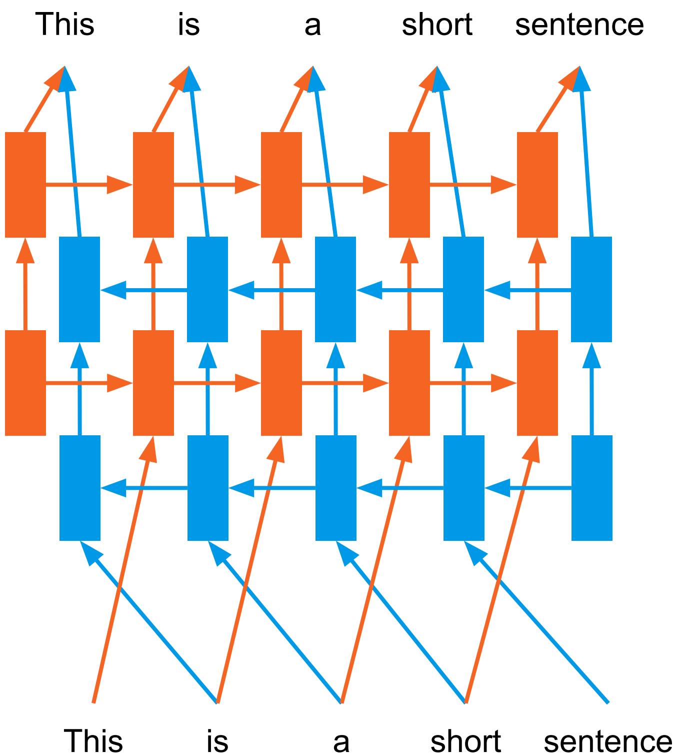
- Translation (CoVe; McCann et al., 2017)
- Language modeling (ELMo; Peters et al., 2018)

A general-purpose sentence encoder (revisited)



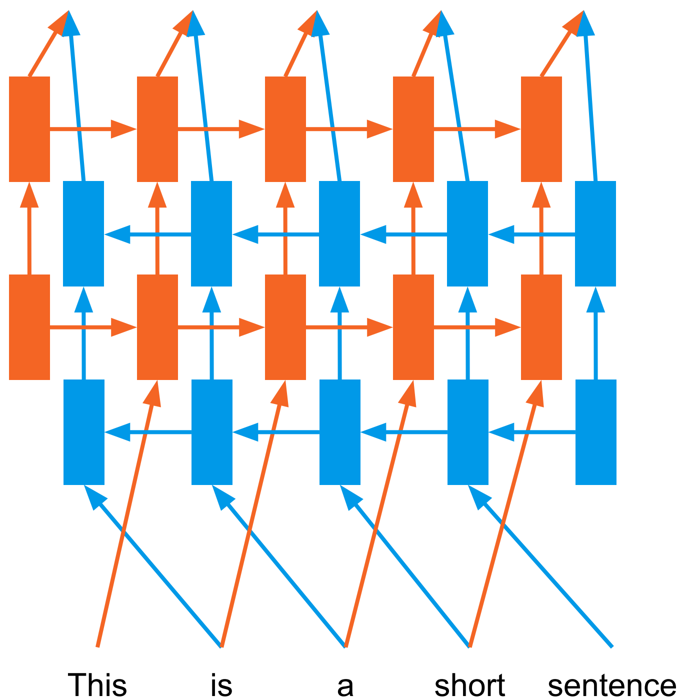
Case Study: ELMo

- Train large forward *and backward* deep LSTM language models.



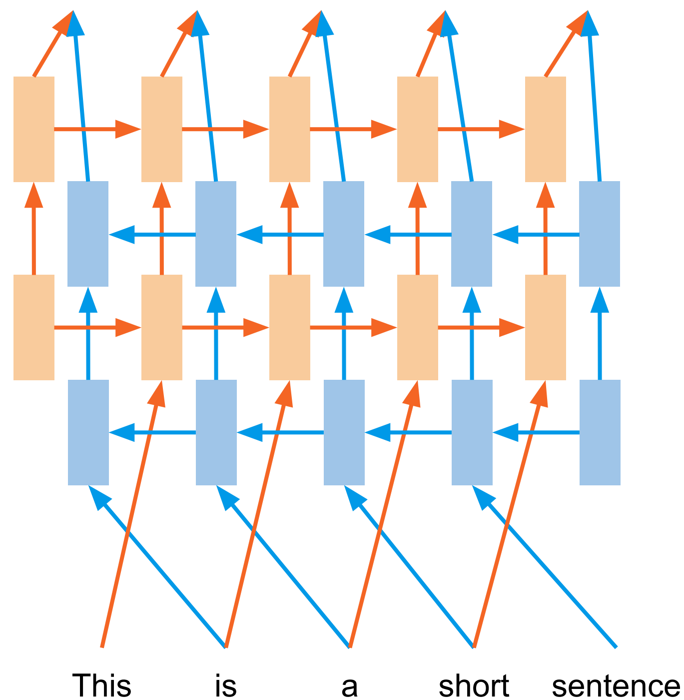
Case Study: ELMo

- At test time, use the hidden states of both language models as inputs to some task-specific model.



Case Study: ELMo

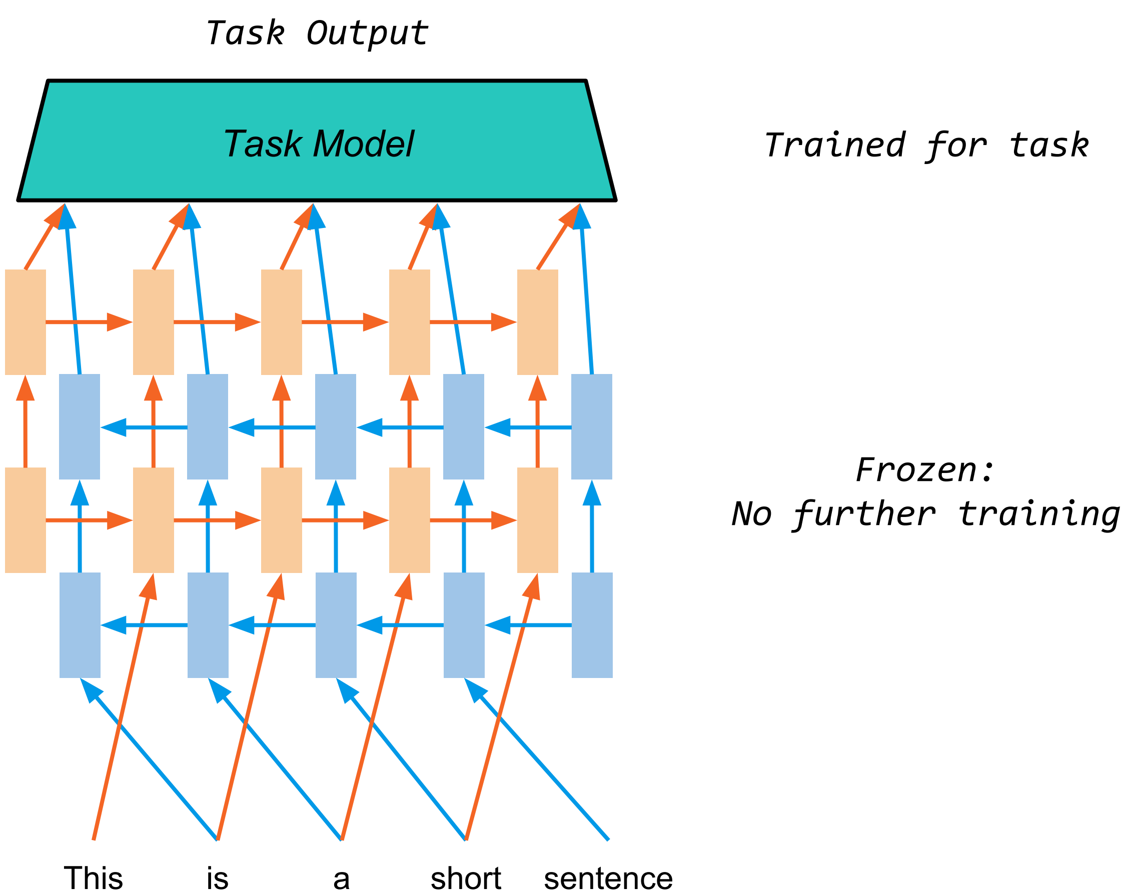
- At test time, use the hidden states of both language models as inputs to some task-specific model.



Frozen:
No further training

Case Study: ELMo

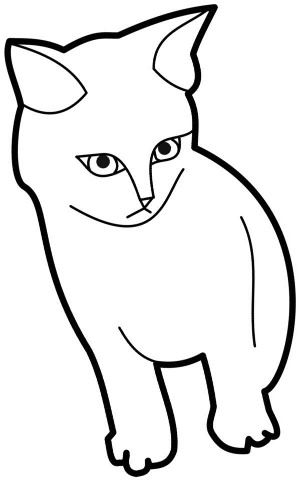
- At test time, use the hidden states of both language models as inputs to some task-specific model.



Results: ELMo

Best paper at NAACL 2018!

Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/relative)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%



Outline

- Background: Sentence-to-vector Encoders
- Recent progress: Newer Encoders
- **Evaluation: GLUE**
- *Very* recent progress: OpenAI
- The JSALT Project

Evaluation:
Beyond $&!#* Vectors

Dataset	Random	GloVe	GloVe+ Char	GloVe+ CoVe-S	GloVe+ CoVe-M	GloVe+ CoVe-L	GloVe+ Char+(
SST-2	84.2						
SST-5	48.6						
IMDb	88.4						
TREC-6	88.9						
TREC-50	81.9						
SNLI	82.3						
SQuAD	65.4						

Task	Previous SOTA		Our baseline
SQuAD	Liu et al. (2017)	84.4	81.1
SNLI	Chen et al. (2017)	88.6	88.0
SRL	He et al. (2017)	81.7	81.4
Coref	Lee et al. (2017)	67.2	67.2
NER	Peters et al. (2017)	91.93 ± 0.19	90.15
SST-5	McCann et al. (2017)	53.7	51.4

This Spring: GLUE

The General Language Understanding Evaluation (GLUE):

An open-ended competition and evaluation platform for sentence representation learning models.

Wang et al. '18

GLUE

GLUE, in short

- Nine sentence understanding tasks based on existing data, varying widely in:
  - Task difficulty
  - Training data volume and degree of training set /test set similarity
  - Language style/genre
  - (...but limited to classification/regression outputs.)
- No restriction on model type—must only be able to accept sentences and sentence pairs as inputs.
- Kaggle-style evaluation platform with private test data.
- Online leaderboard w/ single-number performance metric.
- Auxiliary analysis toolkit.
- 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

Bold = Private

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 Tasks

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

- *Binary acceptability judgments* over strings of English words.
- Extracted from articles, textbooks, and monographs in formal linguistics, with labels from original sources.
- Test examples include some topics/authors not seen at training time.

✓ *The more people you give beer to, the more people get sick.*

* *The more does Bill smoke, the more Susan hates him.*

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

The Stanford Sentiment Treebank (Socher et al. '13)

- *Binary sentiment judgments* over English sentences.
- Derived from IMDB movie reviews, with crowdsourced annotations.

\+ *It's a charming and often affecting journey.*

\- *Unflinchingly bleak and desperate.*

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

The Microsoft Research Paraphrase Corpus (Dolan & Brockett, 2005)

- *Binary paraphrase judgments* over headline pairs.
    - *Yucaipa owned Dominick's before selling the chain to Safeway in 1998 for $2.5 billion.*

      Yucaipa bought Dominick's in 1995 for $693 million and sold it to Safeway for $1.8 billion in 1998.

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

The Semantic Textual Similarity Benchmark (Cer et al., 2017)

- *Regression* over non-expert similarity judgments on sentence pairs (labels in 0–5).
- Diverse source texts.

4.750 *A young child is riding a horse.*
A child is riding a horse.

2.000 *A method used to calculate the distance between stars is 3 Dimensional trigonometry.*
You only need two-dimensional trigonometry if you know the distances to the two stars and their angular separation.

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 (Cer et al., 2017)

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

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

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 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.

neutral *The Old One always comforted Ca'daan, except today.*
Ca'daan knew the Old One very well.

Corpus	Tra					n
			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.

The Question Natural Language Inference Corpus (Rajpurkar et al., 2018/us)

- *Balanced binary classification* for pairs of sentences into *answers question* and *does not answer question.*
- Derived from SQuAD (Rajpurkar et al., 2018), with filters to ensure that lexical overlap features don't perform well.

- *What is the observable effect of W and Z boson exchange?*
  The weak force is due to the exchange of the heavy W and Z bosons.

<table>
<tr><th>Corpus</th><th>|Tr</th><th></th><th></th><th></th><th></th><th>in</th></tr>
<tr><td>CoLA</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>SST-2</td><td></td><td></td><td></td><td></td><td></td><td>movie reviews</td></tr>
<tr><td colspan="7">Similarity and Paraphrase Tasks</td></tr>
<tr><td>MRPC</td><td>3.7k</td><td>408</td><td>1.7k</td><td>paraphrase</td><td>acc./F1</td><td>news</td></tr>
<tr><td>STS-B</td><td>7k</td><td>1.5k</td><td>1.4k</td><td>sentence similarity</td><td>Pearson/Spearman corr.</td><td>misc.</td></tr>
<tr><td>QQP</td><td>364k</td><td>40k</td><td>391k</td><td>paraphrase</td><td>acc./F1</td><td>social QA questions</td></tr>
<tr><td colspan="7">Inference Tasks</td></tr>
<tr><td>MNLI</td><td>393k</td><td>20k</td><td>20k</td><td>NLI</td><td>matched acc./mismatched acc.</td><td>misc.</td></tr>
<tr><td>QNLI</td><td>108k</td><td>5.7k</td><td>5.7k</td><td>QA/NLI</td><td>acc.</td><td>Wikipedia</td></tr>
<tr><td>RTE</td><td>2.5k</td><td>276</td><td>3k</td><td>NLI</td><td>acc.</td><td>misc.</td></tr>
<tr><td>WNLI</td><td>634</td><td>71</td><td>146</td><td>coreference/NLI</td><td>acc.</td><td>fiction books</td></tr>
</table>

The Recognizing Textual Entailment Challenge Corpora (Dagan et al., 2006, etc.)

- *Binary classification* for expert-constructed pairs of sentences into *entailment* and *not entailment* on news and wiki text.
- Training and test data from four annual competitions: RTE1, RTE2, RTE3, and RTE5.

entailment *On Jan. 27, 1756, composer Wolfgang Amadeus Mozart was born in Salzburg, Austria.*
Wolfgang Amadeus Mozart was born in Salzburg.

Corpus	|Tr					in
CoLA						
SST-2						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 Winograd Schema Challenge, recast as NLI (Levesque et al., 2011/us)

- *Binary classification* for expert-constructed pairs of sentences, converted from coreference resolution to NLI.
- Manually constructed to foil superficial statistical cues.
- Using new private test set from corpus creators.

not_entailment *Jane gave Joan candy because she was hungry.*
Jane was hungry.

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

Corpus	|Tr					in
CoLA						
SST-2						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

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 Diagnostic Data

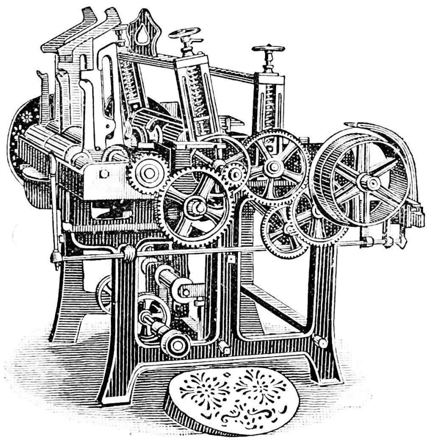
The Diagnostic Data

- Hand-constructed suite of 550 sentence pairs, each made to exemplify at least one of 33 specific phenomena.
- Seed sentences drawn from several genres.
- Each labeled with NLI labels in both directions.

The Diagnostic Data

Tags	Sentence 1	Sentence 2	Fwd	Bwd
Lexical Entailment (Lexical Semantics), Downward Monotone (Logic)	The timing of the meeting has not been set, according to a Starbucks spokesperson.	The timing of the meeting has not been considered, according to a Starbucks spokesperson.	N	E
Universal Quantifiers (Logic)	Our deepest sympathies are with all those affected by this accident.	Our deepest sympathies are with a victim who was affected by this accident.	E	N
Quantifiers (Lexical Semantics), Double Negation (Logic)	I have never seen a hummingbird not flying.	I have never seen a hummingbird.	N	E

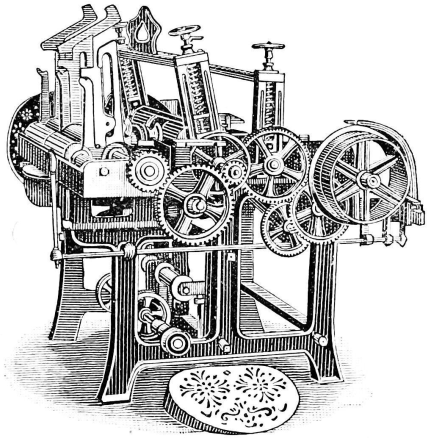
Baselines



Baseline Models

Three model types:

- Existing pretrained **sentence-to-vector encoders**
  - Used as-is, no fine-tuning.
  - Train separate downstream classifiers for each GLUE task.
- Models trained primarily on GLUE tasks
  - Trained either on each task separately (**single-task**) or on all tasks together (**multi-task**)



Model Architecture

- Our architecture:
  - Two-layer BiLSTM (1500D per direction/layer)
  - Optional attention layer for sentence pair tasks with additional shallow BiLSTM (following Seo et al., 2016)
- Input to trained BiLSTM any of:
  - GloVe (840B version, Pennington et al., 2014)
  - CoVe (McCann et al., 2017)
  - ELMo (Peters et al., 2018)
- For multi-task learning, need to balance updates from big and small tasks.
  - Sample data-poor tasks less often, but make larger gradient steps.

Results

Model	Avg	Single Sentence CoLA	SST-2	Similarity and Paraphrase MRPC	QQP	STS-B	Natural Language Inference MNLI	QNLI	RTE	WNLI
				Single-Task Training						
BiLSTM	62.0	15.7	85.9	69.3/79.4	81.7/61.4	66.0/62.8	70.3/70.8	60.8	52.8	62.3
+ELMo	66.2	**35.0**	90.2	69.0/80.8	85.7/65.6	64.0/60.2	72.9/73.4	69.4	50.1	**65.1**
+CoVe	62.4	14.5	88.5	73.4/81.4	83.3/59.4	67.2/64.1	64.5/64.8	64.8	53.5	61.6
+Attn	60.0	15.7	85.9	68.5/80.3	83.5/62.9	59.3/55.8	74.2/73.8	51.9	51.9	55.5
+Attn, ELMo	64.8	**35.0**	90.2	68.8/80.2	**86.5**/**66.1**	55.5/52.5	**76.9**/**76.7**	61.1	50.4	**65.1**
+Attn, CoVe	60.8	14.5	88.5	68.6/79.7	84.1/60.1	57.2/53.6	71.6/71.5	53.8	52.7	64.4
				Multi-Task Training						
BiLSTM	63.5	24.0	85.8	71.9/82.1	80.2/59.1	68.8/67.0	65.8/66.0	71.1	46.8	63.7
+ELMo	64.8	27.5	89.6	76.2/83.5	78.5/57.8	67.0/65.9	67.1/68.0	66.7	55.7	62.3
+CoVe	62.2	16.2	84.3	71.8/80.0	82.0/59.1	68.0/67.1	65.3/65.9	70.4	44.2	**65.1**
+Attn	65.7	0.0	85.0	75.1/**83.7**	84.3/63.6	73.9/71.8	72.2/72.1	82.1	**61.7**	63.7
+Attn, ELMo	**69.0**	18.9	**91.6**	**77.3**/83.5	85.3/63.3	72.8/71.1	75.6/75.9	81.7	61.2	**65.1**
+Attn, CoVe	64.3	19.4	83.6	75.2/83.0	84.9/61.1	72.3/71.1	69.9/68.7	78.9	38.3	**65.1**
				Pre-Trained Sentence Representation Models						
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	75.1	54.1	62.3
Skip-Thought	61.5	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	74.7	53.1	**65.1**
InferSent	64.7	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	79.8	58.0	**65.1**
DisSent	62.1	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	75.2	56.4	**65.1**
GenSen	66.6	7.7	83.1	76.6/83.0	82.9/59.8	**79.3**/**79.2**	71.4/71.3	**82.3**	59.2	**65.1**

Results

Model	Avg	Single Sentence		Similarity and Paraphrase			Natural Language Inference			
		CoLA	SST-2	MRPC	QQP	STS-B	MNLI	QNLI	RTE	WNLI
Single-Task Training										
BiLSTM	62.0	15.7	85.9	69.3/79.4	81.7/61.4	66.0/62.8	70.3/70.8	60.8	52.8	62.3
+ELMo	66.2	**35.0**	90.2	69.0/80.8	85.7/65.6	64.0/60.2	72.9/73.4	69.4	50.1	**65.1**
+CoVe	62.4	14.5	88.5	73.4/81.4	83.3/59.4	67.2/64.1	64.5/64.8	64.8	53.5	61.6
+Attn	60.0	15.7	85.9	68.5/80.3	83.5/62.9	59.3/55.8	74.2/73.8	51.9	51.9	55.5
+Attn, ELMo	64.8	**35.0**	90.2	68.8/80.2	**86.5/66.1**	55.5/52.5	**76.9/76.7**	61.1	50.4	**65.1**
+Attn, CoVe	60.8	14.5	88.5	68.6/79.7	84.1/60.1	57.2/53.6	71.6/71.5	53.8	52.7	64.4
Multi-Task Training										
BiLSTM	63.5	24.0	85.8	71.9/82.1	80.2/59.1	68.8/67.0	65.8/66.0	71.1	46.8	63.7
+ELMo	64.8	27.5	89.6	76.2/83.5	78.5/57.8	67.0/65.9	67.1/68.0	66.7	55.7	62.3
+CoVe	62.2	16.2	84.3	71.8/80.0	82.0/59.1	68.0/67.1	65.3/65.9	70.4	44.2	**65.1**
+Attn	65.7	0.0	85.0	75.1/**83.7**	84.3/63.6	73.9/71.8	72.2/72.1	82.1	**61.7**	63.7
+Attn, ELMo	**69.0**	18.9	**91.6**	**77.3**/83.5	85.3/63.3	72.8/71.1	75.6/75.9	81.7	61.2	**65.1**
+Attn, CoVe	64.3	19.4	83.6	75.2/83.0	84.9/61.1	72.3/71.1	69.9/68.7	78.9	38.3	**65.1**
Pre-Trained Sentence Representation Models										
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	75.1	54.1	62.3
Skip-Thought	61.5	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	74.7	53.1	**65.1**
InferSent	64.7	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	79.8	58.0	**65.1**
DisSent	62.1	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	75.2	56.4	**65.1**
GenSen	66.6	7.7	83.1	76.6/83.0	82.9/59.8	**79.3/79.2**	71.4/71.3	**82.3**	59.2	**65.1**

Results

		Single Sentence		Similarity and Paraphrase			Natural Language Inference			
Model	**Avg**	**CoLA**	**SST-2**	**MRPC**	**QQP**	**STS-B**	**MNLI**	**QNLI**	**RTE**	**WNLI**
Single-Task Training										
BiLSTM	62.0	15.7	85.9	69.3/79.4	81.7/61.4	66.0/62.8	70.3/70.8	60.8	52.8	62.3
+ELMo	66.2	**35.0**	90.2	69.0/80.8	85.7/65.6	64.0/60.2	72.9/73.4	69.4	50.1	**65.1**
+CoVe	62.4	14.5	88.5	73.4/81.4	83.3/59.4	67.2/64.1	64.5/64.8	64.8	53.5	61.6
+Attn	60.0	15.7	85.9	68.5/80.3	83.5/62.9	59.3/55.8	74.2/73.8	51.9	51.9	55.5
+Attn, ELMo	64.8	**35.0**	90.2	68.8/80.2	**86.5**/**66.1**	55.5/52.5	**76.9**/**76.7**	61.1	50.4	**65.1**
+Attn, CoVe	60.8	14.5	88.5	68.6/79.7	84.1/60.1	57.2/53.6	71.6/71.5	53.8	52.7	64.4
Multi-Task Training										
BiLSTM	63.5	24.0	85.8	71.9/82.1	80.2/59.1	68.8/67.0	65.8/66.0	71.1	46.8	63.7
+ELMo	64.8	27.5	89.6	76.2/83.5	78.5/57.8	67.0/65.9	67.1/68.0	66.7	55.7	62.3
+CoVe	62.2	16.2	84.3	71.8/80.0	82.0/59.1	68.0/67.1	65.3/65.9	70.4	44.2	**65.1**
+Attn	65.7	0.0	85.0	75.1/**83.7**	84.3/63.6	73.9/71.8	72.2/72.1	82.1	**61.7**	63.7
+Attn, ELMo	**69.0**	18.9	**91.6**	**77.3**/83.5	85.3/63.3	72.8/71.1	75.6/75.9	81.7	61.2	**65.1**
+Attn, CoVe	64.3	19.4	83.6	75.2/83.0	84.9/61.1	72.3/71.1	69.9/68.7	78.9	38.3	**65.1**
Pre-Trained Sentence Representation Models										
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	75.1	54.1	62.3
Skip-Thought	61.5	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	74.7	53.1	**65.1**
InferSent	64.7	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	79.8	58.0	**65.1**
DisSent	62.1	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	75.2	56.4	**65.1**
GenSen	66.6	7.7	83.1	76.6/83.0	82.9/59.8	**79.3**/**79.2**	71.4/71.3	**82.3**	59.2	**65.1**

Results

Model	Avg	Single Sentence CoLA	SST-2	Similarity and Paraphrase MRPC	QQP	STS-B	Natural Language Inference MNLI	QNLI	RTE	WNLI
				Single-Task Training						
BiLSTM	62.0	15.7	85.9	69.3/79.4	81.7/61.4	66.0/62.8	70.3/70.8	60.8	52.8	62.3
+ELMo	66.2	**35.0**	90.2	69.0/80.8	85.7/65.6	64.0/60.2	72.9/73.4	69.4	50.1	**65.1**
+CoVe	62.4	14.5	88.5	73.4/81.4	83.3/59.4	67.2/64.1	64.5/64.8	64.8	53.5	61.6
+Attn	60.0	15.7	85.9	68.5/80.3	83.5/62.9	59.3/55.8	74.2/73.8	51.9	51.9	55.5
+Attn, ELMo	64.8	**35.0**	90.2	68.8/80.2	**86.5**/**66.1**	55.5/52.5	**76.9**/**76.7**	61.1	50.4	**65.1**
+Attn, CoVe	60.8	14.5	88.5	68.6/79.7	84.1/60.1	57.2/53.6	71.6/71.5	53.8	52.7	64.4
				Multi-Task Training						
BiLSTM	63.5	24.0	85.8	71.9/82.1	80.2/59.1	68.8/67.0	65.8/66.0	71.1	46.8	63.7
+ELMo	64.8	27.5	89.6	76.2/83.5	78.5/57.8	67.0/65.9	67.1/68.0	66.7	55.7	62.3
+CoVe	62.2	16.2	84.3	71.8/80.0	82.0/59.1	68.0/67.1	65.3/65.9	70.4	44.2	**65.1**
+Attn	65.7	0.0	85.0	75.1/**83.7**	84.3/63.6	73.9/71.8	72.2/72.1	82.1	**61.7**	63.7
+Attn, ELMo	**69.0**	18.9	**91.6**	**77.3**/83.5	85.3/63.3	72.8/71.1	75.6/75.9	81.7	61.2	**65.1**
+Attn, CoVe	64.3	19.4	83.6	75.2/83.0	84.9/61.1	72.3/71.1	69.9/68.7	78.9	38.3	**65.1**
				Pre-Trained Sentence Representation Models						
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	75.1	54.1	62.3
Skip-Thought	61.5	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	74.7	53.1	**65.1**
InferSent	64.7	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	79.8	58.0	**65.1**
DisSent	62.1	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	75.2	56.4	**65.1**
GenSen	66.6	7.7	83.1	76.6/83.0	82.9/59.8	**79.3**/**79.2**	71.4/71.3	**82.3**	59.2	**65.1**

Results

Model	Avg	Single Sentence CoLA	SST-2	Similarity and Paraphrase MRPC	QQP	STS-B	Natural Language Inference MNLI	QNLI	RTE	WNLI
Single-Task Training										
BiLSTM	62.0	15.7	85.9	69.3/79.4	81.7/61.4	66.0/62.8	70.3/70.8	60.8	52.8	62.3
+ELMo	66.2	**35.0**	90.2	69.0/80.8	85.7/65.6	64.0/60.2	72.9/73.4	69.4	50.1	**65.1**
+CoVe	62.4	14.5	88.5	73.4/81.4	83.3/59.4	67.2/64.1	64.5/64.8	64.8	53.5	61.6
+Attn	60.0	15.7	85.9	68.5/80.3	83.5/62.9	59.3/55.8	74.2/73.8	51.9	51.9	55.5
+Attn, ELMo	64.8	**35.0**	90.2	68.8/80.2	**86.5**/**66.1**	55.5/52.5	**76.9**/**76.7**	61.1	50.4	**65.1**
+Attn, CoVe	60.8	14.5	88.5	68.6/79.7	84.1/60.1	57.2/53.6	71.6/71.5	53.8	52.7	64.4
Multi-Task Training										
BiLSTM	63.5	24.0	85.8	71.9/82.1	80.2/59.1	68.8/67.0	65.8/66.0	71.1	46.8	63.7
+ELMo	64.8	27.5	89.6	76.2/83.5	78.5/57.8	67.0/65.9	67.1/68.0	66.7	55.7	62.3
+CoVe	62.2	16.2	84.3	71.8/80.0	82.0/59.1	68.0/67.1	65.3/65.9	70.4	44.2	**65.1**
+Attn	65.7	0.0	85.0	75.1/**83.7**	84.3/63.6	73.9/71.8	72.2/72.1	82.1	**61.7**	63.7
+Attn, ELMo	**69.0**	18.9	**91.6**	**77.3**/83.5	85.3/63.3	72.8/71.1	75.6/75.9	81.7	61.2	**65.1**
+Attn, CoVe	64.3	19.4	83.6	75.2/83.0	84.9/61.1	72.3/71.1	69.9/68.7	78.9	38.3	**65.1**
Pre-Trained Sentence Representation Models										
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	75.1	54.1	62.3
Skip-Thought	61.5	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	74.7	53.1	**65.1**
InferSent	64.7	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	79.8	58.0	**65.1**
DisSent	62.1	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	75.2	56.4	**65.1**
GenSen	66.6	7.7	83.1	76.6/83.0	82.9/59.8	**79.3**/**79.2**	71.4/71.3	**82.3**	59.2	**65.1**

Results on Diagnostic Data (MNLI classifier)

		Coarse-Grained				Fine-Grained					
Model	**All**	**LS**	**PAS**	**L**	**K**	**UQuant**	**MNeg**	**2Neg**	**Coref**	**Restr**	**Down**
Single-Task Training											
BiLSTM	21	25	24	16	16	70	53	4	21	-15	**12**
+ELMo	20	20	21	14	17	70	20	**42**	33	-26	-3
+CoVe	21	19	23	20	18	71	47	-1	33	-15	8
+Attn	25	24	30	20	14	50	47	21	**38**	-8	-3
+Attn, ELMo	**28**	**30**	**35**	**23**	14	**85**	20	**42**	33	-26	-3
+Attn, CoVe	24	29	29	18	12	77	50	1	18	-1	**12**
Multi-Task Training											
BiLSTM	19	16	22	16	17	71	35	-8	26	**0**	8
+ELMo	19	15	21	17	**21**	70	**60**	15	26	**0**	**12**
+CoVe	17	15	21	14	16	50	31	-8	25	-15	**12**
+Attn	25	23	32	19	16	58	26	-5	28	-1	-20
+Attn, ELMo	23	24	30	17	13	78	27	37	30	-15	-20
+Attn, CoVe	20	16	25	15	17	78	37	14	31	-15	8
Pre-Trained Sentence Representation Models											
CBoW	9	6	13	5	10	3	0	13	28	-15	-11
Skip-Thought	12	2	23	11	9	61	6	-2	30	-15	0
InferSent	18	20	20	15	14	77	50	-20	15	-15	-9
DisSent	16	16	19	13	15	70	43	-11	20	-36	-09
GenSen	20	28	26	14	12	78	57	2	21	-15	**12**

Results on Diagnostic Data (MNLI classifier)

Model	All	Coarse-Grained				Fine-Grained					
		LS	PAS	L	K	UQuant	MNeg	2Neg	Coref	Restr	Down
					Single-Task Training						
BiLSTM	21	25	24	16	16	70	53	4	21	-15	**12**
+ELMo	20	20	21	14	17	70	20	**42**	33	-26	-3
+CoVe	21	19	23	20	18	71	47	-1	33	-15	8
+Attn	25	24	30	20	14	50	47	21	**38**	-8	-3
+Attn, ELMo	**28**	**30**	**35**	**23**	14	**85**	20	**42**	33	-26	-3
+Attn, CoVe	24	29	29	18	12	77	50	1	18	-1	**12**
					Multi-Task Training						
BiLSTM	19	16	22	16	17	71	35	-8	26	**0**	8
+ELMo	19	15	21	17	**21**	70	**60**	15	26	**0**	**12**
+CoVe	17	15	21	14	16	50	31	-8	25	-15	**12**
+Attn	25	23	32	19	16	58	26	-5	28	-1	-20
+Attn, ELMo	23	24	30	17	13	78	27	37	30	-15	-20
+Attn, CoVe	20	16	25	15	17	78	37	14	31	-15	8
					Pre-Trained Sentence Representation Models						
CBoW	9	6	13	5	10	3	0	13	28	-15	-11
Skip-Thought	12	2	23	11	9	61	6	-2	30	-15	0
InferSent	18	20	20	15	14	77	50	-20	15	-15	-9
DisSent	16	16	19	13	15	70	43	-11	20	-36	-09
GenSen	20	28	26	14	12	78	57	2	21	-15	**12**

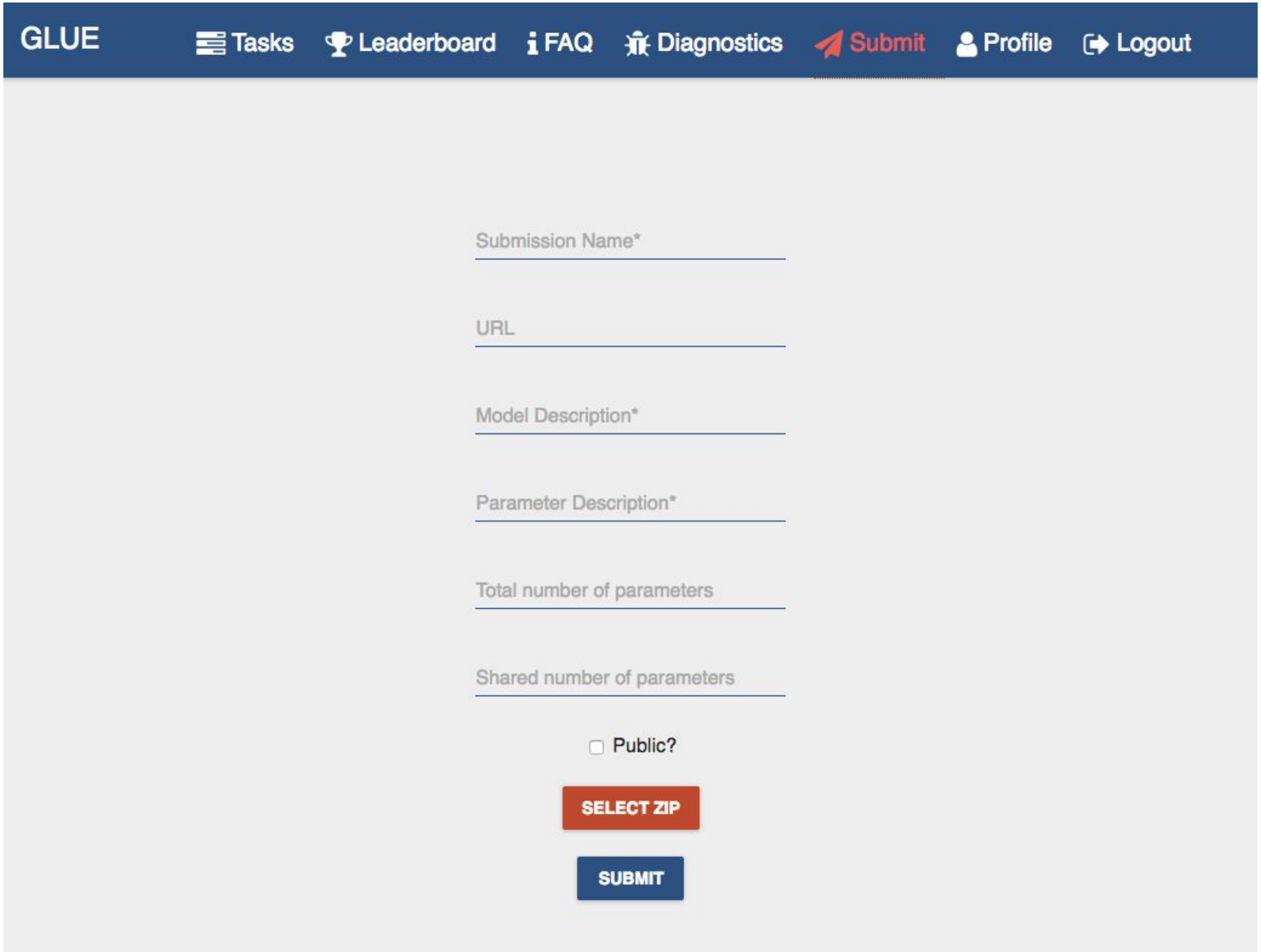
Results on Diagnostic Data (MNLI classifier)

Model	Coarse-Grained					Fine-Grained					
	All	LS	PAS	L	K	UQuant	MNeg	2Neg	Coref	Restr	Down
Single-Task Training											
BiLSTM	21	25	24	16	16	70	<u>53</u>	4	21	-15	**<u>12</u>**
+ELMo	20	20	21	14	17	70	20	**<u>42</u>**	33	-26	-3
+CoVe	21	19	23	20	<u>18</u>	71	47	-1	33	-15	8
+Attn	25	24	30	20	14	50	47	21	**<u>38</u>**	-8	-3
+Attn, ELMo	**<u>28</u>**	**<u>30</u>**	**<u>35</u>**	**<u>23</u>**	14	**<u>85</u>**	20	**<u>42</u>**	33	-26	-3
+Attn, CoVe	24	29	29	18	12	77	50	1	18	<u>-1</u>	**<u>12</u>**
Multi-Task Training											
BiLSTM	19	16	22	16	17	71	35	-8	26	**<u>0</u>**	8
+ELMo	19	15	21	17	**<u>21</u>**	70	**<u>60</u>**	15	26	**<u>0</u>**	**<u>12</u>**
+CoVe	17	15	21	14	16	50	31	-8	25	-15	**<u>12</u>**
+Attn	<u>25</u>	23	<u>32</u>	<u>19</u>	16	58	26	-5	28	-1	-20
+Attn, ELMo	23	<u>24</u>	30	17	13	<u>78</u>	27	<u>37</u>	30	-15	-20
+Attn, CoVe	20	16	25	15	17	<u>78</u>	37	14	<u>31</u>	-15	8
Pre-Trained Sentence Representation Models											
CBoW	9	6	13	5	10	3	0	<u>13</u>	28	<u>-15</u>	-11
Skip-Thought	12	2	23	11	9	61	6	-2	<u>30</u>	<u>-15</u>	0
InferSent	18	20	20	<u>15</u>	14	77	50	-20	15	<u>-15</u>	-9
DisSent	16	16	19	13	<u>15</u>	70	43	-11	20	-36	-09
GenSen	<u>20</u>	<u>28</u>	<u>26</u>	14	12	<u>78</u>	<u>57</u>	2	21	<u>-15</u>	**<u>12</u>**

Limitations

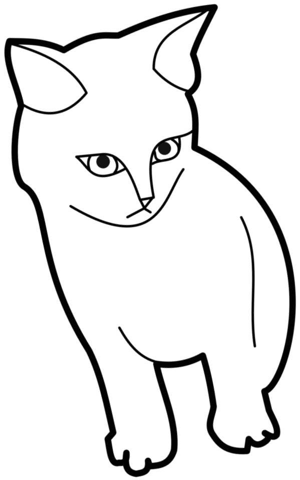
- GLUE is built only on English data.
  - Sentence representation learning may look quite different in lower-resource languages!
- GLUE does not evaluate text *generation*, and uses only small amounts of context.
  - Isolates the problem of extracting sentence meaning, but avoids other hard parts of NLP.
- GLUE uses naturally occurring and crowdsourced data.
  - Models trained on the GLUE training set generally acquire biases and world knowledge that we may not want them to.
  - Models that reflect these biases may do better on GLUE.

http://gluebenchmark.com

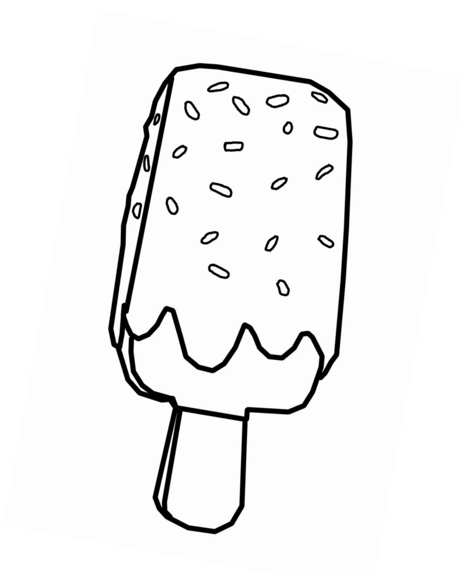


http://gluebenchmark.com

PRIMARY						AUXILIARY								
Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI
1	GLUE Baselines	BiLSTM+ELMo+Attn		68.9	18.9	91.6	77.3/83.5	72.8/71.1	83.5/63.3	75.6	75.9	81.7	61.2	65.1
		GenSen		66.6	7.7	83.1	76.6/83.0	79.3/79.2	82.9/59.8	71.4	71.3	82.3	59.2	65.1
		Single Task BiLSTM+ELMo		66.2	35.0	90.2	69.0/80.8	64.0/60.2	85.7/65.6	72.9	73.4	69.4	50.1	65.1
		BiLSTM+Attn		65.7	0.0	85.0	75.1/83.7	73.9/71.8	84.3/63.6	72.2	72.1	82.1	61.7	63.7
		BiLSTM+ELMo		64.9	27.5	89.6	76.2/83.5	67.0/65.9	78.5/57.8	67.1	68.0	66.7	55.7	62.3
		Single Task BiLSTM+ELMo+Att		64.8	35.0	90.2	68.8/80.2	55.5/52.5	86.5/66.1	76.9	76.7	61.1	50.3	65.1
		InferSent		64.7	4.5	85.1	74.1/81.2	75.9/75.3	81.7/59.1	66.1	65.7	79.8	58.0	65.1
		BiLSTM+CoVe+Attn		64.3	19.4	83.6	75.2/83.0	72.3/71.1	84.9/61.1	69.9	68.7	78.9	38.3	65.1
		BiLSTM		63.5	24.0	85.8	71.9/82.1	68.8/67.0	80.2/59.1	65.8	66.0	71.1	46.8	63.7
		Single Task BiLSTM+CoVe		62.4	14.5	88.5	73.4/81.4	67.2/64.1	83.3/59.4	64.5	64.8	64.8	53.5	61.6
		BiLSTM+CoVe		62.2	16.2	84.3	71.8/80.0	68.0/67.1	82.0/59.1	65.3	65.9	70.4	44.2	65.1

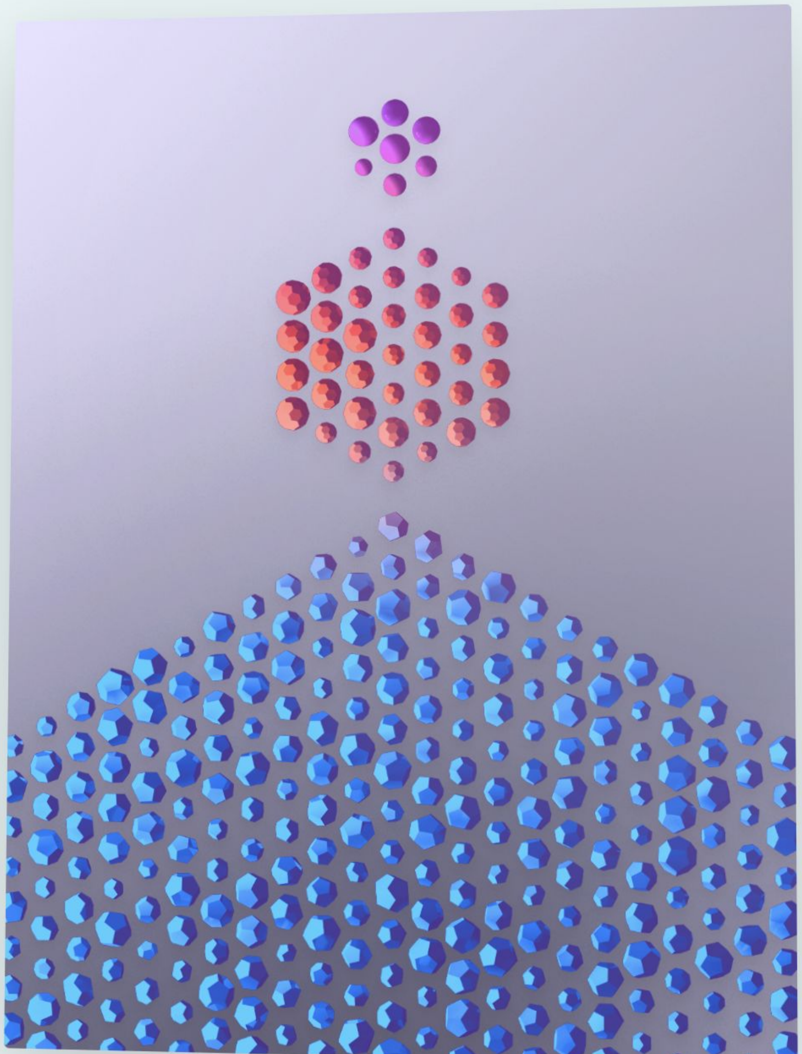


Outline



- Background: Sentence-to-vector Encoders
- Recent progress: Newer Encoders
- Evaluation: GLUE
- ***Very* recent progress: OpenAI**
- The JSALT Project

First Submission: OpenAI

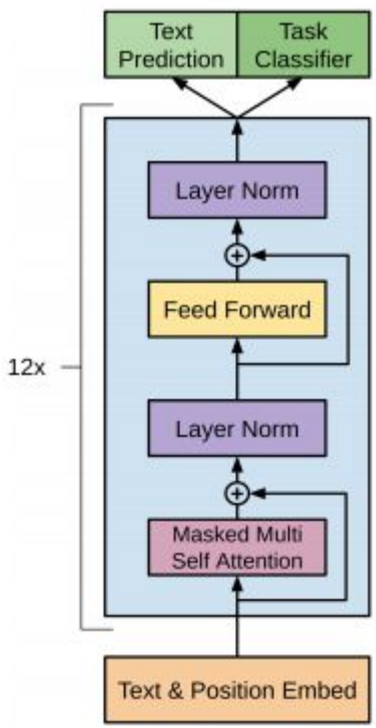


JUNE 11, 2018

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

The OpenAI Model



- Same basic idea as ELMo, but many small differences (and many open quesions!)
- Trained as a language model.
    - ... but not bidirectional.
- *Transformer* encoder architecture:
    - No RNNs, just many layers of self-attention.
- Trained on running text, not sentences in isolation.
- Trained on fiction, not news.
- Slightly larger (90 => 116m parameters)
- Unlike ELMo, entire network is fine-tuned for each task.
    - Generally helpful, may be harmful for very data-poor tasks.

Results

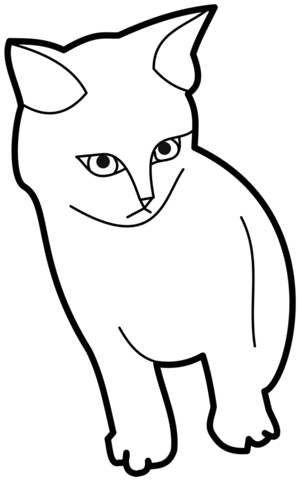
PRIMARY									AUXILIARY					
Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI
1	Alec Radford	Singletask Pretrain Transformer		72.8	45.4	91.3	75.7/82.3	82.0/80.0	88.5/70.3	82.1	81.4	88.1	56.0	53.4
2	GLUE Baselines	BiLSTM+ELMo+Attn		68.9	18.9	91.6	77.3/83.5	72.8/71.1	83.5/63.3	75.6	75.9	81.7	61.2	65.1

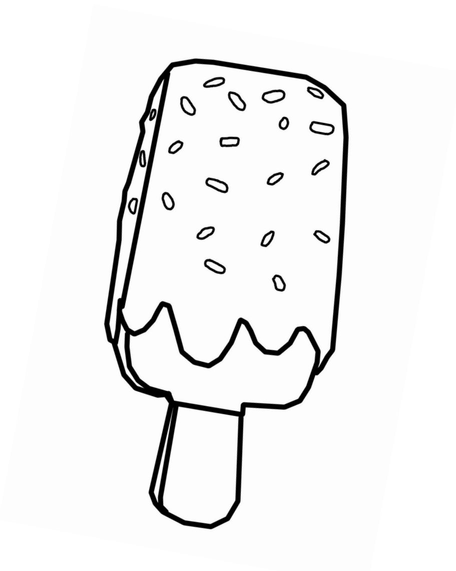
- 4% GLUE score improvement.
- Big improvements on 6 of 9 tasks.
- On analysis data, improvements concentrated in logical reasoning and predicate-argument structure.
    - Less change in world knowledge and lexical semantics.

GLUE: Conclusions

- Sentence representation learning is a hard open problem.
- GLUE offers some tools to evaluate sentence representation learning models:
  - Broad sample of training set sizes, genres, task formats, and degrees of difficulty.
  - Private test sets ensure fairness.
  - Minimal constraints on model design.
  - Automatic linguistic analysis.
- Multi-task learning models with ELMo outperform simple single-task baselines, but don't do well in absolute terms.

Outline





- Background: Sentence-to-vector Encoders
- Recent progress: Newer Encoders
- Evaluation: GLUE
- *Very* recent progress: OpenAI
- **The JSALT Project**

The JSALT Project

General goal:
Understand what it'll take to build sentence representations for human-level NLU, focusing on GLUE.

- What does language model training teach you about language?
- What should we expect to learn by simply scaling up, and what requires new methods?
- Are there training objectives that can teach you what language modeling doesn't?
- What kinds of knowledge are most important for task performance?

Thanks!