RNNs & ELMO

Based on materials from

CS 685, Spring 2021

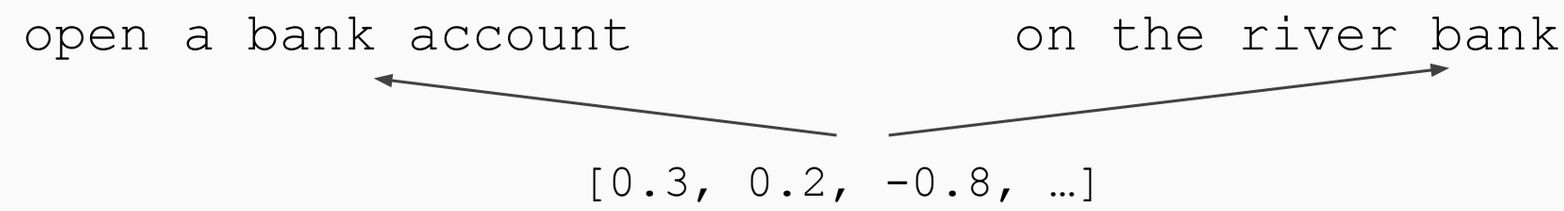
Advanced Topics in Natural Language Processing

http://brenocon.com/cs685

https://people.cs.umass.edu/~brenocon/cs685_s21/

Contextual Representations

- **Problem**: Word embeddings are applied in a context free manner

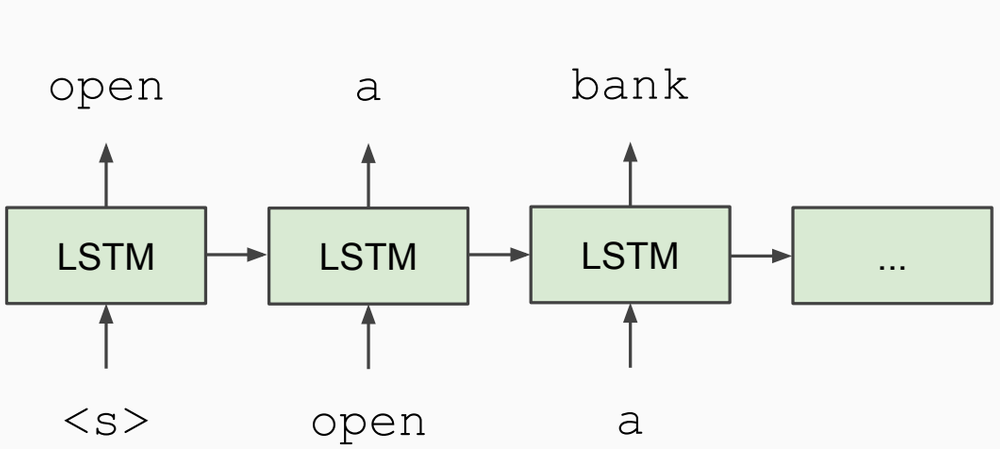


- **Solution**: Train *contextual* representations on text corpus

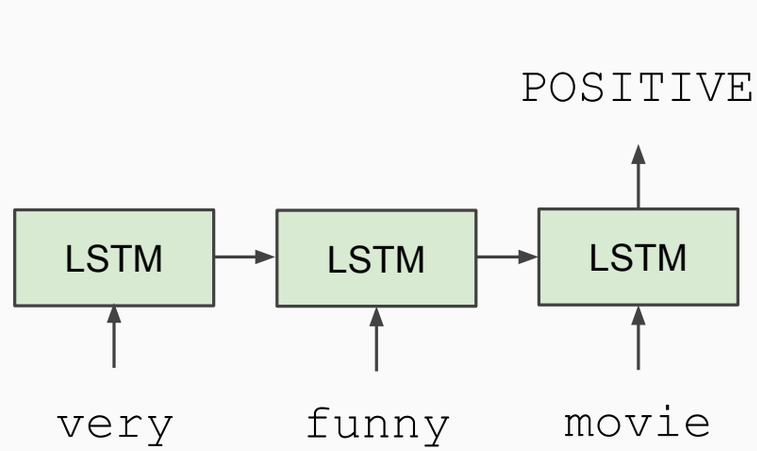
History of Contextual Representations

- *Semi-Supervised Sequence Learning*, Google, 2015

Train LSTM Language Model



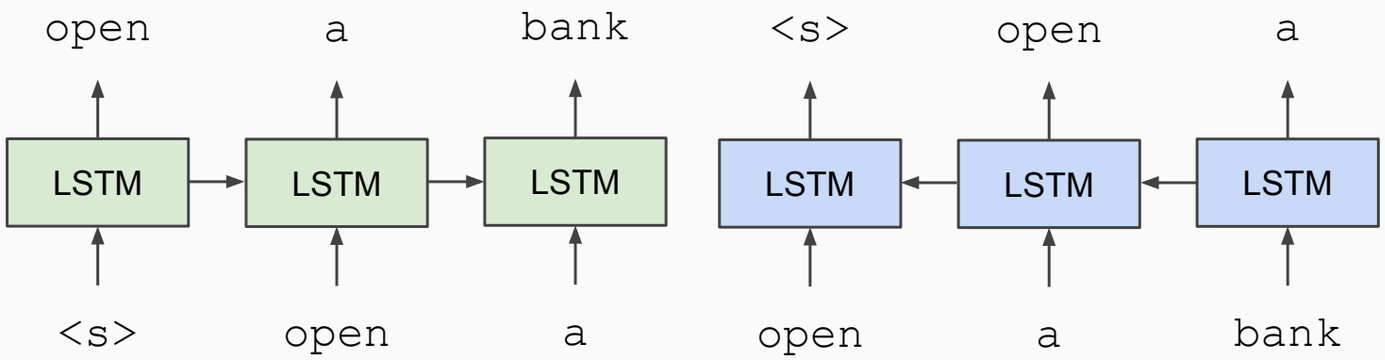
Fine-tune on Classification Task



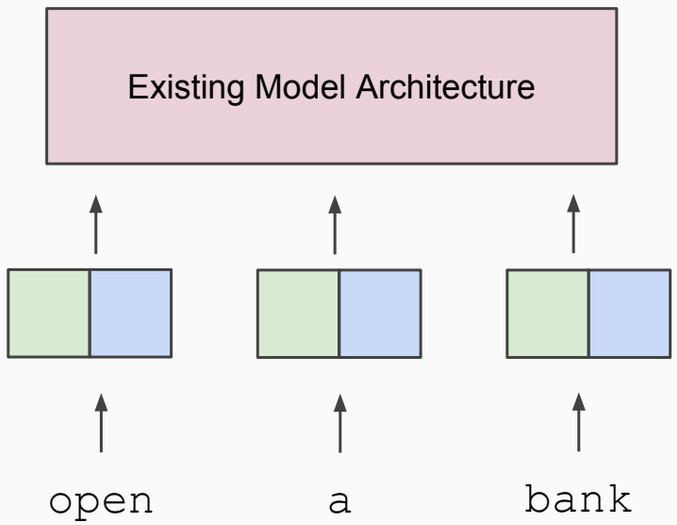
History of Contextual Representations

- *ELMo: Deep Contextual Word Embeddings*, AI2 & University of Washington, 2017

Train Separate Left-to-Right and Right-to-Left LMs



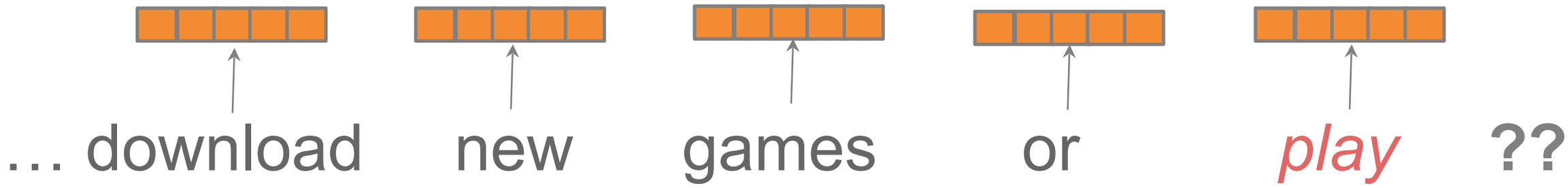
Apply as "Pre-trained Embeddings"



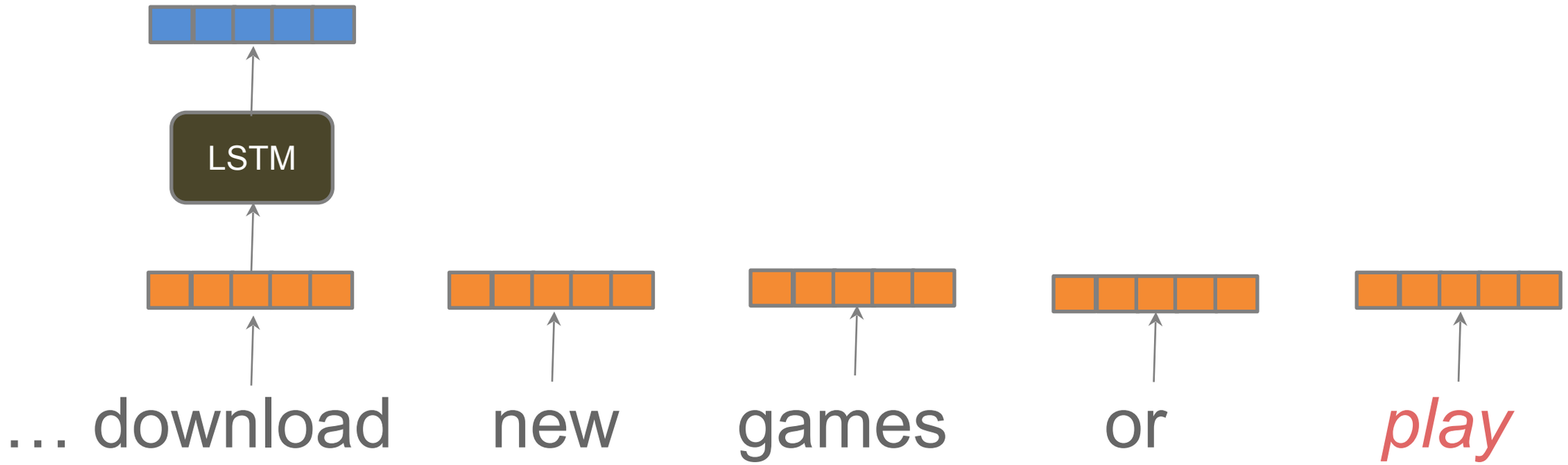
Deep bidirectional language model

… download new games or *play* ??

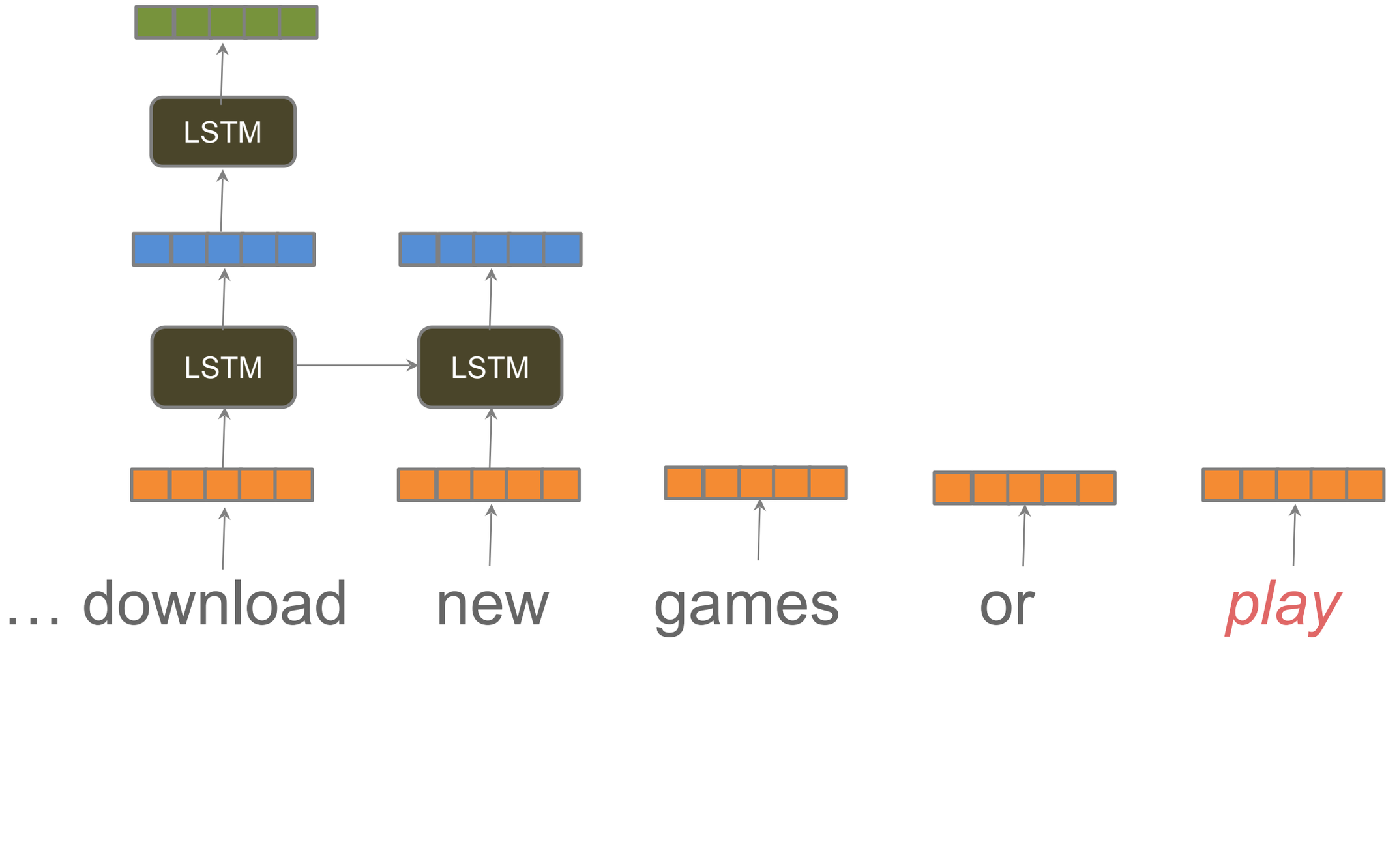
Deep bidirectional language model



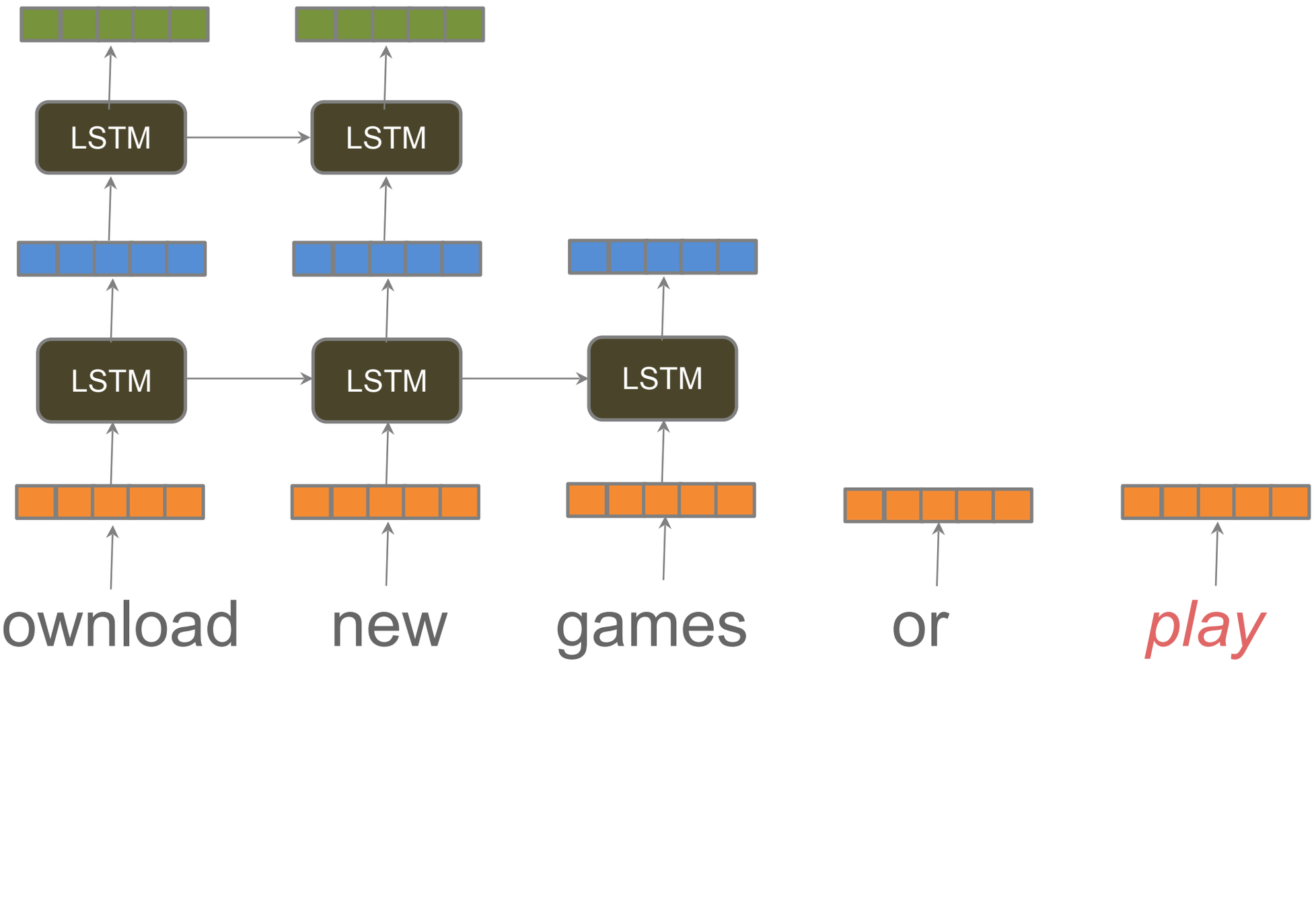
Deep bidirectional language model



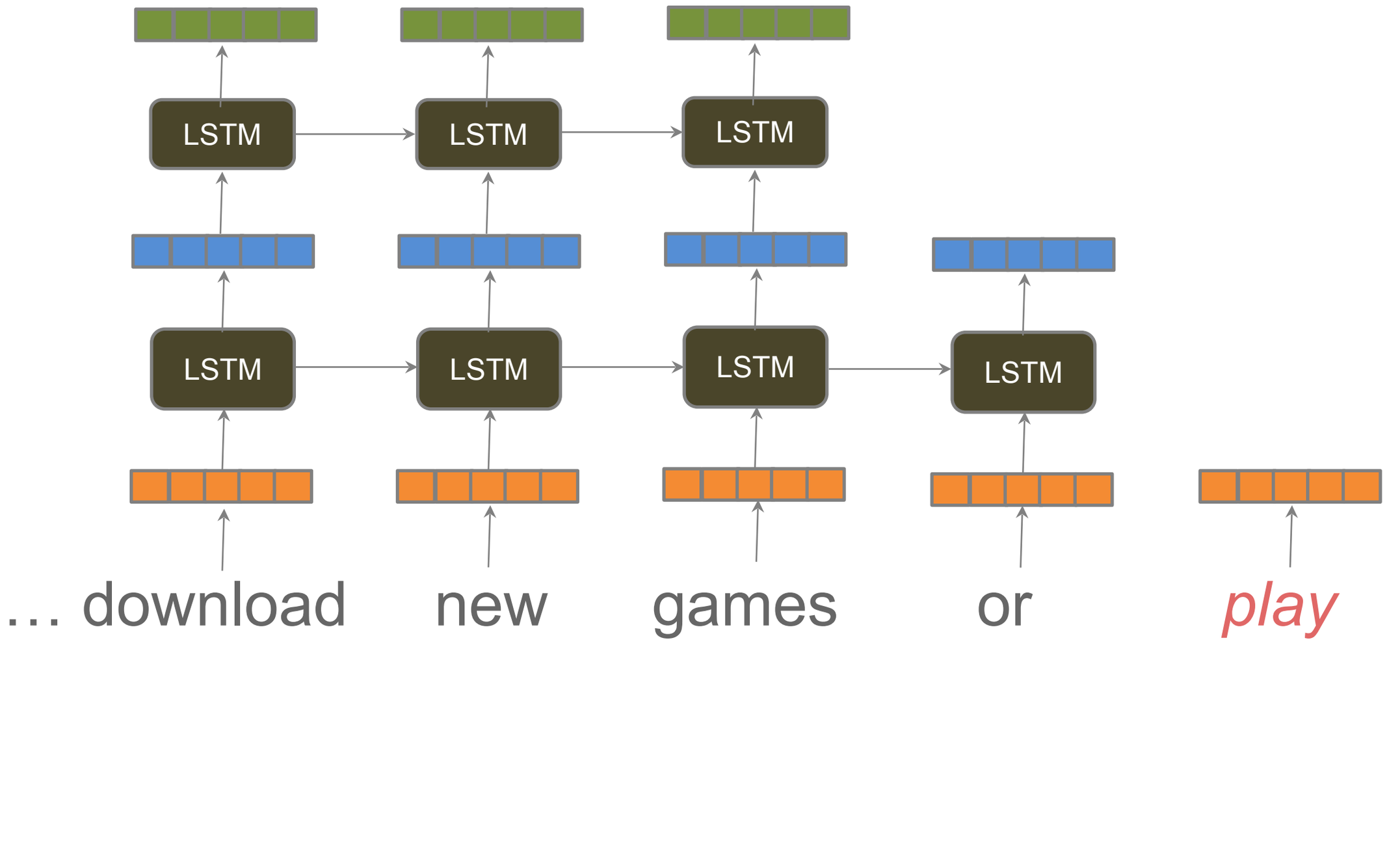
Deep bidirectional language model



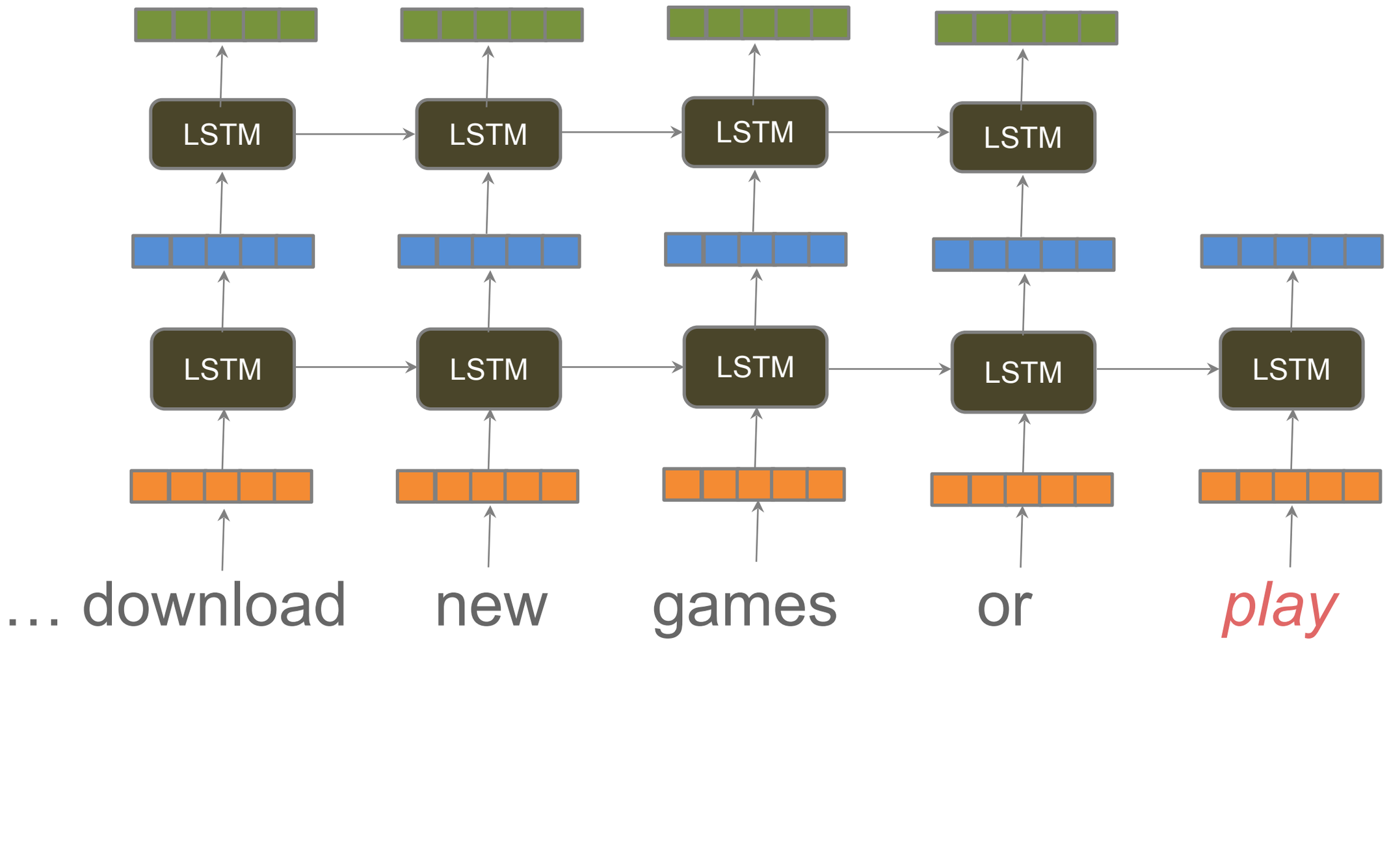
Deep bidirectional language model



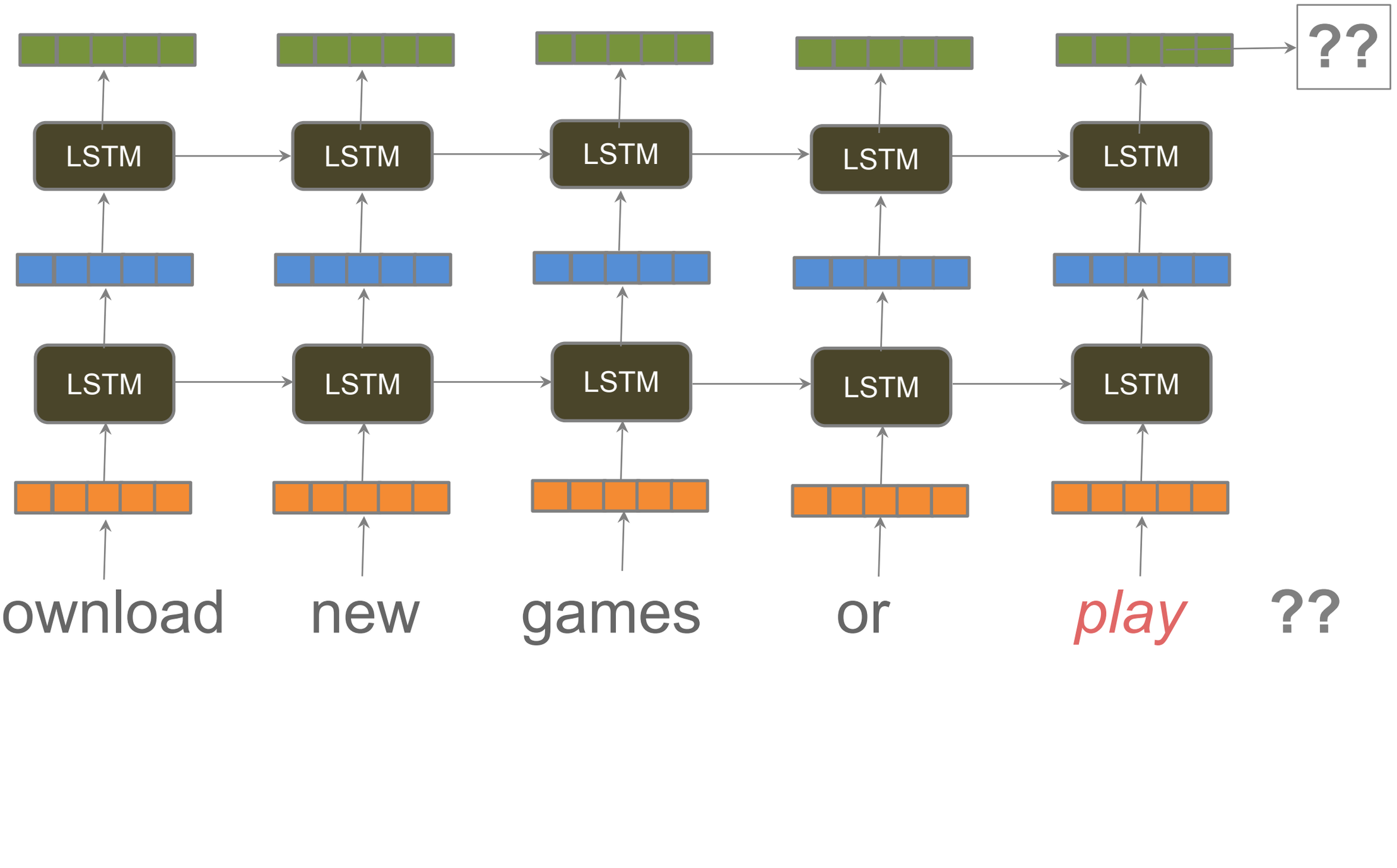
Deep bidirectional language model



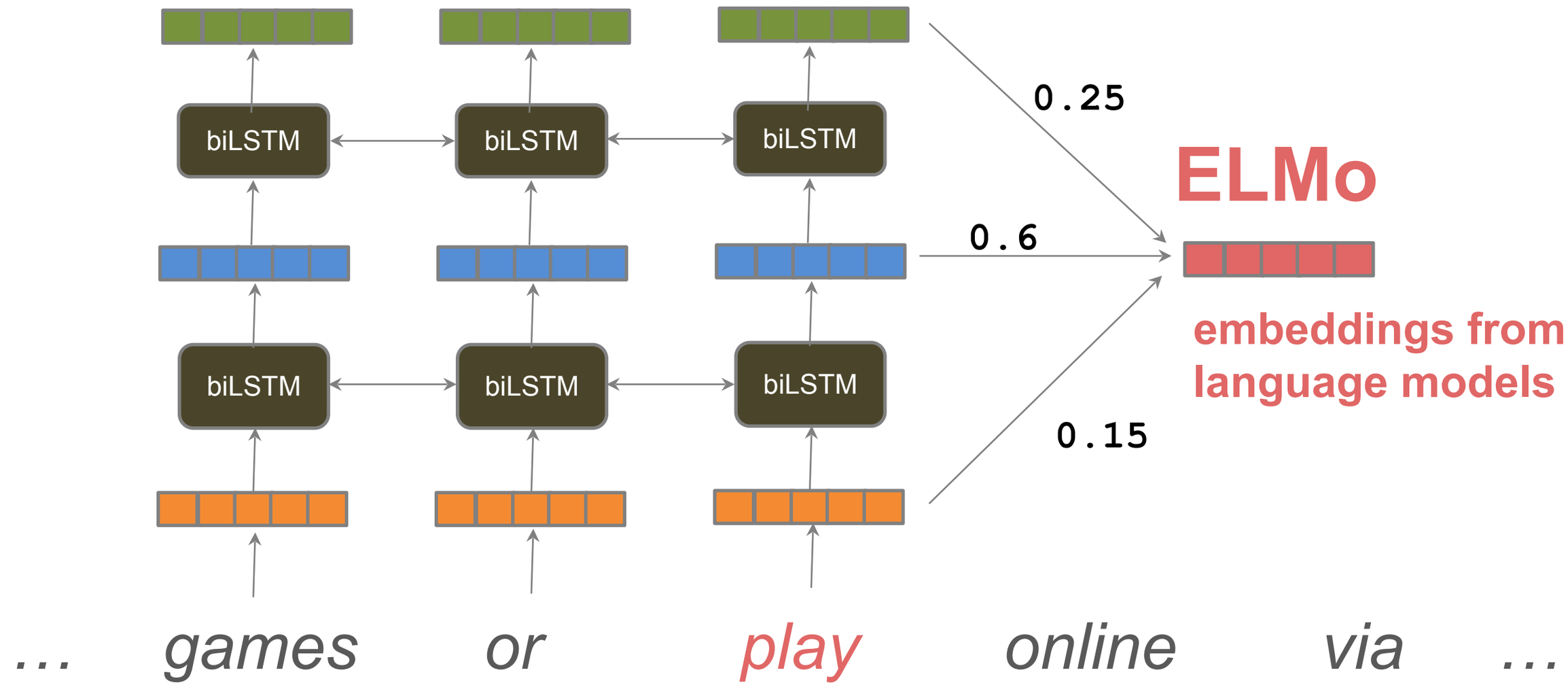
Deep bidirectional language model



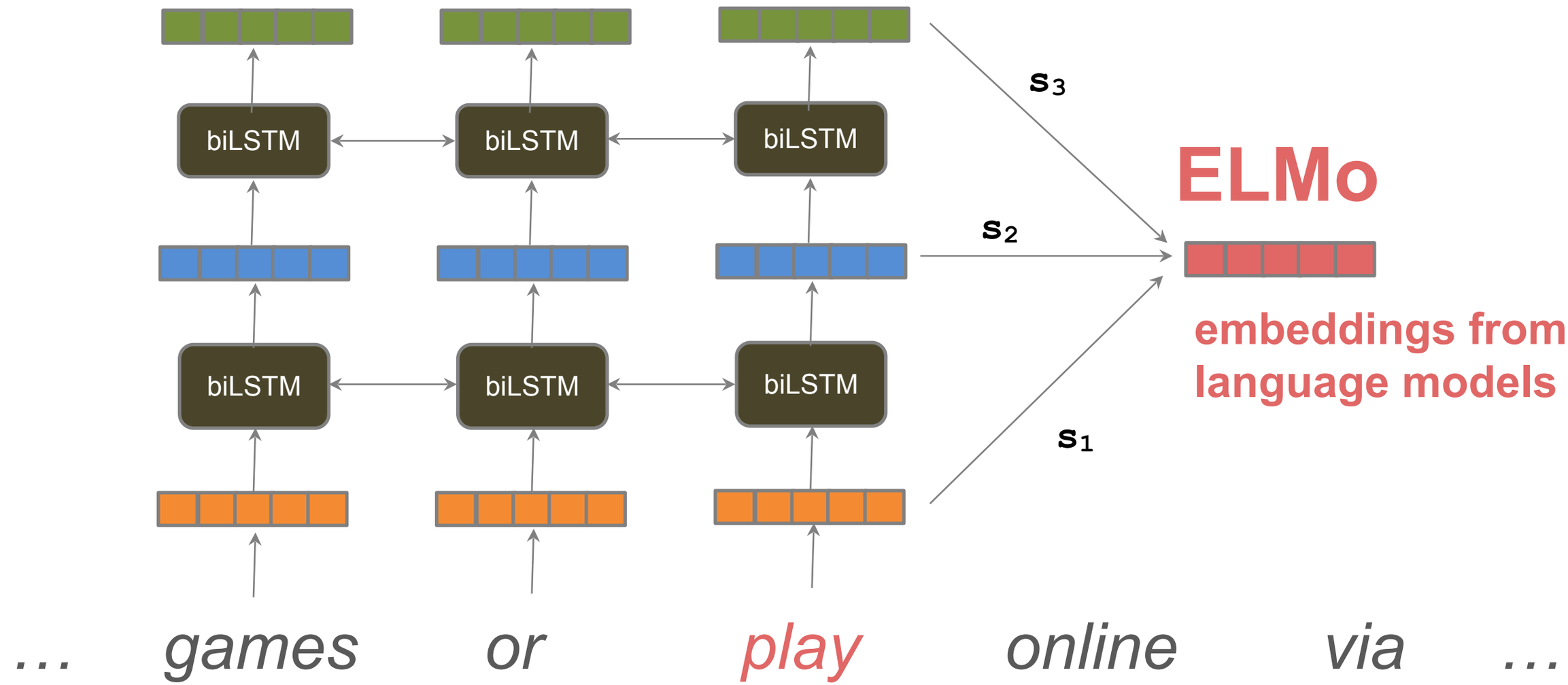
Deep bidirectional language model



Use all layers of language model



Learned task-specific combination of layers

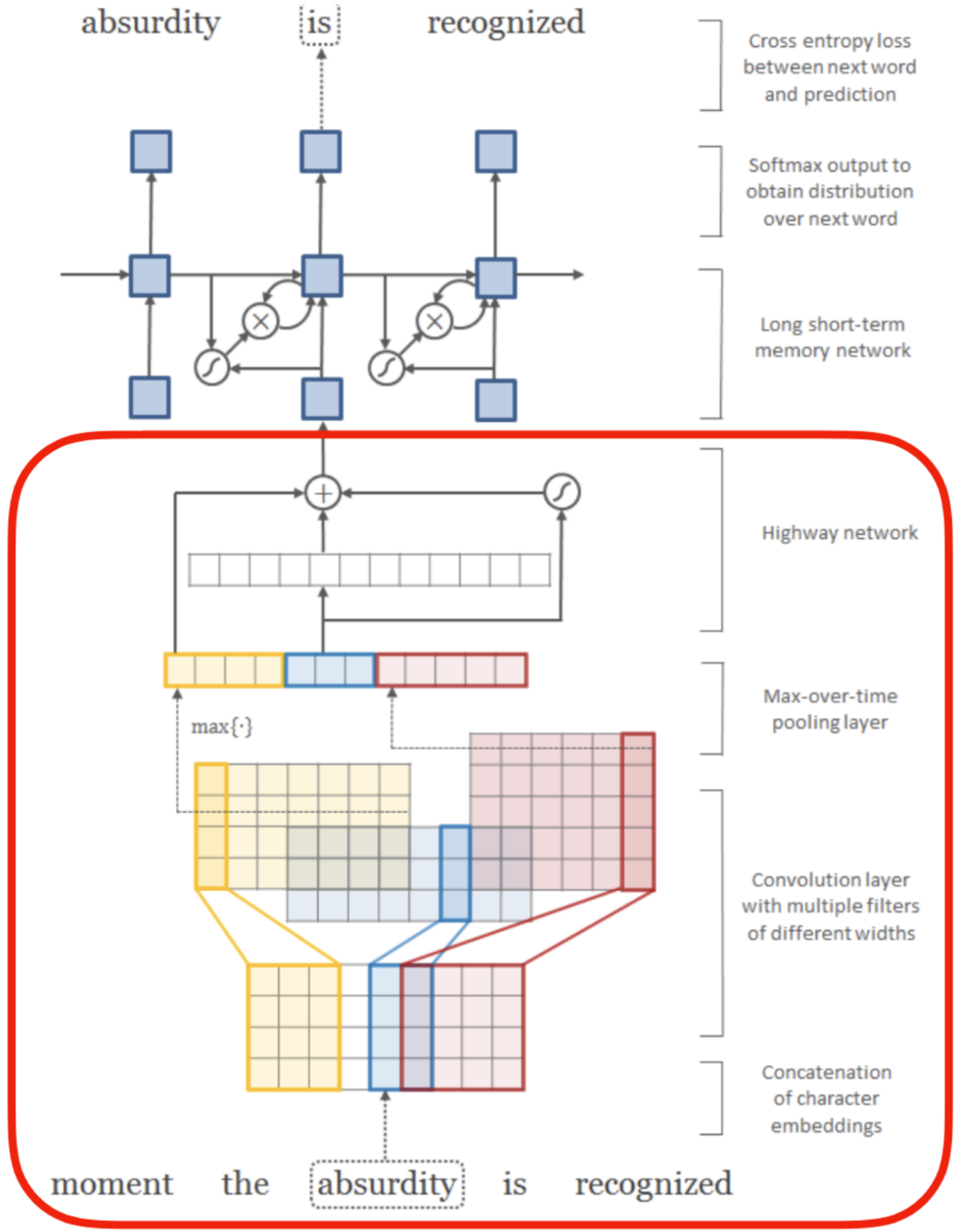


ELMo's token representations

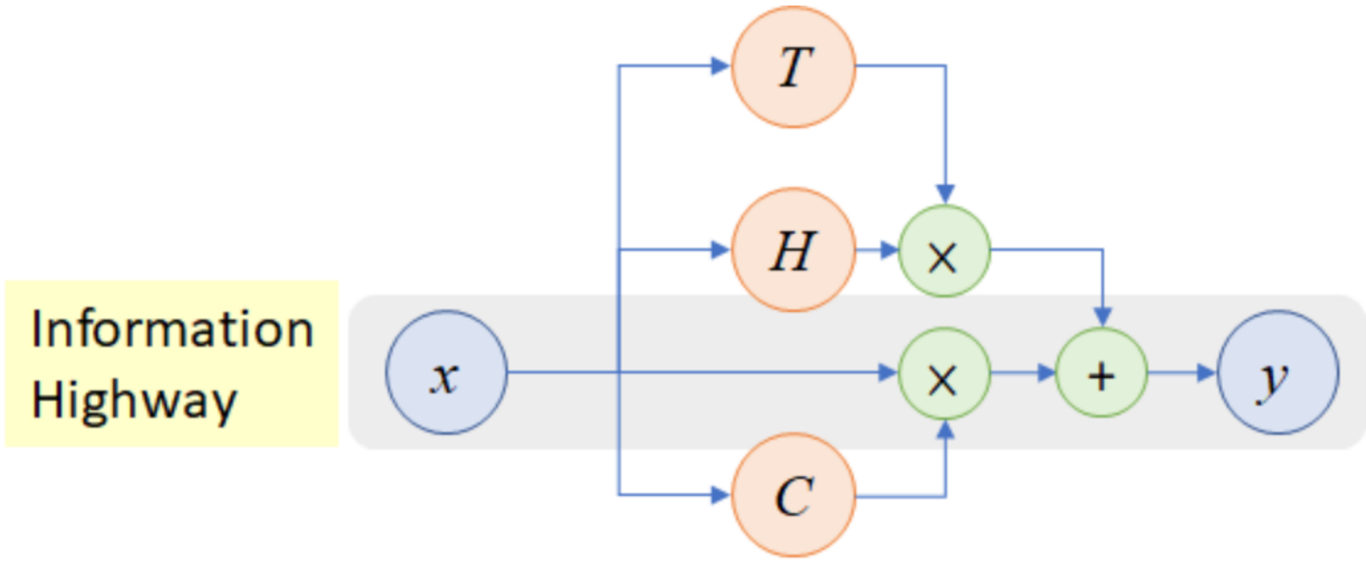
The input token representations are purely character-based: a character CNN, followed by linear projection to reduce dimensionality

"2048 character n-gram convolutional filters with two highway layers, followed by a linear projection to 512 dimensions"

Advantage over using fixed embeddings:
no UNK tokens, any word can be represented


absurdity
is
recognized
Cross entropy loss between next word and prediction
Softmax output to obtain distribution over next word
Long short-term memory network
Highway network
Max-over-time pooling layer
max{·}
Convolution layer with multiple filters of different widths
Concatenation of character embeddings
moment
the
absurdity
is
recognized

1.2. Highway Network



Highway Circuit

- In highway network, two non-linear transforms *T* and *C* are introduced:

$$\mathbf{y} = H(\mathbf{x}, \mathbf{W_H}) \cdot T(\mathbf{x}, \mathbf{W_T}) + \mathbf{x} \cdot C(\mathbf{x}, \mathbf{W_C}).$$

- where ***T* is the Transform Gate** and **C is the Carry Gate.**
- In particular, ***C* = 1 - *T***:

$$\mathbf{y} = H(\mathbf{x}, \mathbf{W_H}) \cdot T(\mathbf{x}, \mathbf{W_T}) + \mathbf{x} \cdot (1 - T(\mathbf{x}, \mathbf{W_T})).$$

The biLM produces $2L + 1$ intermediate representations:

$$\begin{aligned} R_k =& \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} \\ =& \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\} \end{aligned}$$

where $\mathbf{h}_{k,0}^{LM} = \mathbf{x}_k^{LM}$ 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.

ELMo: A task specific combination of these features:

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}.$$

where $\mathbf{s}^{task}$ are softmax-normalized weights and γ^{task} is a scaling parameter.

The biLM produces $2L + 1$ intermediate representations:

$$\begin{aligned} R_k =& \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} \\ =& \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\} \end{aligned}$$

where $\mathbf{h}_{k,0}^{LM} = \mathbf{x}_k^{LM}$ 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.

ELMo: A task specific combination of these features:

layer weights

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM} .$$

where $\mathbf{s}^{task}$ are softmax-normalized weights and γ^{task} is a scaling parameter.

Contextual representations

ELMo representations are **contextual** – they depend on the entire sentence in which a word is used.

how many different embeddings does ELMo compute for a given word?

ELMo improves NLP 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%