Deep Contextualized Word Representations

Matthew E. Peters, Mark Neumann, Mohit Iyyer,
Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer

Presented by: Tanya Goyal

Overview

- Objective: Provide “contextual” embeddings for words.
- Bi-directional LSTM with language modelling objective.
- Show experiments on down-stream tasks.

Model

1) Language Modelling Objective:

Forward Language Model: $$p(t_1, t_2, .. t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k-1} \dots t_1)$$

Backward Language Model: $$p(t_1, t_2, \dots t_N) = \prod_{k=1}^{N} p(t_k \mid t + k + 1 ... t_N)$$

2) biLM model jointly maximises the log likelihood of the forward and backward directions.

Model

- Character Level model to deal with OOV words.
  (Denoted by $h_{k,0}^{LM}$)

- Stacked LSTM layers with softmax layer used to predict next token.
  At each position k, each LSTM layer outputs $\overrightarrow{h}_{k,j}^{LM}$ where j = 1,2..L
  The top layer LSTM output, $\overrightarrow{h}_{k,L}^{LM}$ is used to predict the next token t_{k+1}

- Forward and backward LM share the token representation and softmax parameters.

Model

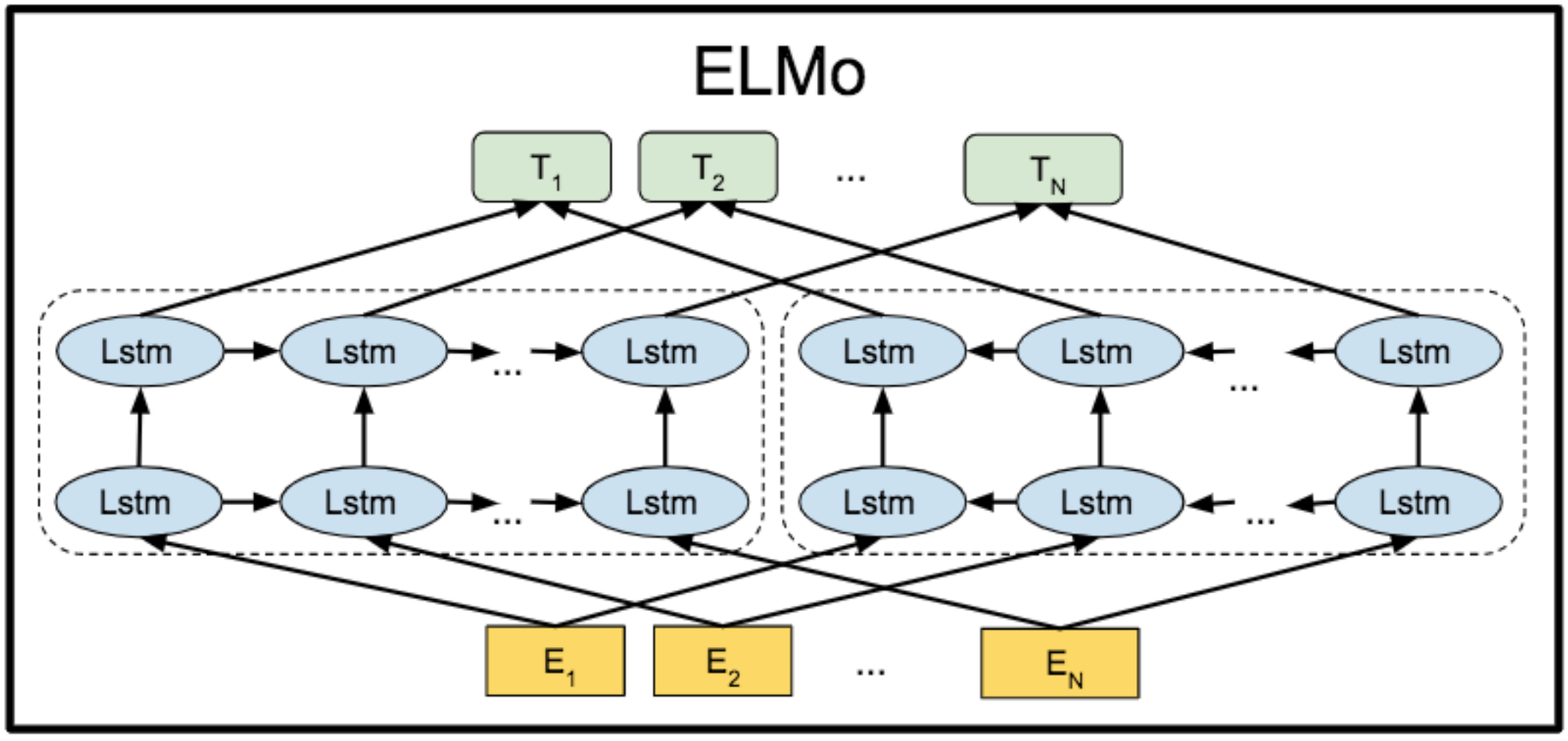


Image taken from the Bert paper

Inclusion in downstream tasks

For every token t_k, a L-layer biLM computes 2L + 1 representations.

1) $h_{k,0}^{LM}$: context-independent from the token layer
2) $\overrightarrow{h}_{k,j}^{LM}$, $\overleftarrow{h}_{k,j}^{LM}$: context-dependant from each layer.

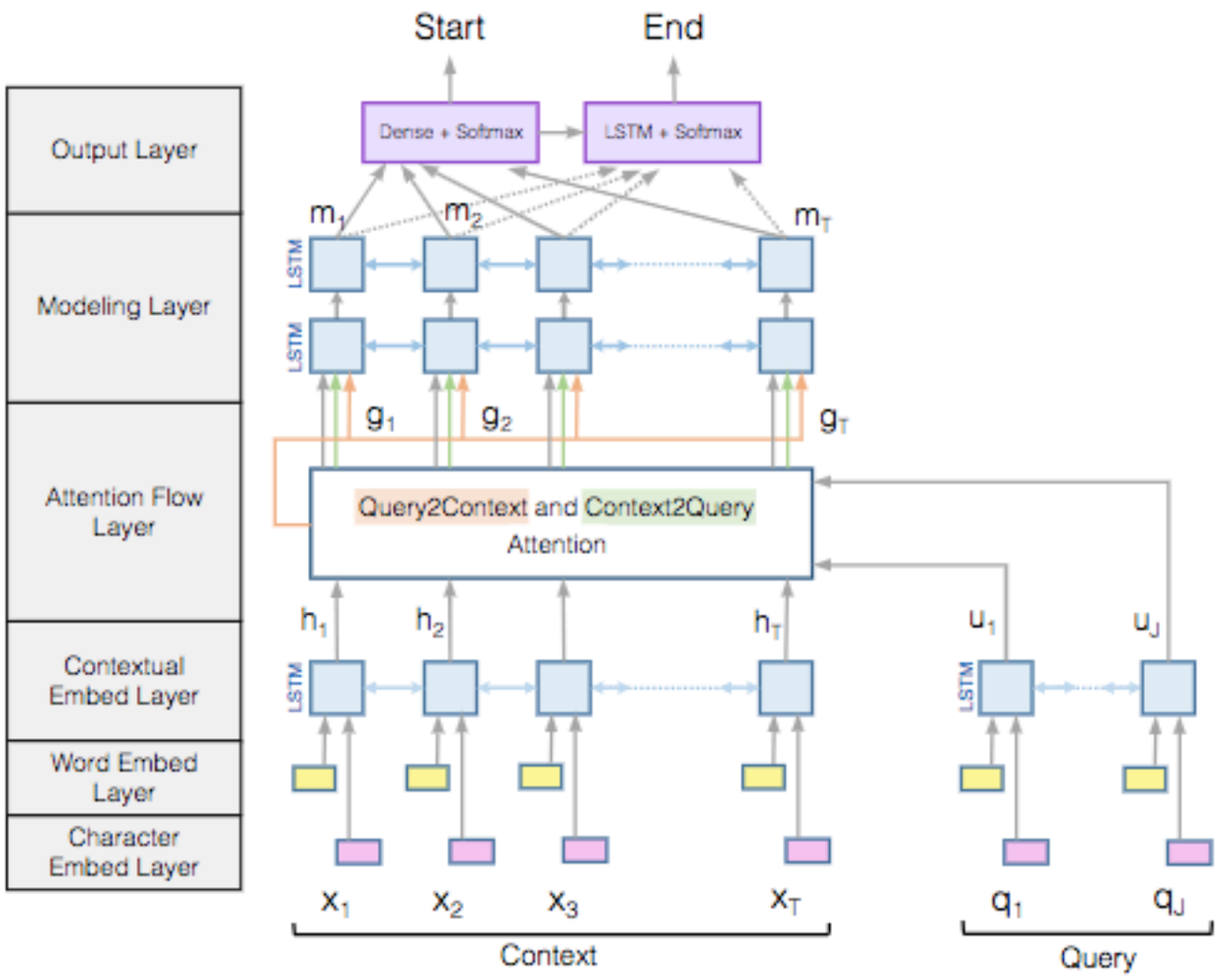
Compute a task specific weighted vector

$$ELMO_k = \gamma^{task} \sum_{j=0}^{L} s_j^{task} h_{k,j}^{LM}$$

Context Independent input : x_k

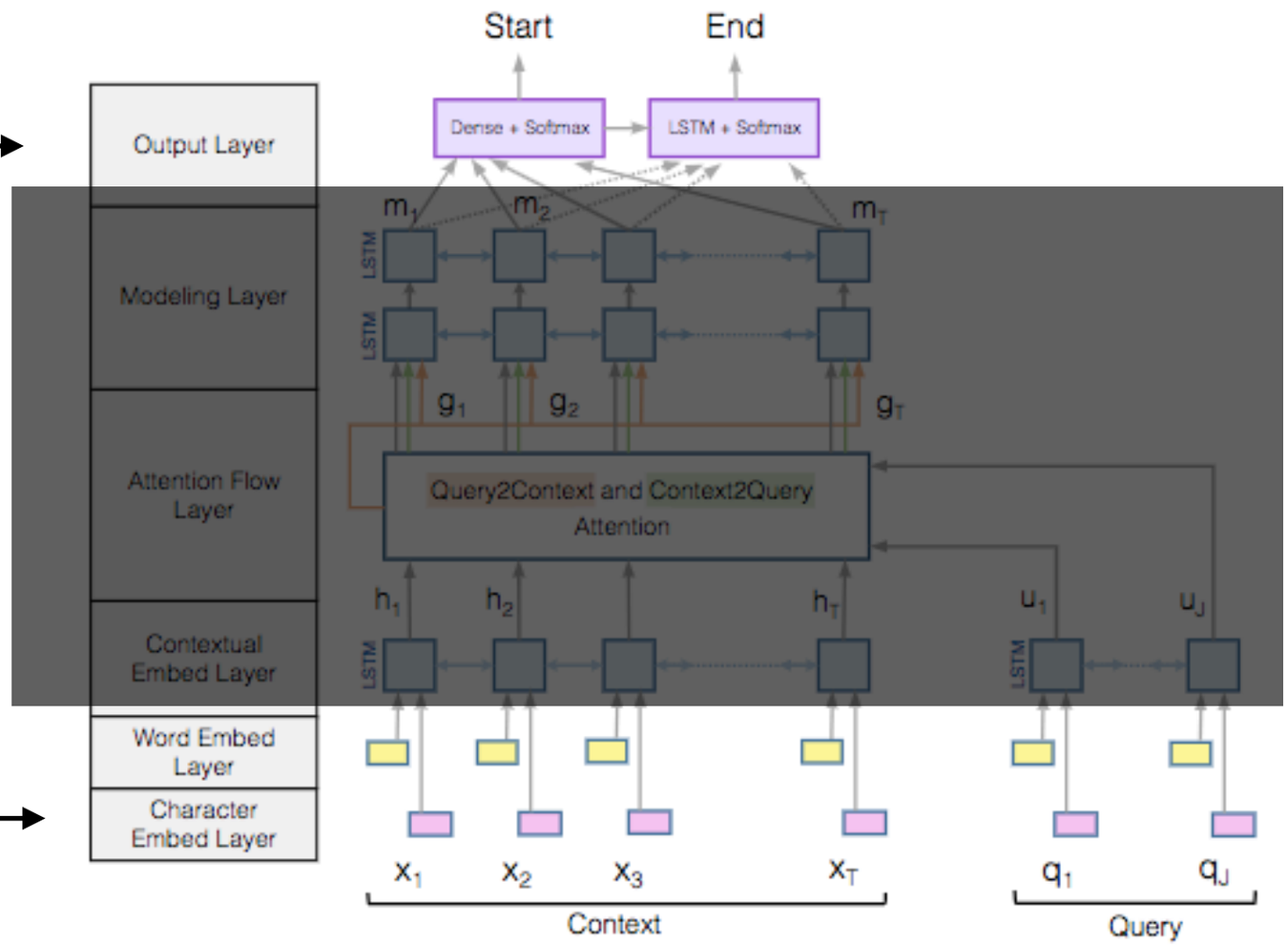
ELMO input : [x_k, $ELMO_k$]

Inclusion in downstream tasks (Example)



Inclusion in downstream tasks (Example)

Elmo here →



Elmo here →

Results

Elmo Embeddings lead to improvement in all tasks!

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%

Layer weighting is effective

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**

Results

POS and Word Sense Disambiguation

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

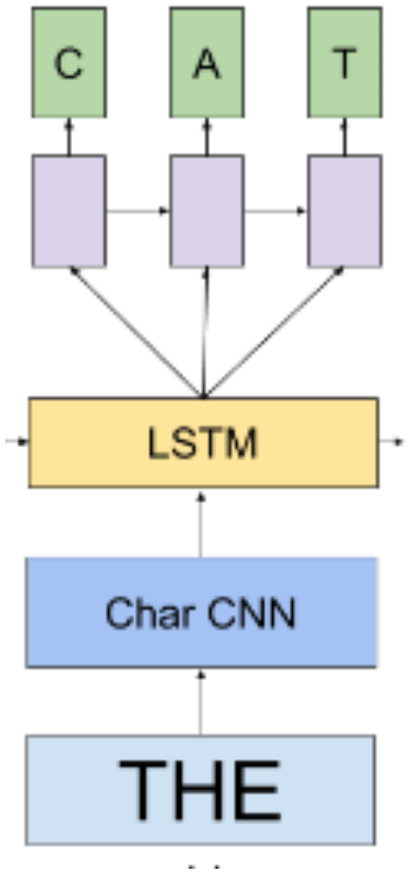
Pros

#1 Very strong results

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%

Beats the state of the art across tasks and architectures!

#2 Character based embeddings



Takes cares of the oov problem.

#3 Ablation studies to justify their design decisions

Layer weighting scheme:

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**

Adding to input and output layers:

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

#4 Show that different layers capture different information

Word Sense Disambiguation (Semantics)

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*

POS tagging (Syntax)

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