Lecture 8-3: ELMo

Pilsung Kang
School of Industrial Management Engineering
Korea University

ELMo: Embeddings from Language Models

Peters et. al (2018)

- Pre-trained word representations
  - ✓ A key component in many neural language understanding models
- High quality representations should ideally model
  - ✓ Complex characteristics of word use (e.g., syntax and semantics)
  - ✓ How these uses vary across linguistic contexts (i.e., to model polysemy)

http://jalammar.github.io/illustrated-bert/

ELMo: Embeddings from Language Models

Peters et. al (2018)

- GloVe vs. ELMo

Example

GloVe mostly learns *sport*-related context

	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 .

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.

ELMo can distinguish the word sense based on the context

https://www.slideshare.net/shuntaroy/a-review-of-deep-contextualized-word-representations-peters-2018

ELMo: Embeddings from Language Models

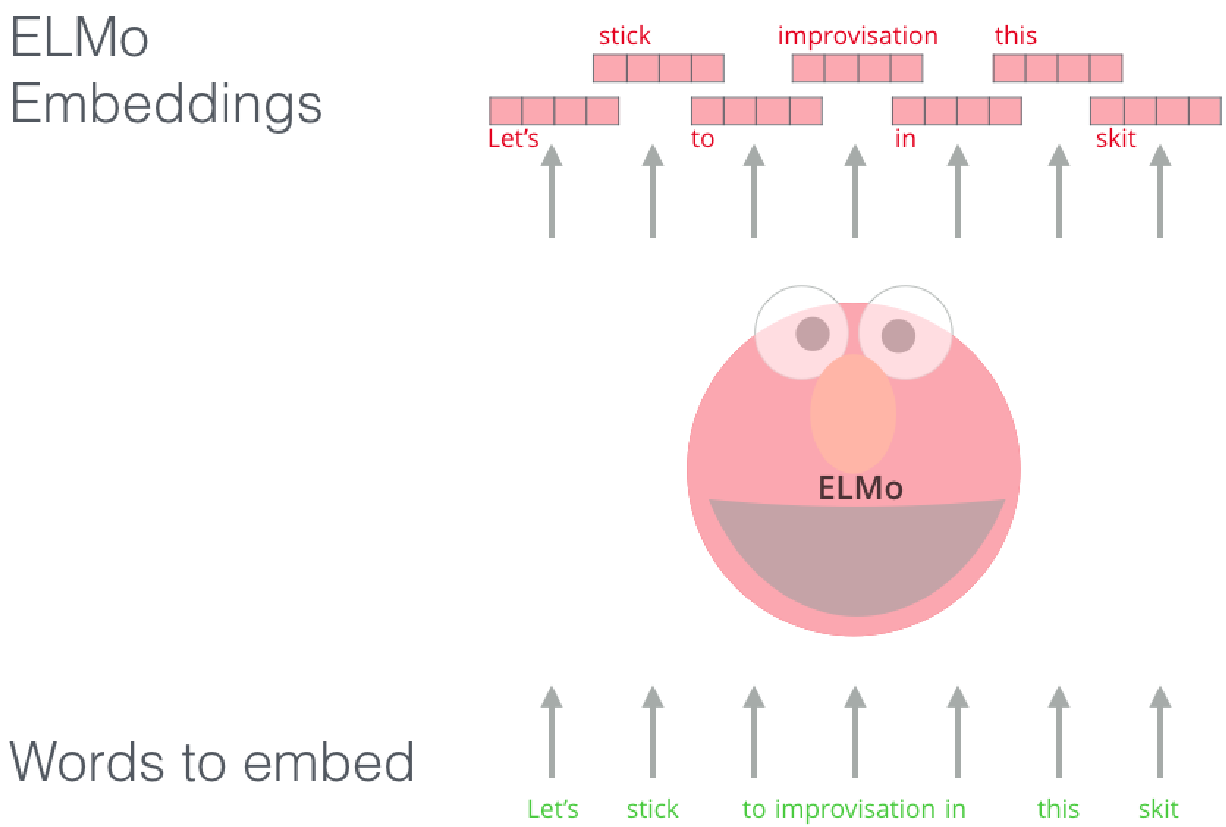
Peters et. al (2018)

- ELMo
  - ✓ Each token is assigned a representation that is a function of the entire input sentence
  - ✓ Use vectors derived from a bidirectional LSTM that is trained with a coupled language model (LM) objective on a large text corpus
- Features
  - ✓ ELMo representations are deep in the sense that they are a function of all of the internal layers of the biLM
    - a linear combination of the vectors stacked above each input word for each end task is learned, which markedly improves performance over just using the top LSTM layer
    - This allows for very rich word representations
      - Higher-level LSTM states captures context-dependent aspects of word meaning
      - Lower-level state model aspects of syntax

ELMo: Embeddings from Language Models

Peters et. al (2018)

- Graphical illustration
  - ✓ ELMo looks at the entire sentence before assigning each word in it an embedding

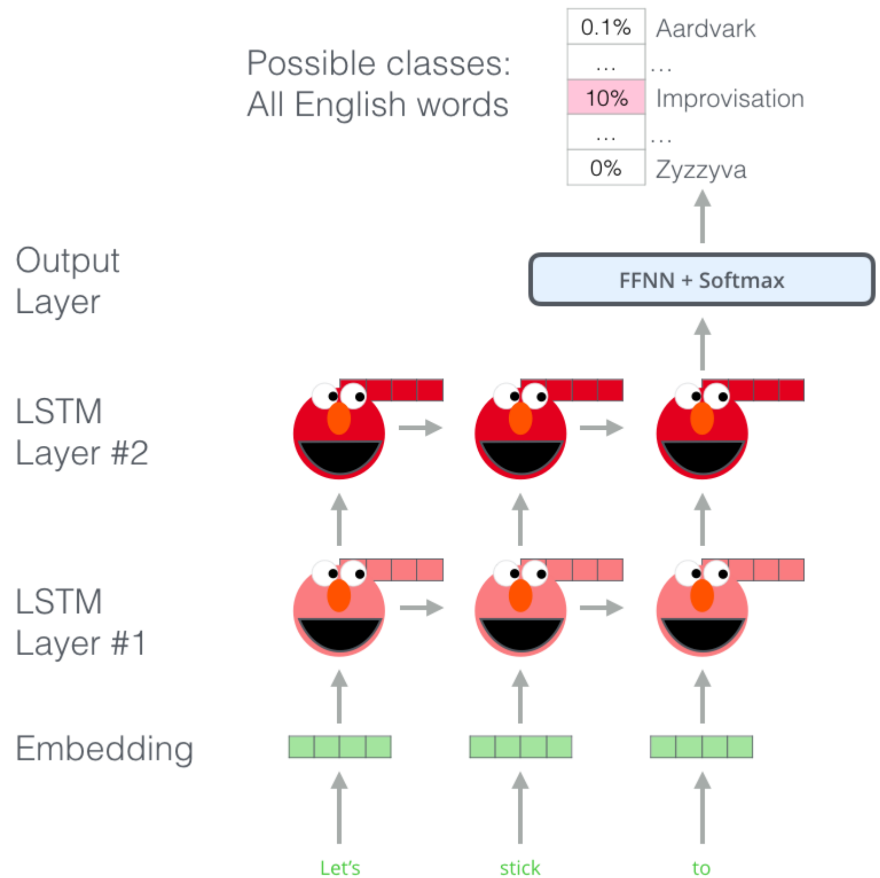


http://jalammar.github.io/illustrated-bert/

ELMo: Embeddings from Language Models

Peters et. al (2018)

- Graphical illustration
  - ✓ ELMo gained its language understanding from being trained to predict the next word in a sequence of words - a task called Language Modeling

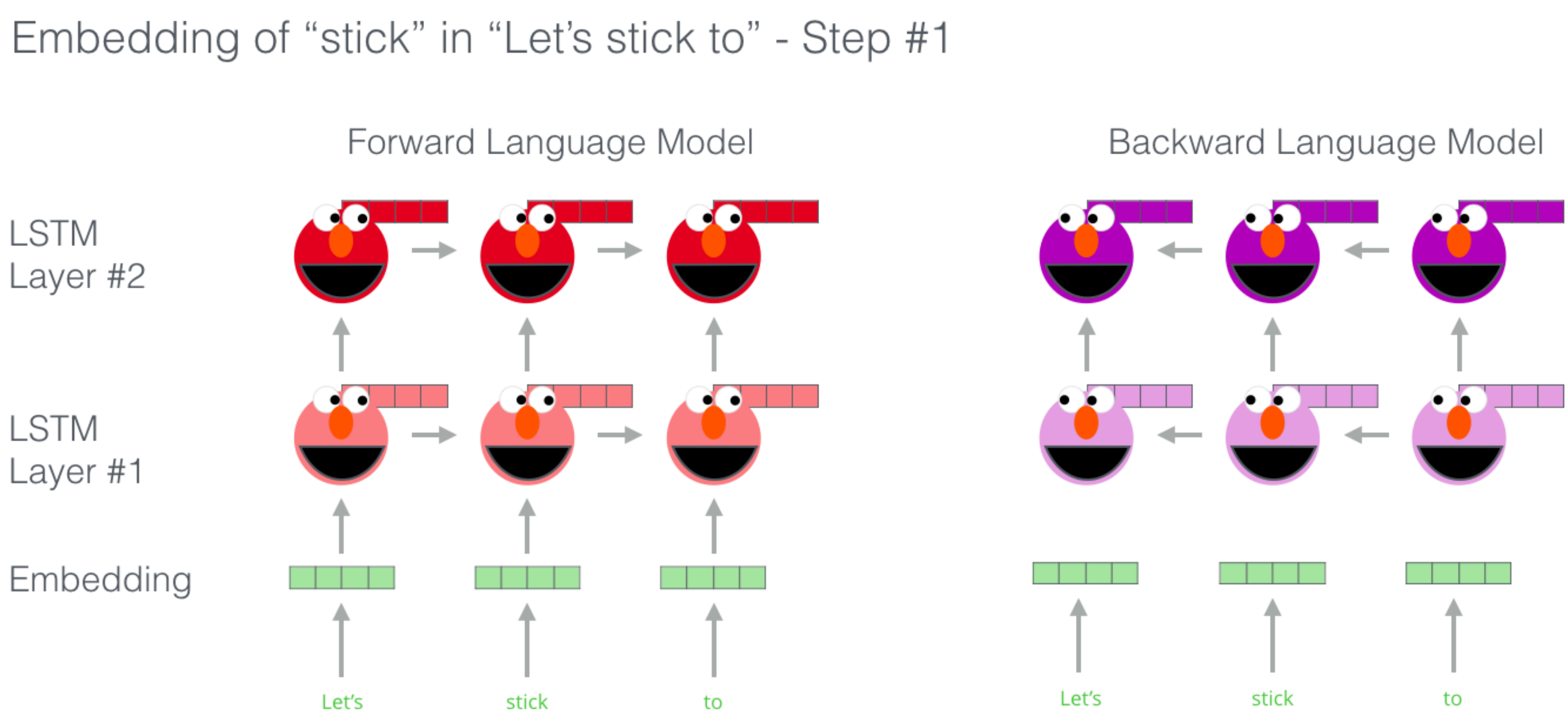


http://jalammar.github.io/illustrated-bert/

ELMo: Embeddings from Language Models

Peters et. al (2018)

- Graphical illustration
  - ✓ ELMo actually goes a step further and trains a bi-directional LSTM – so that its language model doesn't only have a sense of the next word, but also the previous word.

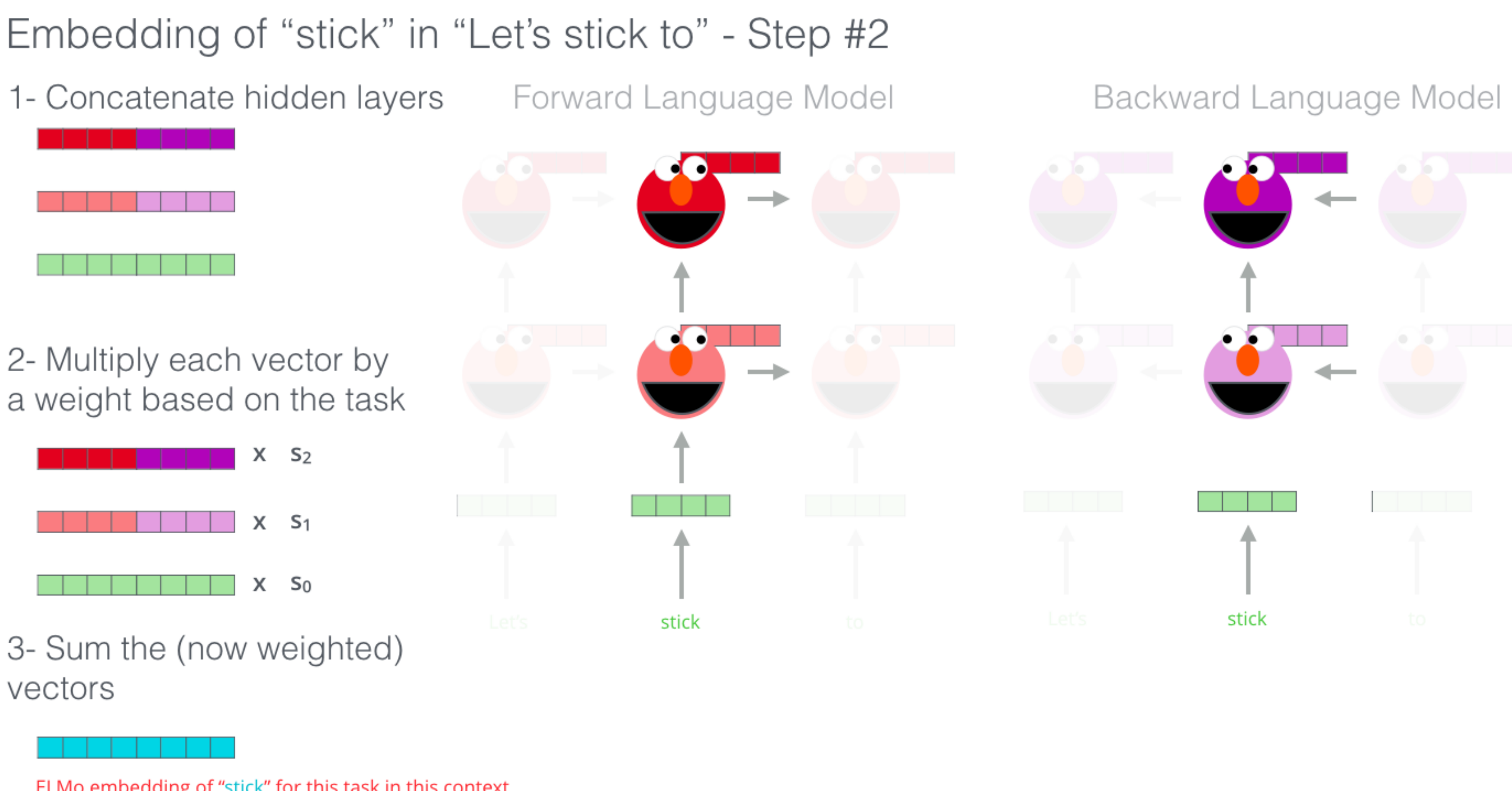


http://jalammar.github.io/illustrated-bert/

ELMo: Embeddings from Language Models

Peters et. al (2018)

- Graphical illustration
  - ✓ ELMo comes up with the contextualized embedding through grouping together the hidden states (and initial embedding) in a certain way (concatenation followed by weighted summation)



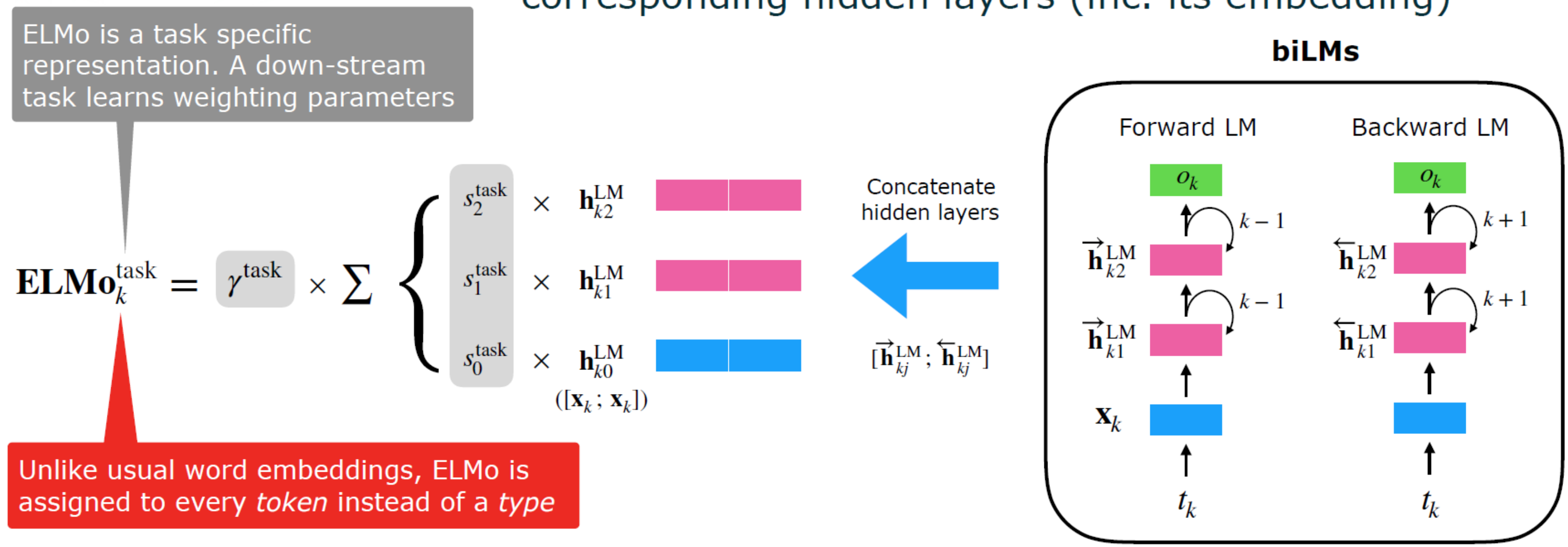
http://jalammar.github.io/illustrated-bert/

ELMo: Embeddings from Language Models

Peters et. al (2018)

- ELMo for downstream task

ELMo represents a word t_k as a linear combination of corresponding hidden layers (inc. its embedding)



https://www.slideshare.net/shuntaroy/a-review-of-deep-contextualized-word-representations-peters-2018

ELMo: Embeddings from Language Models

Peters et. al (2018)

- Mathematical demonstration: Bidirectional language models
  - ✓ Given a sequence of N tokens (t_1, t_2, …, t_N), a forward language model computes the probability of the sequence by modeling probability of token t_k given the history (t_1, t_2, …, t_{k-1})

$$p(t_1, t_2, ..., t_N) = \prod_{k=1}^{N} (t_k | t_1, t_2, ..., t_{k-1})$$

  - ✓ Neural language models compute a context-independent token representation x_k^{LM} (via token embeddings or a CNN over characters) then pass it through L layers of forward LSTMs
  - ✓ At each position k, each LSTM layer outputs a context-dependent representation $\overrightarrow{h}_{k,j}^{LM}$ where j = 1, …, L
  - ✓ The top layer LSTM output, $\overrightarrow{h}_{k,L}^{LM}$ is used to predict the next token t_{k+1} with a Softmax layer

ELMo: Embeddings from Language Models

Peters et. al (2018)

- Mathematical demonstration: Bidirectional language models
  - ✓ A backward LM is similar to a forward LM, except it runs over the sequence in reverse, predicting the previous token given the future context

$$p(t_1, t_2, ..., t_N) = \prod_{k=1}^{N} (t_k | t_{k+1}, t_{k+2}, ..., t_N)$$

  - ✓ Each backward LSTM layer j in an L layer deep model producing representations $\overleftarrow{h}_{k,j}^{LM}$ of t_k given $(t_{k+1}, \ldots, t_N)$

ELMo: Embeddings from Language Models

Peters et. al (2018)

- Mathematical demonstration: Bidirectional language models
  - ✓ Jointly maximizes the log likelihood of the forward and backward directions

$$\sum_{k=1}^{N} \Big(\log p(t_k | t_1, ..., t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k | t_{k+1}, ..., t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \Big)$$

  - Θ_x, Θ_s : tied token representation & softmax layer parameters
  - Separated parameters for the LSTMs in each direction

ELMo: Embeddings from Language Models

Peters et. al (2018)

- ELMo
  - ✓ A task specific combination of the intermediate layer representations in the biLM
  - ✓ For each token t_k, a l-layer biLM computes a set of 2L+1 representations

$$R_k = \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} | j = 1, ..., L\} = \{\mathbf{h}_{k,j}^{LM}, | j = 0, ..., L\}$$

    - where $\mathbf{h}_{k,0}^{LM}$ is the token layer and $\mathbf{h}_{k,j}^{LM} = [\overrightarrow{\mathbf{h}}_{k,j}^{LM}; \overleftarrow{\mathbf{h}}_{k,j}^{LM}]$ for each biLSTM layer
  - ✓ For inclusion in a downstream model, ELMo collapses all layers in R into a single vector

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \underbrace{\gamma^{task}}_{\text{allows task model to scale the entire ELMo vector}} \sum_{j=0}^{L} \underbrace{s_j^{task}}_{\text{softmax-normalized weights}} \mathbf{h}_{k,j}^{LM}$$

ELMo: Embeddings from Language Models

Peters et. al (2018)

- Natural language inference (NLI) task
  - ✓ Classify two given sentence to one of the three classes: entailment, contradiction, neutral
    - Examples (https://nlp.stanford.edu/projects/snli/)

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

ELMo: Embeddings from Language Models

Peters et. al (2018)

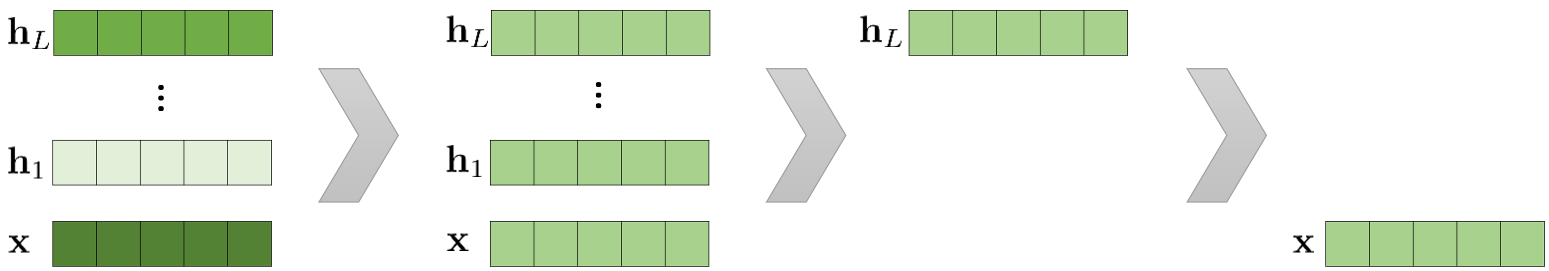
- Performances

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 $\pm$ 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 $\pm$ 0.19	90.15	92.22 $\pm$ 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 $\pm$ 0.5	3.3 / 6.8%

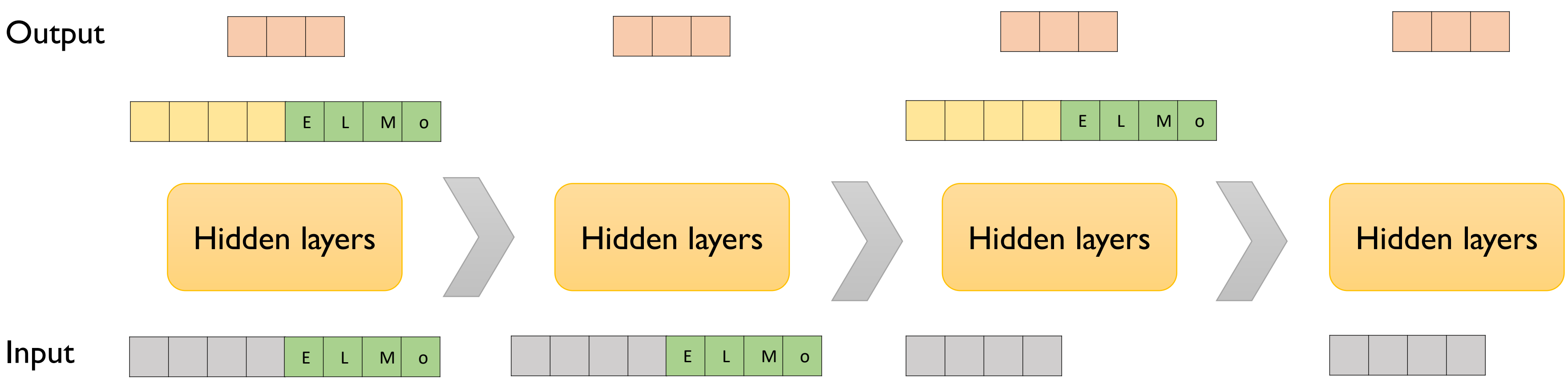
ELMo: Embeddings from Language Models

Peters et. al (2018)

- Analysis: Alternate layer weighting scheme



- Analysis: Where to include ELMo?



ELMo: Embeddings from Language Models

Peters et. al (2018)

- Analysis: What information is captured by the biLM's representation?
  - ✓ Disambiguating the meaning of words using their context

	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 .


ANY
questions?