# Deep Contextualized **Word Representations**

Giovanni Luca Favuzzi
TU-B @ MAR 4.033
03/07/2019

# Intro

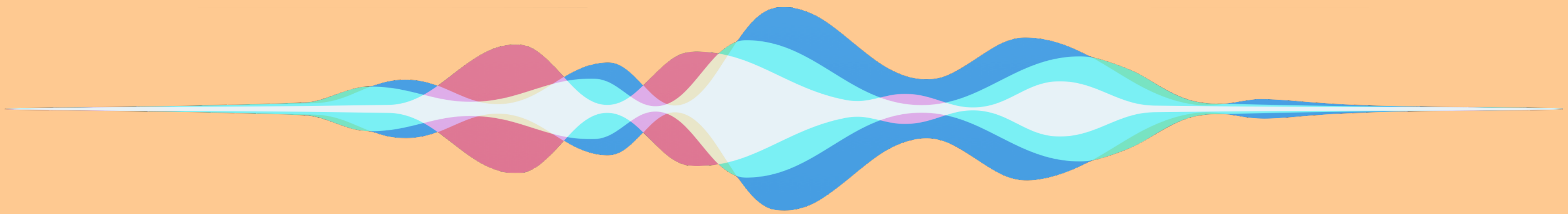
NLP

# Definition

# Word Embedding

Language models that map **words** onto **vectors**

# Word Embedding

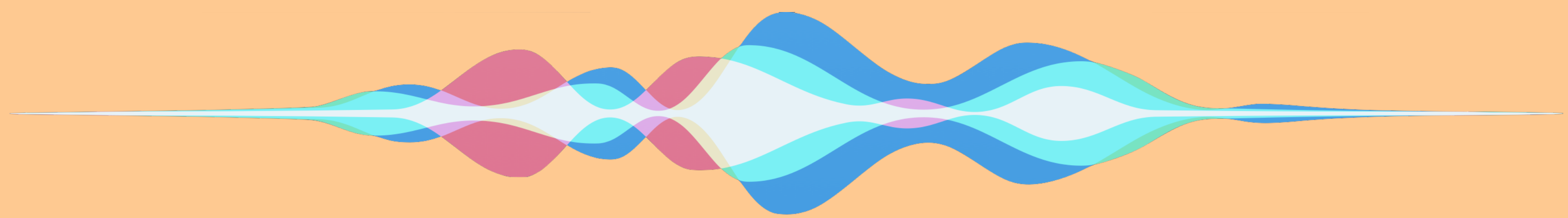


ciao

Word
**Embedding**

$$\mathbf{x} \in \mathbf{R}^d$$

# Word Embedding



ciao

Word
**Embedding**

**syntactic**
**semantic**

**polysemy**

$$\mathbf{x} \in \mathbf{R}^d$$

# ELMo

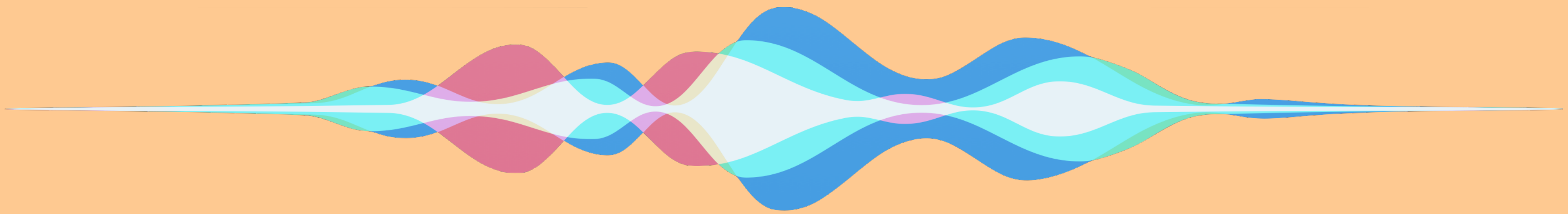
**E**mbeddings

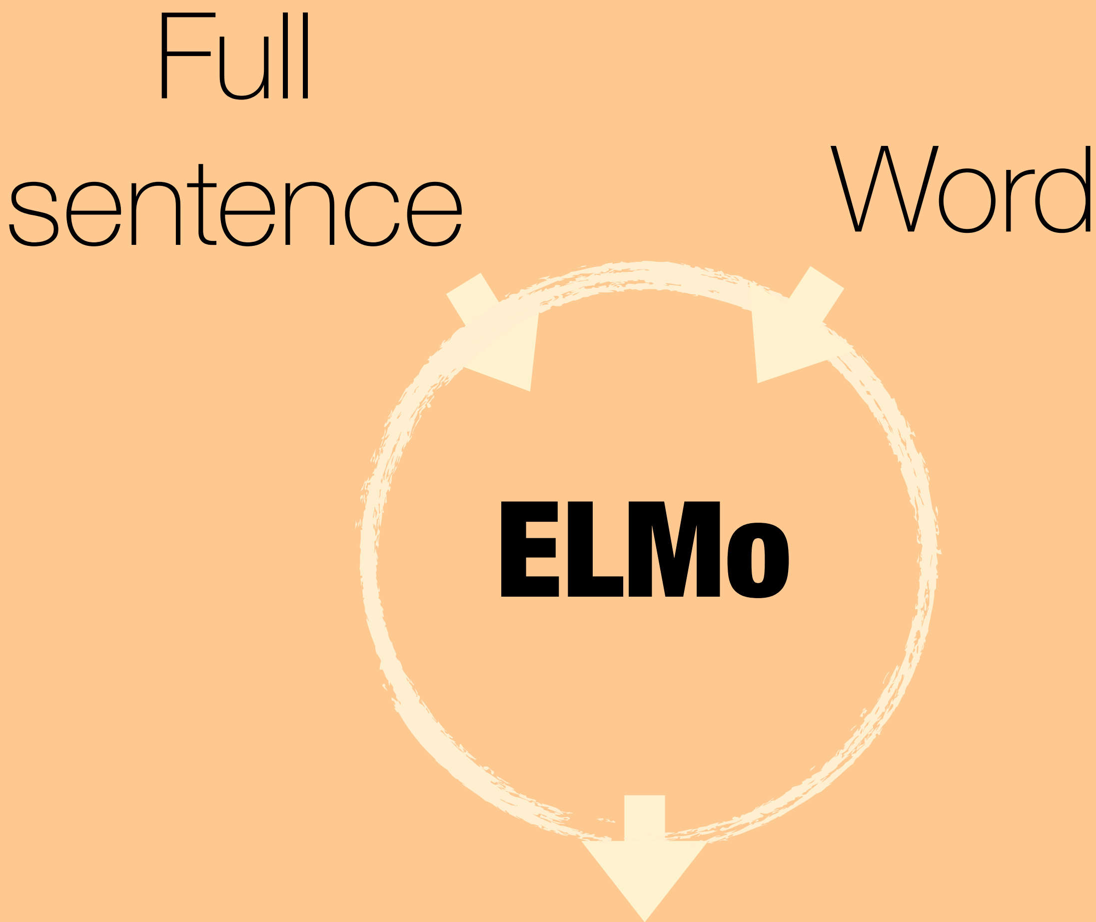
**for**

**L**anguage

**Mo**dels

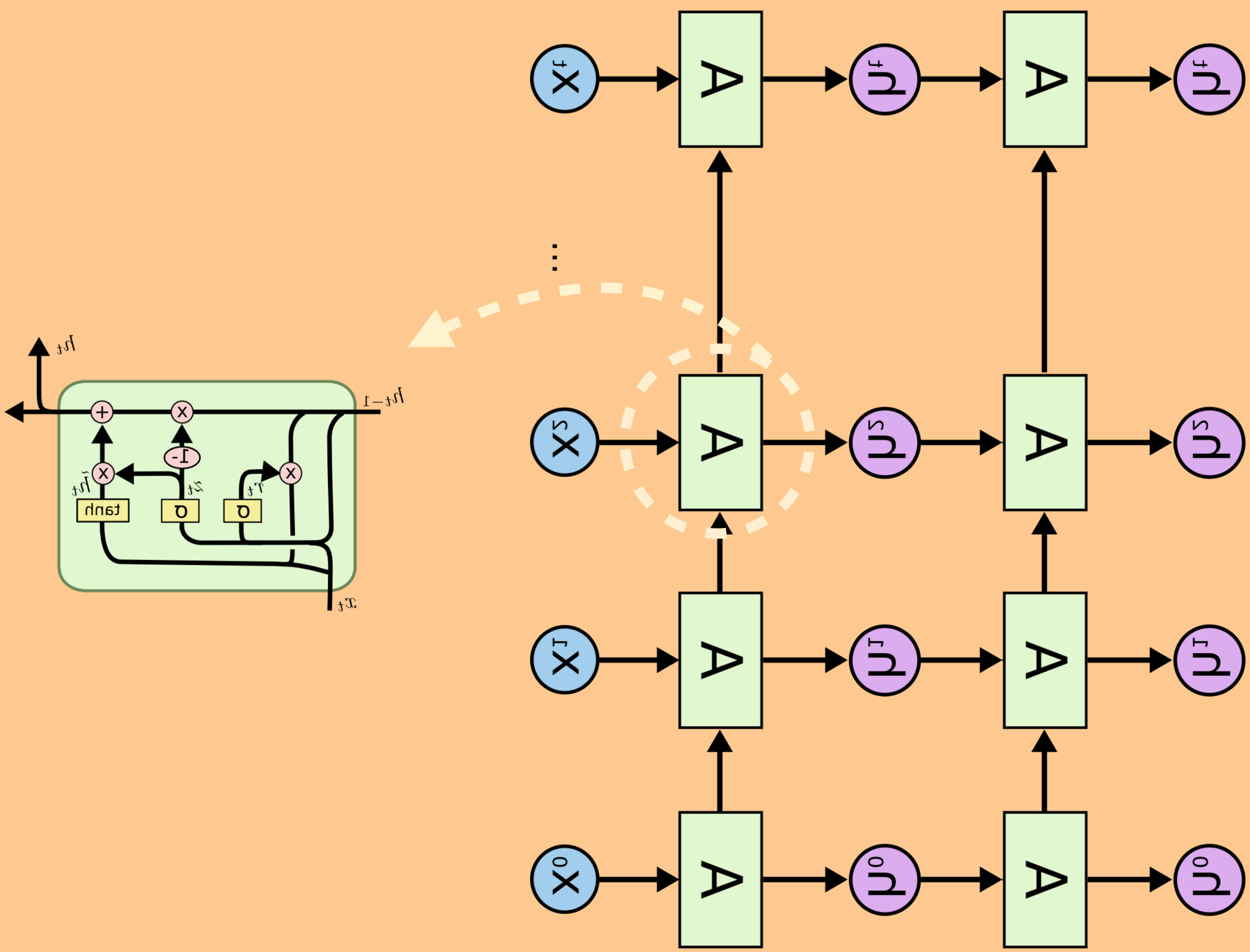
# ELMo






$$\mathbf{x} \in \mathbf{R}^d$$

# ELMo

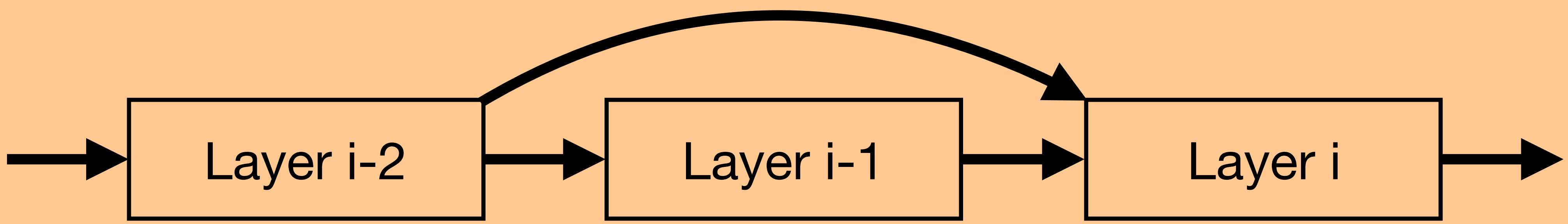


# Definition

# Residual Connection

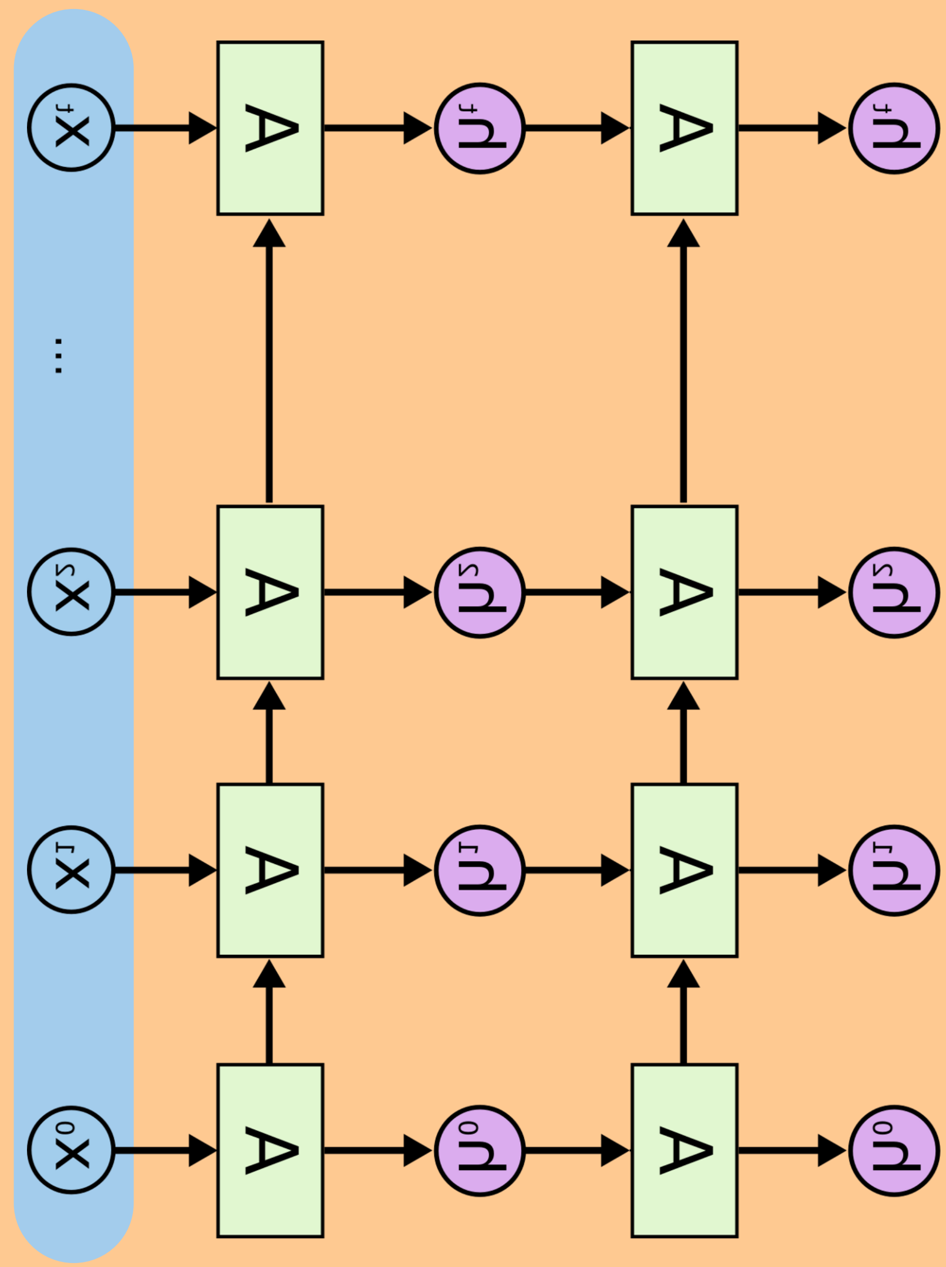
https://en.wikipedia.org/wiki/Residual_neural_network

Residual connections are **short-cuts** that jump over some **layers**.



# ELMo

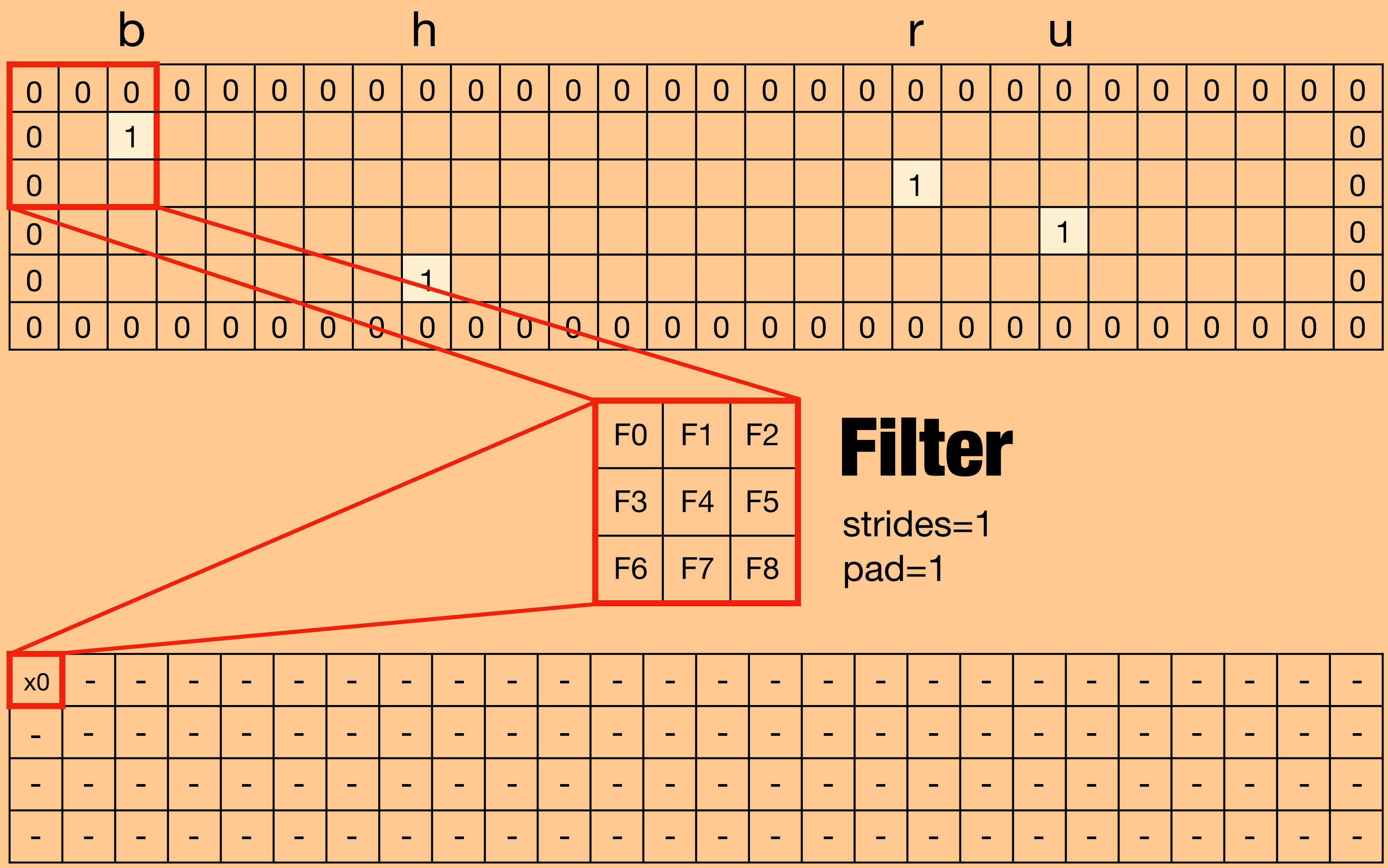
Character
Convolution
Embeddings



# Character Convolutions

$\mathbf{t}_i =$ bruh

**1-Hot**

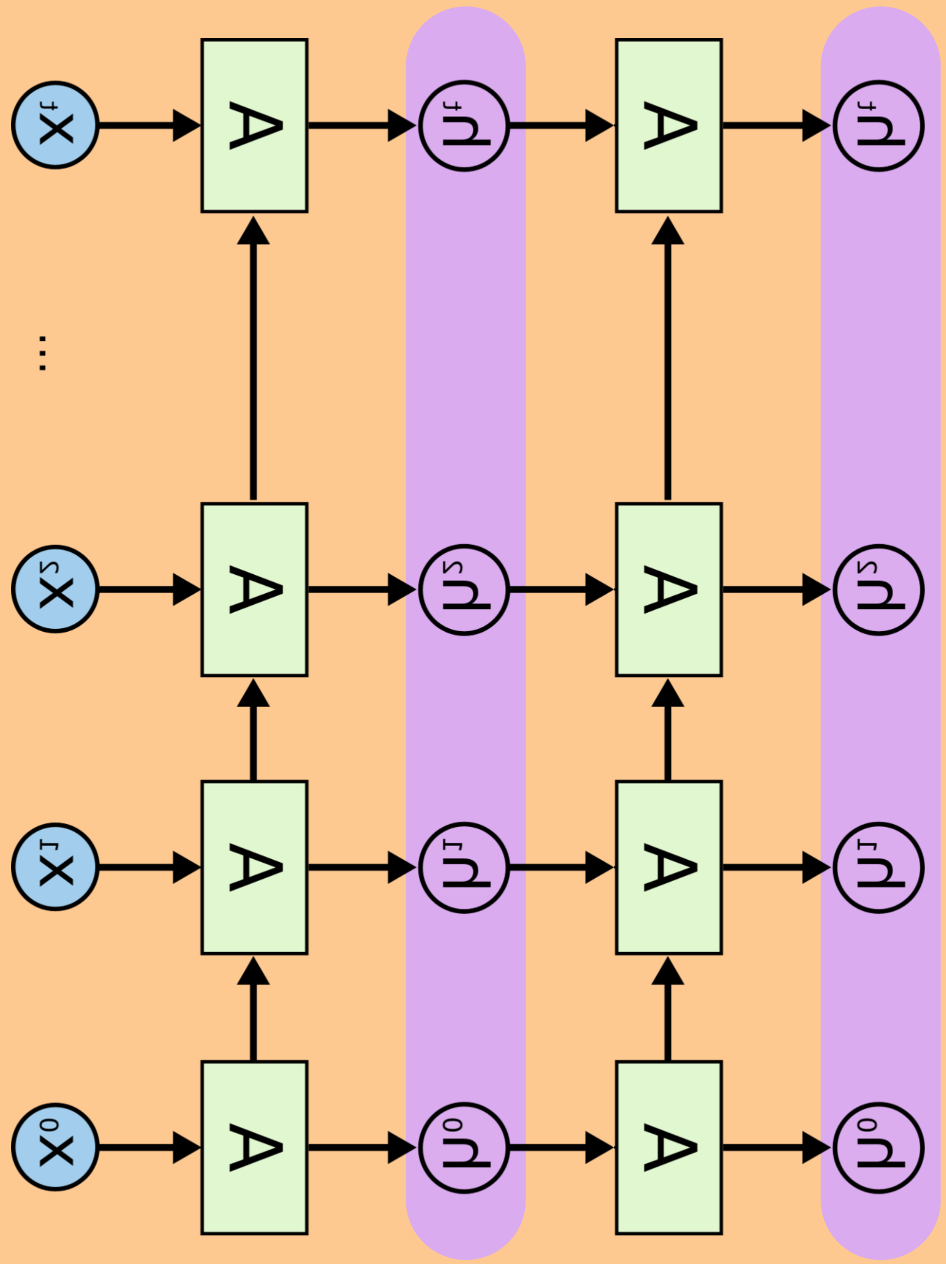



**Filter**

strides=1
pad=1

$$c_F(\mathbf{t}_i) = \mathbf{x}_i$$

# ELMo



Intra-Layer Representations

# Definition

# Language Model

https://en.wikipedia.org/wiki/Language_model

A statistical language model is a **probability** distribution **over** sequences of **words**.

# Bi-Language Model

This is a full sentence

$t_1$ $t_2$ ... $t_{N=5}$

[1] $P(t_1 \ldots t_N) = \prod_{i}^{N} P(t_i \mid t_1 \ldots t_{i-1})$ **Forward**

[2] $P(t_1 \ldots t_N) = \prod_{i}^{N} P(t_i \mid t_{i+1} \ldots t_N)$ **Backward**

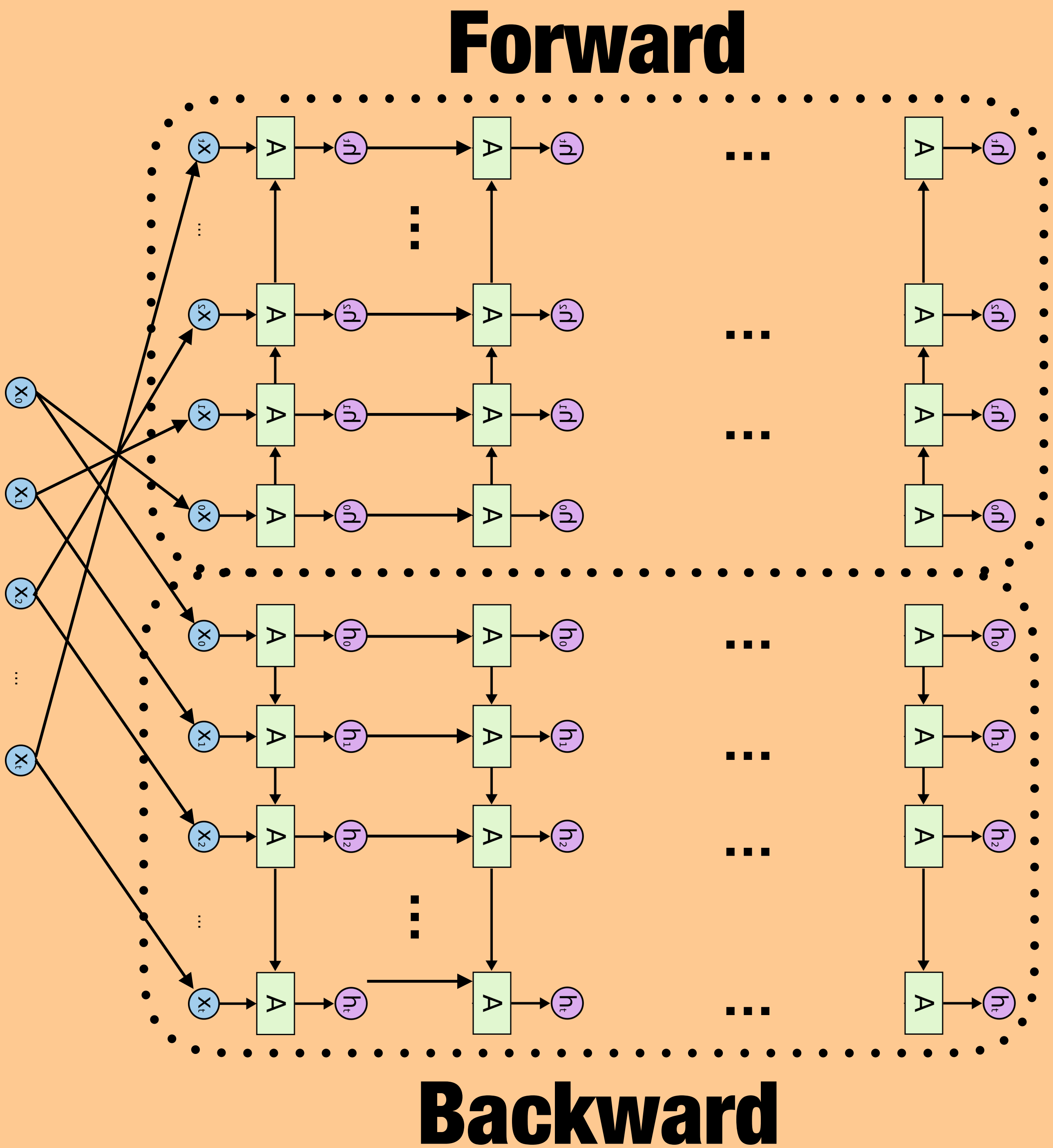
$\log\left\{[1] \cdot [2]\right\}$

## Logarithmic Likelihood

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

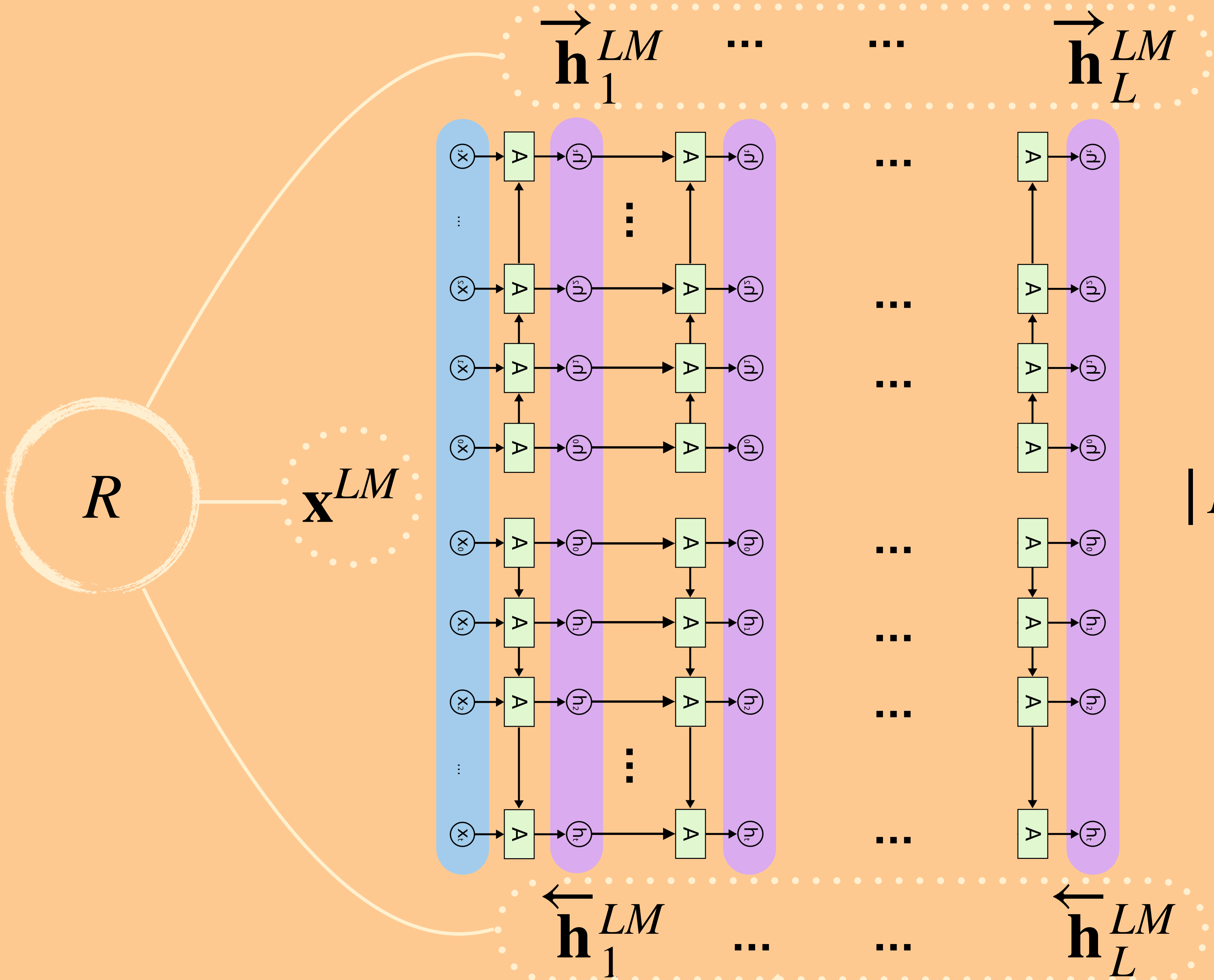
# Neural Architecture

Input




Output

# Embeddings




$$|R| = 2 \cdot L + 1$$

# Embedding Interpolation

Task-Specific
Scalar

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

Softmax
Normalization
Coefficients

# Task Specific Integration

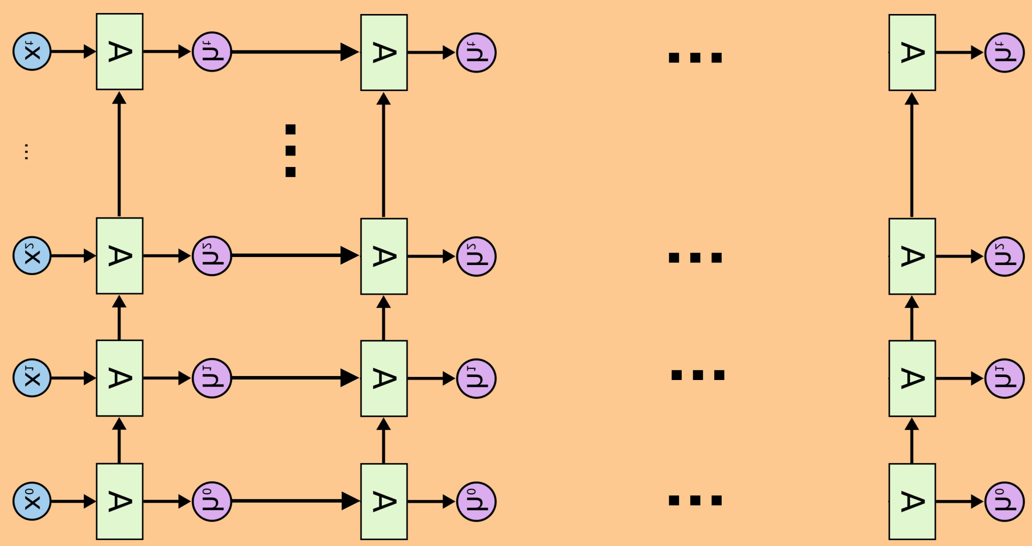
**Framework**

1. Run the pre-trained biLM
2. Record all layer representations
3. Learn the optimal interpolation

# Optimal Interpolation

## Original Model

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}$$



$$\mathbf{O} = \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_N \end{bmatrix}$$

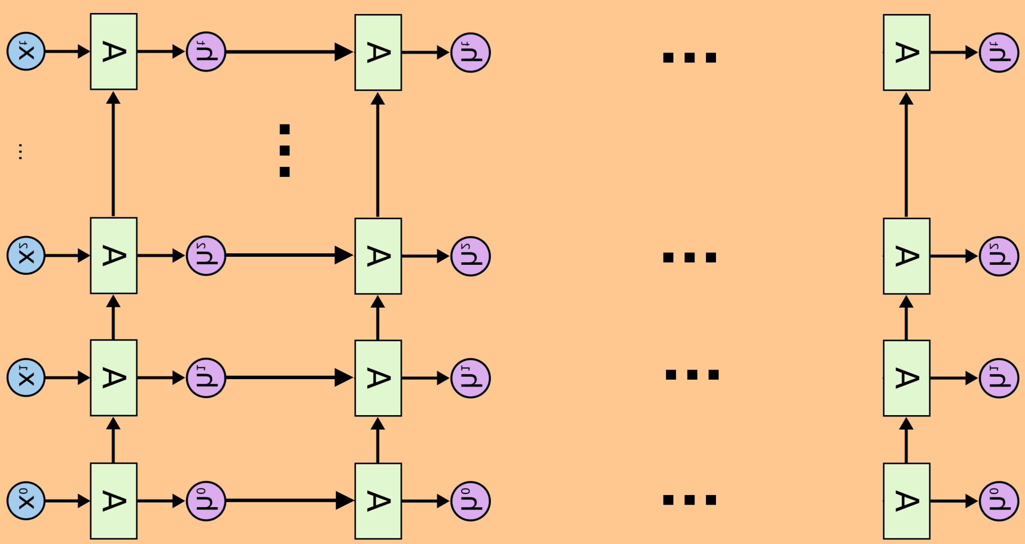
# Optimal Interpolation

## Enhanced Model

$$[\mathbf{x}_1, \mathbf{ELMO}_1^{task}]$$

$$\vdots$$

$$[\mathbf{x}_N, \mathbf{ELMO}_N^{task}]$$



$$[\mathbf{h}_1^L, \mathbf{ELMO}_1^{task}]$$

$$\vdots$$

$$[\mathbf{h}_N^L, \mathbf{ELMO}_N^{task}]$$

# Features

## ELMo

# Layers = 2 → Residual Connections from first to second layer
# Units = 4096
# Dimension Projections = 512
Rich Dropout & L2 Regularization

## Character Convolutions

# n-gram CC Filters = 2048
2 Highway Layers
Linear Projection: 512 features

# Definition

# Projection Layer

https://stackoverflow.com/questions/37889914/what-is-a-projection-layer-in-the-context-of-neural-networks

Projection layers map discrete word **indices** of an n-gram to a **continuous vector** space.

n-gram

projection matrix

projection

$$\begin{array}{rl} w_{j-n+1} & [0\ \ 1\ \ 0] \\ w_{j-n+2} & [1\ \ 0\ \ 0] \\ & \quad\vdots \\ w_{j-1} & [0\ \ 0\ \ 1] \end{array} \odot \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} = \begin{bmatrix} w_{12} \\ w_{22} \\ w_{11} \\ w_{21} \\ w_{13} \\ w_{23} \end{bmatrix}$$

# Evaluation

| 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% |

# Evaluation

| 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% |

# Evaluation

| 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% |

# Evaluation

| 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% |

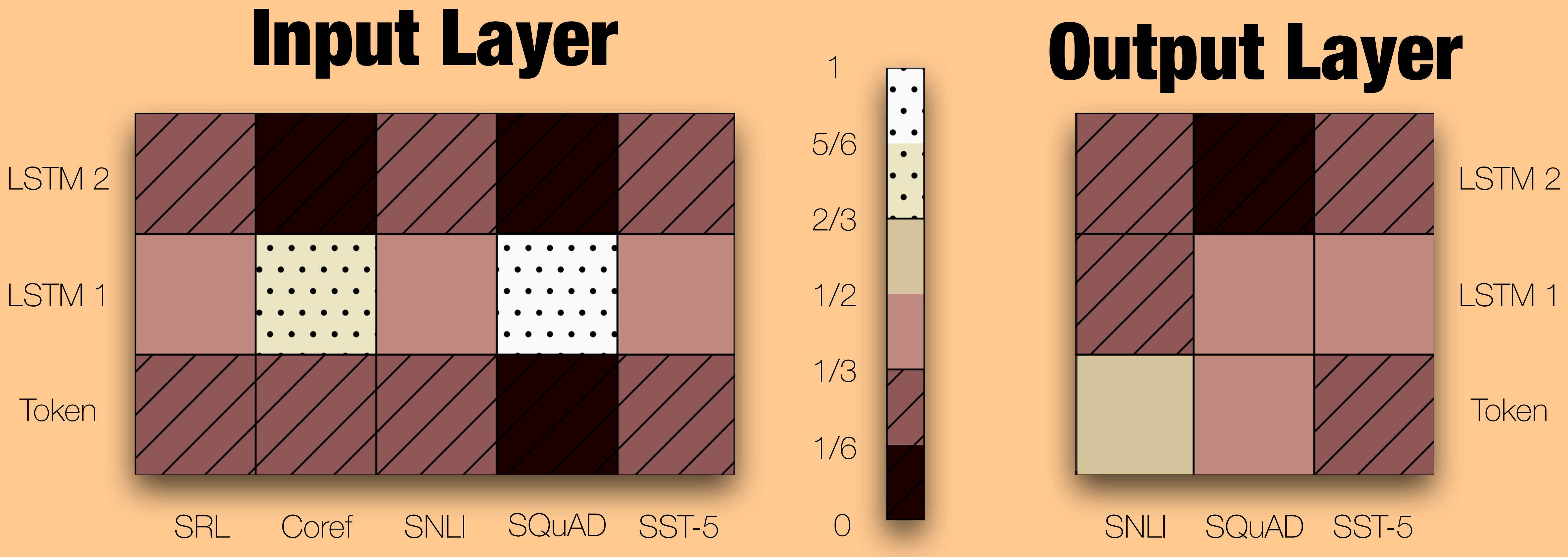
# Evaluation

| 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% |

# Evaluation

| 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% |

# Interpolation Weights



# Outro

Thanks for your attention

**Questions?**