Deep Contextualized Word representation

ELMo, (peters2018deep)

Hassan Alhuzali

University of British Columbia

NLP@CS reading group, 2018

Outline

Outline

Background

What have been done so far!

- Pretrained word representations (Mikolov et al., 2013; Pennington et al., 2014) are a key component in many neural language processing models.

Background

What have been done so far!

- Pretrained word representations (Mikolov et al., 2013; Pennington et al., 2014) are a key component in many neural language processing models.
- However, word embeddings compress all contexts into a single vector.

Background

What have been done so far!

- Pretrained word representations (Mikolov et al., 2013; Pennington et al., 2014) are a key component in many neural language processing models.
- However, word embeddings compress all contexts into a single vector.
- For example:
  - The commercial bank" vs "the river bank"

Background

What have been done so far!

- Pretrained word representations (Mikolov et al., 2013; Pennington et al., 2014) are a key component in many neural language processing models.
- However, word embeddings compress all contexts into a single vector.
- For example:
  - The commercial bank" vs "the river bank"
  - is ..... in (interested or interesting)
  - is ..... to (interested or interesting)

Outline

Elmo

Contributions

- Introduce a new type of deep contextualized word representation that models:

Elmo

Contributions

- Introduce a new type of deep contextualized word representation that models:
  - complex characteristics of word use (e.g., syntax and semantics).
  - how these uses vary across linguistic contexts (i.e., to model word senses).

Elmo

Contributions

- Introduce a new type of deep contextualized word representation that models:
  - complex characteristics of word use (e.g., syntax and semantics).
  - how these uses vary across linguistic contexts (i.e., to model word senses).
- Learn a representation for each word given context (Contextualized repres).

Outline

BiLM

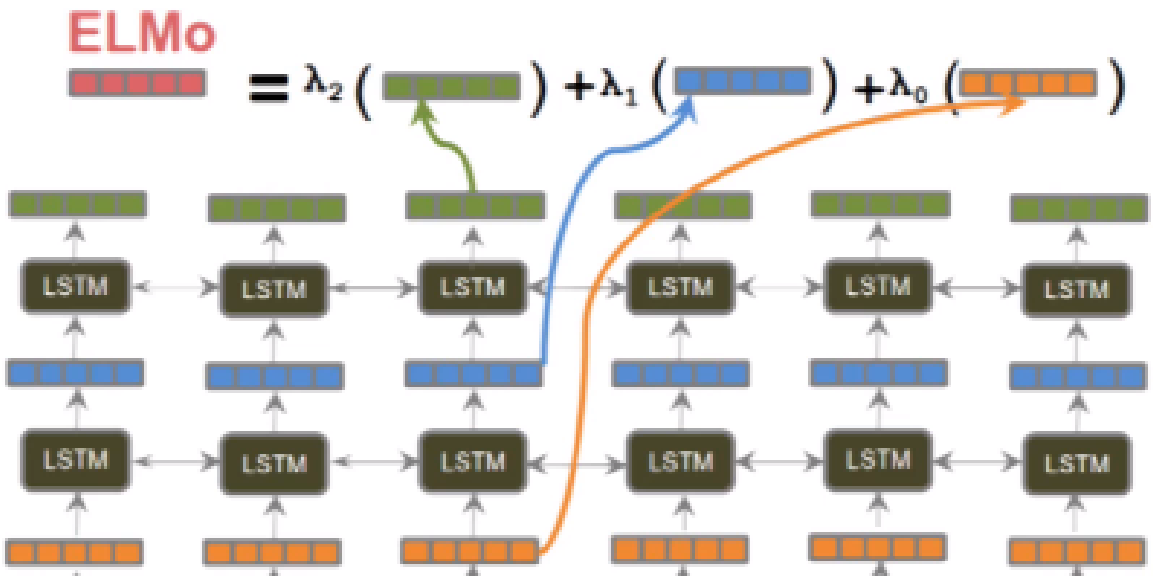
- For each token tk, biLM computes a set of 2L+1 representations.

$$R_k = \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\}$$

- ELMo collapses all layers in R into a single vector.

Outline

(ELMo) Representation



Outline

(BiLM) Arch

Elmo Architecture:

- uses 2-layer BiLSTM with 4096 units.

(BiLM) Arch

Elmo Architecture:

- uses 2-layer BiLSTM with 4096 units.
- also incorporated a size of 2048 character-level CNN.

(BiLM) Arch

Elmo Architecture:

- uses 2-layer BiLSTM with 4096 units.
- also incorporated a size of 2048 character-level CNN.
- pretrained 10 epochs over a 1 billion word corpus.

Outline

ELMO for downstream tasks

- Pass the ELMo representation into multiple layers of the supervised model.
  - Input layer
  - Output Layer
  - Both

ELMO for downstream tasks

- Pass the ELMo representation into multiple layers of the supervised model.
  - Input layer
  - Output Layer
  - Both
- Interestingly, Elmo is very simple model that can be combined with any type of supervised model.

Outline

Intrinsic (POS vs WSD)

Model	Acc.
Collobert et al. (2011)	97.3
Ma and Hovy (2016)	97.6
Ling et al. (2015)	**97.8**
CoVe, First Layer	*93.3*
CoVe, Second Layer	92.8
biLM, First Layer	*97.3*
biLM, Second Layer	96.8

Intrinsic (POS vs WSD)

Model	Acc.
Collobert et al. (2011)	97.3
Ma and Hovy (2016)	97.6
Ling et al. (2015)	**97.8**
CoVe, First Layer	*93.3*
CoVe, Second Layer	92.8
biLM, First Layer	*97.3*
biLM, Second Layer	96.8

Model	F_1
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	**70.1**
CoVe, First Layer	59.4
CoVe, Second Layer	*64.7*
biLM, First layer	67.4
biLM, Second layer	*69.0*

Outline

Textual entailment & Sentiment Analysis

TA:

Def

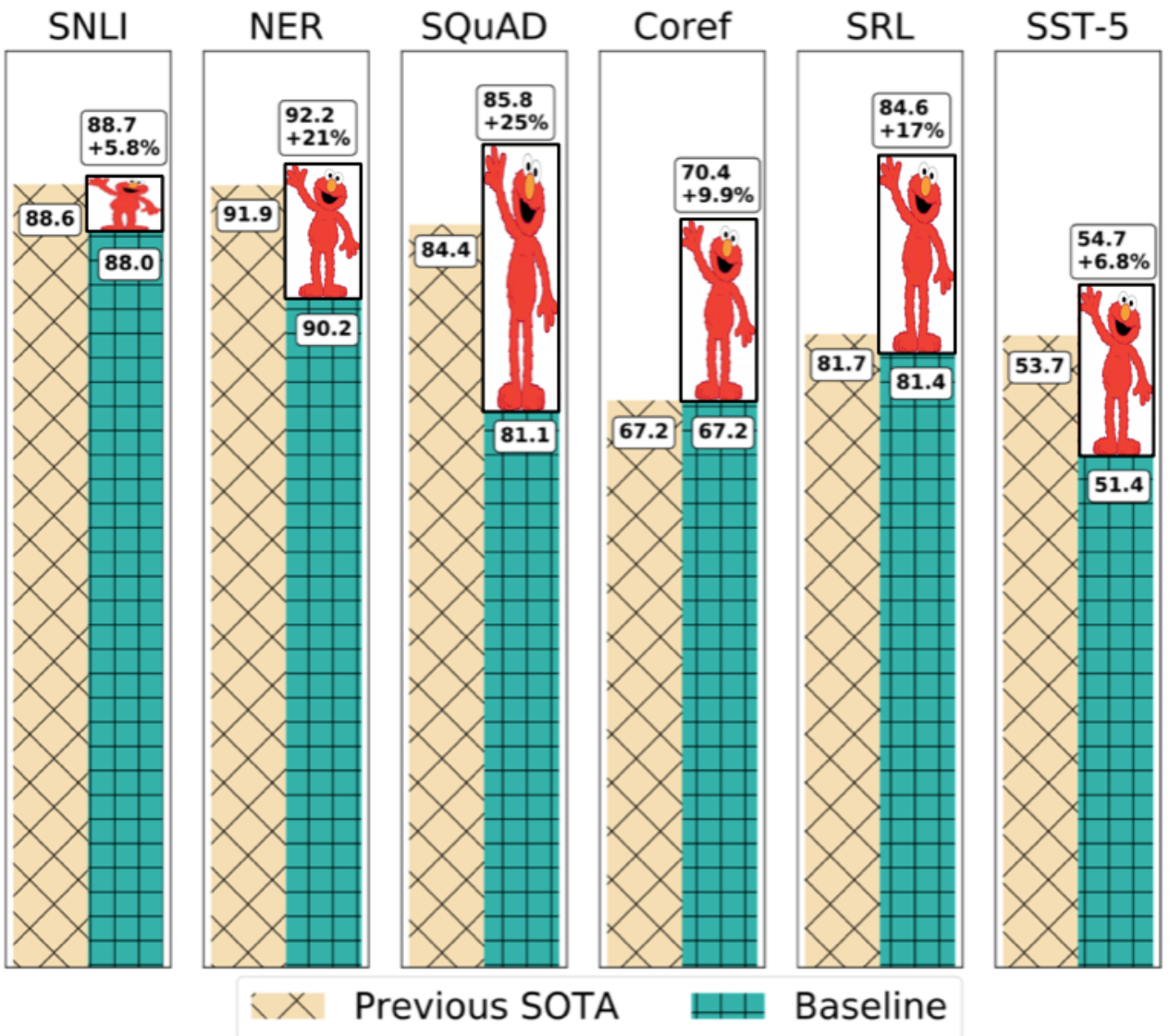
Textual entailment is the task of determining whether a hypothesis is true, given a premise.

- A known corpus for this task is "The Stanford Natural Language Inference (SNLI)" (Bowman et al., 2015), which provides approximately 550K hypothesis/premise pairs.

SA:

- Many corpora are available for SA task, but in this work, they use (SST-5: Socher et al., 2013)

Results


SNLI
NER
SQuAD
Coref
SRL
SST-5
88.7
+5.8%
88.6
88.0
92.2
+21%
91.9
90.2
85.8
+25%
84.4
81.1
70.4
+9.9%
67.2
67.2
84.6
+17%
81.7
81.4
54.7
+6.8%
53.7
51.4
Previous SOTA
Baseline

Outline

Alternate layer weighting schemes

- To assess whether it's enough to include the representation of the last layer or the representation of all layers.

Task	Baseline	Last Only	All layers	
			λ=1	λ=0.001
SQuAD	80.8	84.7	85.0	**85.2**
SNLI	88.1	89.1	89.3	**89.5**
SRL	81.6	84.1	84.6	**84.8**

Outline

Where to include ELMo?

- Previous work tends to add learned representation at the input layer, such as pre-trained embedding.
- However, in this work, it's shown that it's useful to add learned representation at both the input-level and output-level.

Task	Input Only	Input & Output	Output Only
SQuAD	85.1	**85.6**	84.8
SNLI	88.9	**89.5**	88.7
SRL	**84.7**	84.3	80.9

Outline

Sample efficiency

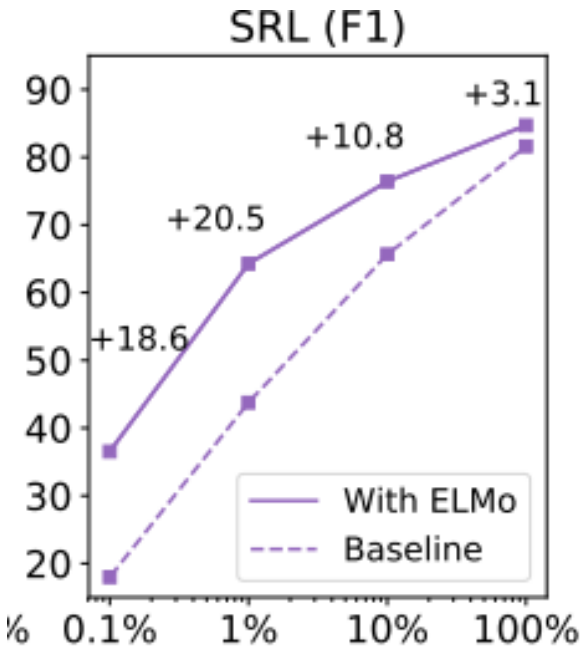
- # epochs:
  - The SRL model reaches best F1 after 400 epochs of training.
  - ELMo, the model exceeds the baseline maximum at epoch 10.

Sample efficiency

- # epochs:
  - The SRL model reaches best F1 after 400 epochs of training.
  - ELMo, the model exceeds the baseline maximum at epoch 10.
- Sample of training data:
  - ELMO obtaines the same score as SRL model using only 1% of the training data compared to 10% for SRL model.

Sample efficiency

- # epochs:
  - The SRL model reaches best F1 after 400 epochs of training.
  - ELMo, the model exceeds the baseline maximum at epoch 10.
- Sample of training data:
  - ELMO obtaines the same score as SRL model using only 1% of the training data compared to 10% for SRL model.



Summary

- Introduced ELMo paper that:

Summary

- Introduced ELMo paper that:
  - learns deep context-dependent representations from biLMs.

Summary

- Introduced ELMo paper that:
  - learns deep context-dependent representations from biLMs.
  - LM objective forces network to learn how syntax and semantic vary across contexts.

Summary

- Introduced ELMo paper that:
  - learns deep context-dependent representations from biLMs.
  - LM objective forces network to learn how syntax and semantic vary across contexts.
  - ELMo model also shows large improvements when it's applied to a broad range of NLP tasks.

Summary

- Introduced ELMo paper that:
  - learns deep context-dependent representations from biLMs.
  - LM objective forces network to learn how syntax and semantic vary across contexts.
  - ELMo model also shows large improvements when it's applied to a broad range of NLP tasks.
- For more Info about implementation: visit below website `https://allennlp.org/elmo`