Contextualized Embeddings

COSC 6336: Natural Language Processing
Spring 2020

Some content in these slides has been adapted from Greg Durrett and NAACL 2018 presentation by Peters et al.

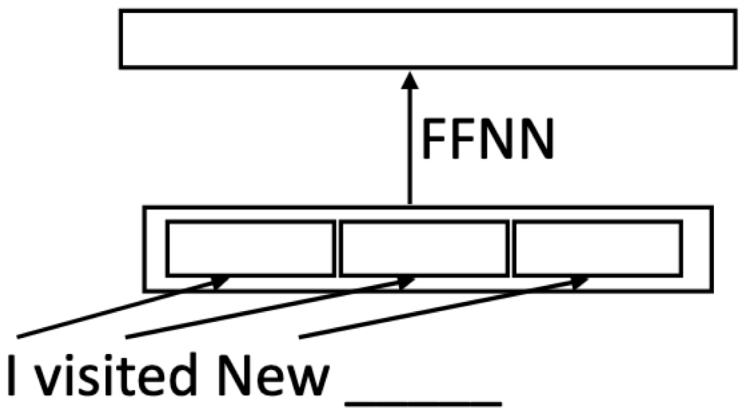
UNIVERSITY of HOUSTON

Today's Lecture

- ★ Neural Language Models
- ★ Word Embeddings with Context
  - ★ ELMo
- ★ Task-Specific Fine Tuned Embeddings

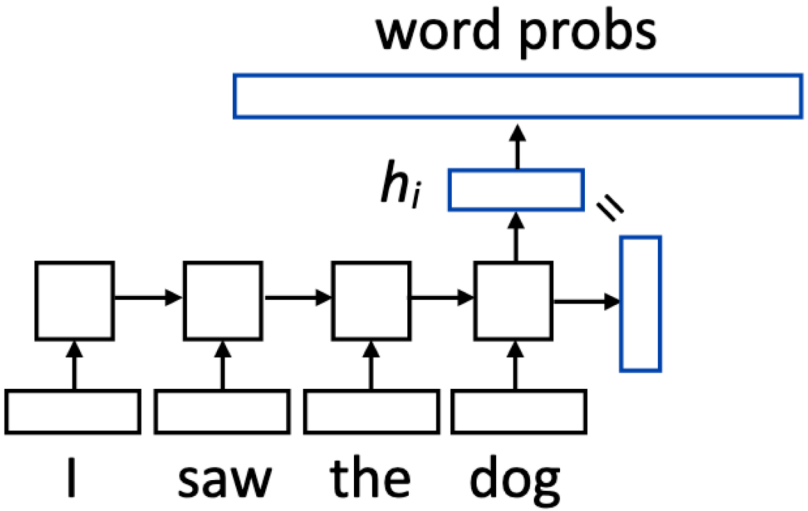
Neural Language Models

- Mnih and Hinton (2003)
- FFNN looking at context



$$P(w_i|w_{i-n},\ldots,w_{i-1})$$

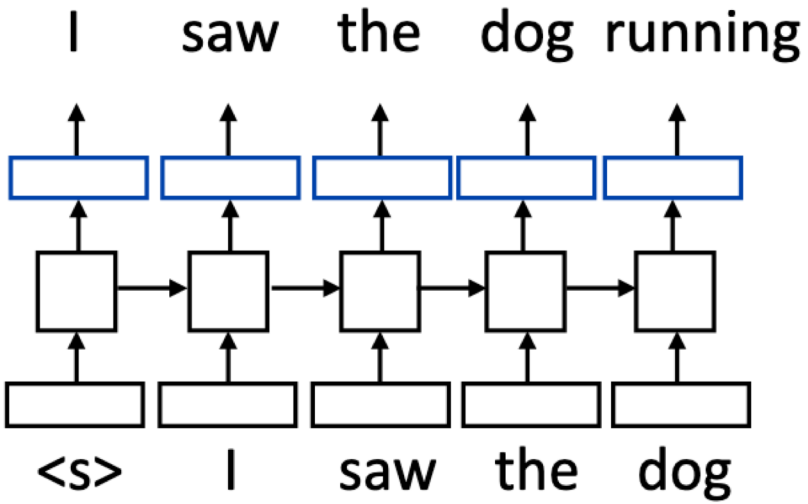
RNN Language Modeling



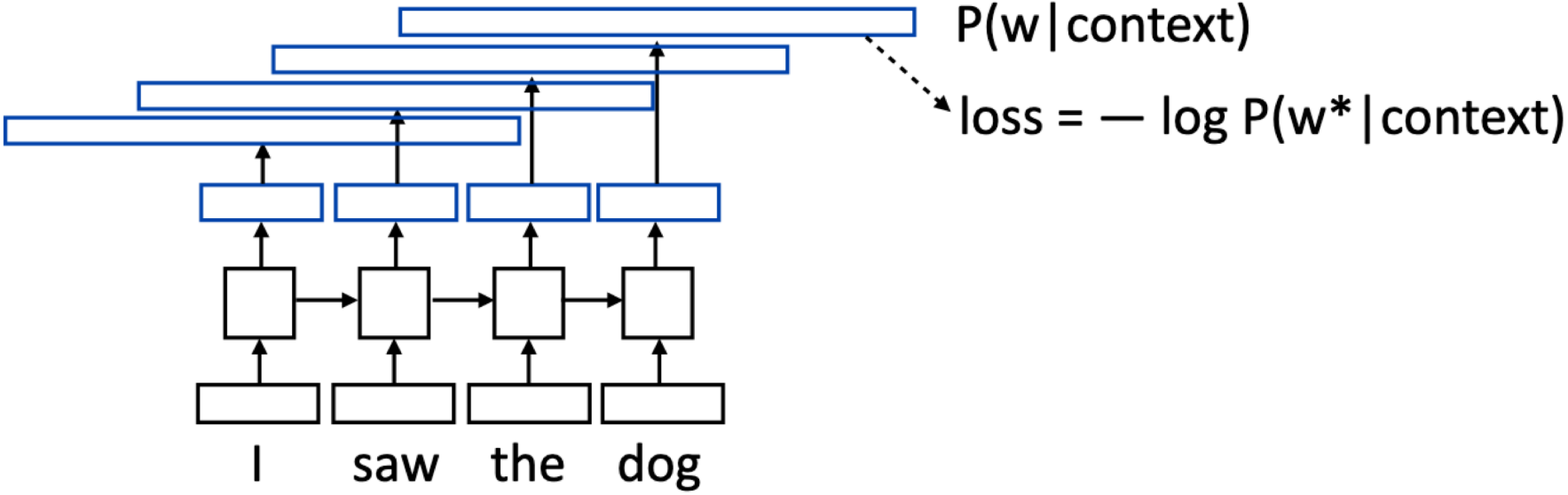
$$P(w|\text{context}) = \text{softmax}(W\mathbf{h}_i)$$

- W is a (vocab size) x (hidden size) matrix

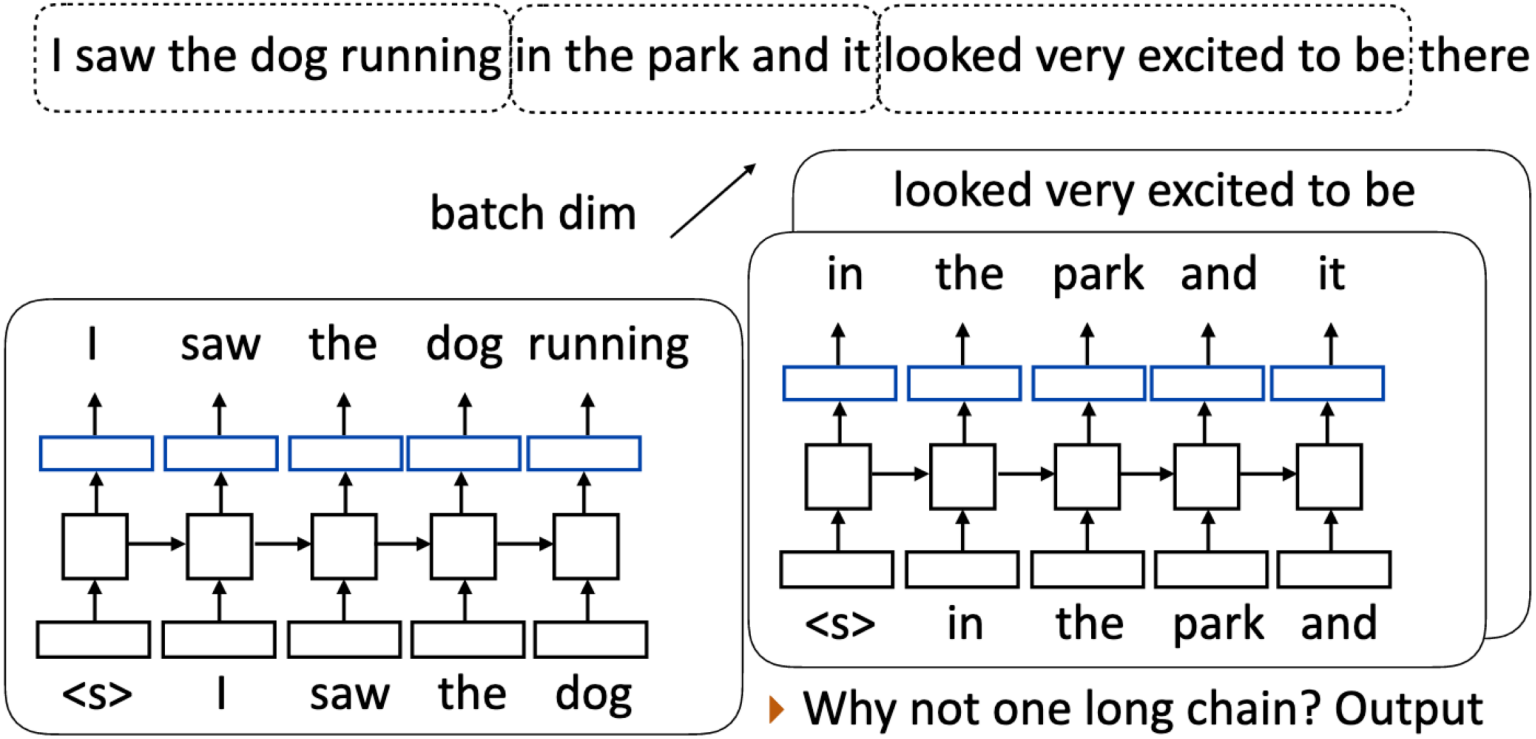
Training RNN LMs



Training RNN LMs

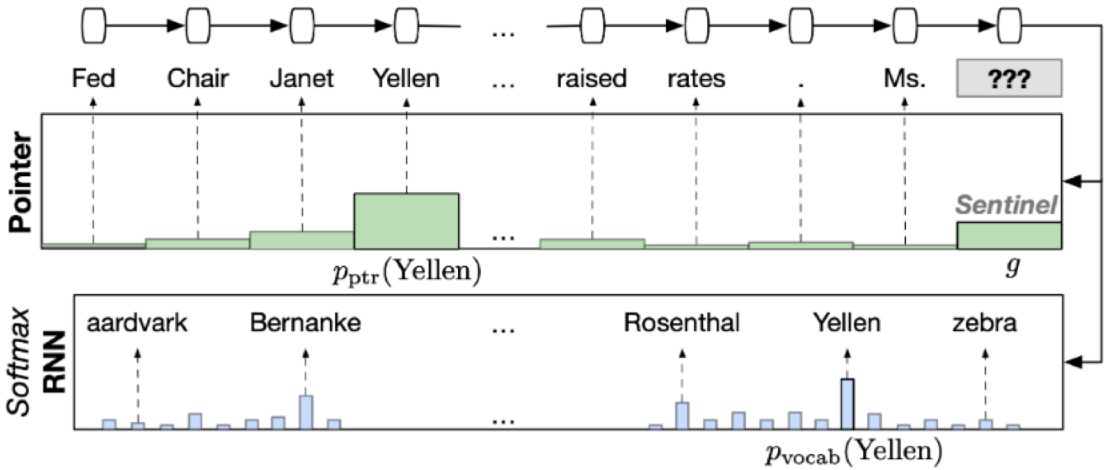


Batched LM Training



- Why not one long chain? Output

Limitations of LSTM LMs



$$p(\text{Yellen}) = g\, p_{\text{vocab}}(\text{Yellen}) + (1-g)\, p_{\text{ptr}}(\text{Yellen})$$

Merity et al. (2016)

Contextualized Embeddings

Recall from word2vec, fasttext and GloVe

- Pretrained word representations that are fixed
- Different senses of the word are conflated into single vectors

ELMo

- Intuition: language models can allow us to form useful word representations, similar to word2vec
- Take a powerful language model, train it on large amounts of data, then use those representations in downstream tasks

Compute contextual vector:

$$c_k = f(w_k \mid w_1, \ldots, w_n) \in \mathbb{R}^N$$

f(play | Elmo and Cookie Monster play a game .)

$\neq$

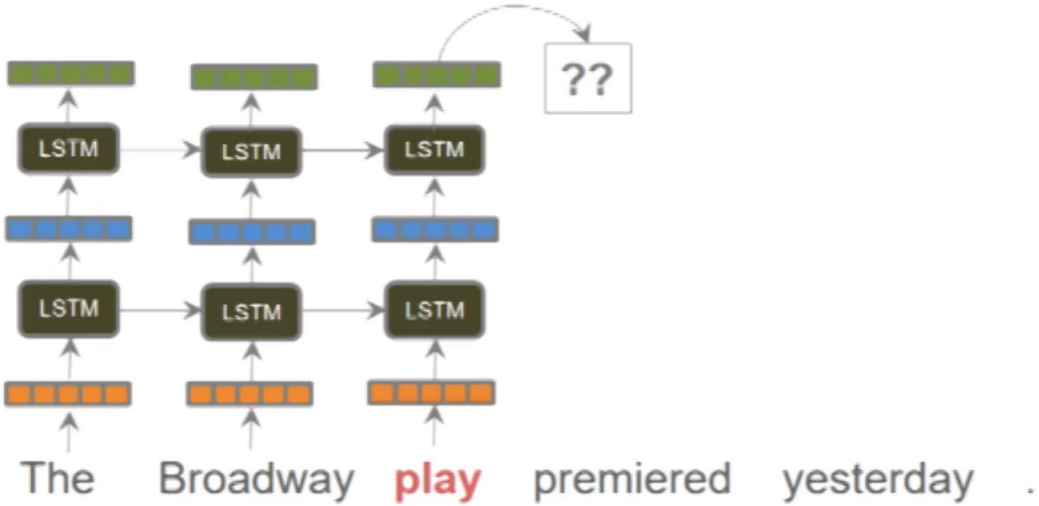
f(play | The Broadway play premiered yesterday .)

Key Ideas in ELMo

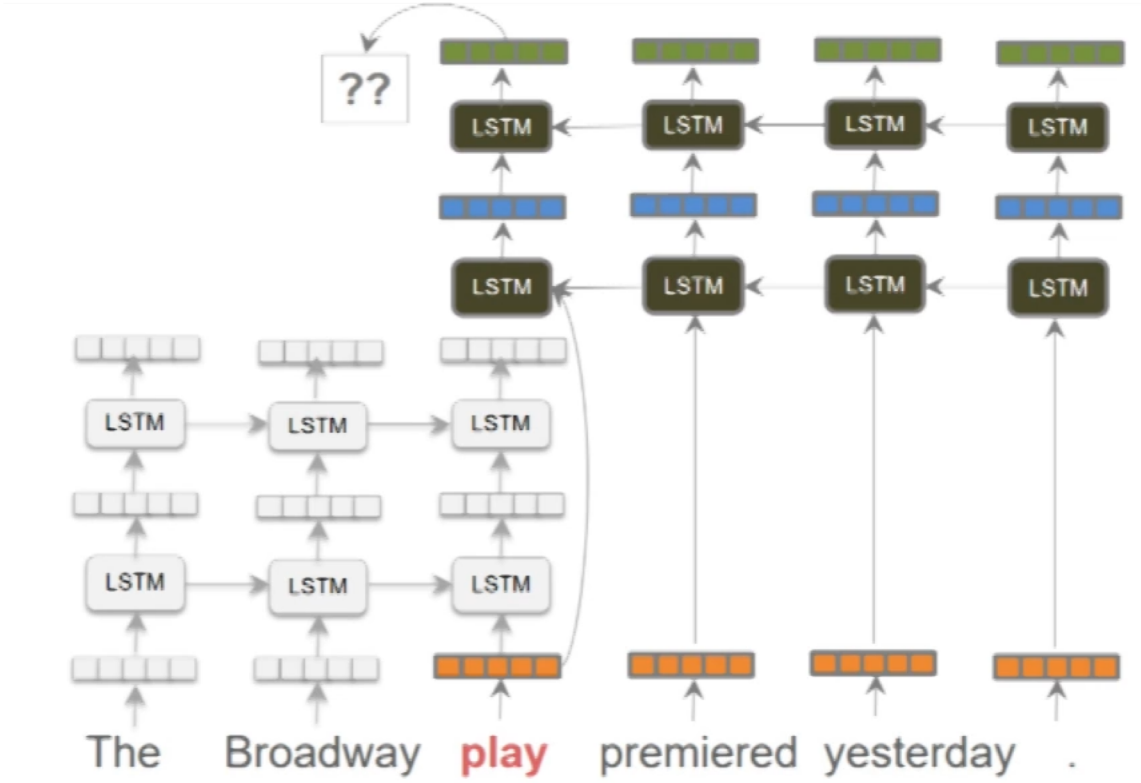
Neural LMs embed the left context of a word.

We can introduce a bidirectional LM to embed left and right context.

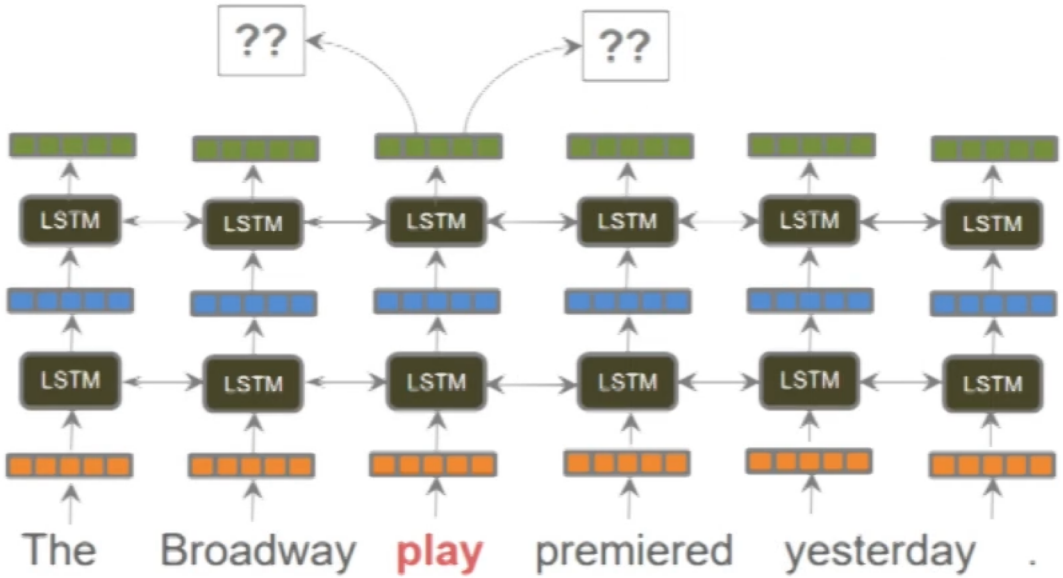
ELMo: Intuition



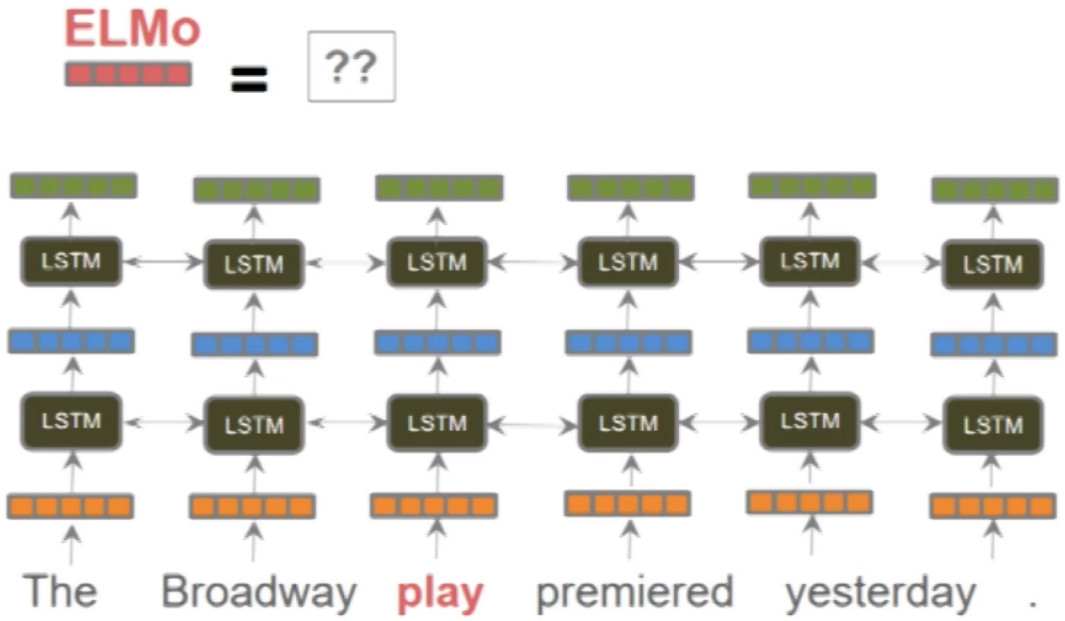
ELMo: Intuition



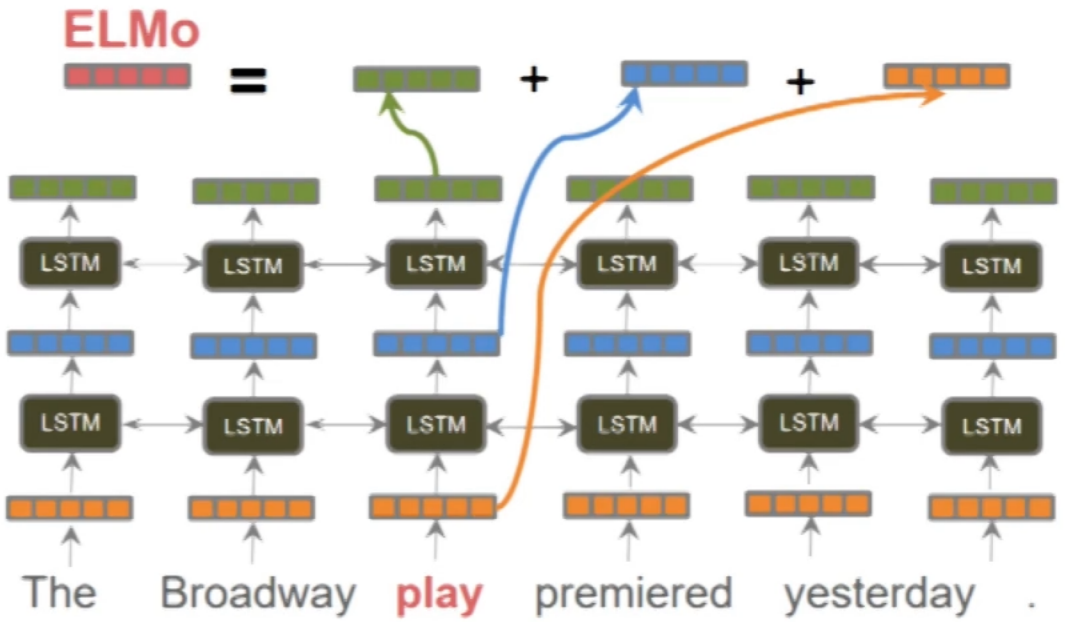
ELMo: Intuition



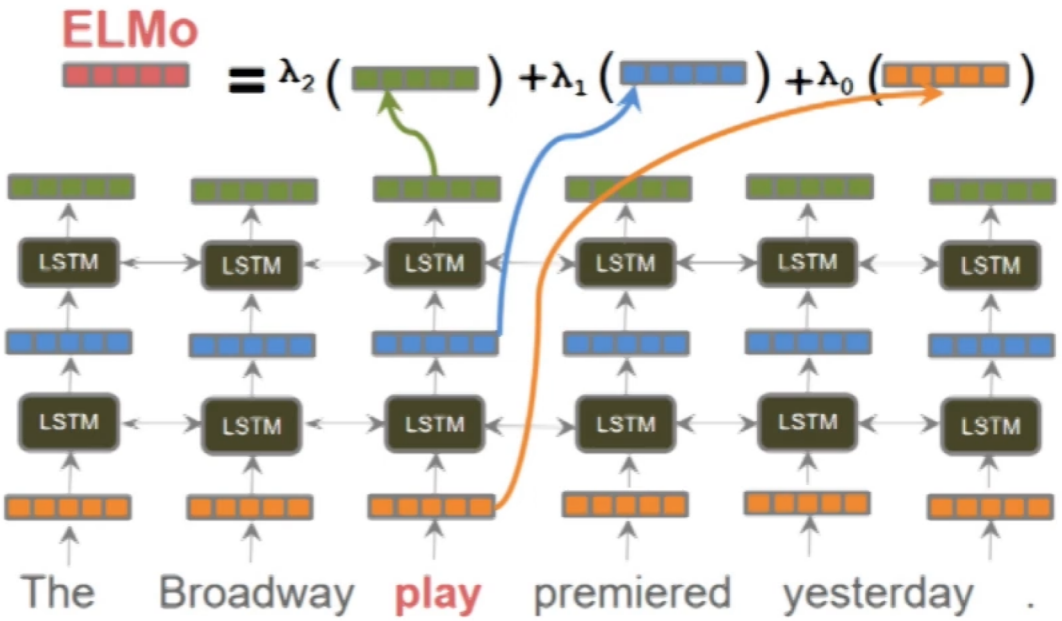
ELMo: Intuition



ELMo: Intuition



ELMo: Intuition



ELMo Properties

- Unsupervised
- Contextual
- Deep
- Character-based
- Versatile

Why ELMo works

- Language models allow to learn syntax and semantic knowledge
- Deep LM allows the end model to decide which representations to use

What is ELMo Learning?

From each layer of the ELMo model, attempt to predict something: POS tags, word senses, etc.

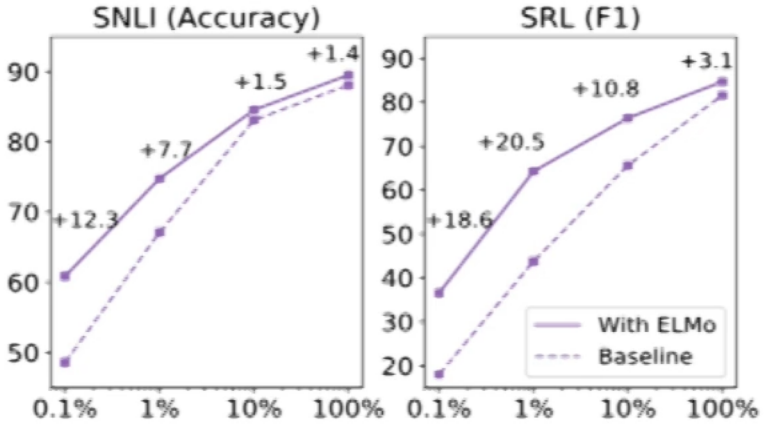
Model	F_1
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	**70.1**
CoVe, First Layer	59.4
CoVe, Second Layer	*64.7*
biLM, First layer	67.4
biLM, Second layer	*69.0*

Table 5: All-words fine grained WSD F_1. For CoVe and the biLM, we report scores for both the first and second layer biLSTMs.

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

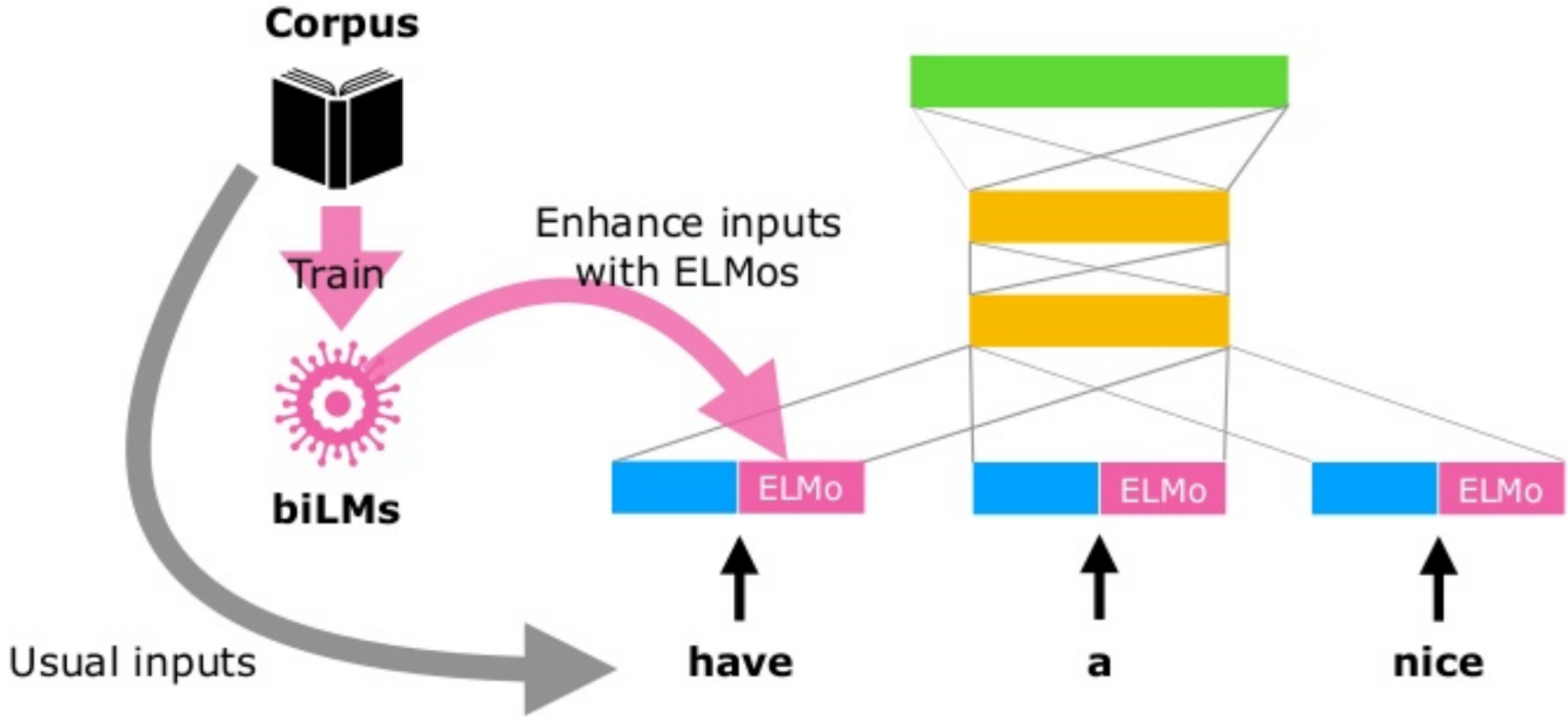
Table 6: Test set POS tagging accuracies for PTB. For CoVe and the biLM, we report scores for both the first and second layer biLSTMs.

Notes on training with ELMo



ELMo enhanced models are remarkably sample efficient

ELMo can be integrated to almost all neural NLP tasks with simple concatenation to the embedding layer



How to use ELMo

- Frozen embeddings: update the weights of your network but keep ELMo's parameters frozen

- Fine-tuning: backpropagate all the way into ELMo when training your model

Results: Frozen ELMo

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%

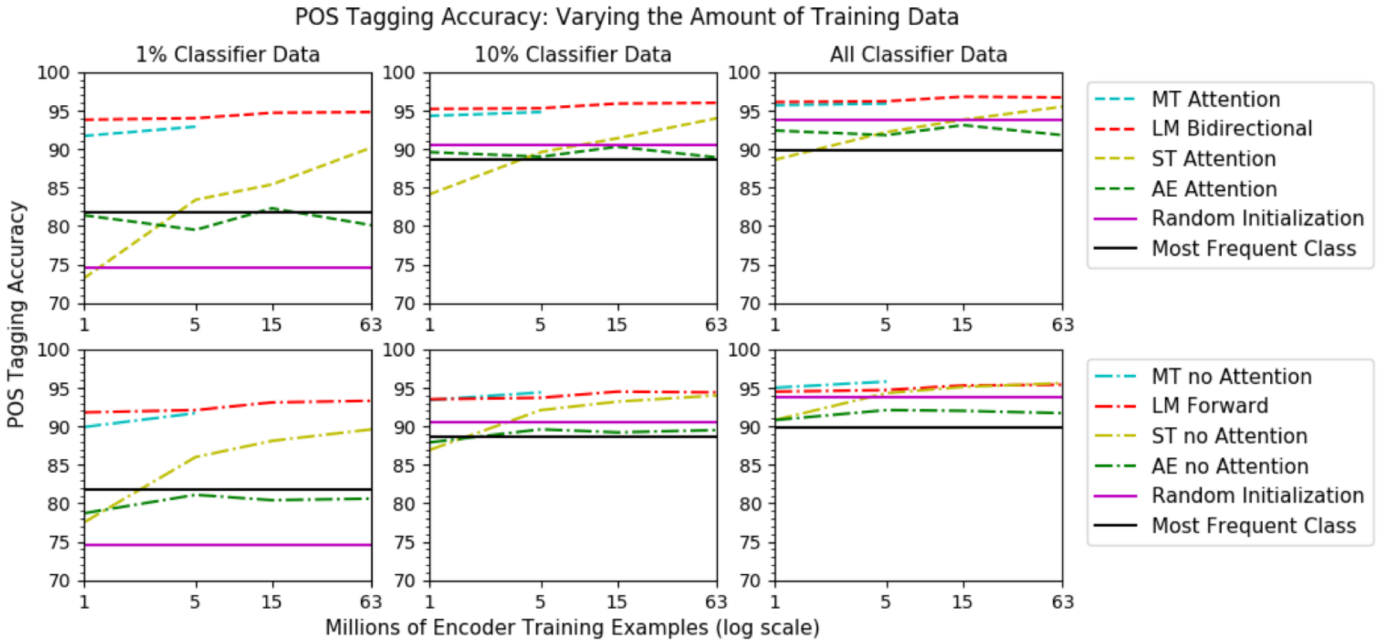
Recommendations on using ELMo

- From Peters, Ruder & Smith (2019)

Conditions			Guidelines
Pretrain	Adapt.	Task	
Any	❄️	Any	Add many task parameters
Any	🔥	Any	Add minimal task parameters ⚠️ Hyper-parameters
Any	Any	Seq. / clas.	❄️ and 🔥 have similar performance
ELMo	Any	Sent. pair	use ❄️
BERT	Any	Sent. pair	use 🔥

Why is LM a good objective?

- Zhang and Bowman (2019)



Summary

- Fixed word-based representations are suboptimal
- Contextualized representations leverage the entire input sequence
- The learning objective in LMs (predicting next word given context) allows models to learn many “useful” facts of language.
- In general, deep contextualized representations are versatile and can help improve results in many tasks