

Lec 11: Text generation with RNN



hsyi@kisti.re.kr

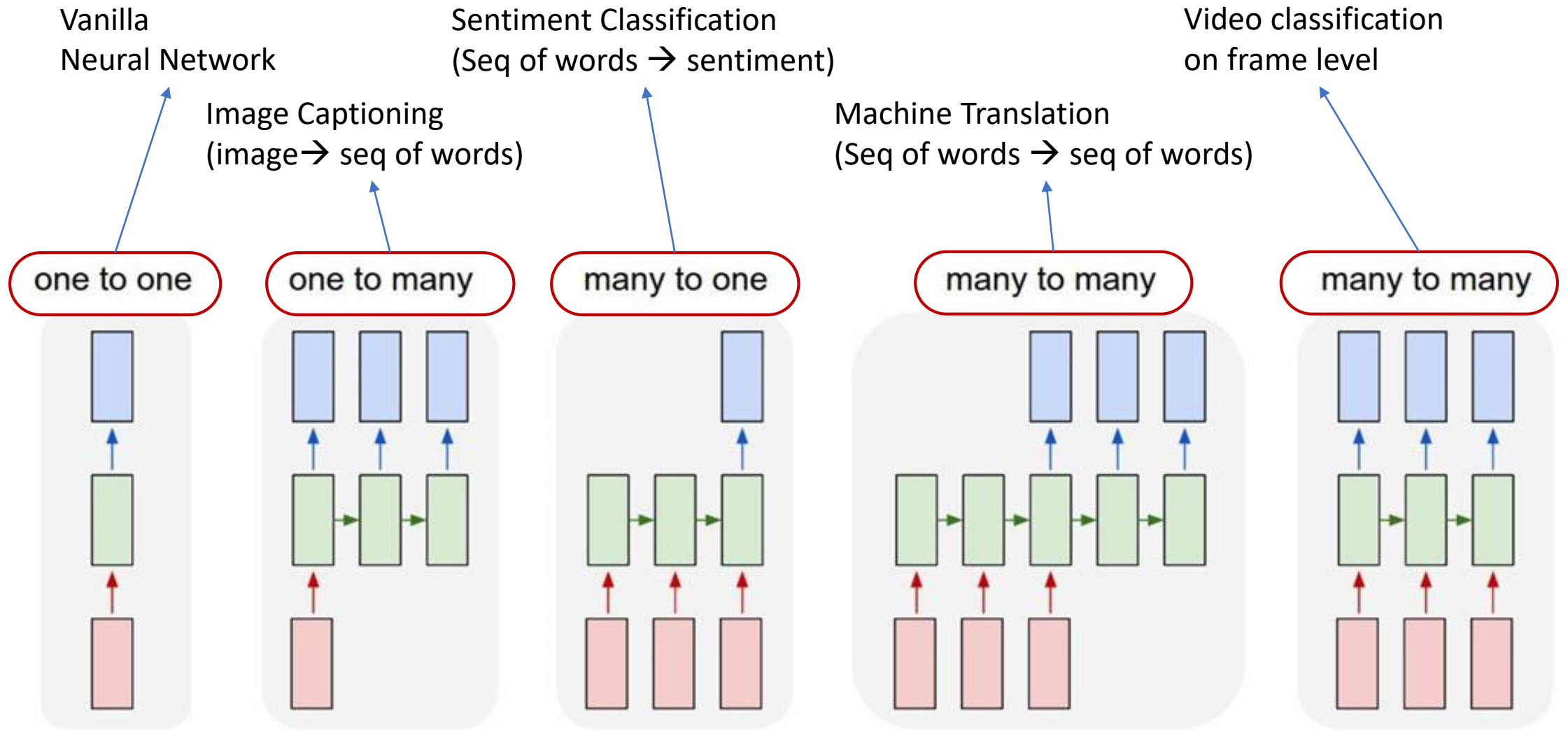
Hongsuk Yi (이홍석)



- ❖ Introduction to Recurrent Neural Network
 - ✓ Simple RNN, BPTT, Memory Cell
 - ✓ Code: Implementing an RNN with Keras
- ❖ Introduction to Long-Short Term Memory
 - ✓ Cell state, LSTM, and GRU, and Applications
 - ✓ A Visual Guide to Recurrent Layers in Keras
 - ✓ Code: A simple LSTM layers
- ❖ **Text generation with RNN**
 - ✓ Tokenizer, Character-Level Language model
 - ✓ Code: Alice's Adventures in Wonderland
- ❖ Sequence to Sequence Learning model with RNN
 - ✓ Introduction to Seq2Seq and Attention model
 - ✓ Code: Character-Level Neural Machine Translation

Review the last class:

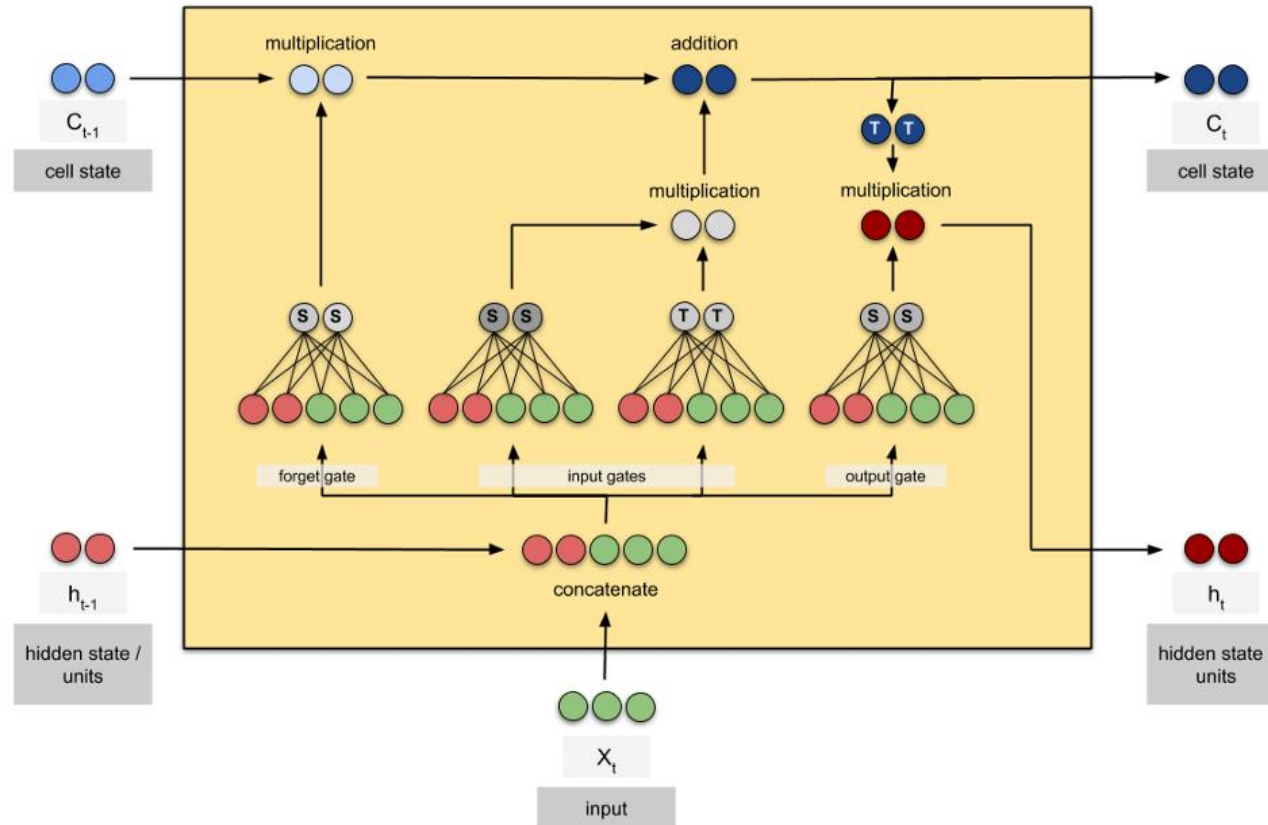
Long-Short Term Memory and RNN



- ❖ RNN의 기억 상태 (메모리, 혹은 Hidden State)를 제어하기 위해 '셀 상태' 를 도입
 - ✓ Cell State를 도입후 4개의 게이트로 RNN의 장기 메모리를 다룰 수 있는 장단기 메모리로 만듦

추가된
Cell State

RNN 메모리

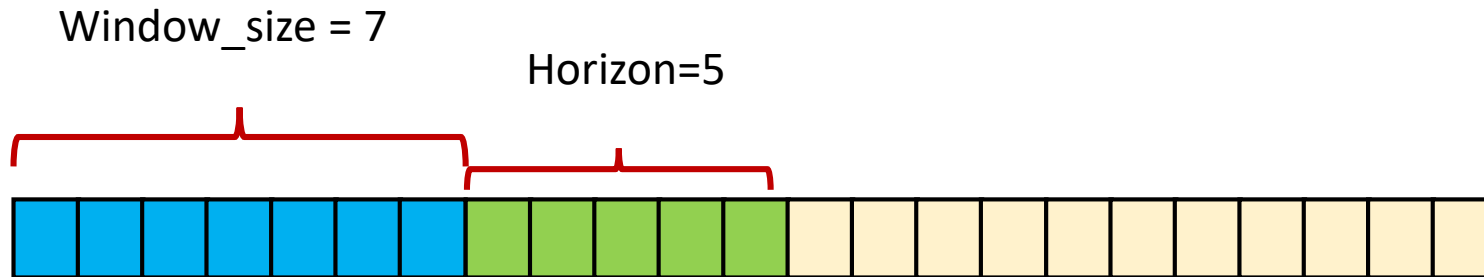


❖ Two terms are really important in the type of forecasting model

- ✓ **Window Size** : The number of timesteps we take to predict into the future
- ✓ **Horizon** : The number of timesteps ahead into the future we predict.

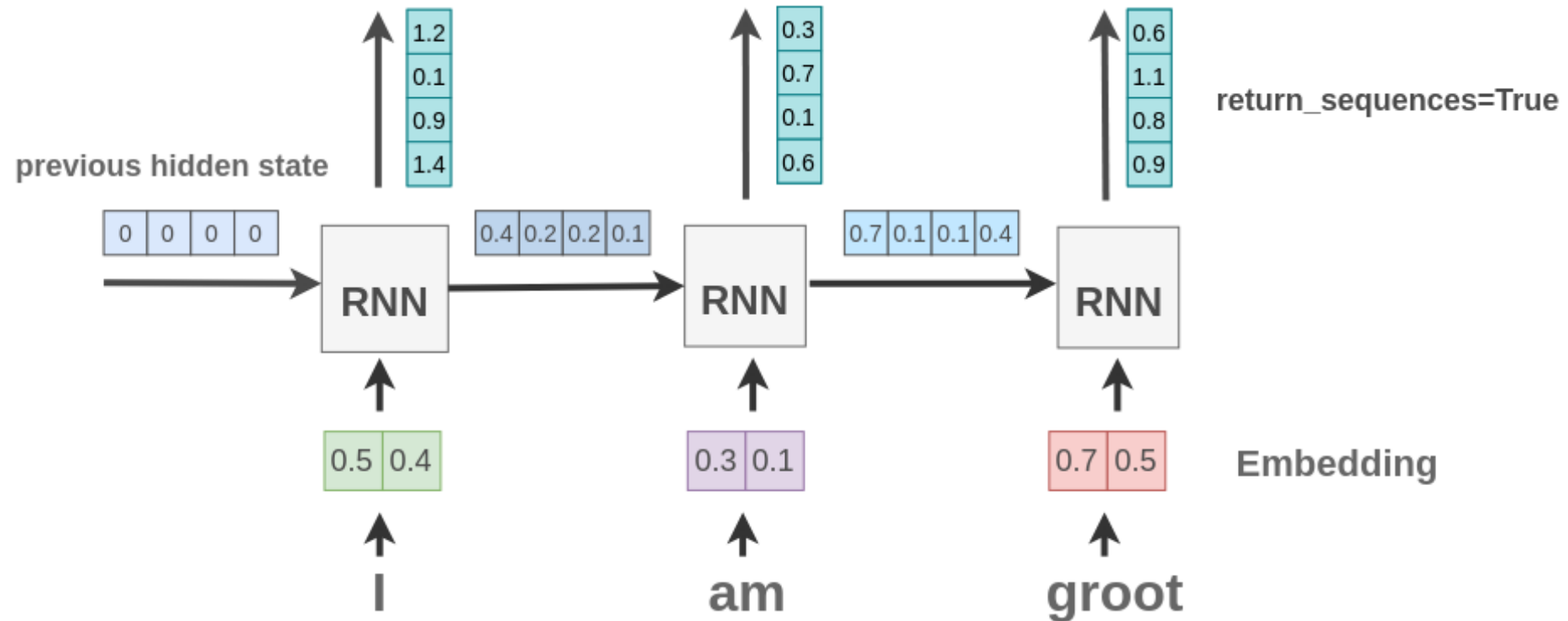
```
model.add(SimpleRNN(5, input_shape=(7, 1), return_sequences=True))
```

→ 자연어 처리에서는 임베딩 차원이 된다.



- ❖ The output from each unfolded RNN cell is returned instead of only the last cell.

```
model.add(SimpleRNN(4, input_shape=(3, 2), return_sequences=True))
```



Text Tockenizer

❖ Token : Language elements that we can't share anymore

- ✓ Word tokenization divides sentences based on spacing as follows.



❖ Tokenizer : work to input text data into the neural network.

- ✓ The preprocessing process that converts it into an appropriate form through encoding

❖ we can use two more techniques

- ✓ one-hot encoding
- ✓ we can use unique numbers to represent words in a vocabulary.
 - In the case of text data, an embedding layer is basically used.

❖ A simple example of one-hot encoding






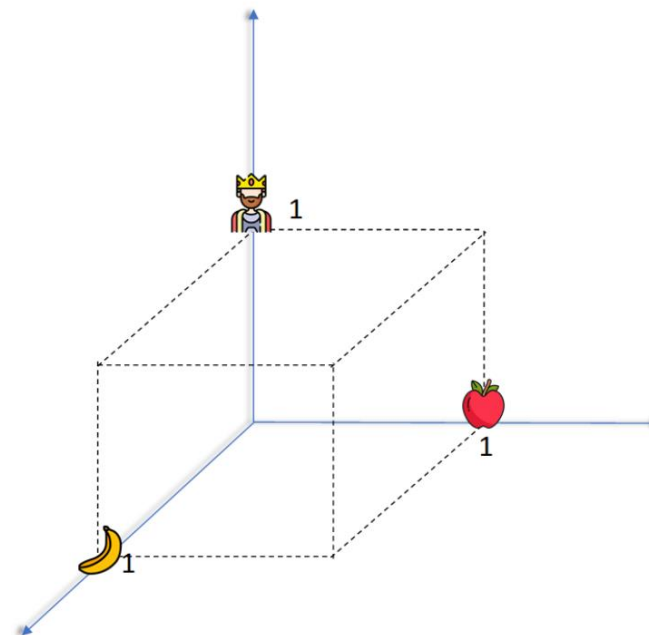
Index: 0 1 2 3 ... 99998 99999



Index: 0 1 2 3 ... 99998 99999

- ❖ Word embeddings are basically a form of word representation that bridges the human understanding of language to that of a machine.

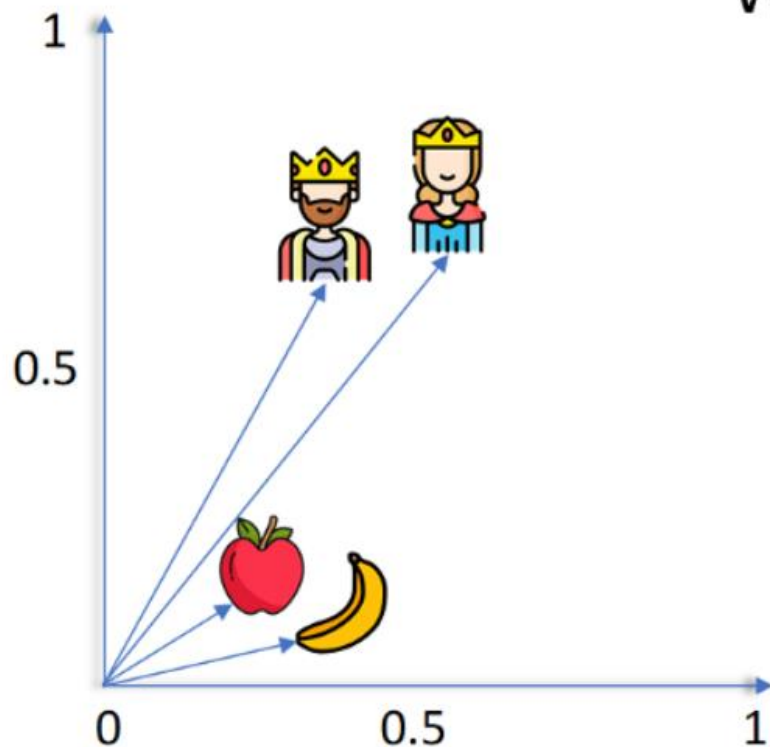
	1	0	0
Index:	0	1	2
	0	1	0
Index:	0	1	2
	0	0	1
Index:	0	1	2



❖ Representations of text in an n-dimensional space

- ✓ where two similar words are represented by almost similar vectors that are very closely placed in a vector space.

Word embeddings



0.25

0.16



0.33

0.10



0.29

0.68



0.51

0.71

❖ **vocab_size = 30**

✓ Usually the vocabulary size is thousands

❖ **seq_length = 5**

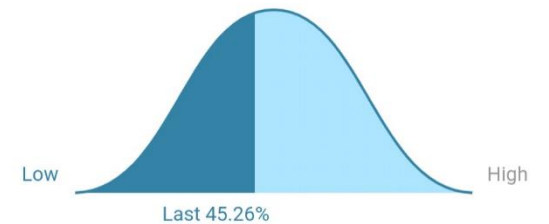
✓ Usually, sentences consist of more than five words

"This is a small vector"

[7 0 6 1 28]

Your Korean Vocabulary Size is:

3173



Last 45.26%
Your vocabulary size is like that of a 8-year-old child in Korea.

❖ vocab_size = 30, seq_length = 5

"This is a small vector"

[7 0 6 1 28]

Word: *This*

Position in some
dictionary: 7

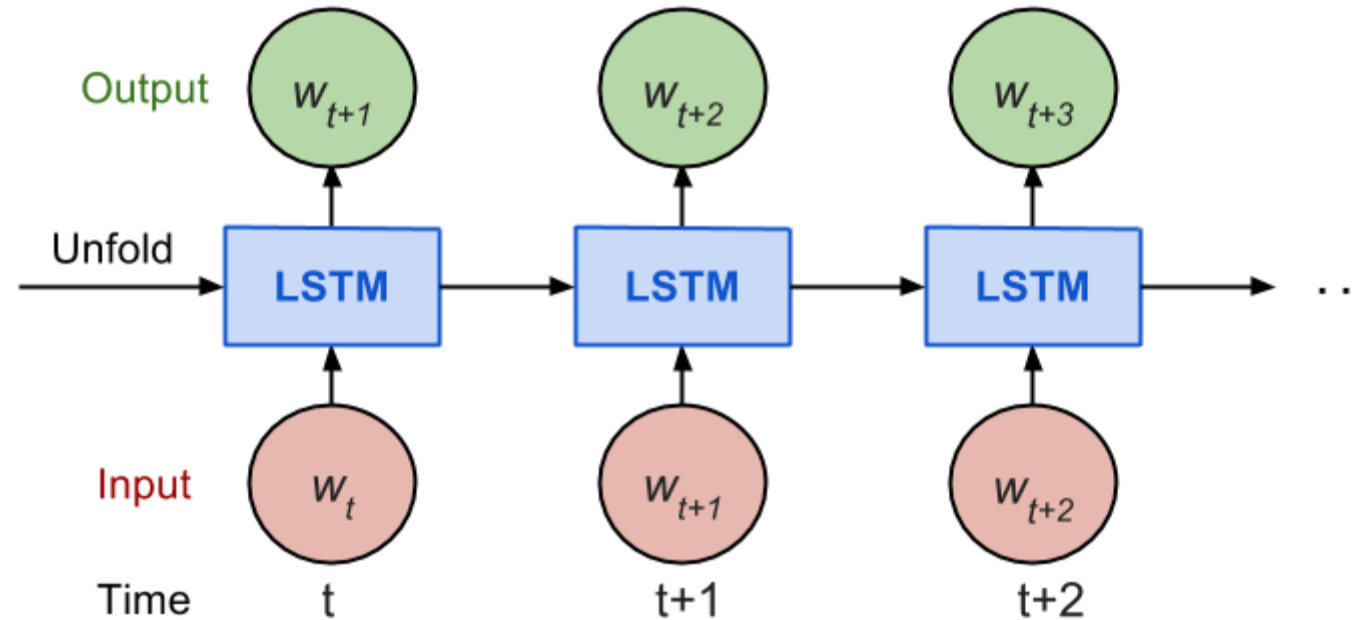
10

Word "Embedding"
for *This*

<i>This</i>	<i>is</i>	<i>a</i>	<i>small</i>	<i>vector</i>
0.1	0.3	1.1	0.0	0.5
0.7	0.1	1.2	0.1	0.4
0.4	0.1	0.9	0.0	0.5
1.1	1.2	0.0	1.1	1.0
0.3	0.1	0.2	2.3	1.0
0.4	0.8	0.4	1.2	1.1
0.7	0.9	0.6	0.0	0.1
0.9	0.1	0.0	0.1	0.5
2.1	0.0	0.7	1.1	0.7
0.0	0.1	0.9	0.1	0.7

Character-level language model

❖ architecture of the model

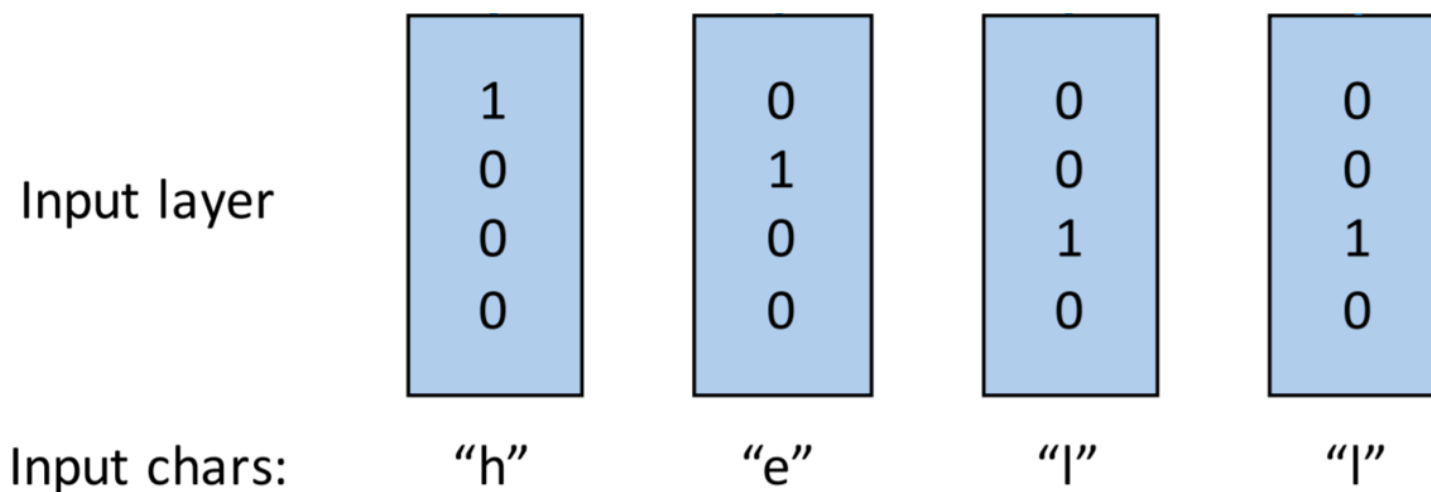


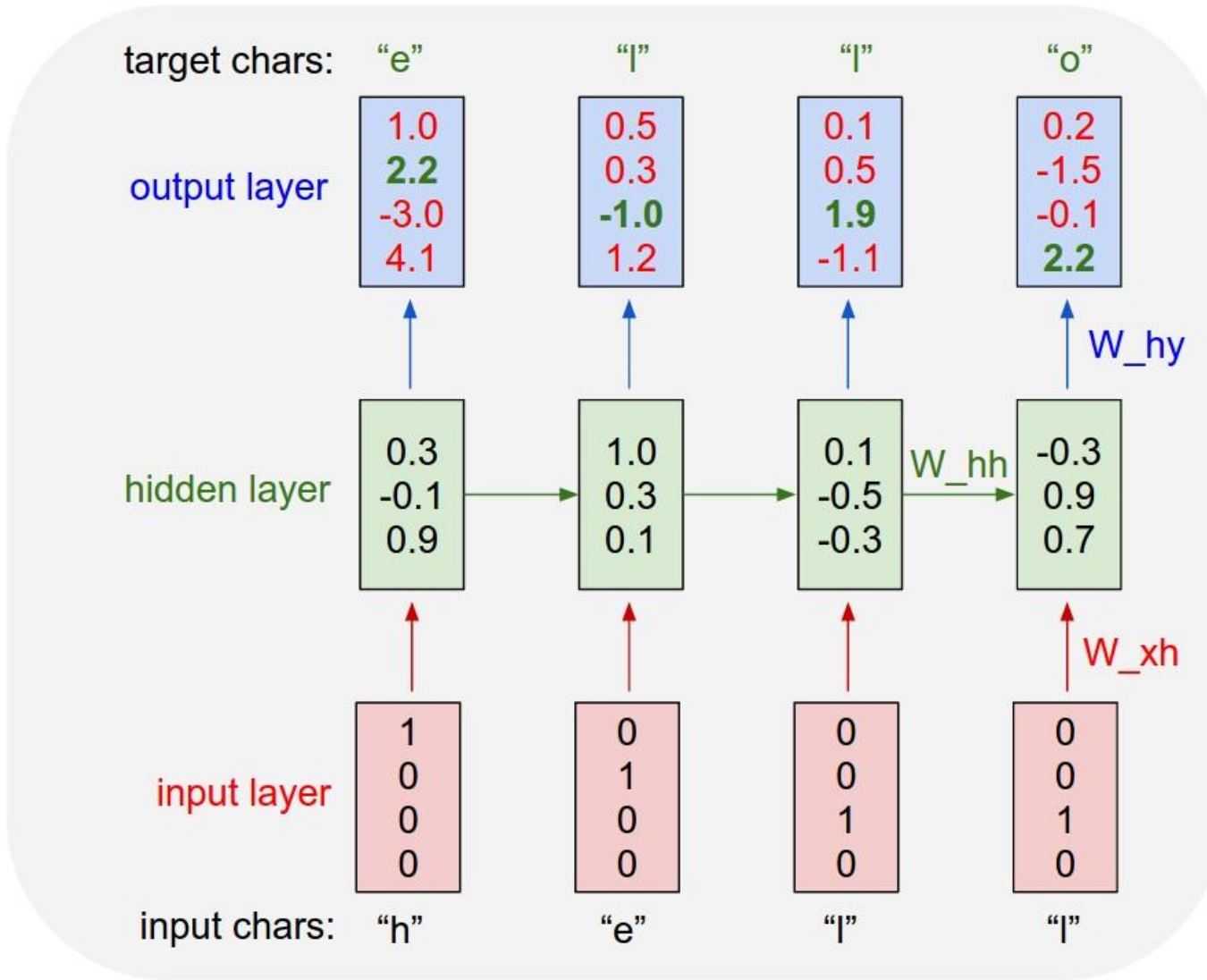
❖ Andrej Karpathy blog

✓ <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

❖ Character-level language model : we only had a vocabulary of "hell": [h,e,l,o]

✓ Encode each character into a vector using 1-of-k encoding



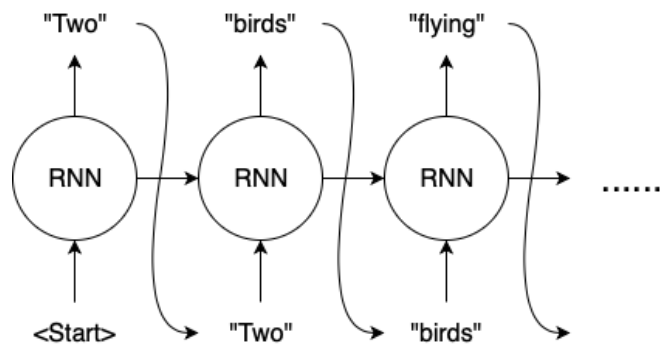


We want the green numbers to be high and red numbers to be low.

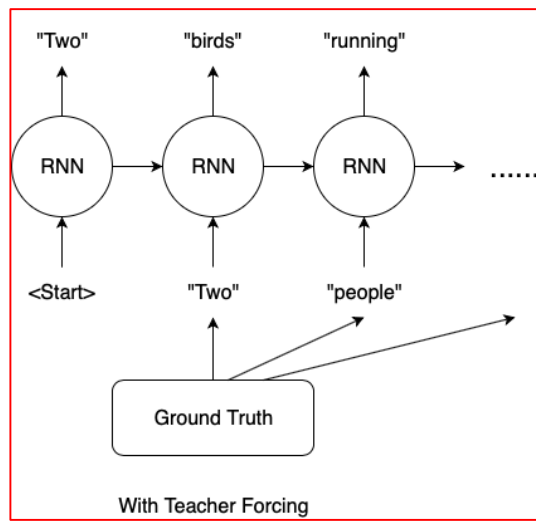


Teacher forcing

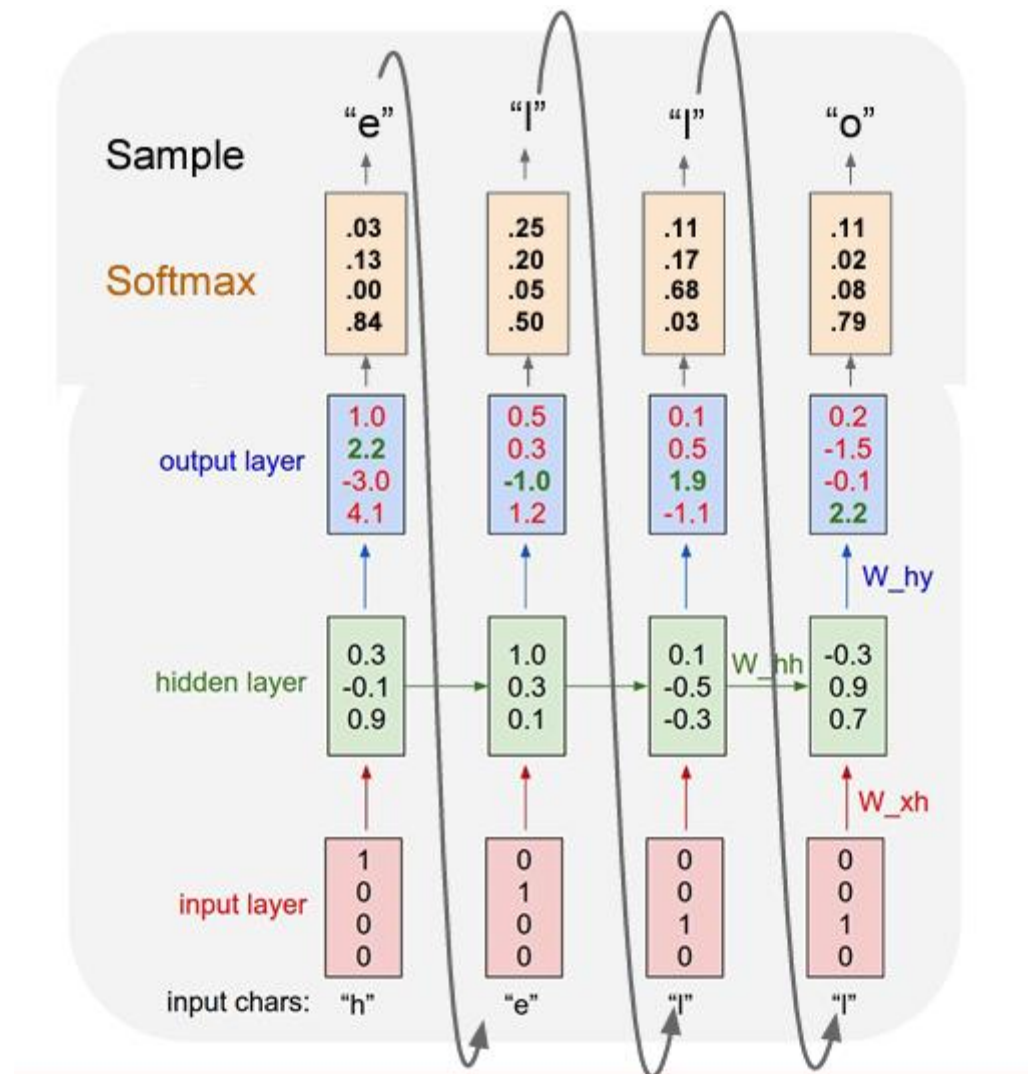
Teacher forcing : Character-level language model



Without Teacher Forcing



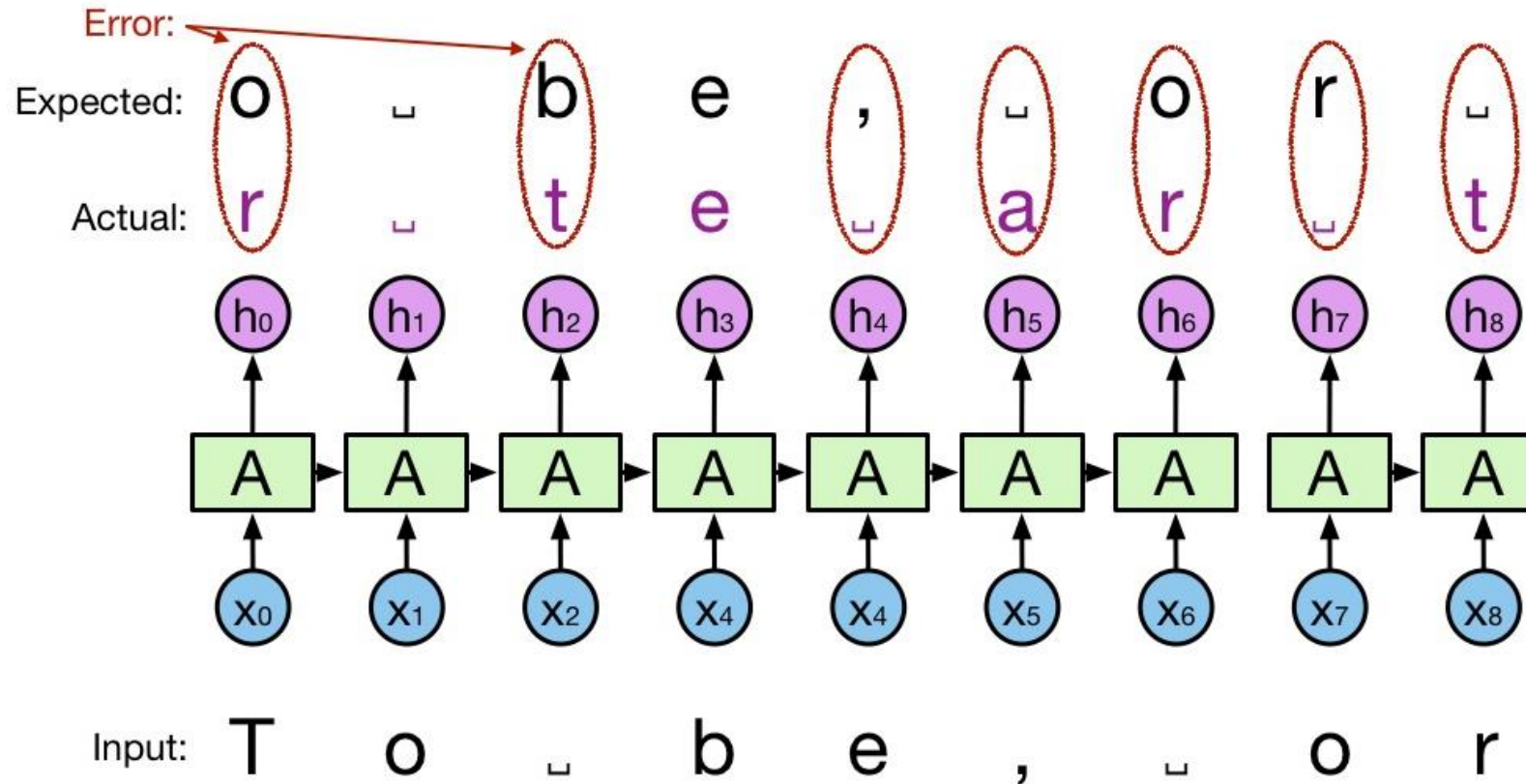
With Teacher Forcing



Generating Text Using a Char-RNN

Prediction of the next character

To _ be, _ or _ not _ to _ b _



❖ Downloading the data from Andrej Karpathy's Char-RNN project :

```
shakespeare_url = "https://raw.githubusercontent.com/karpathy/char-rnn/master/data/tinyshakespeare/input.txt"
```

```
print(shakespeare_text[:248])
```

First Citizen:

Before we proceed any further, hear me speak.

All:

Speak, speak.

First Citizen:

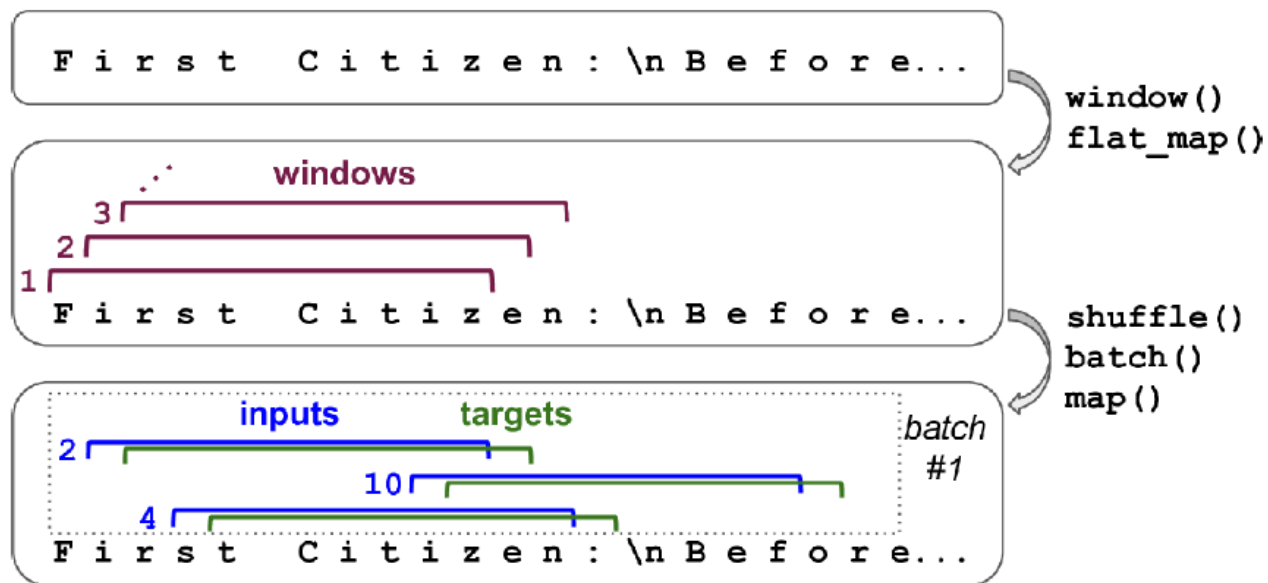
You are all resolved rather to die than to famish?

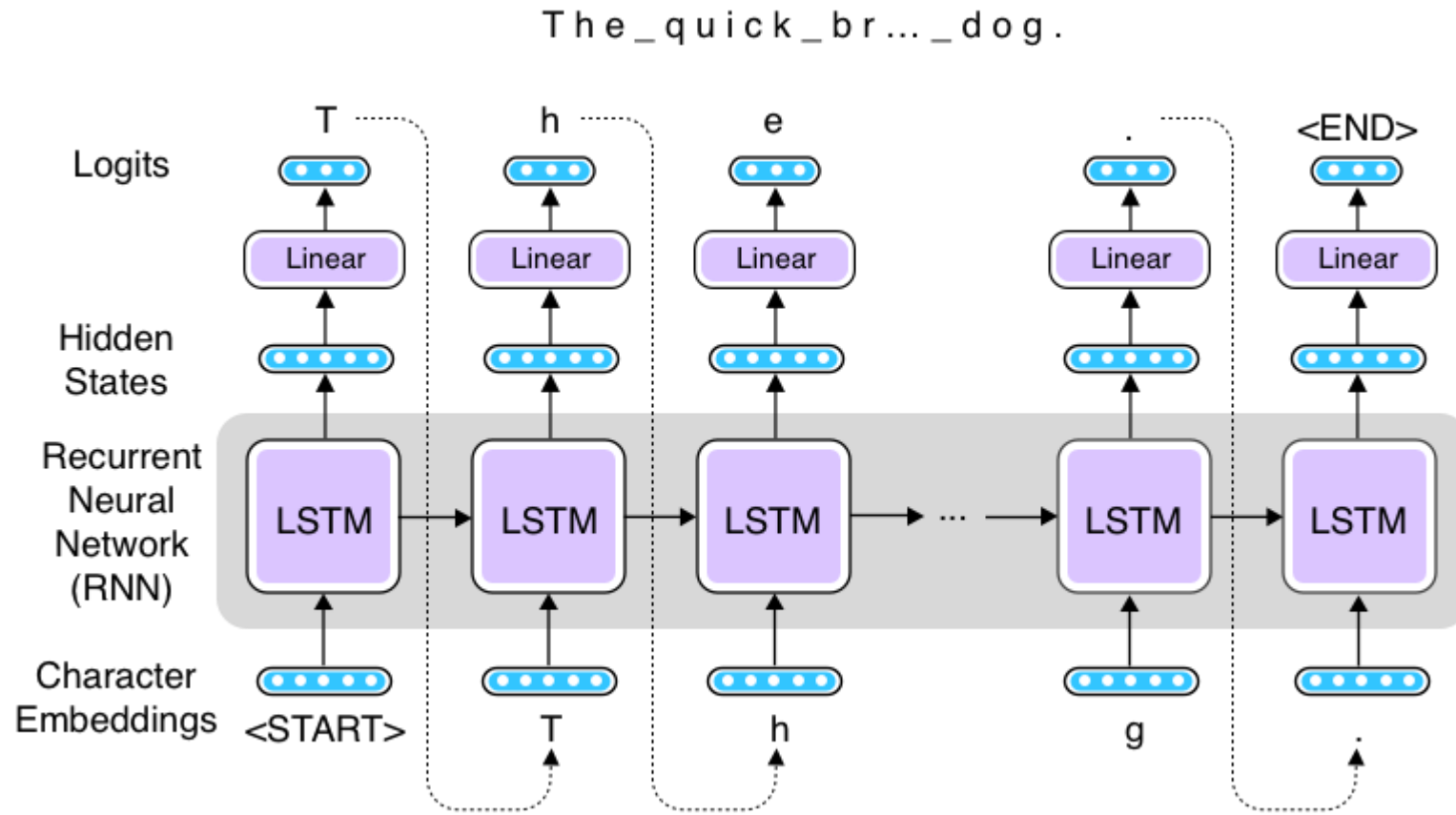
All:

Resolved. resolved.

First Citizen:

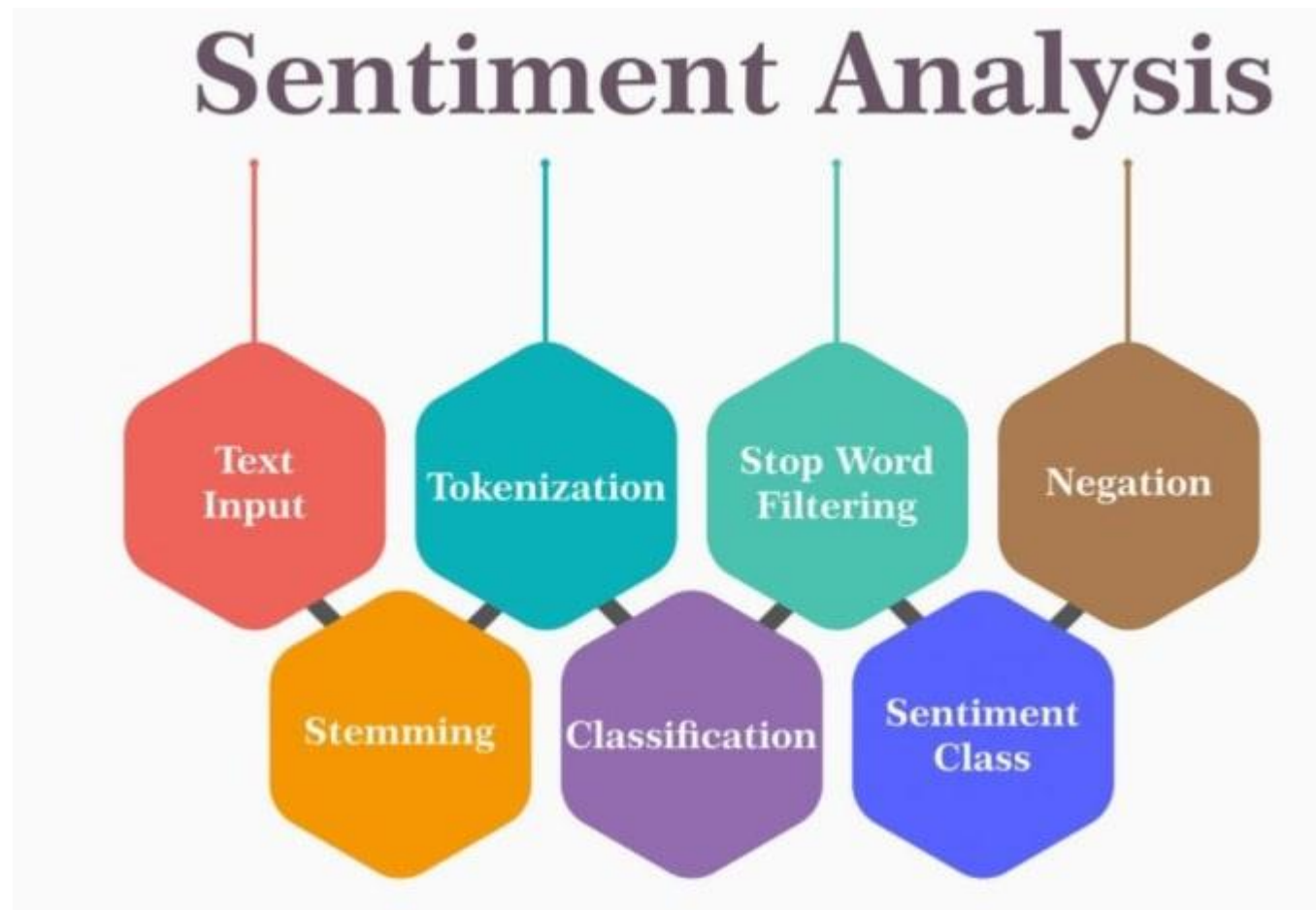
First, you know Caius Marcius is chief enemy to the people





<source> <http://www.realworldnlpbook.com/blog/training-a-shakespeare-reciting-monkey-using-rl-and-seqgan.html>

Text Classification Using RNN



❖ Text and Target

✓ Positive, Negative, Neutral, Happy, Sad



The screenshot shows the IMDb page for 'The Hunger Games: Mockingjay - Part 2 (2015)'. The movie poster is on the left. The right side contains the title, year, and a red-bordered box highlighting the rating section. The rating is 7.1/10 from 38,869 users. Below the rating is a synopsis and the director's name, Francis Lawrence.

IMDb Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist

The Hunger Games: Mockingjay - Part 2 (2015)
12A | 137 min | Adventure, Sci-Fi | 19 November 2015 (UK)

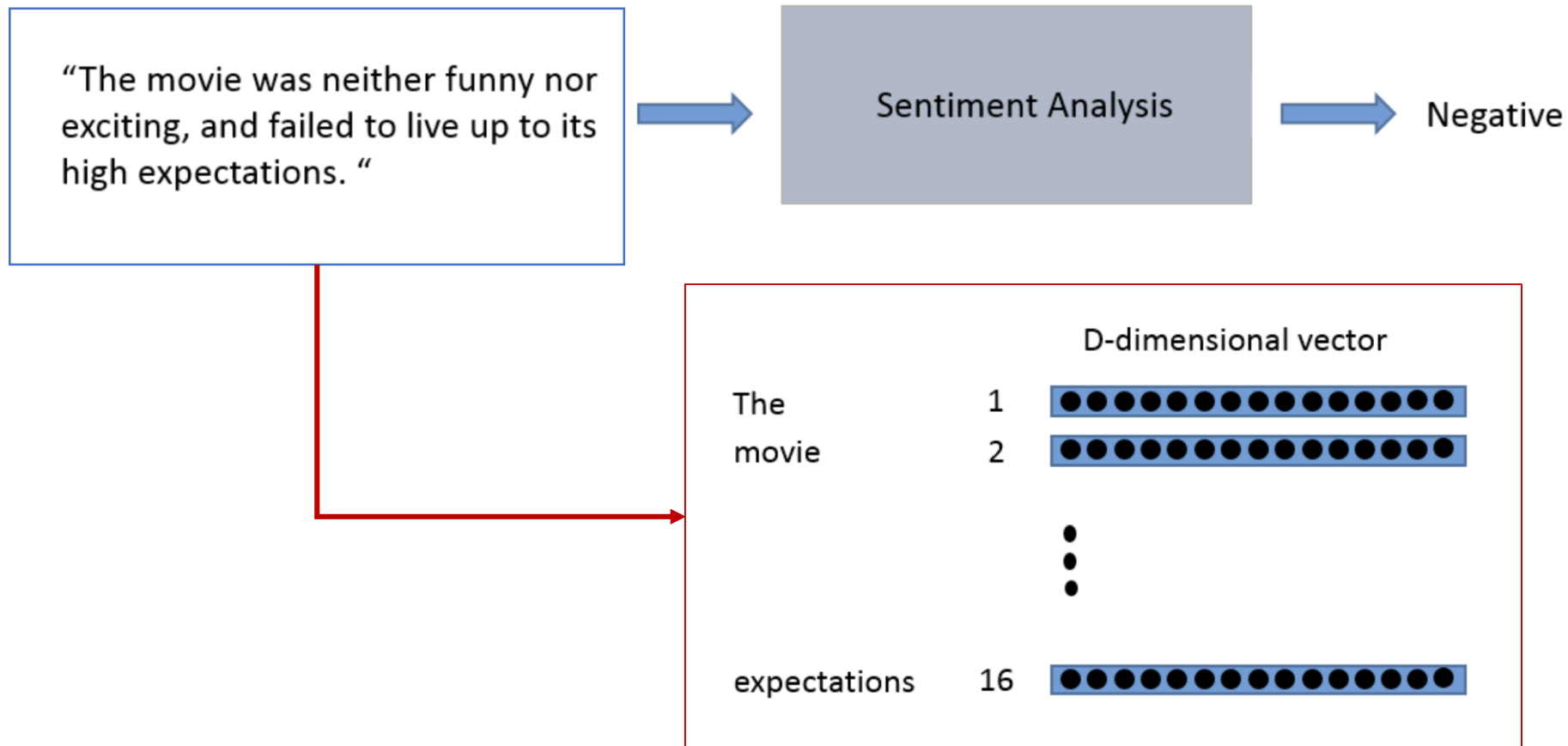
Your rating: ★★★★★★ ★★ -/10
7.1 Ratings: 7.1/10 from [38,869 users](#) Metascore: 65/100
Reviews: 178 user | 286 critic | [4 from Metacritic.com](#)

As the war of Panem escalates to the destruction of other districts by the Capitol, Katniss Everdeen, the reluctant leader of the rebellion, must bring together an army against President Snow, while all she holds dear hangs in the balance.

Director: [Francis Lawrence](#)
Writers: [Peter Craig](#) (screenplay), [Danny Strong](#)

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The filming tec...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
5	Probably my all-time favorite movie, a story o...	positive
6	I sure would like to see a resurrection of a u...	positive
7	This show was an amazing, fresh & innovative i...	negative
8	Encouraged by the positive comments about this...	negative

- ❖ Convert each word in the sentence to a vector



- ❖ The output of a Word2Vec model is called an embedding matrix
 - This embedding matrix will contain vectors for every distinct word in the training corpus.

English Wikipedia Corpus

The Annual Reminder continued through July 4, 1969. This final Annual Reminder took place less than a week after the June 28 Stonewall riots, in which the patrons of the Stonewall Inn, a gay bar in Greenwich Village, fought against police who raided the bar. Rodwell received several telephone calls threatening him and the other New York participants, but he was able to arrange for police protection for the chartered bus all the way to Philadelphia. About 45 people participated, including the deputy mayor of Philadelphia and his wife. The dress code was still in effect at the Reminder, but two women from the New York contingent broke from the single-file picket line and held hands. When Kameny tried to break them apart, Rodwell furiously denounced him to onlooking members of the press. Following the 1969 Annual Reminder, there was a sense, particularly among the younger and more radical participants, that the time for silent picketing had passed. Dissent and dissatisfaction had begun to take new and more emphatic forms in society.^[1] The conference passed a resolution drafted by Rodwell, his partner Fred Sargeant, Broidy and Linda Rhodes to move the demonstration from July 4 in Philadelphia to the last weekend in June in New York City, as well as proposing to "other organizations throughout the country... suggesting that they hold parallel demonstrations on that day" to commemorate the Stonewall riot.



Word2Vec



Embedding Matrix

D-dimensional vector

aardvark



apple

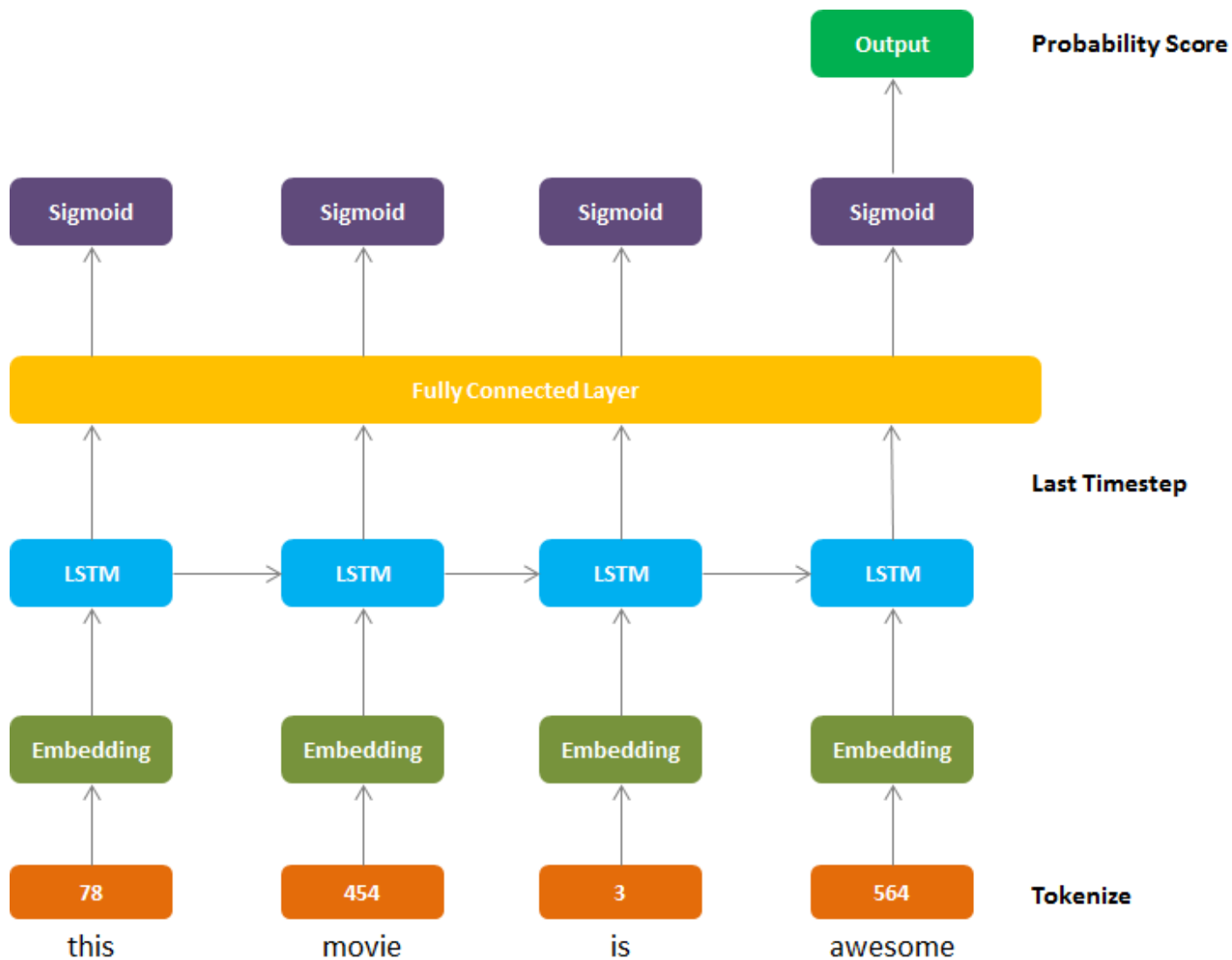


⋮

zoo



Many-to-One Sequence Model



2022

Korea Institute of Science
and Technology Information

TRUST
KISTIL

