Neural LMs and Transformers

Sandeep Soni

10/08/2024

Replication

- Objective: Learn by replicating a paper
- Bonus: replicate and then extend
- Example: Replicate gender bias paper from Garg et. al. on a different dataset

Sí o no, ¿què penses?
Catalonian Independence and Linguistic Identity on Social Media

Ian Stewart* **and Yuval Pinter*** **and Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA, USA
{istewart6, uvp, jacobe}@gatech.edu

Abstract

Political identity is often manifested in language variation, but the relationship between the two is still relatively unexplored from a quantitative perspective. This study examines the use of Catalan, a language local to the semi-autonomous region of Catalonia in Spain, on Twitter in discourse related to the 2017 independence referendum. We corroborate prior findings that pro-independence tweets are more likely to include the local language than anti-independence tweets. We also find that Catalan is used more often in referendum-related discourse than in other contexts, contrary to prior findings on language variation. This suggests a strong role for the Catalan language in the expression of Catalonian political identity.

this setting, we apply the methodology used by Shoemark et al. (2017) in the context of the 2014 Scottish independence referendum to a dataset of tweets related to the Catalonian referendum. We use the phenomenon of *code-switching* between Catalan and Spanish to pursue the following research questions in order to understand the choice of language in the context of the referendum:

1. Is a speaker's stance on independence strongly associated with the rate at which they use Catalan?
2. Does Catalan usage vary depending on whether the discussion topic is related to the referendum, and on the intended audience?

For the first question, our findings are similar

Choose your own adventure

- Objective: Learn by doing something new
- Bonus: novelty of idea
- Example: Measure storytelling in music

Where Do People Tell Stories Online?
Story Detection Across Online Communities

Maria Antoniak[♣] **Joel Mire**[♢] **Maarten Sap**[♢♣] **Elliott Ash**[♠] **Andrew Piper**[♡]

[♣]Allen Institute for AI [♢]Carnegie Mellon University [♠]ETH Zürich [♡]McGill University

Abstract

Story detection in online communities is a challenging task as stories are scattered across communities and interwoven with non-storytelling spans within a single text. We address this challenge by building and releasing the StorySeeker toolkit, including a richly annotated dataset of 502 Reddit posts and comments, a detailed codebook adapted to the social media context, and models to predict storytelling at the document and span level. Our dataset is sampled from hundreds of popular English-language Reddit communities ranging across 33 topic categories, and it contains fine-grained expert annotations, including binary story labels, story spans, and event spans. We eval

The mods removed my post last week, very frustrating. Anyway, my major is in Information Science and I'm entering my senior year. [I began school in CS, but then I switched to the iSchool because I discovered that the topics were more interesting for me.] I know I shouldn't worry about this, but I feel like my IS degree could hurt my chances of getting into a CS graduate program. I thought you all might have input about my options.

Table 1: A motivating example that shows event and [story] spans and illustrates the difficulty of determining story boundaries and event sequences.

telling, i.e., what is a story and what is not a story, is a difficult task that the field of narratology has been concerned with for decades (Bal and Van Bo-

Grading criteria

- Sketch out a (detailed) proposal
- Identify the data sources, point out the challenges, give a timeline for completion, layout an evaluation plan
- Discuss with me!

Story so far

- Language model is a probabilistic, generative model over words.
- Used to estimate P(x), if x is a sequence of linguistic units

N-gram language models

$$P(x) = \prod_i P(x_i)$$

unigram

$$P(x) = \prod_i P(x_i \mid x_{i-1})$$

bigram

$$P(x) = \prod_i P(x_i \mid x_{i-2}, x_{i-1})$$

trigram

Can we do better than N-gram language models?

Generative Vs Discriminative

	Naive Bayes	Logistic Regression
Type of classifier	Generative	Discriminative
Model	$P(x,y)$	$P(y \mid x)$
Objective	Generate the data and label jointly	Predict the label from the data

Language modeling

- Instead of modeling P(x), model P(w|c).
- Language modeling as a self-supervised learning task

Language modeling

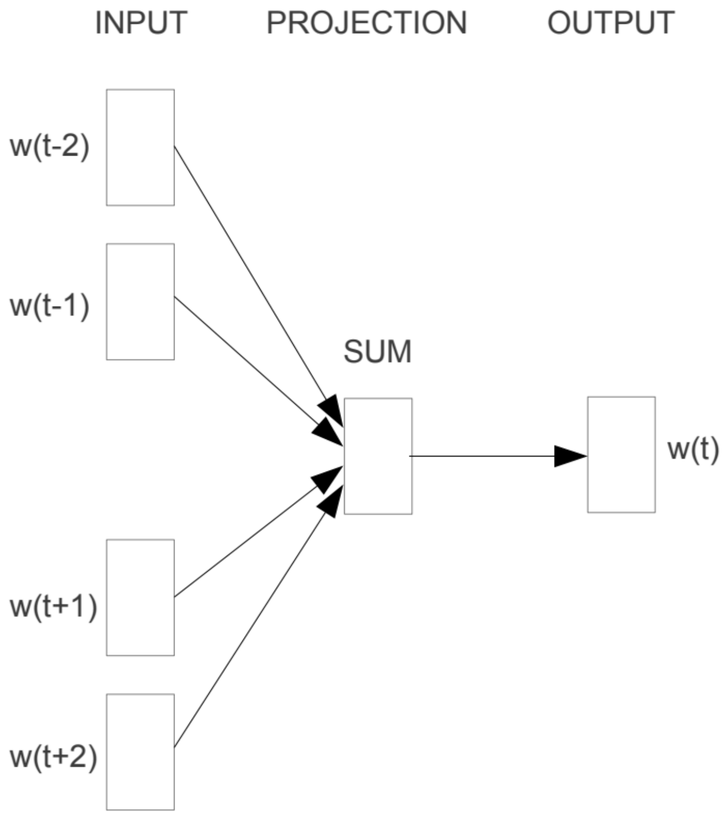
Parametrize $P(w|c) = \frac{\exp(\beta_w v_c)}{\sum_{w'} \exp(\beta_{w'} v_c)}$, where β_w is a dense vector for the word and v_c is the dense vector for context

Word2Vec (CBoW)

- $$P(w \mid c) = \frac{\exp(\beta_w v_c)}{\sum_{w'} \exp(\beta_{w'} v_c)}$$

- In CBoW model of word2vec, w is a word and c are words on the left and right of w

- v_c was calculated as a sum of output vectors



Mikolov et. al. 2013

CBoW-style LM

At every position m:

CBoW-style LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

CBoW-style LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

$$\mathbf{h}_m = \text{SUM}(\mathbf{x}_m, \mathbf{h}_{m-1})$$

CBoW-style LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

$$\mathbf{h}_m = \text{SUM}(\mathbf{x}_m, \mathbf{h}_{m-1})$$

Context vector

CBoW-style LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

$$\mathbf{h}_m = \text{SUM}(\mathbf{x}_m, \mathbf{h}_{m-1})$$

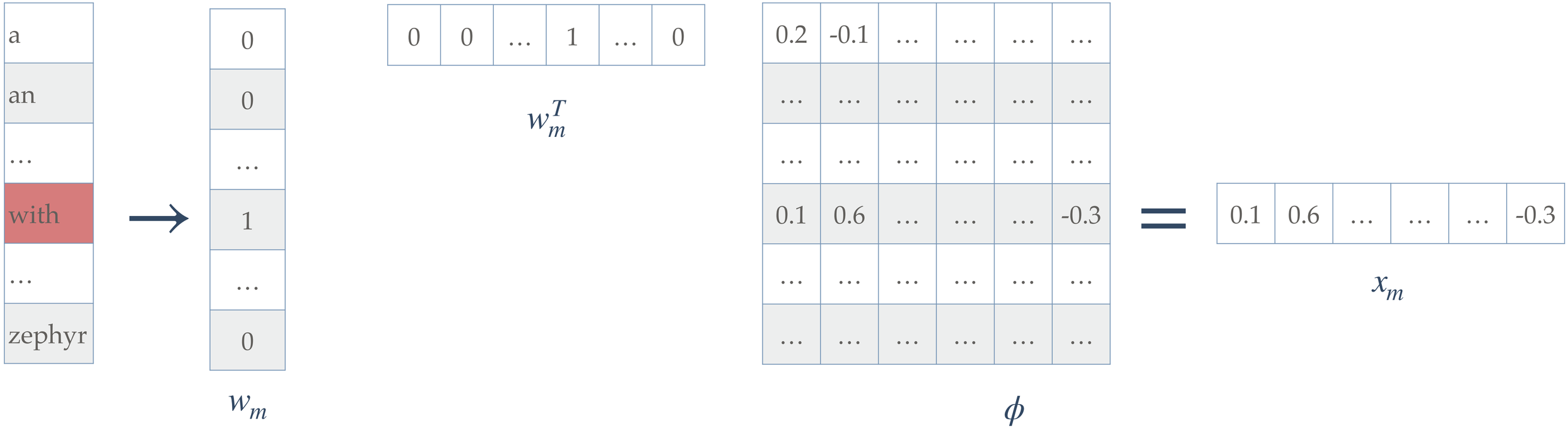
Context vector

$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$

Lookup

May the Force be with ____

$$x_m = \text{Lookup}(\phi, w_m)$$



CBoW-style LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

$$\mathbf{h}_m = \text{SUM}(\mathbf{x}_m, \mathbf{h}_{m-1})$$

Context vector

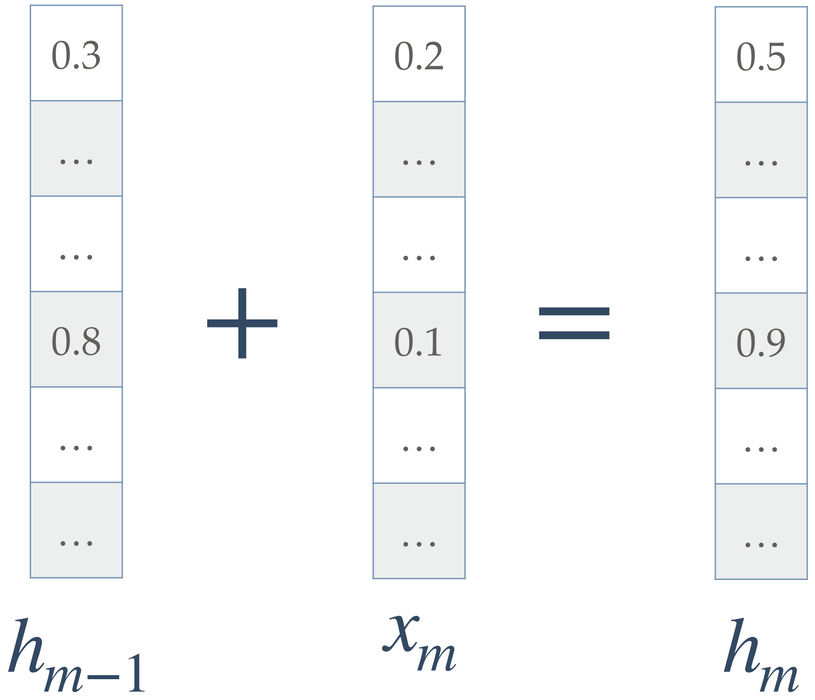
$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$

SUM

Context

May the Force be with ___

$$h_m = \text{SUM}(x_m, h_{m-1})$$

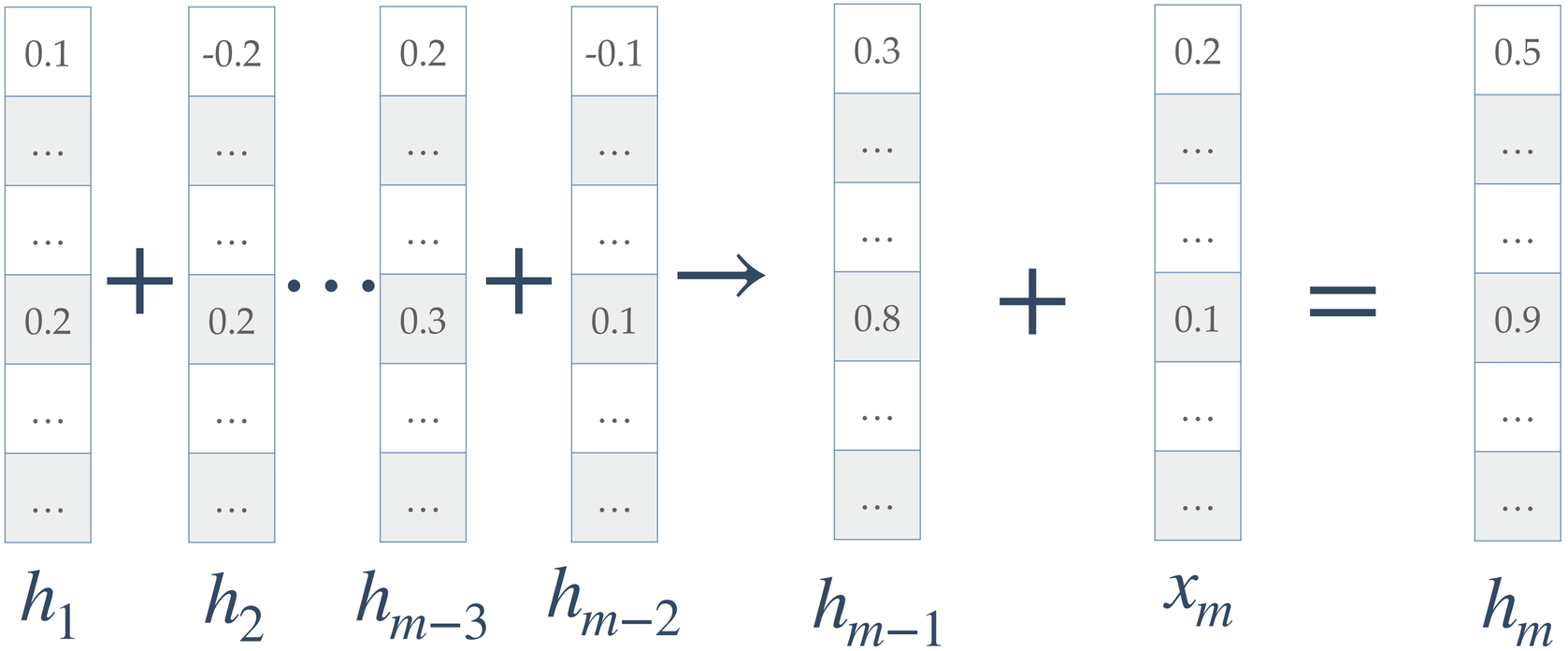


SUM

Context

May the Force be with ___

$$h_m = \text{SUM}(x_m, h_{m-1})$$



CBoW-style LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

$$\mathbf{h}_m = \text{SUM}(\mathbf{x}_m, \mathbf{h}_{m-1})$$

Context vector

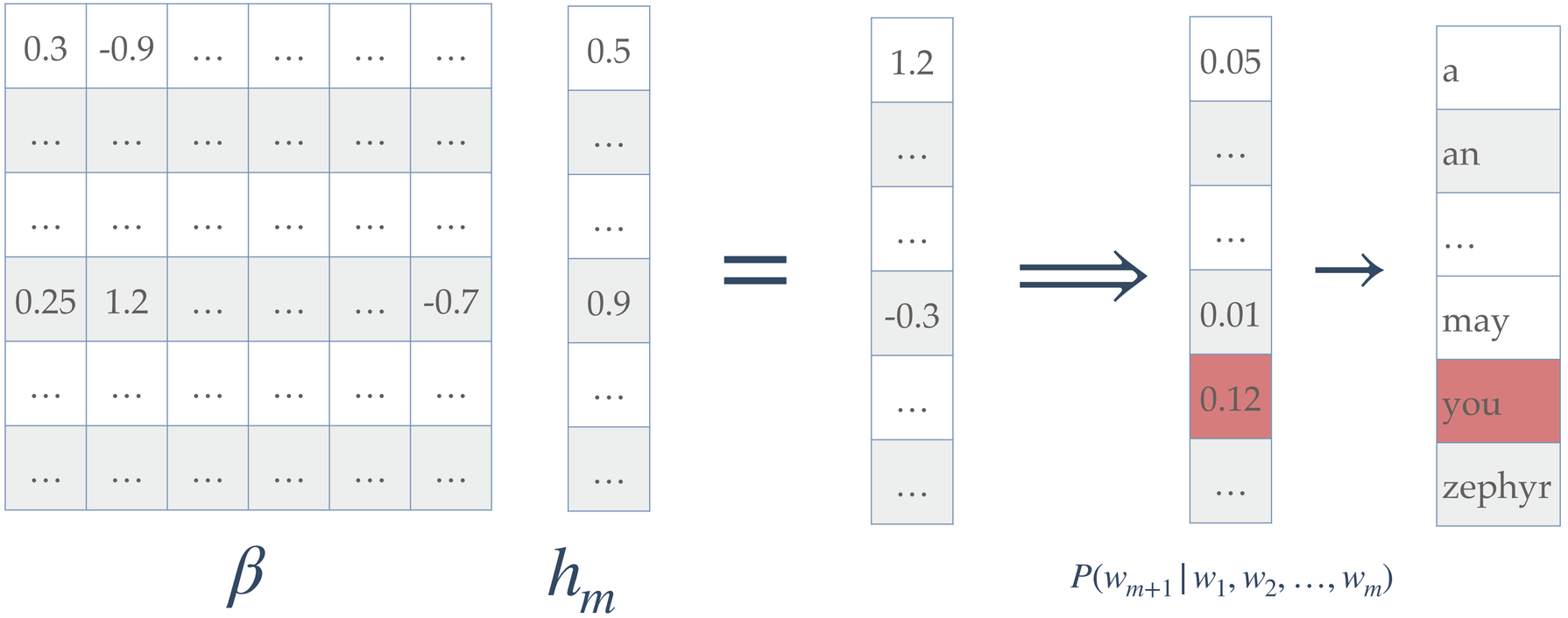
$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$

Softmax

Context: May the Force be with

Word: you

$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$



CBoW-style LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

$$\mathbf{h}_m = \text{SUM}(\mathbf{x}_m, \mathbf{h}_{m-1})$$

Context vector

$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$

Is there a better way to compute the vector representation of the context?

CBoW-style LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

$$\mathbf{h}_m = \text{SUM}(\mathbf{x}_m, \mathbf{h}_{m-1})$$

Context vector

$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$

Recurrent Neural Network LM

At every position m:

Recurrent Neural Network LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

Recurrent Neural Network LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

$$\mathbf{h}_m = \text{RNN}(\mathbf{x}_m, \mathbf{h}_{m-1})$$

Recurrent Neural Network LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

$$\mathbf{h}_m = \text{RNN}(\mathbf{x}_m, \mathbf{h}_{m-1})$$

Context vector

Recurrent Neural Network LM

At every position m:

$$\mathbf{x}_m = \text{Lookup}(\phi, w_m)$$

$$\mathbf{h}_m = \text{RNN}(\mathbf{x}_m, \mathbf{h}_{m-1})$$

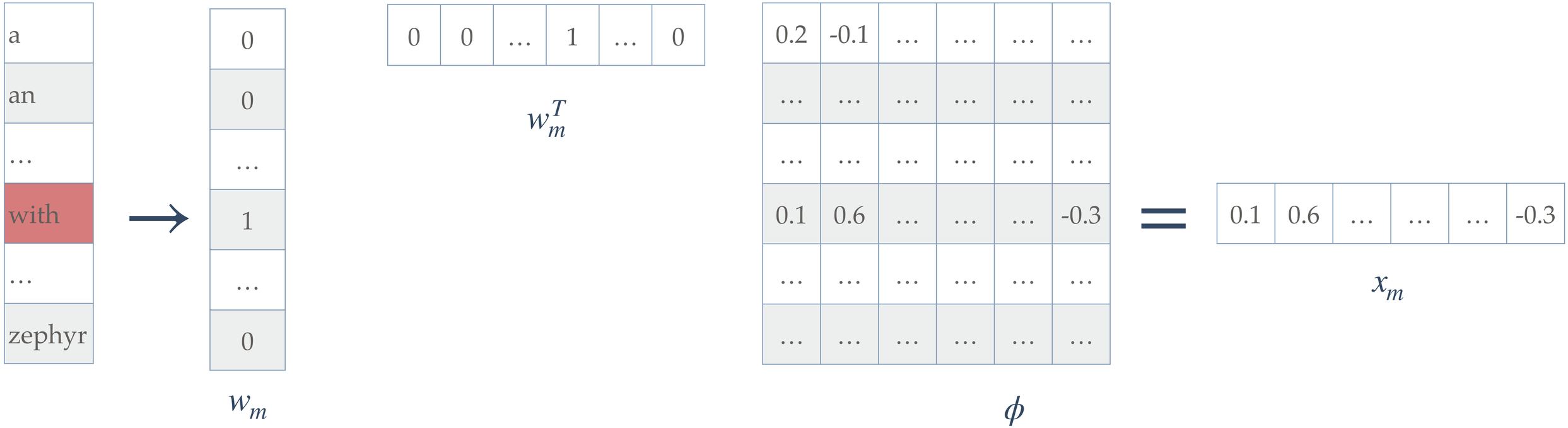
Context vector

$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$

Lookup

May the Force be with ____

$$x_m = \text{Lookup}(\phi, w_m)$$



Recurrent Neural Network LM

At every position m:

$$x_m = \text{Lookup}(\phi, w_m)$$

$$h_m = \text{RNN}(x_m, h_{m-1})$$

$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$

Recurrent Neural Network

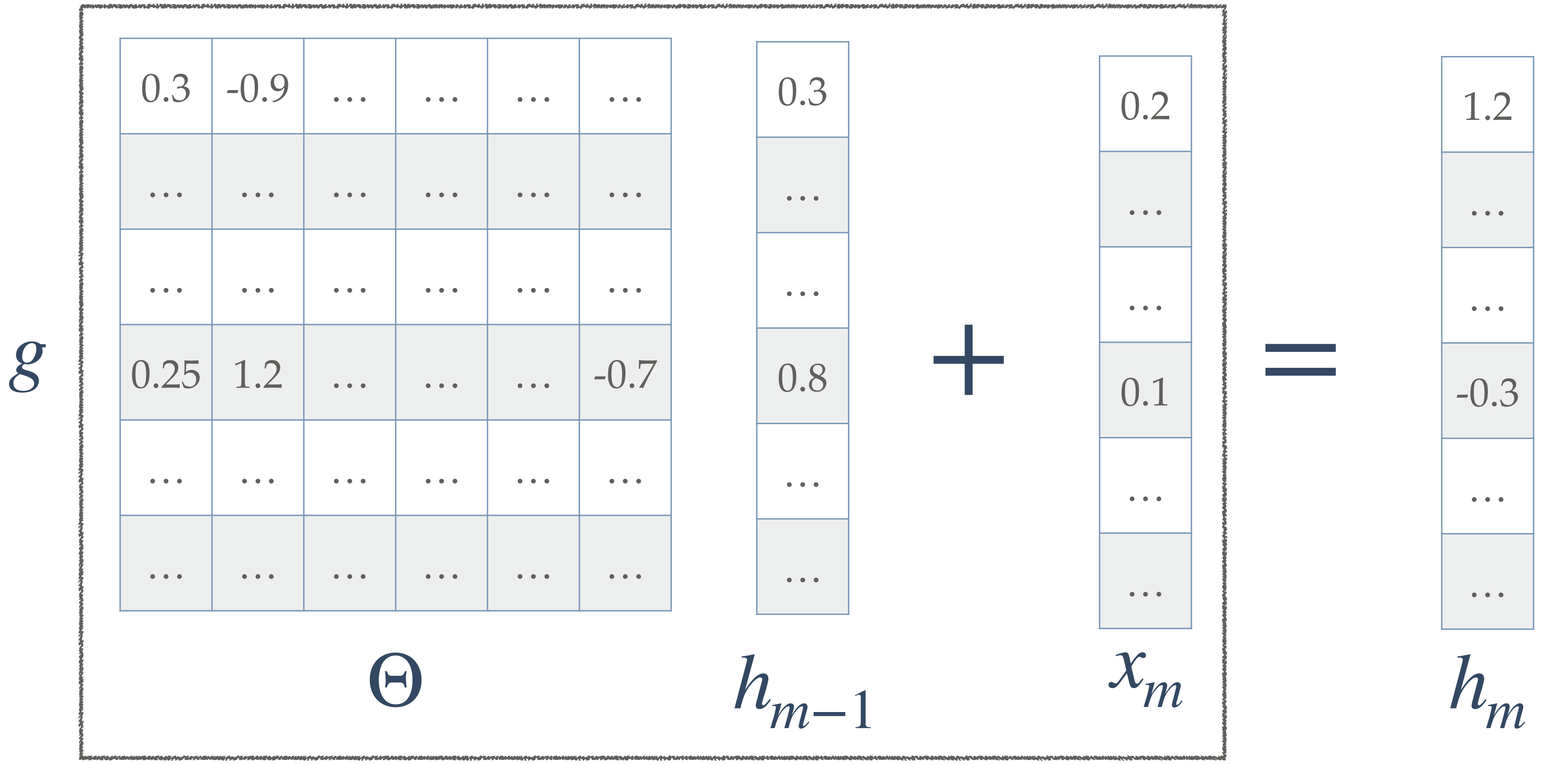
Context

May the Force be with ___

$$h_m = \text{RNN}(x_m, h_{m-1})$$

$$h_m = g(\Theta h_{m-1} + x_m)$$

Elman unit



Recurrent Neural Network LM

At every position m:

$$x_m = \text{Lookup}(\phi, w_m)$$

$$h_m = \text{RNN}(x_m, h_{m-1})$$

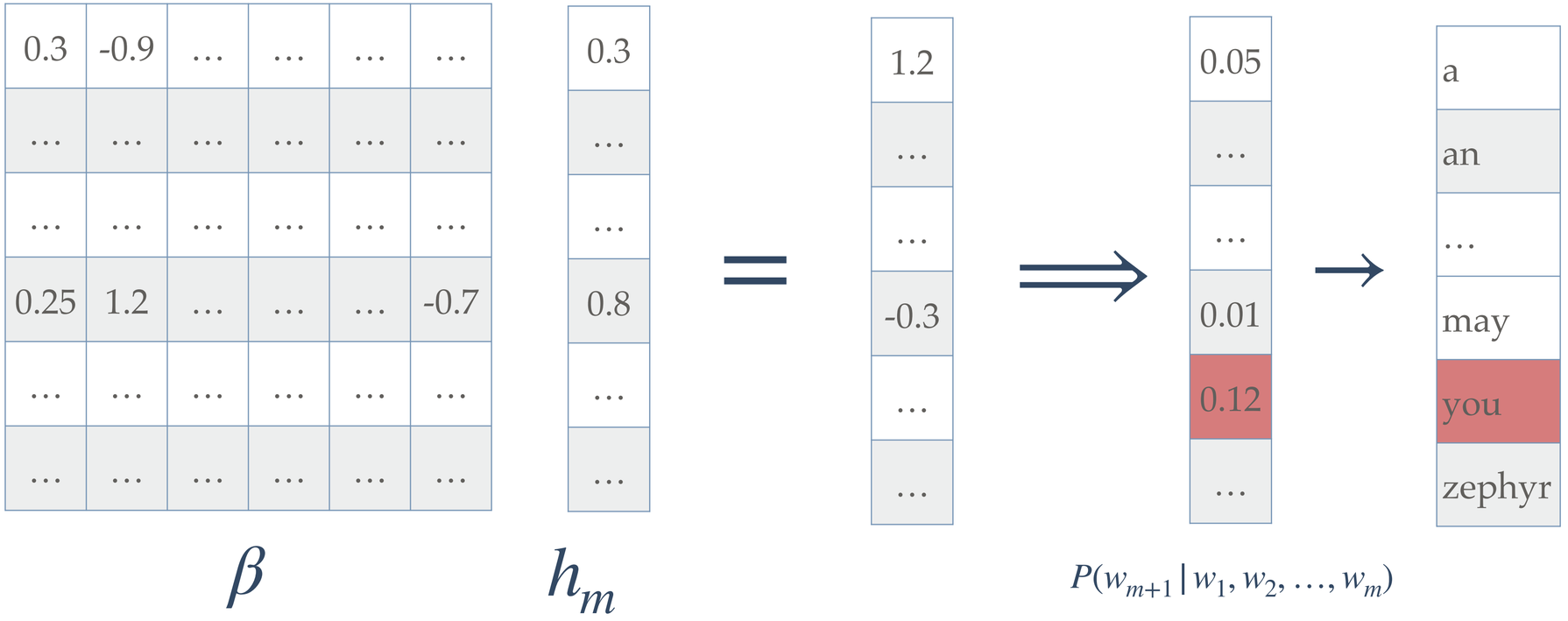
$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$

Softmax

Context: May the Force be with

Word: you

$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$



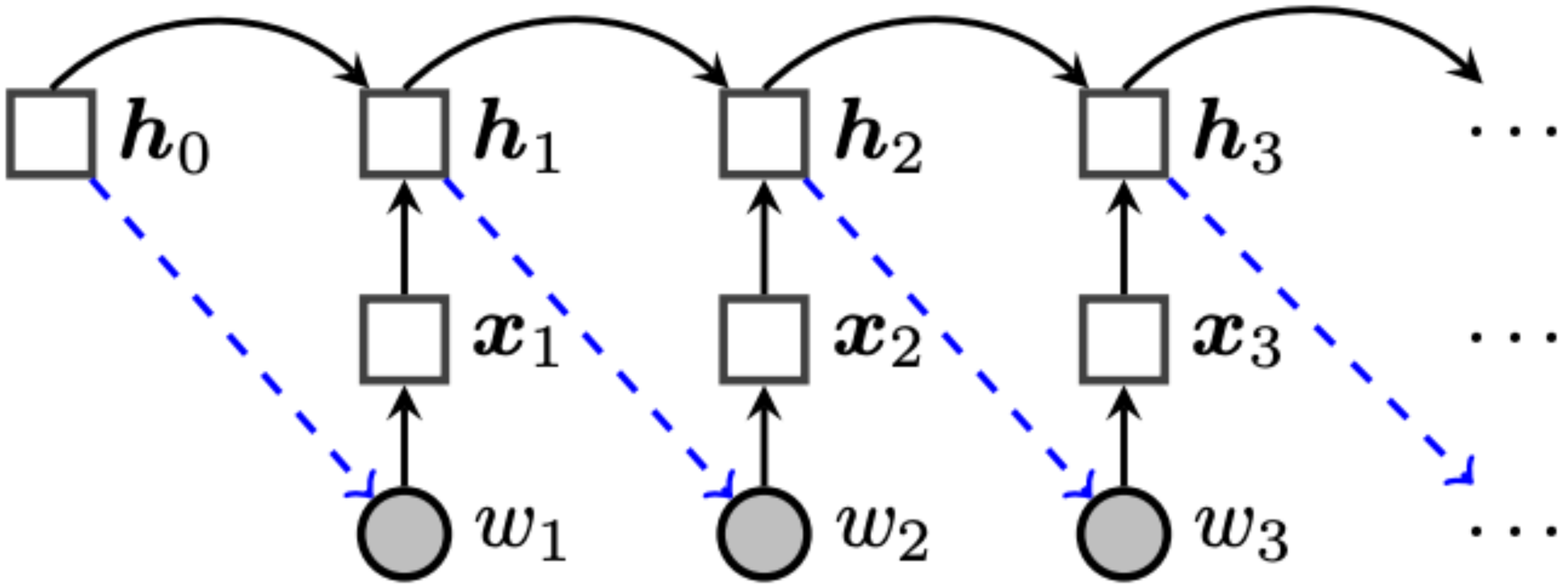


Figure taken from Eisenstein 2018

RNN LM

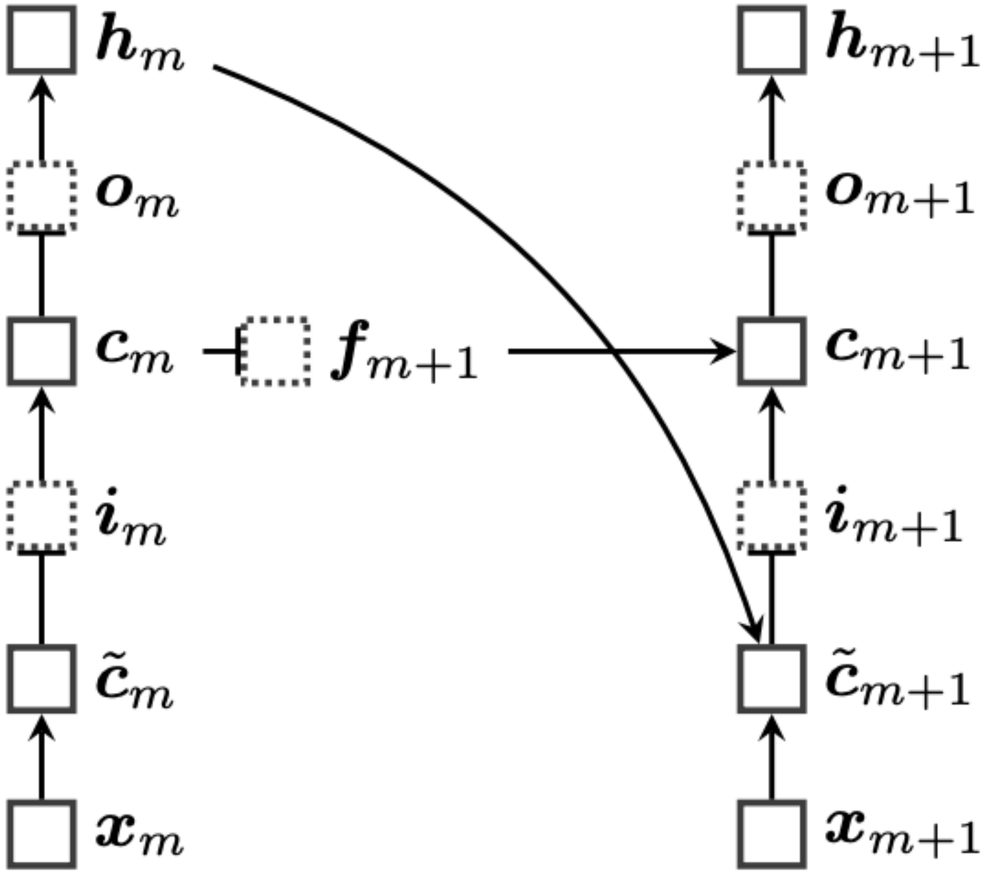
At every position m:

$$x_m = \text{Lookup}(\phi, w_m)$$

$$h_m = \text{LSTM}(x_m, h_{m-1})$$

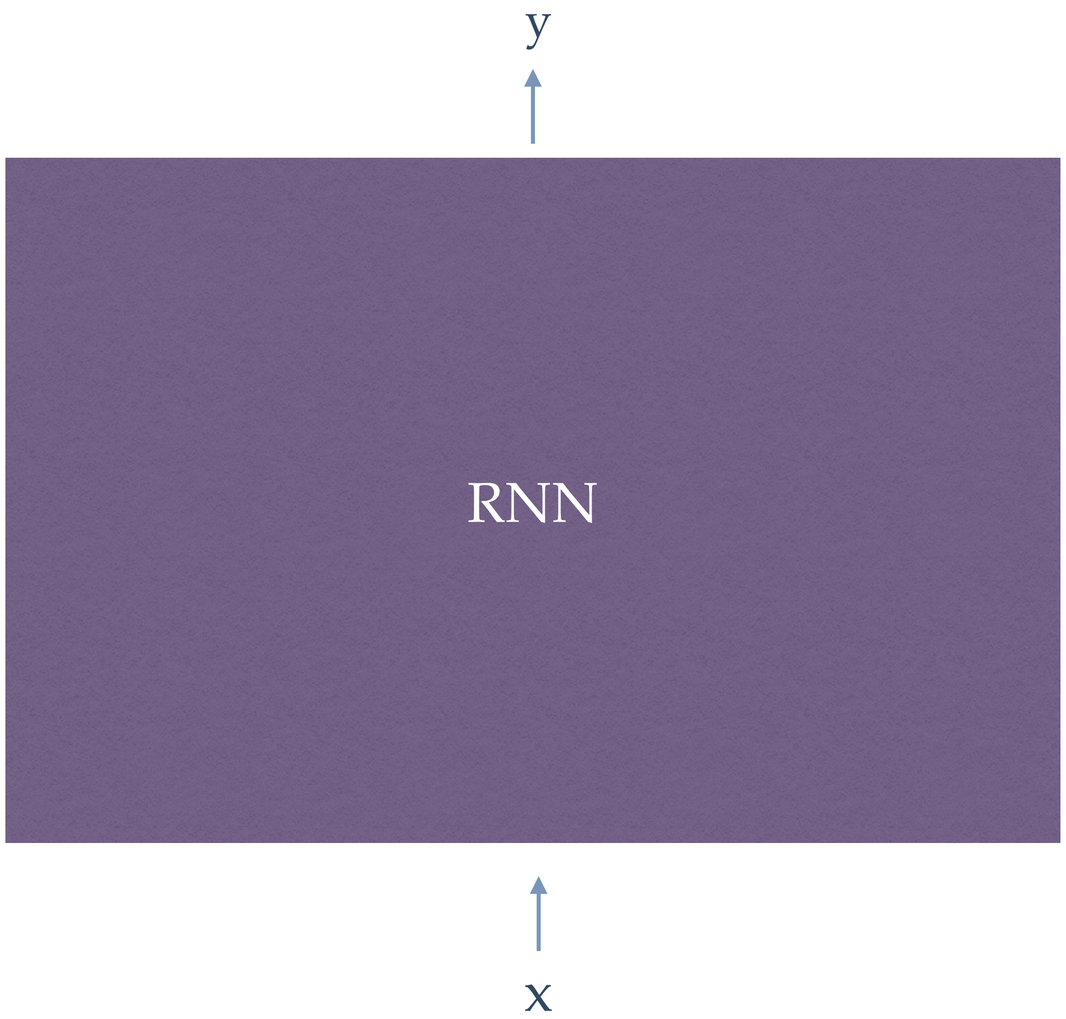
$$P(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \text{softmax}(\beta_{w_m}, \mathbf{h}_m)$$

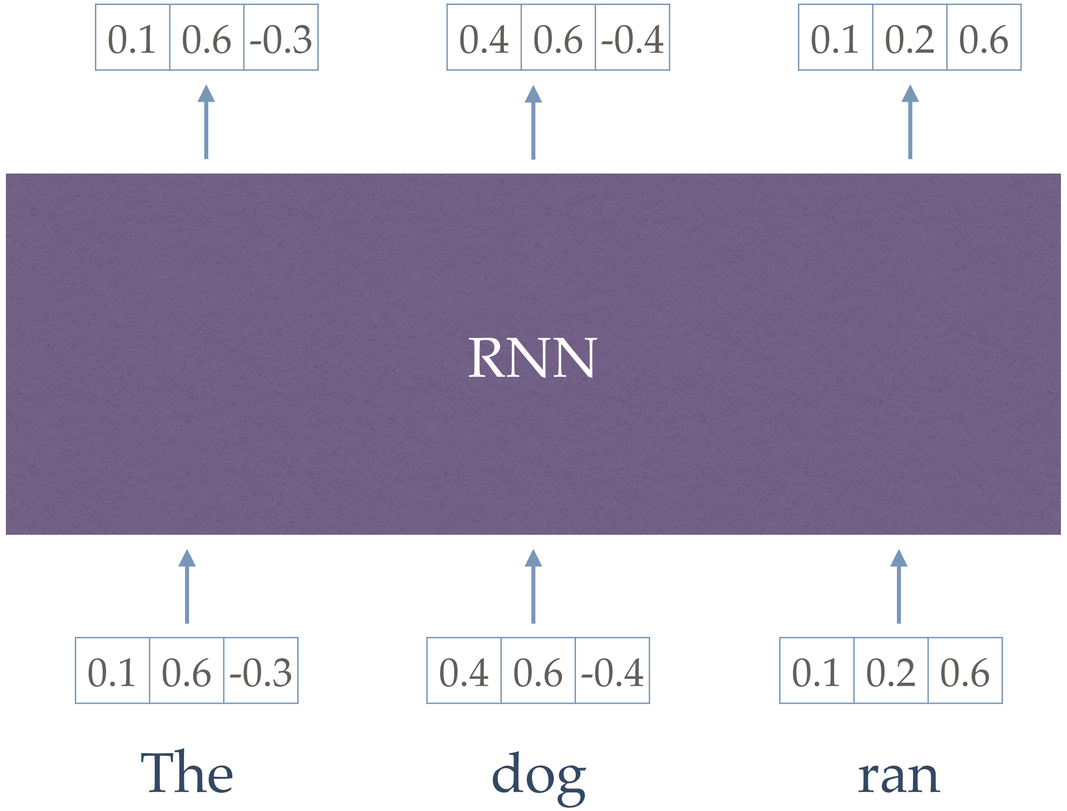
Long short-term Memories (LSTM)



- x -> h through gates
- Control information propagation over long distances
- Downweight unimportant contexts in the past

RNN

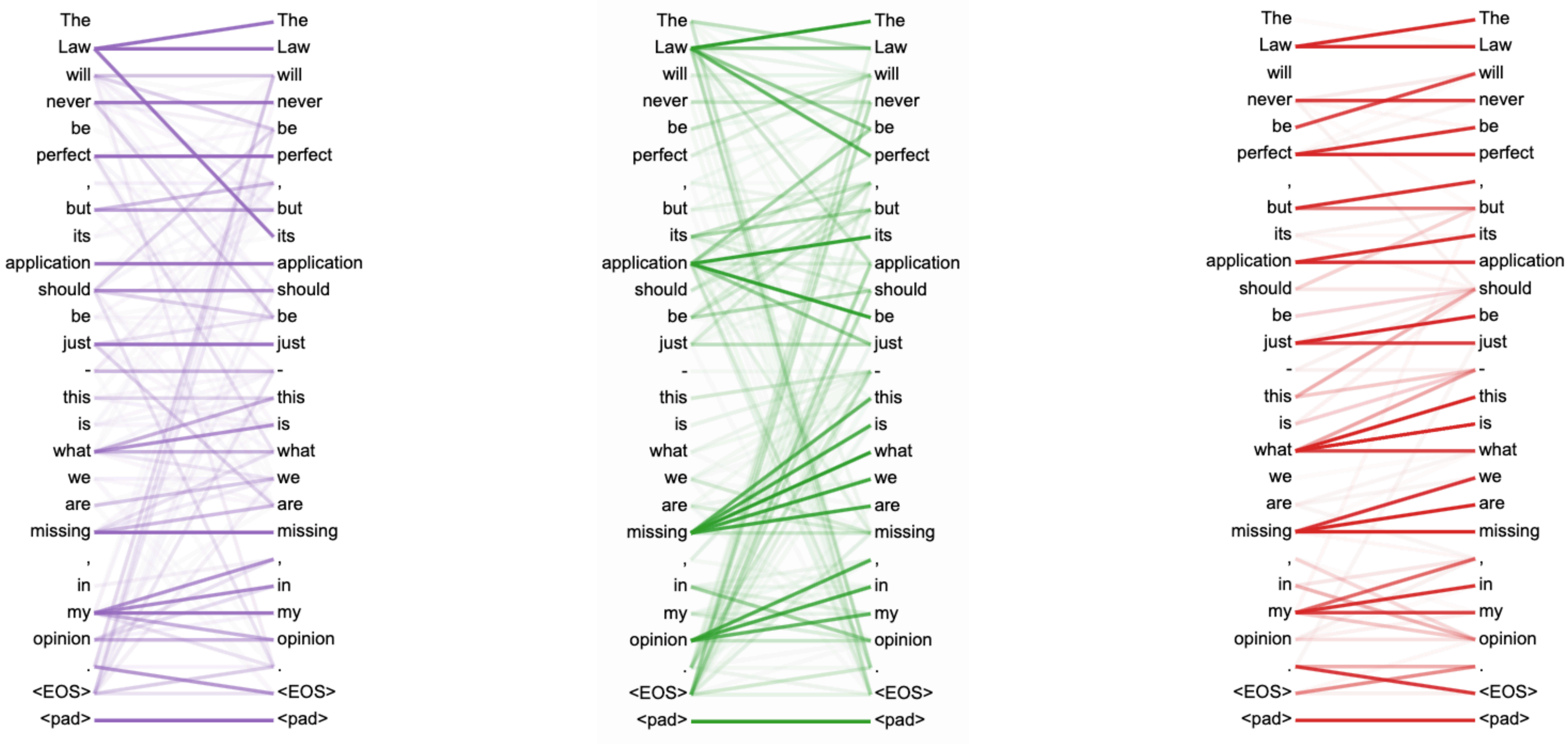

y
RNN
x


0.1 0.6 -0.3
0.4 0.6 -0.4
0.1 0.2 0.6
RNN
0.1 0.6 -0.3
0.4 0.6 -0.4
0.1 0.2 0.6
The
dog
ran

Can we do better?

- Still unresolved:
  - Ability to lookahead
  - Let the model determine which part of the context should be attended or disregarded?
  - Avoid sequential dependency

Attention: Linguistic Intuition



Source: Vaswani et. al. (2017)

Transformer: Setup

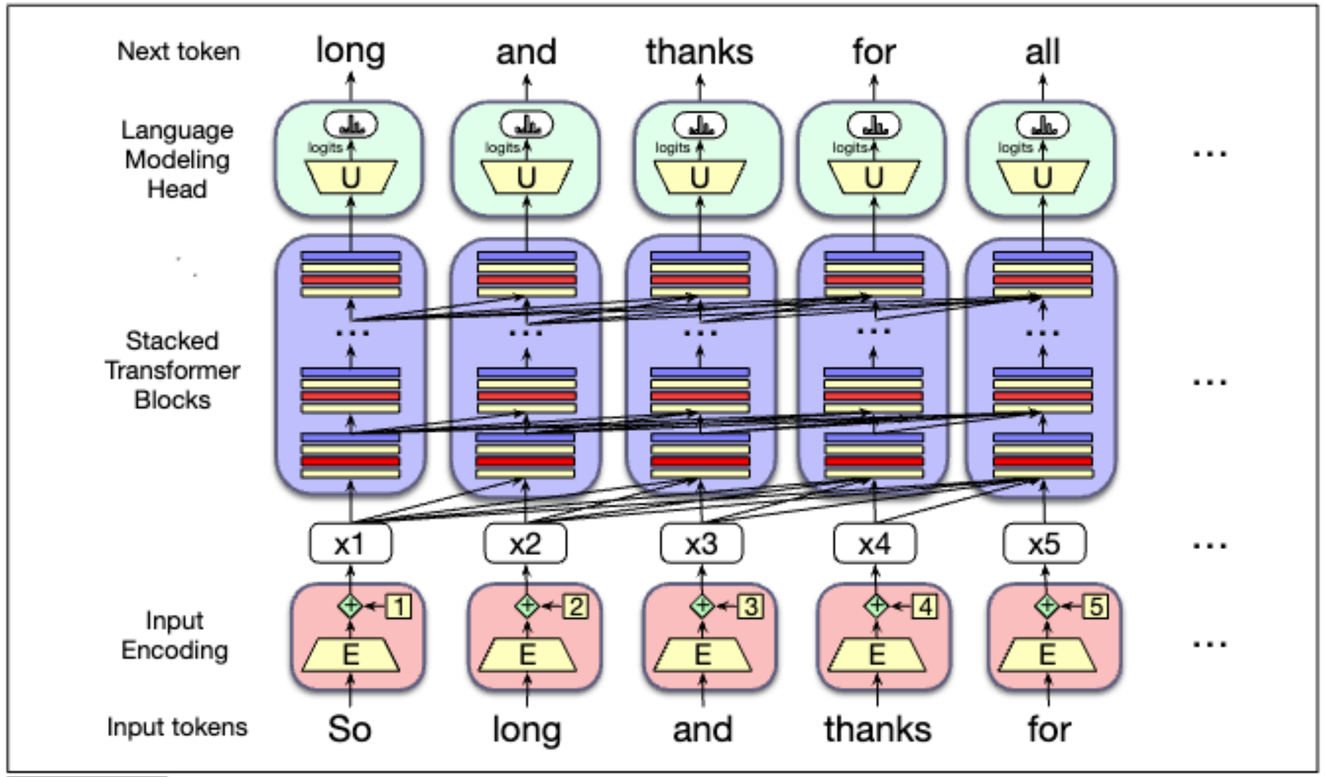


Figure 9.1 The architecture of a (left-to-right) transformer, showing how each input token get encoded, passed through a set of stacked transformer blocks, and then a language model head that predicts the next token.

Transformer: Architecture

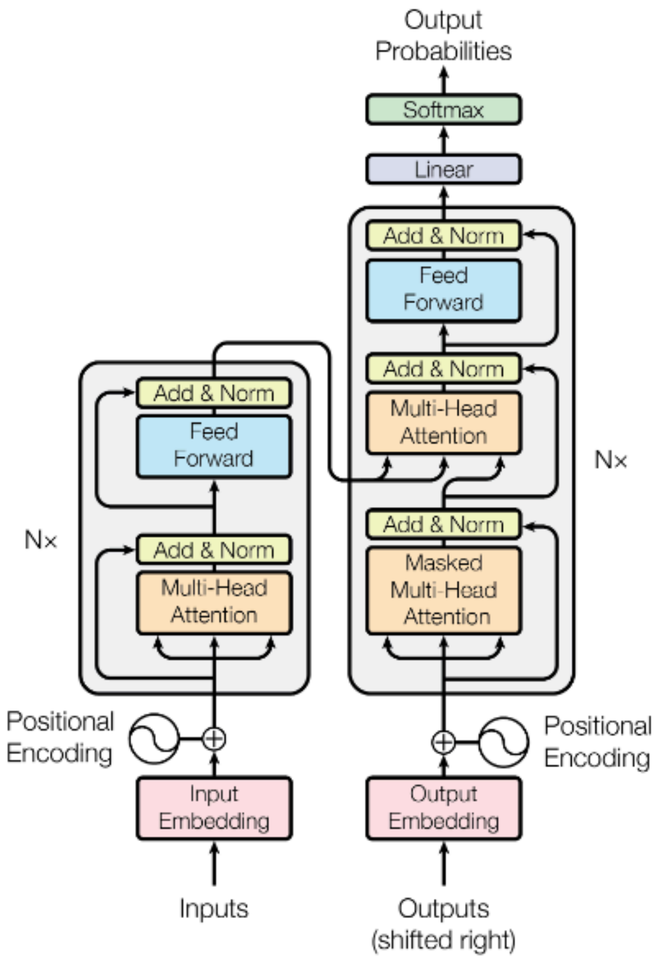
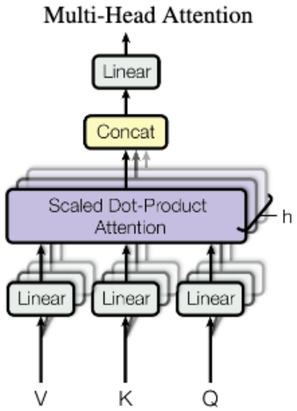
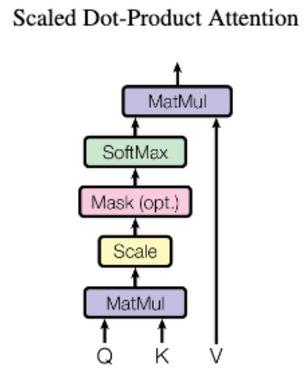


Figure 1: The Transformer - model architecture.





Vaswani et. al. 2017

Attention

- Key Idea: Learn to score the dependency of each linguistic unit in a sequence with every other linguistic unit in the same sequence

Simplest Form of Attention

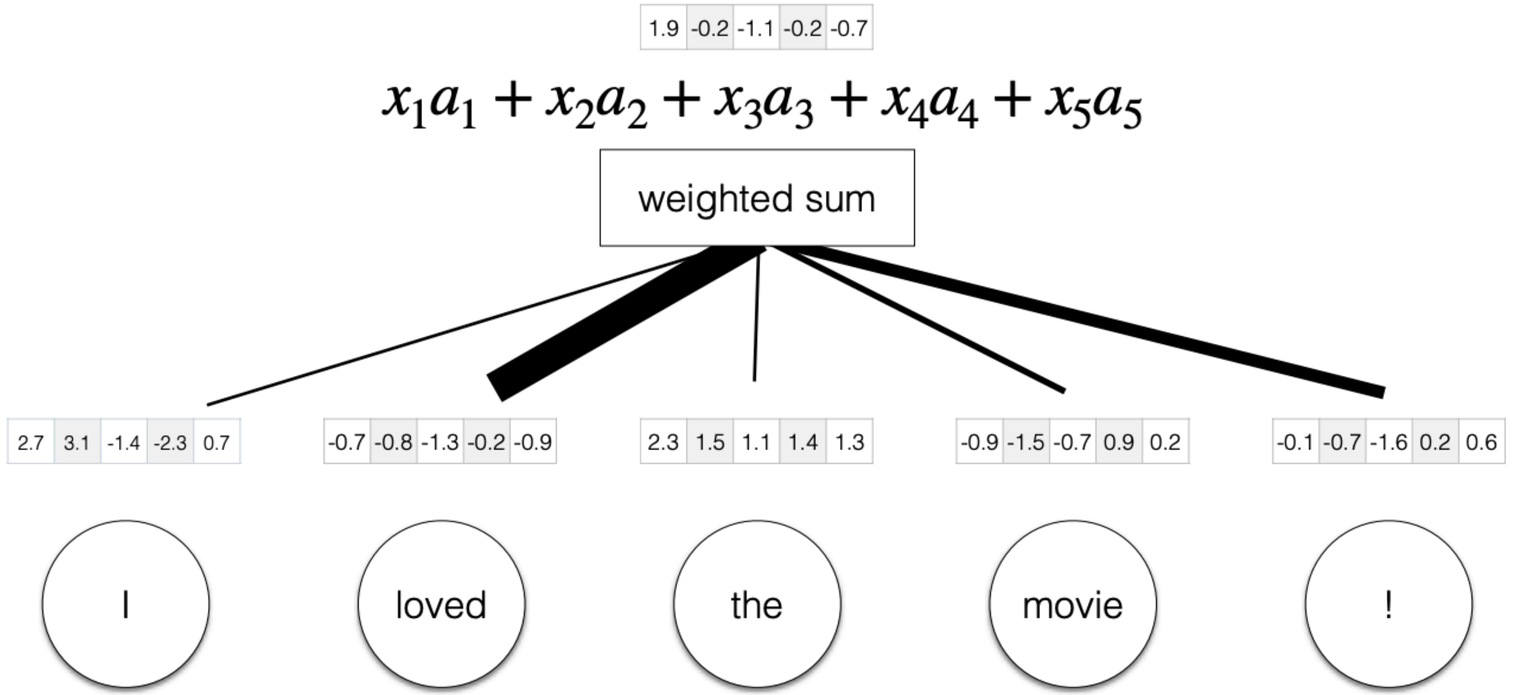
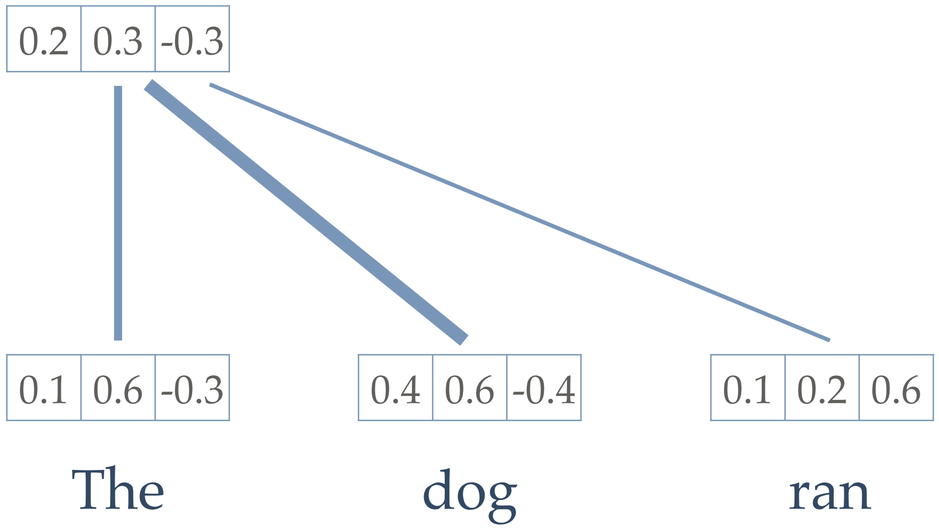


Figure taken from David Bamman's class slides

$$r_i = v^\top x_i$$

$$a = \text{softmax}(r)$$

Self-Attention



- Create a new token vector as a weighted sum of input token vectors
- Weights are the importance or attention to other tokens

0.2	0.3	-0.3	Value
0.1	0.2	-0.3	Key
0.3	0.1	-0.5	Query
0.1	0.6	-0.3	Previous value
The			

$v \in \mathbb{R}^{768} = eW_V$

$k \in \mathbb{R}^{96} = eW_K$

$q \in \mathbb{R}^{96} = eW_Q$

$e \in \mathbb{R}^{768}$

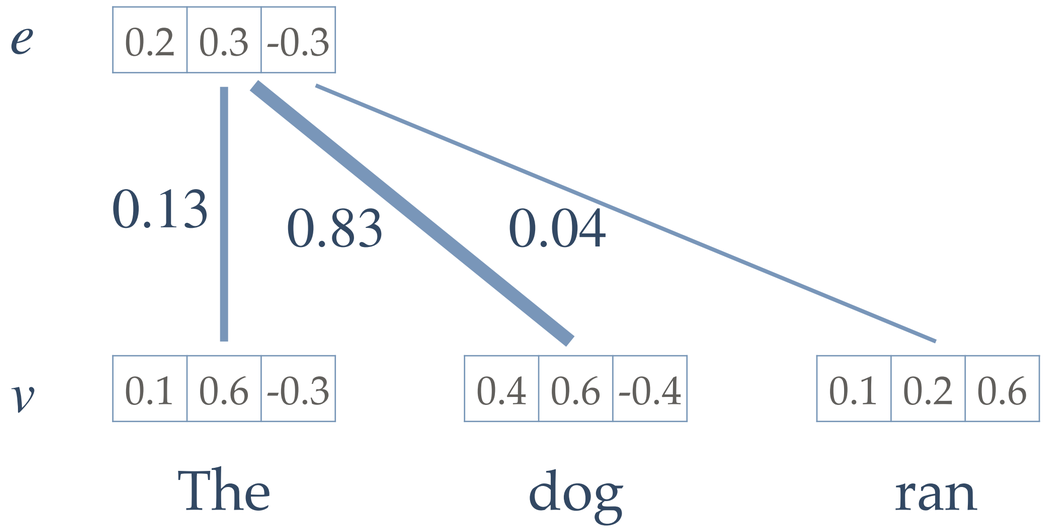
$W_V \in \mathbb{R}^{768\times768}$

$W_K \in \mathbb{R}^{768\times96}$

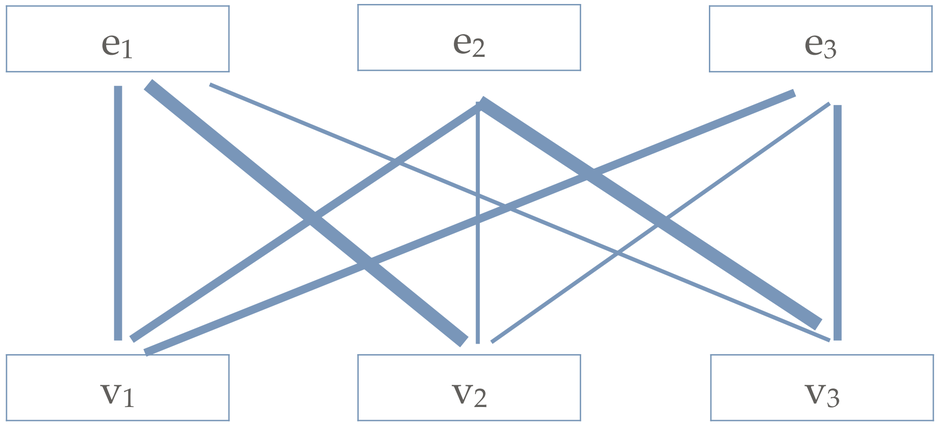
$W_Q \in \mathbb{R}^{768\times96}$

0.13	0.83	0.04	Attention weights	$a_{1j} = \text{softmax}(\text{dot}(q_1, k_j))$
0.3	1.9	−0.7	Compatibility scores	$\text{dot}(q_1, k_j)$

0.1 | 0.6 | -0.3	0.4 | 0.6 | -0.4	0.1 | 0.2 | 0.6
The	dog	ran



- Query and key vectors are used to calculate compatibility or attention scores.
- Attention weights are used to combine value vectors to create a new vector



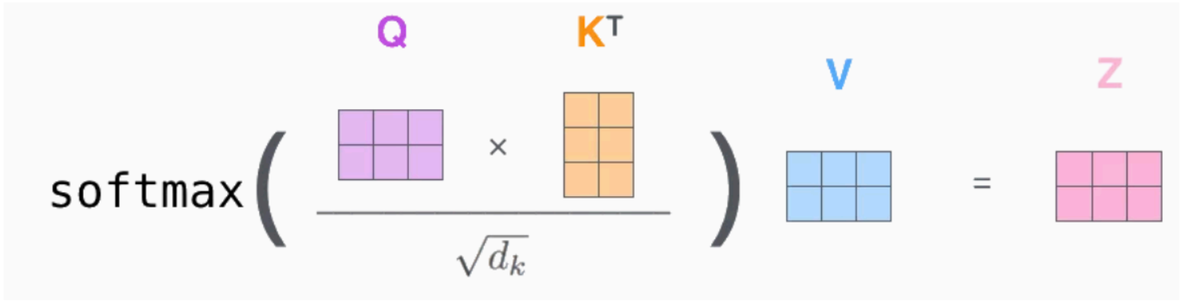
- Compatibility for all the pairs in a sequence is calculated

Self Attention

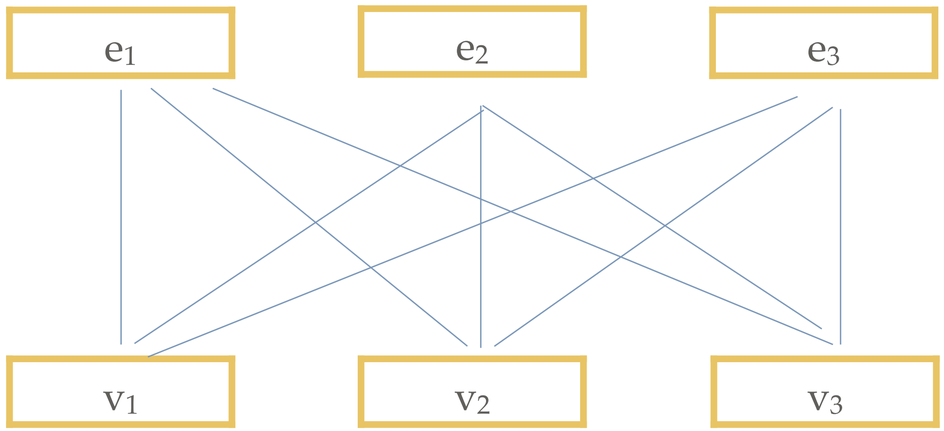
query $Q = XW^Q$

key $K = XW^K$

value $V = XW^V$

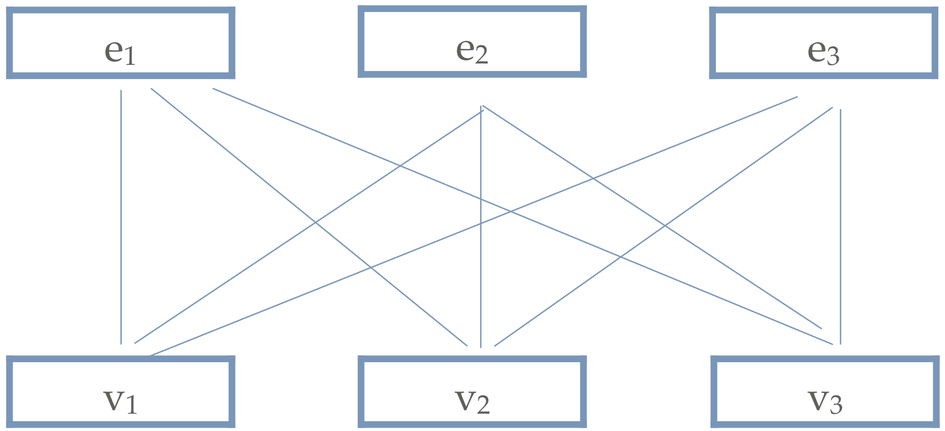


Multi-head Attention



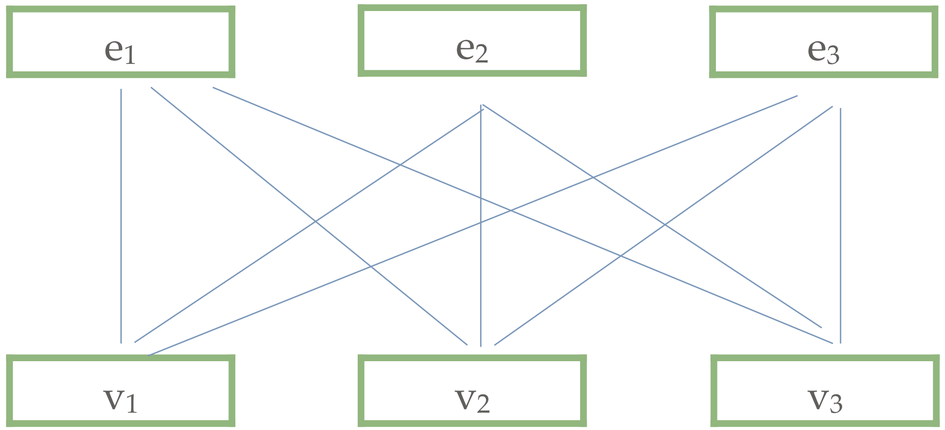
- Each attention head can have its own query, key, and value vectors such that every attention head gives one view of the data

Multi-head Attention



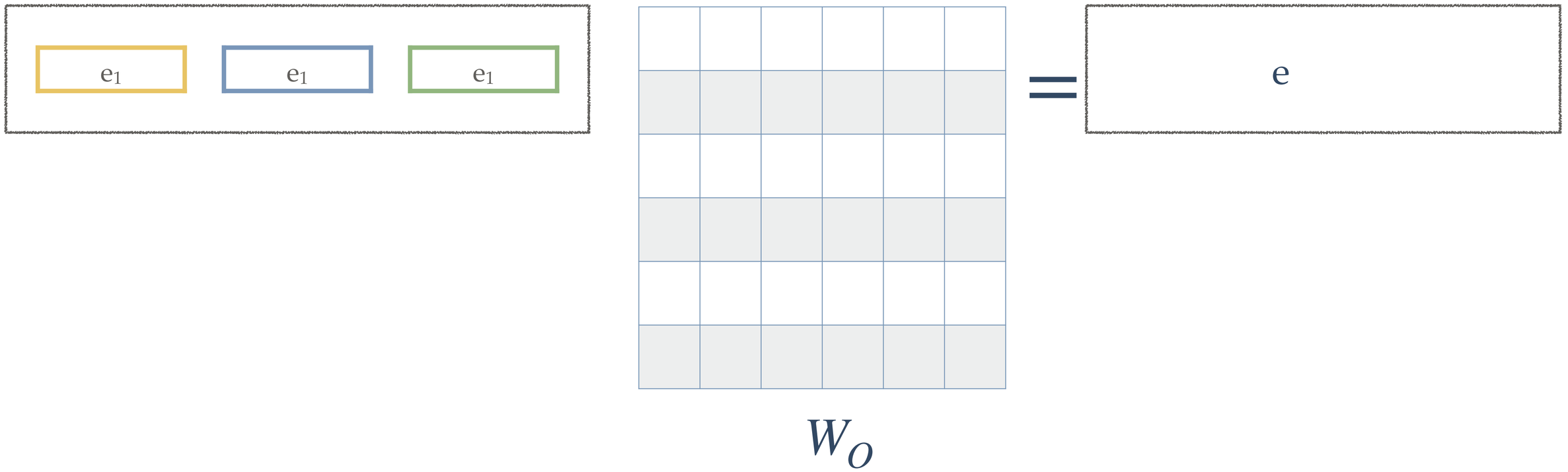
- Each attention head can have its own query, key, and value vectors such that every attention head gives one view of the data

Multi-head Attention

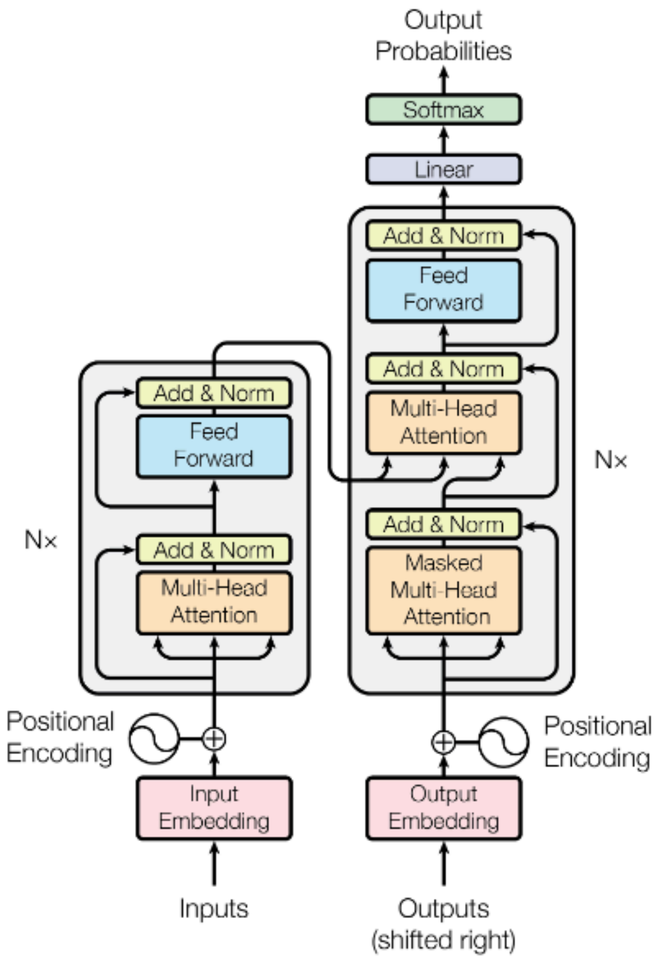


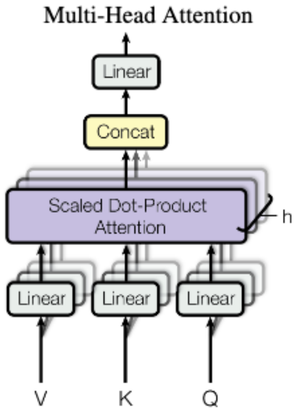
- Each attention head can have its own query, key, and value vectors such that every attention head gives one view of the data

Multi-head Attention



Transformer: Architecture





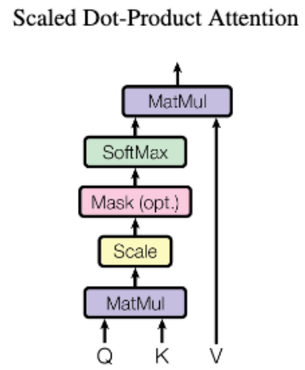
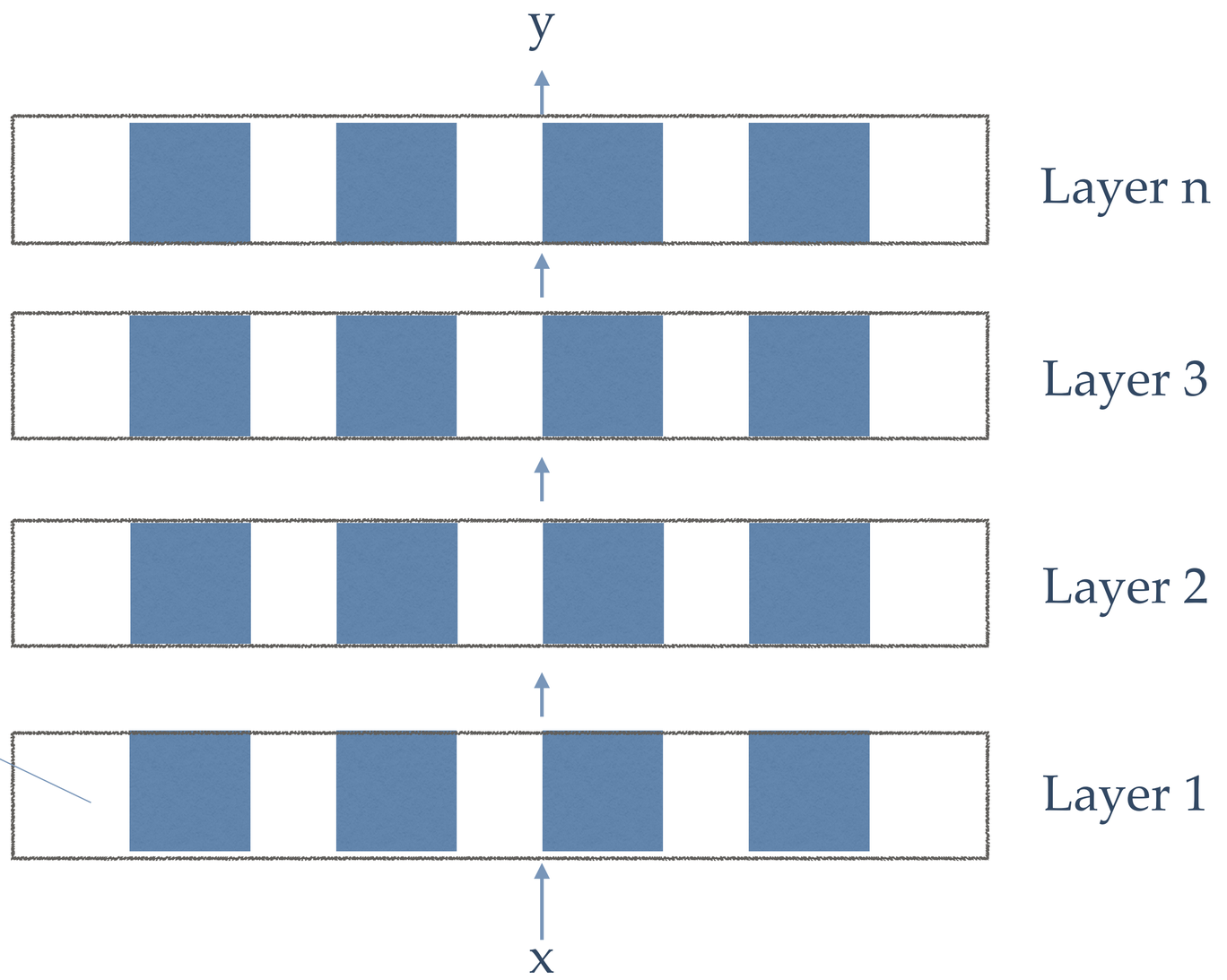


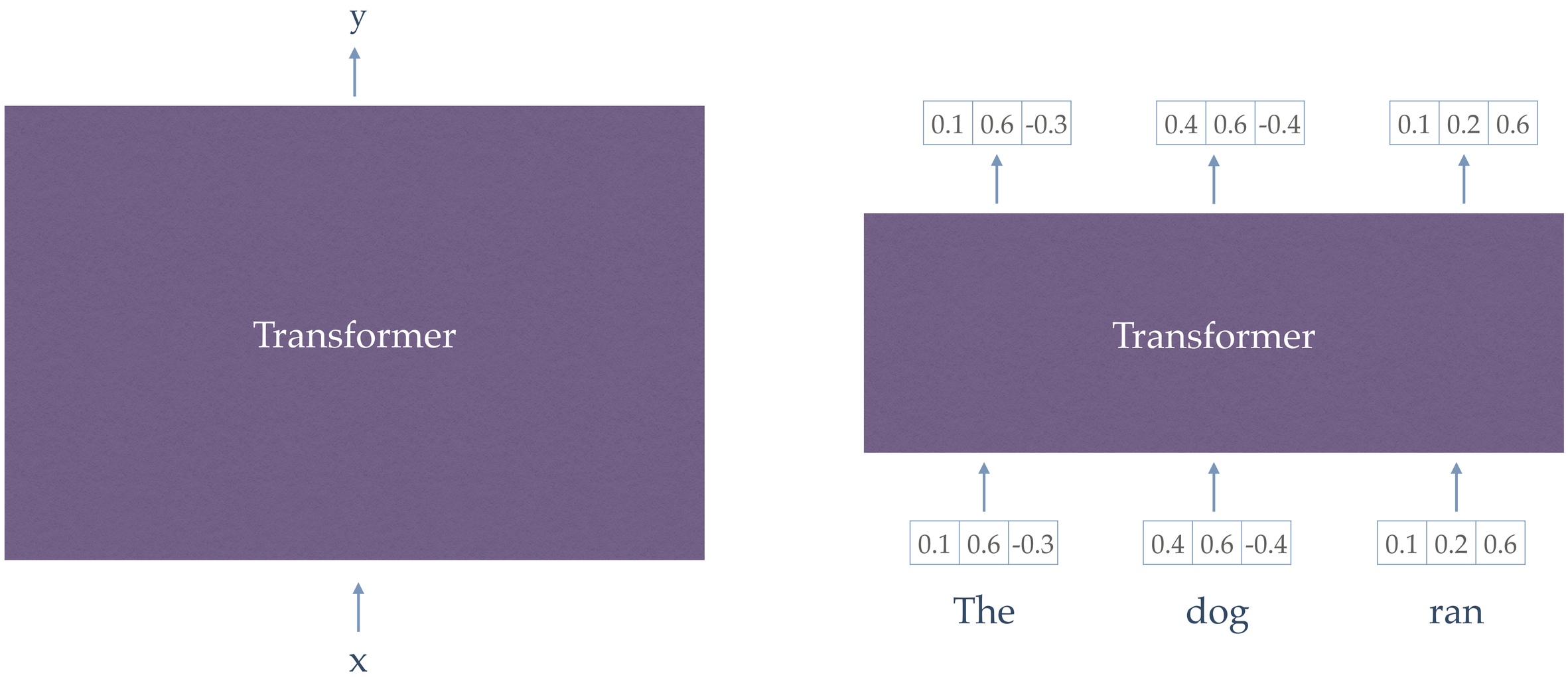
Figure 1: The Transformer - model architecture.

Vaswani et. al. 2017



Self attention units

Transformers



Summary

- Contemporary LLMs (e.g., BERT, GPT, etc) are based on the transformer architecture.
- Many layers, high dimensional vector representations, and large corpora for pretraining are all hallmarks of contemporary LLMs.

In class

- lm exploration