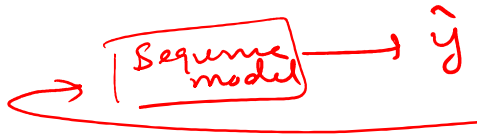
Models for Sequences

Sequence Modeling tasks



- Classify sequences $x_1, \ldots, x_n \rightarrow y$.
  - Sentiment classification, question intent classification, normal versus abnormal traffic
- Next word in a sequence $x_1, \ldots, x_n \rightarrow x_{\{n+1\}}$
  - Language modelling, forecasting.
- Label per token in sequence $x_1, \ldots, x_n \rightarrow y_1, \ldots, y_n$
  - POS, Named entity recognition
- Sequence prediction tasks $x_1, \ldots, x_n \rightarrow y_1, \ldots, y_m$
  - Translation, conversation assistant, speech recognition

x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9
(The Idlis in Maddu mess are to die for!)
y = positive

x_1 x_2 x_3 x_4?
I woke up ___
$y \in$ {booktitle, author name, None}

x: Have you read Goldberg's last Conjecture
x_1 x_2 x_3 x_4 x_5 x_6
None None None Booktitle Booktitle Booktitle

More examples

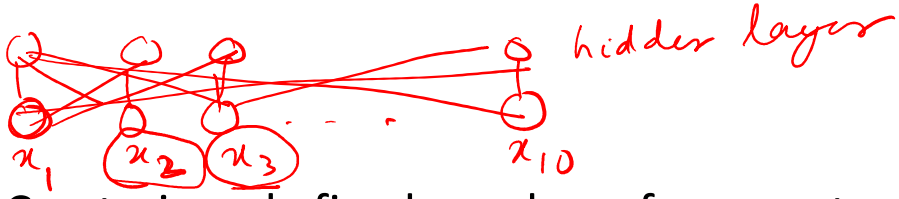
- Forecasting (Time Series)

$y_1, y_2 \cdots y_t$ outputs eg: demand for an item

$x_1, x_2 \cdots x_t$ inputs

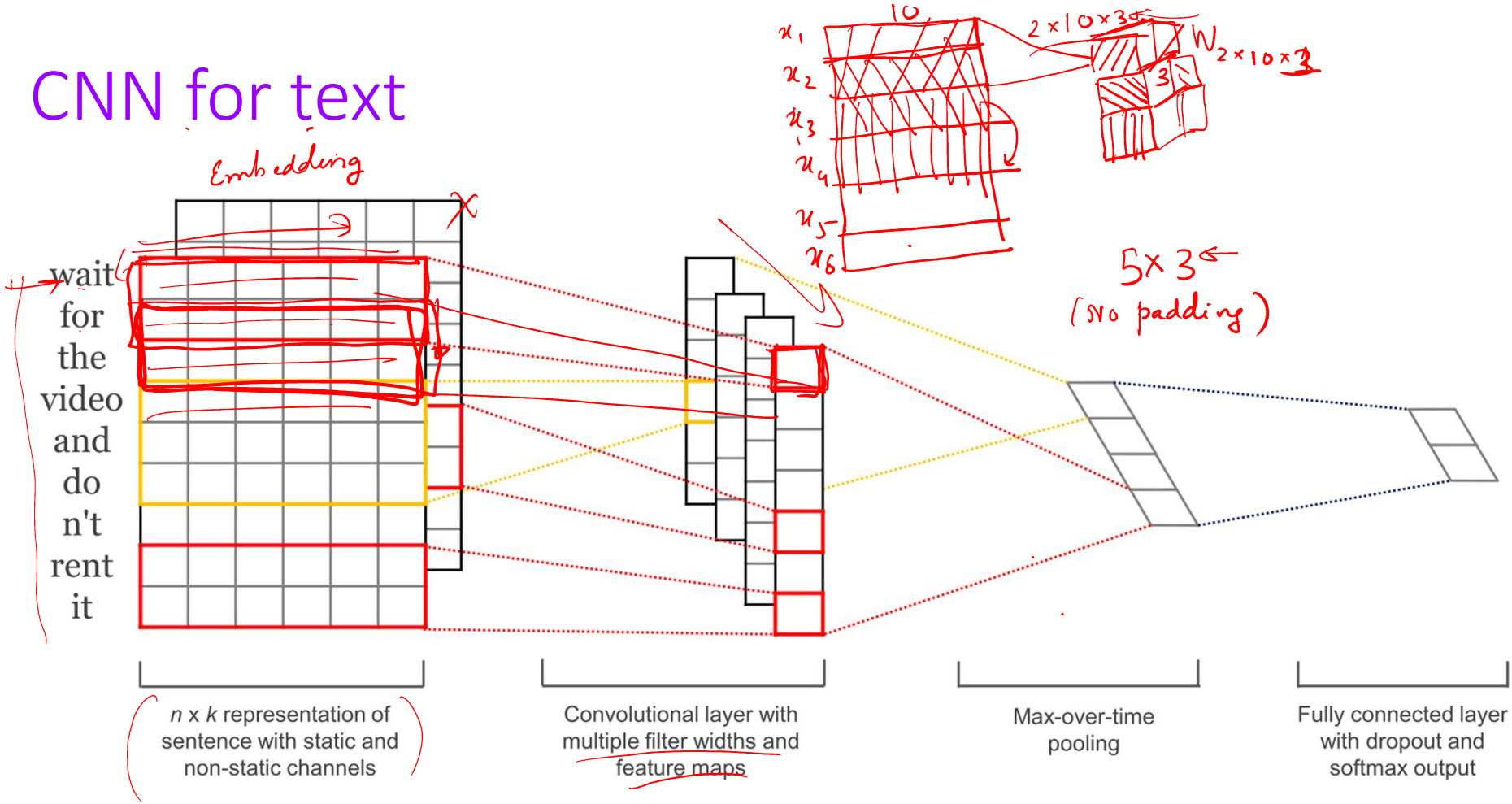
x_t : [day of the week, holiday?, temperature, month]

Why existing network architectures will not work



- Feed forward networks
  - Sequences can be of variable length. Can train only fixed number of parameters.
  - Even when sequences were of fixed length, do we really need to train a parameter per input token?
- CNNs
  - Designed to capture local patterns whereas in sequences interaction may be non-local. E.g.
    - He, after a vigorous workout, is enjoying a delicious breakfast with his friends.

CNN for text


wait
for
the
video
and
do
n't
rent
it
Embedding
5x3
(No padding)
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

RNN: Recurrent Neural Network

- A model to process variable length 1-D input
- In CNN, each hidden output is a function of corresponding input and some immediate neighbors.
- In RNN, each output is a function of a 'state' summarizing all previous inputs and current input. State summary computed recursively.
- RNN allows deeper, longer range interaction among ~~parameters~~ input tokens than CNNs for the same cost.

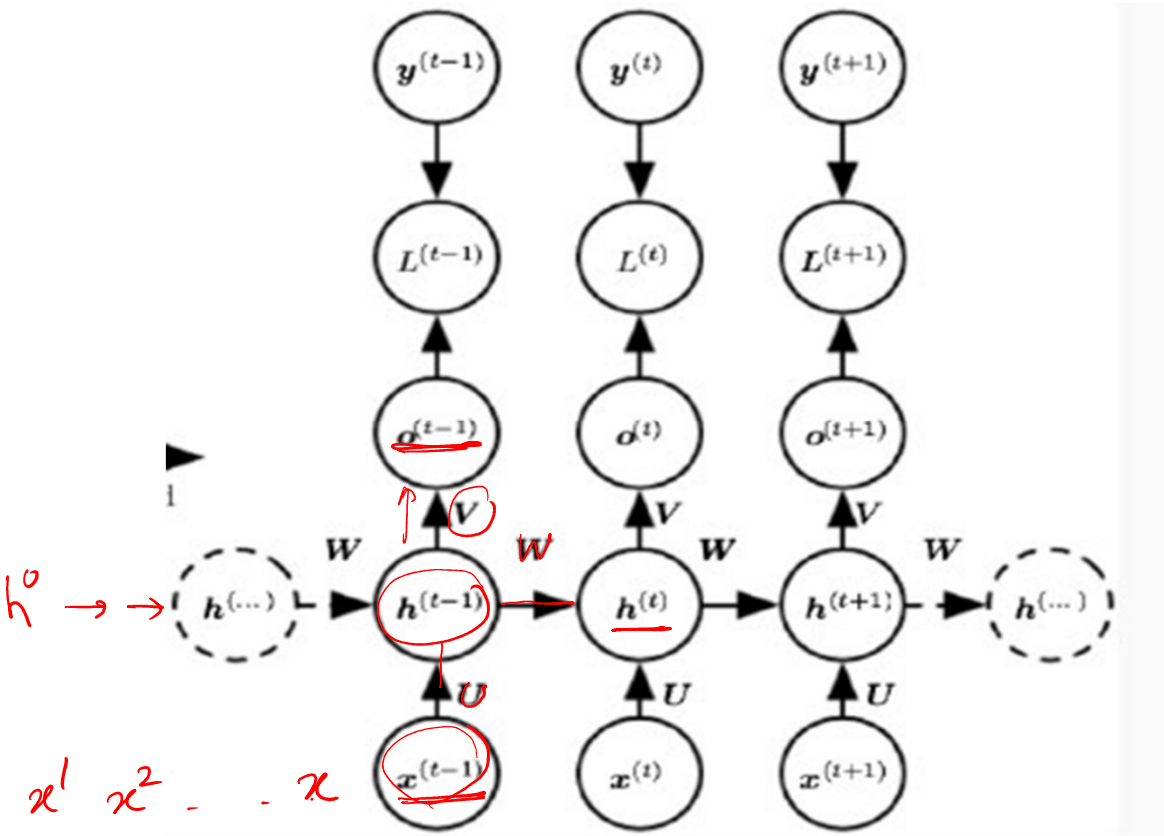
RNNs: Basic type

Notation:

- State to be denoted as h_t or z_t
- Input to RNN can be x_t or y_t

$$h^t = \sigma(b + Wh^{t-1} + Ux^t)$$

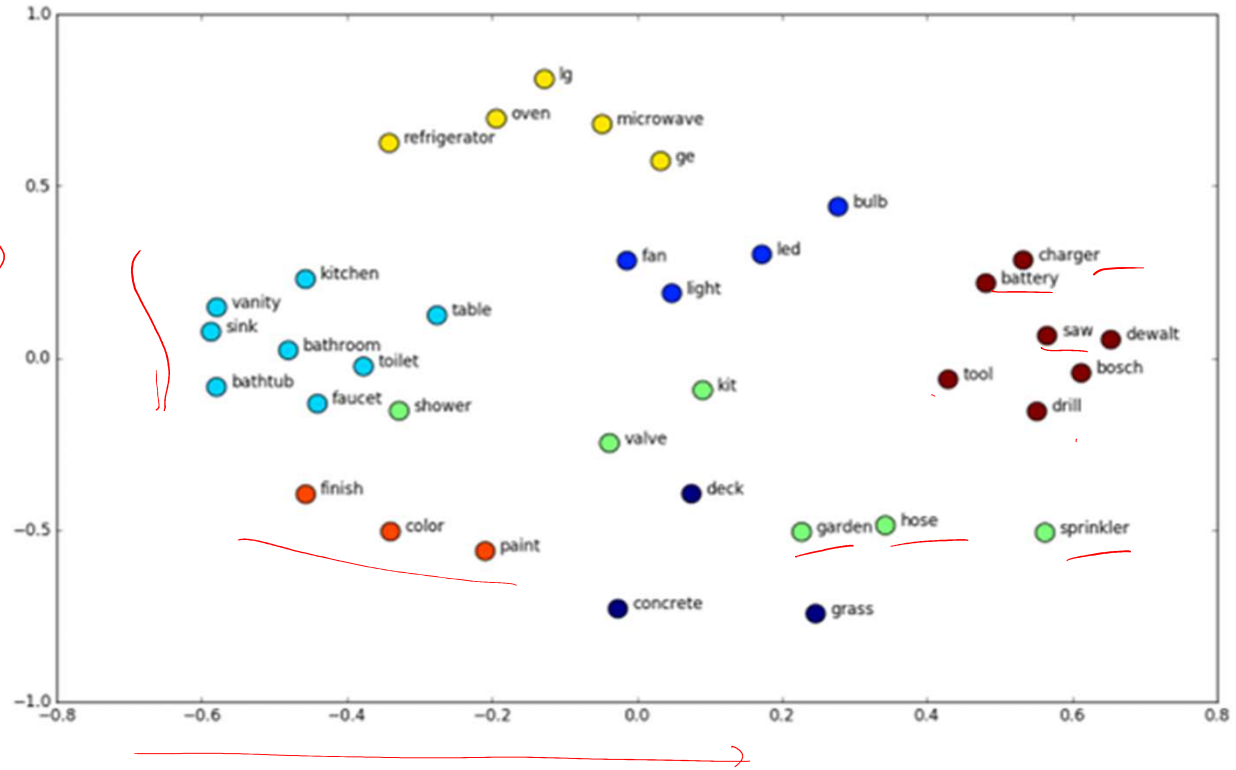
$$o^t = c + Vh^t,$$



Input: Text, Task: Text classification

- Input sequence: $x_1, x_2, \ldots, x_n$
- Each x_t instead of a real-valued vector is a discrete word.
- The task is to predict a label for this text
  - Example: Spam or Not-spam
  - Sentiment or star rating of a review
  - Topic of a document
  - Intent of a query
- How to convert a discrete word into a real-value for input into a RNN?
- How to convert a variable length sentence as a fixed-dimensional vector

Word-embeddings: Represent each word as a point in a multi-dimensional space..

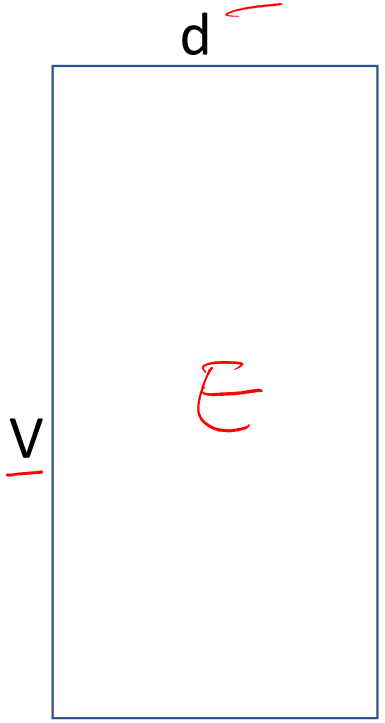


Similar words are close together in that space.

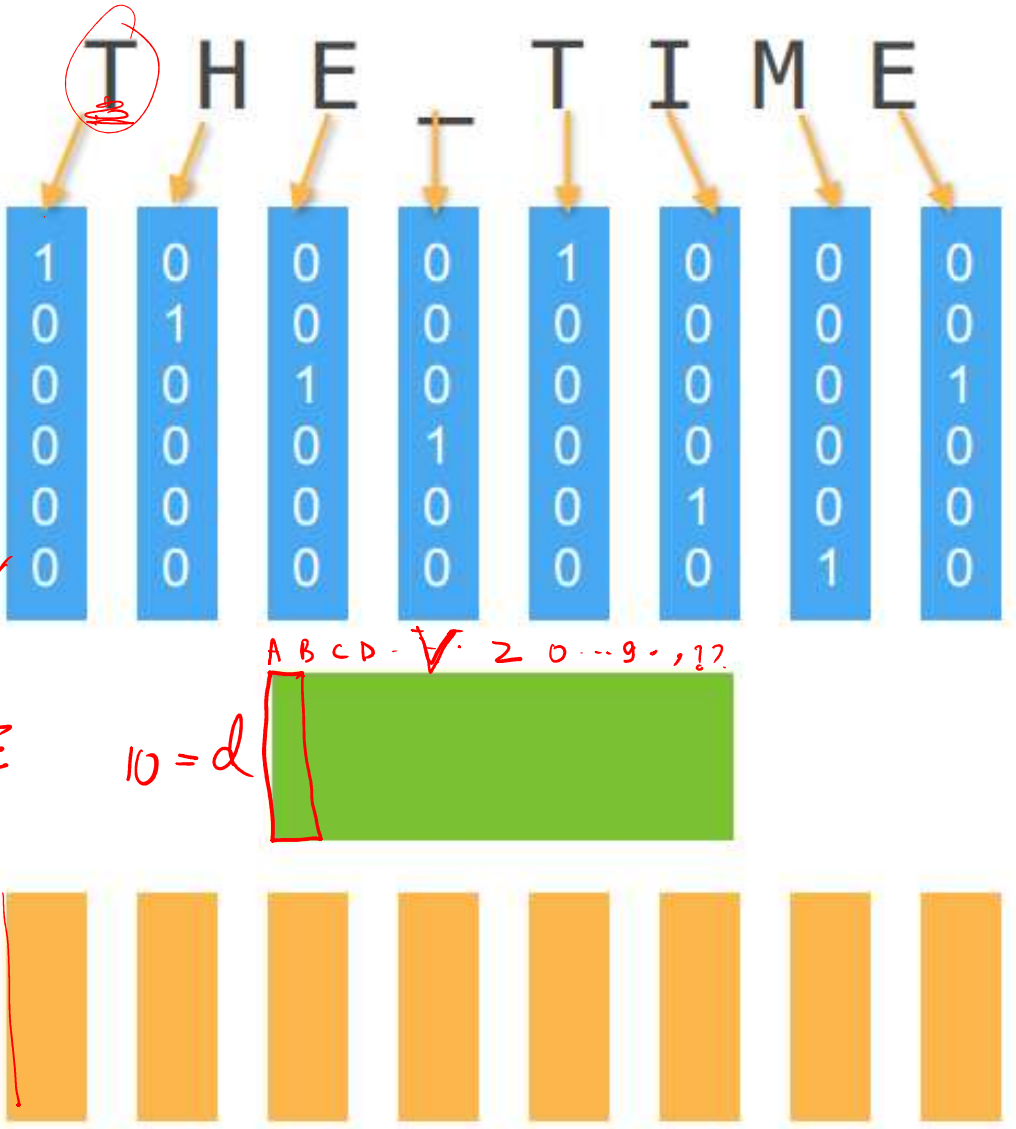
Src: http://suriyadeepan.github.io/

Word embeddings

- Choose the embedding size: d
  - Typical value for words: 128
- Choose the vocabulary size: V
  - Typical size: 30 thousand.
  - Rest are broken into subwords or UNKs.
- Embedding matrix: E
- If an input word x_t is a one-hot vector of size V, then embedding is obtained as $E^T x_t$



Input Encoding



Canonical Vectors $\mathbf{v}$

$|V| = 50$

26 + 10 + punctuation.

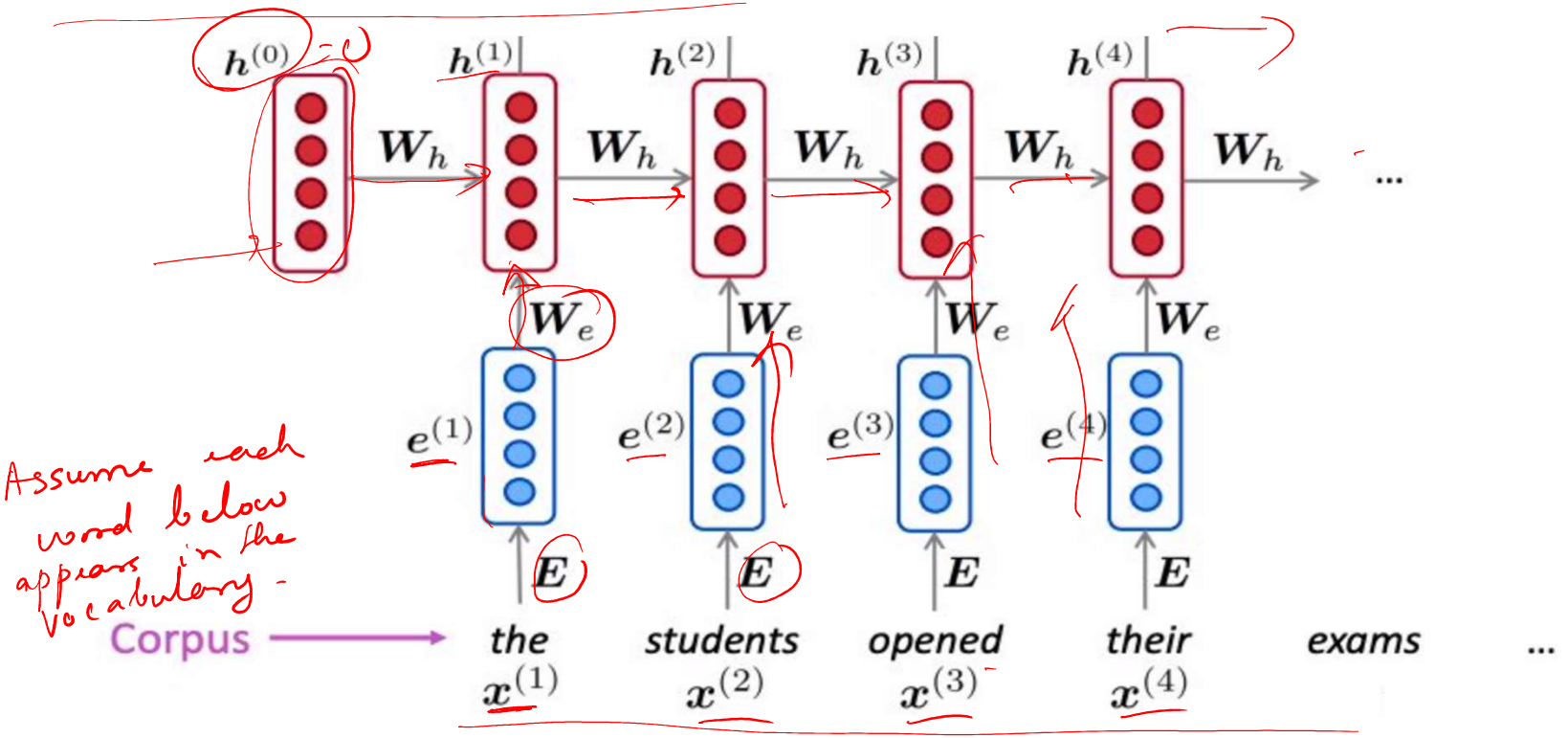
Embedding Matrix ~~$\mathbf{W}$~~ E

Embedded Vectors $\mathbf{v}'$

10

aws

RNNs on sentences

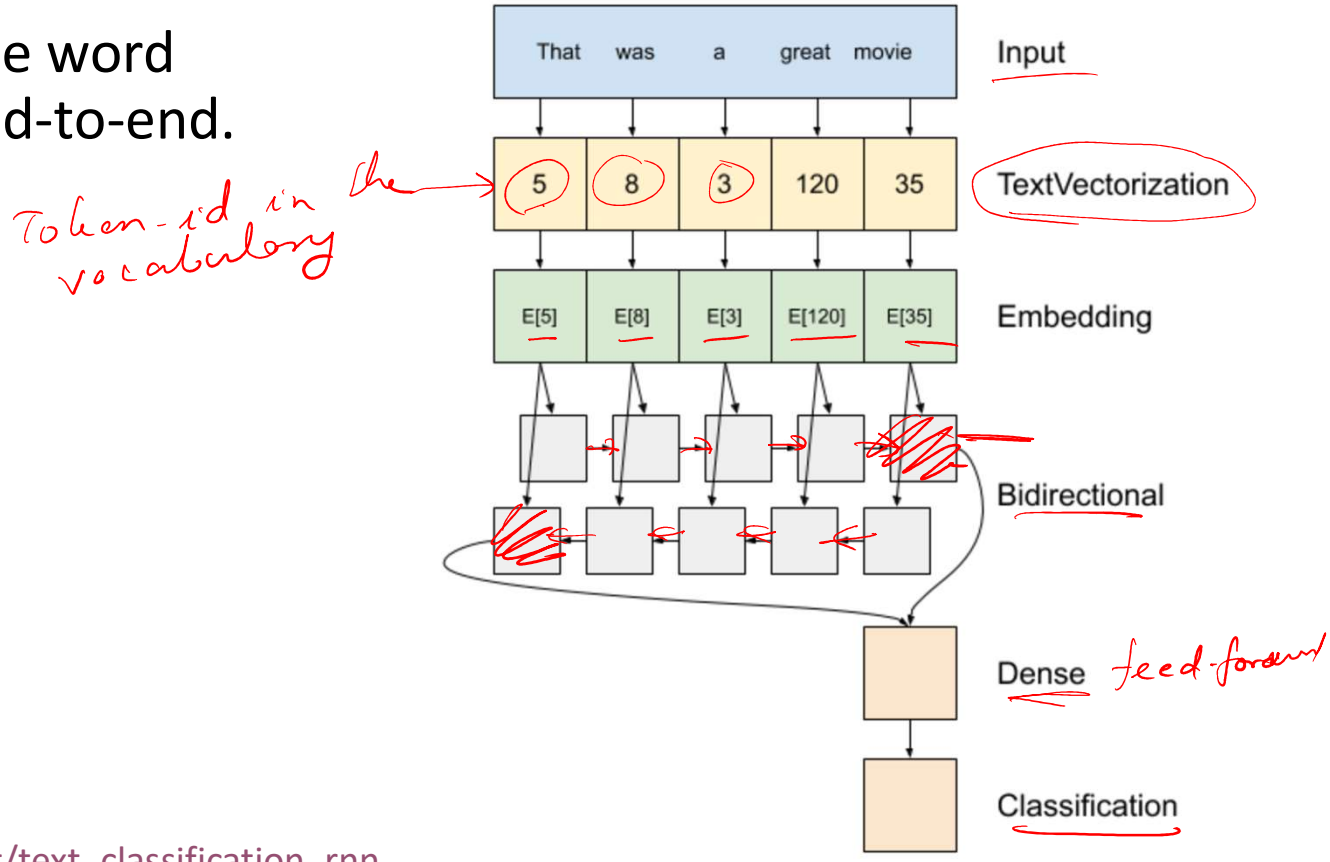


The last state (h) is the sentence encoding that converts variable length sentences into a fixed dimensional vector

Image src: https://towardsdatascience.com/introduction-to-rnns-sequence-to-sequence-language-translation-and-attention-fc43ef2cc3fd

Text classification with a bidirectional RNN.

All parameters including the word embeddings are trained end-to-end.



https://www.tensorflow.org/tutorials/text/text_classification_rnn

Gated RNNs

- Gates control which part of the long past is used for current prediction
- Gates also allow forgetting of part of the state
- LSTM: Long Short Term Memory, one of the most successful gated RNNs [Not in syllabus]
- Gated Recurrent Units (GRU)
  - [Not in syllabus]
- An excellent introductions here:
  - http://colah.github.io/posts/2015-08-Understanding-LSTMs/
  - http://blog.echen.me/2017/05/30/exploring-lstms/
  - https://www.tensorflow.org/tutorials/text/text_generation