

Natural Language Processing

Contextualized word embeddings

Marco Kuhlmann

Department of Computer and Information Science

Contextualized embeddings

- In standard word embeddings, each word is assigned a single word vector, independently of its context.
- Such a model cannot account for **polysemy**, the phenomenon that one and the same word may have multiple meanings.

The children *play* in the park. The *play* premiered yesterday.

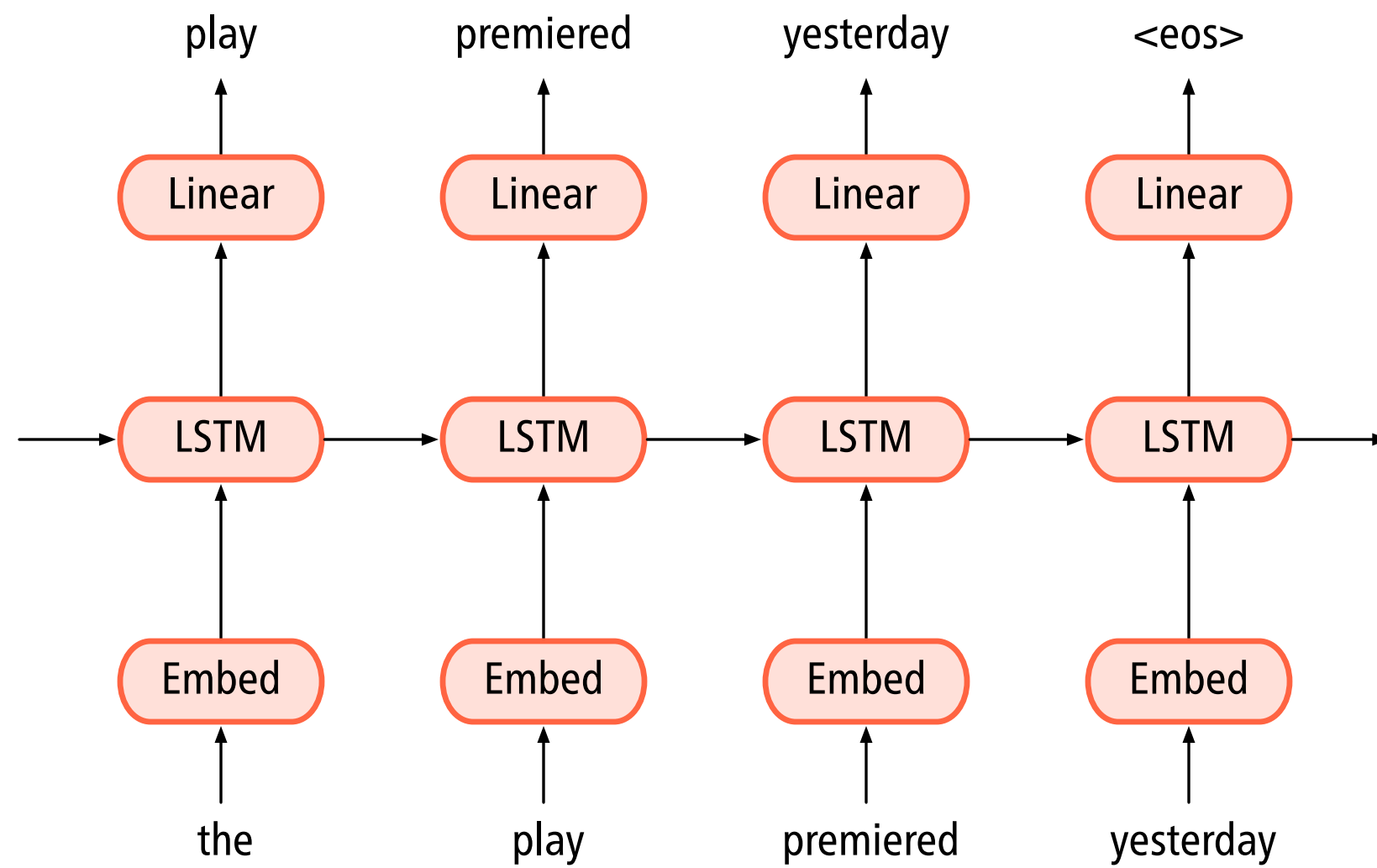
- In **contextualized embeddings**, each token is assigned a representation that depends on its context.

ELMo – Embeddings from Language Models

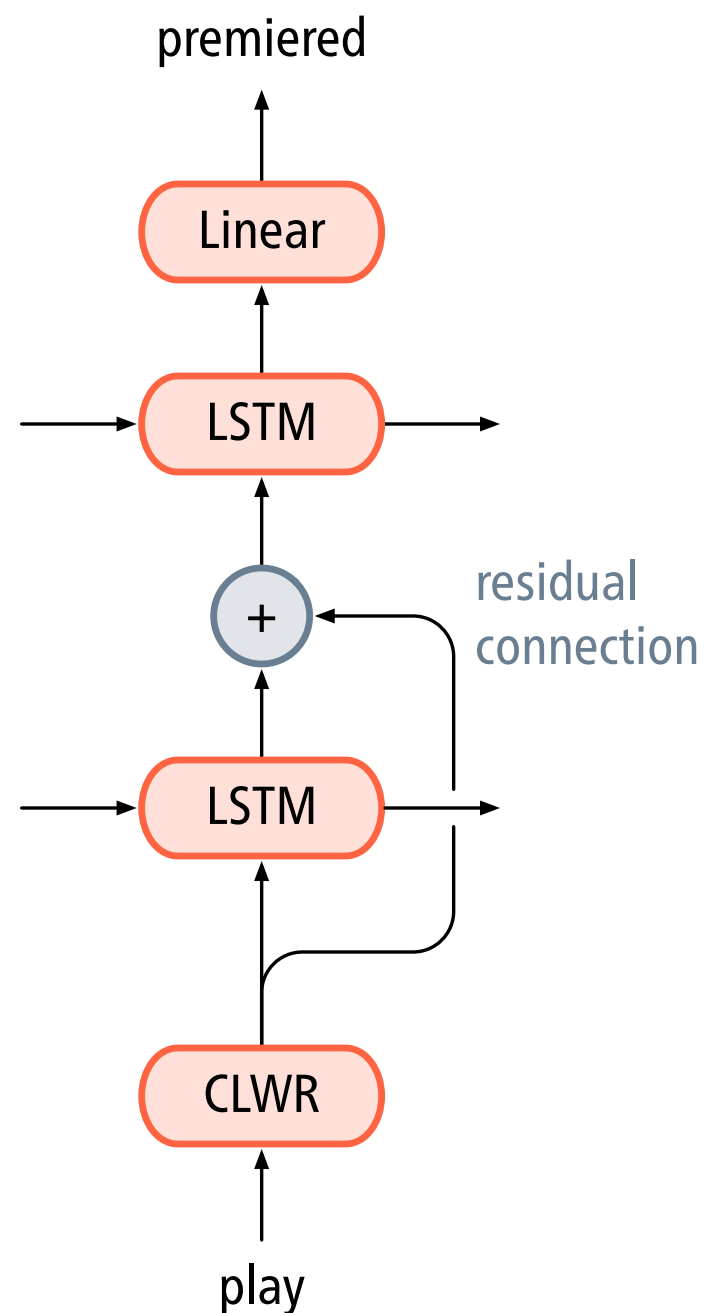
- A token is represented as a task-specific, weighted sum of representations derived from a bidirectional language model.
weights are learned for a specific task
- The basic ELMo model is frozen after pre-training and can complement or replace a standard word embedding layer.
- However, it is often beneficial to fine-tune a pre-trained ELMo model on task-specific data.



LSTM language model

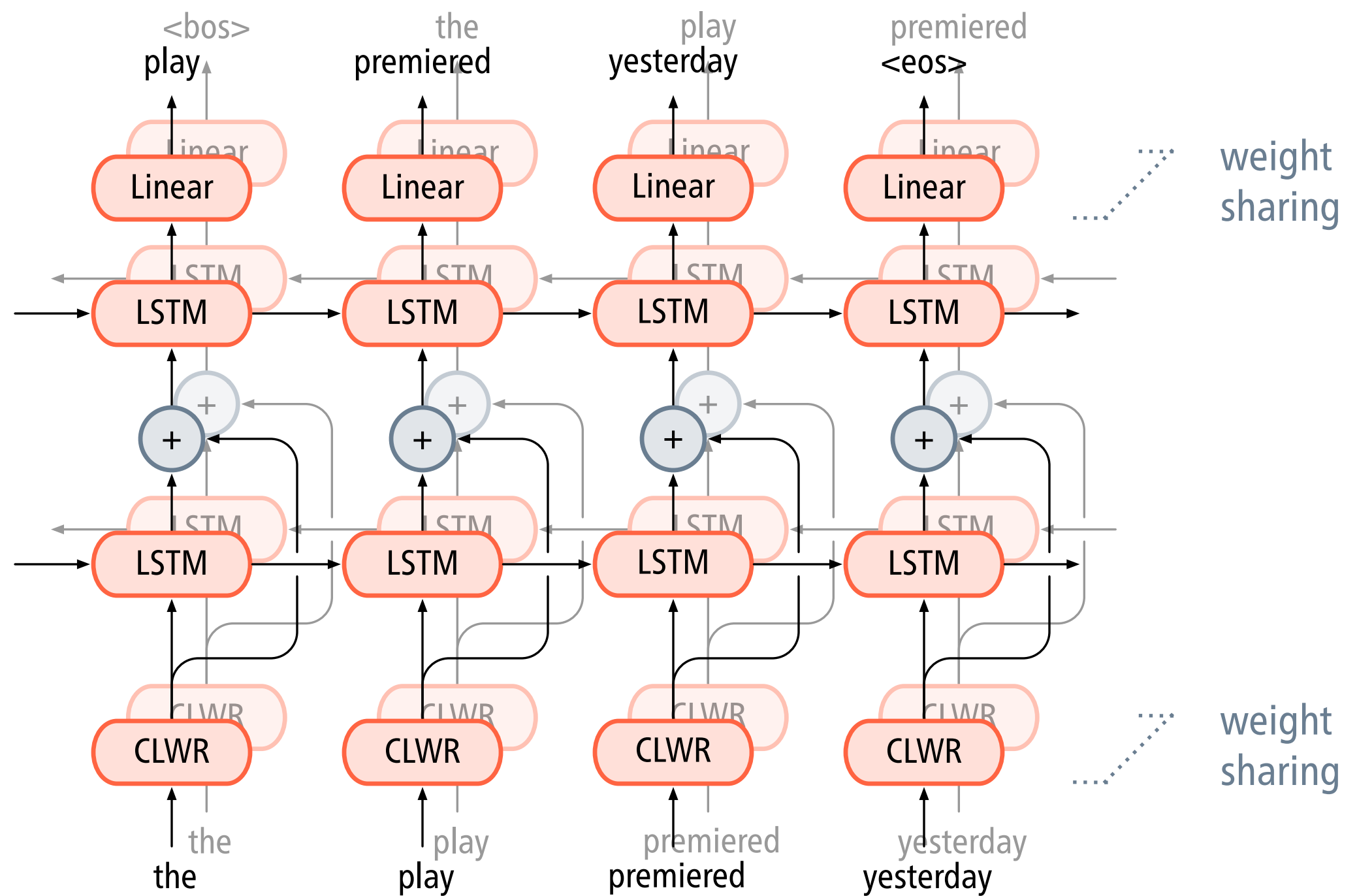


ELMo architecture



- two bidirectional LSTM layers with a residual connection between the layers
- context-insensitive word representation using character convolutions
- final softmax layer computes a probability distribution over the next tokens

Bidirectional language model

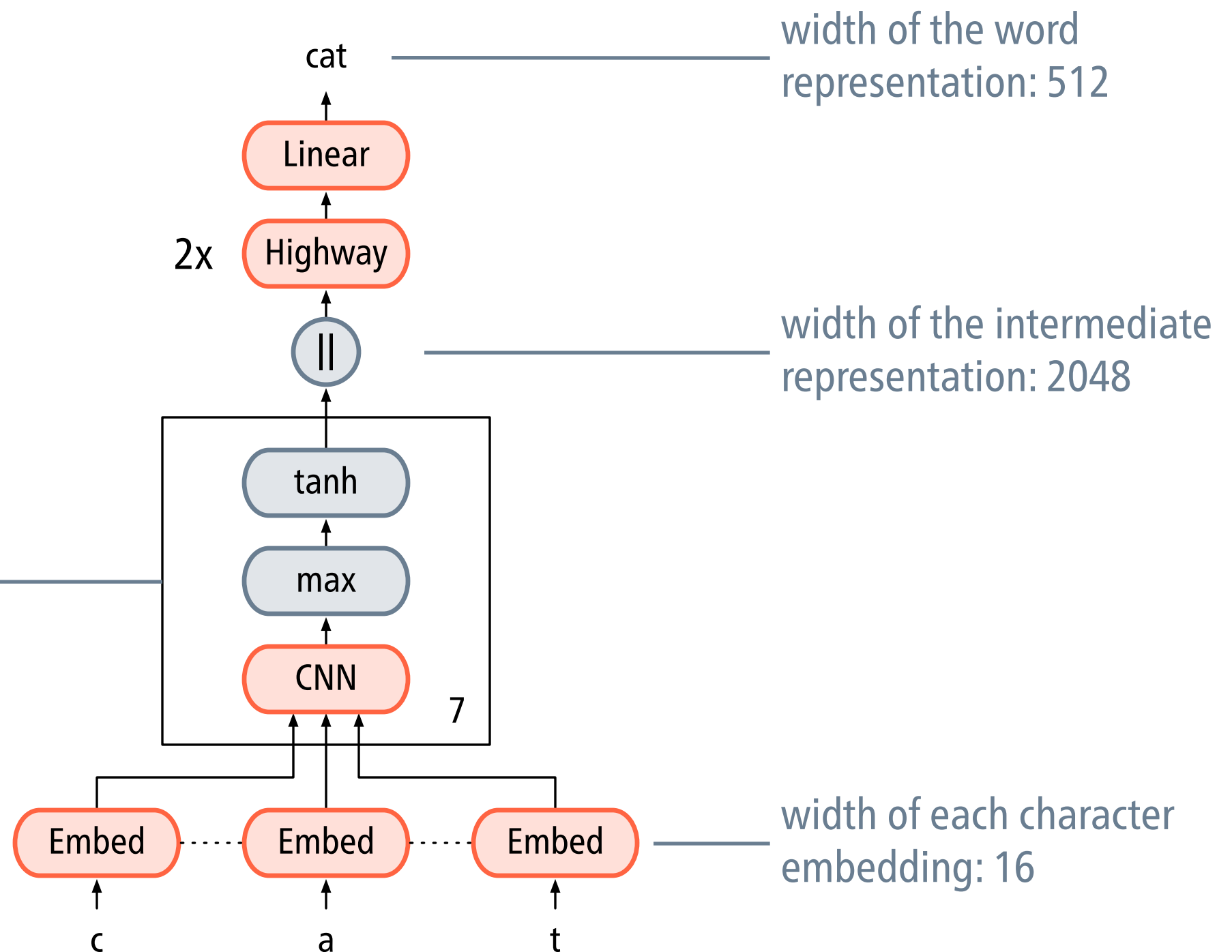


Word representations in ELMo

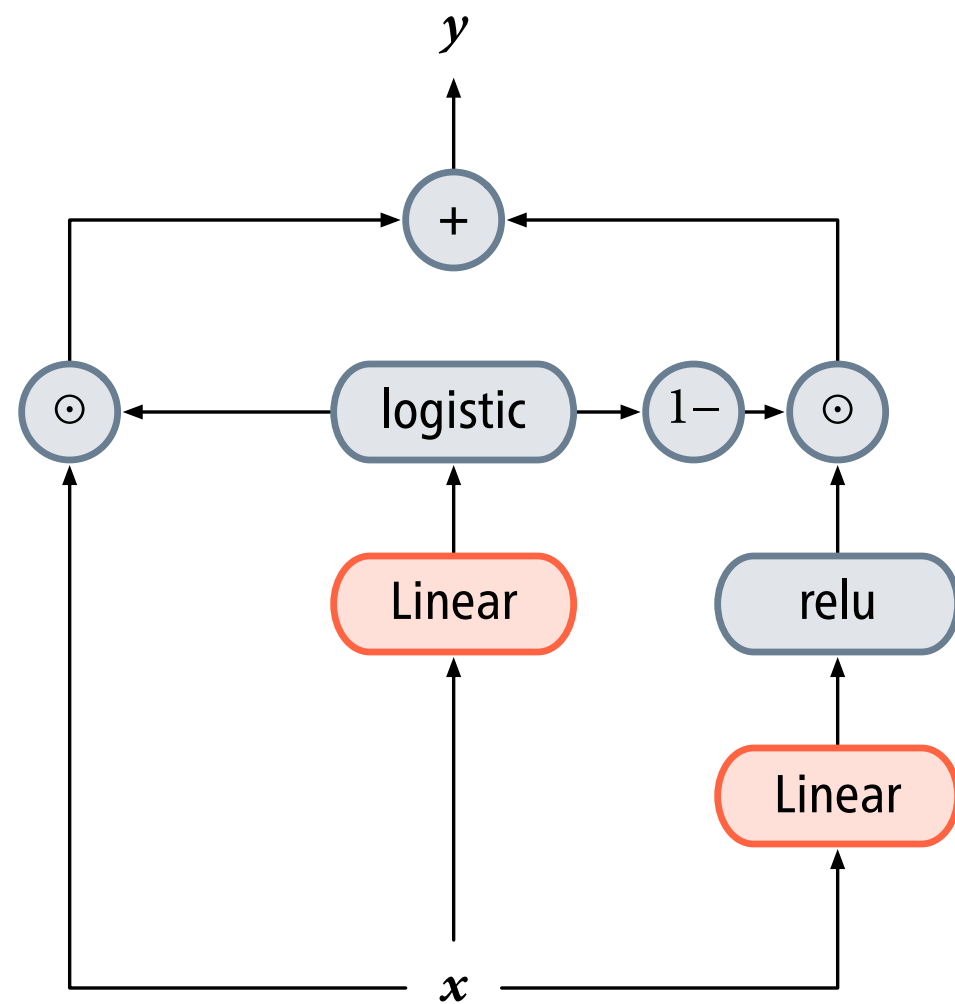
Filter specification

Width	Channels
1	32
2	32
3	64
4	128
5	256
6	512
7	1024

[Peters et al. \(2018\)](#)



Highway layers



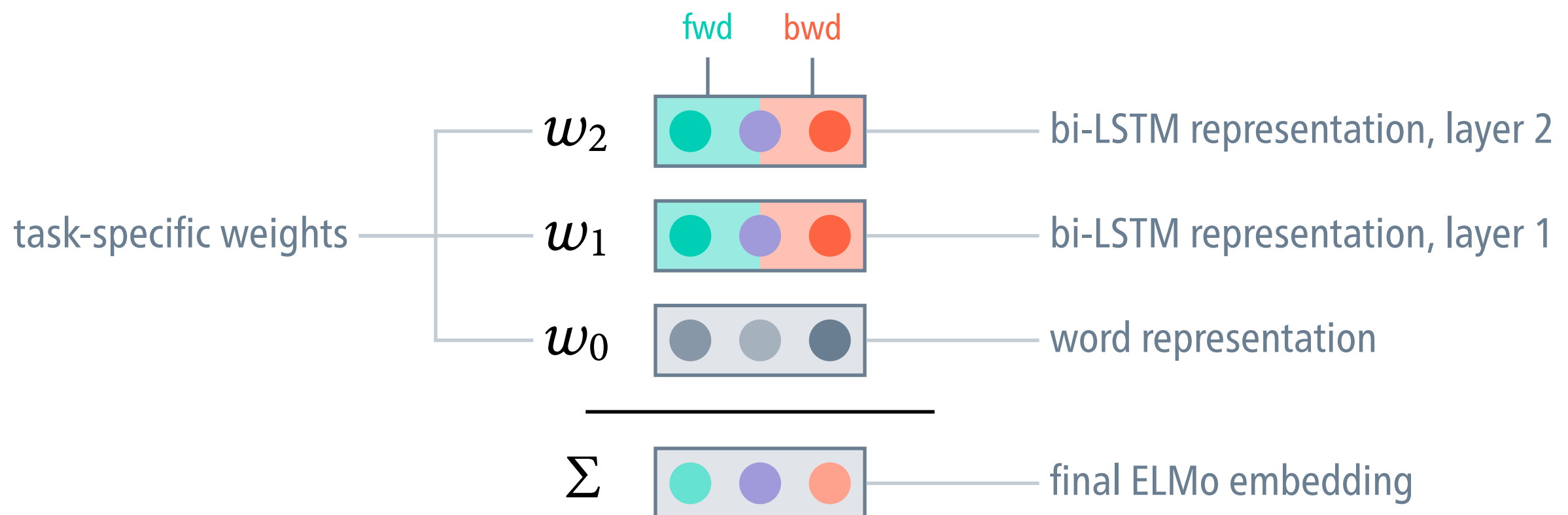
A **highway layer** computes a gated combination of a linear and a non-linear transformation of its input:

$$y = g \odot x + (1 - g) \odot f(xA)$$

where f is an element-wise non-linearity (such as ReLU) and $g = \sigma(xB)$ is an element-wise gate.

ELMo – Embeddings from Language Models

ELMo is a task-specific weighted sum of the intermediate representations in the bidirectional language model.



Relative improvements by using ELMo embeddings

Task	Baseline	+ ELMo	Relative increase
Question answering (SQuAD)	81.1	85.8	24.9%
Coreference resolution (Coref)	67.2	70.4	9.8%
Sentiment analysis (SST-5)	51.4	54.7	6.8%
Textual entailment (SNLI)	88.0	88.7	5.8%