Neural LMs and Transformers

Sandeep Soni

02/29/2024

Proposal guidelines

- Three options (survey, replication, choose your own adventure)
- Proposal should be two pages and include at least 5 references
- One submission per group

Survey

- Objective: Learn everything about a specific topic by reading papers
- Bonus: Some data-driven experiments
- Example: Survey LLMs and copyright use

Survey of Computational Approaches to Lexical Semantic Change Detection

Nina Tahmasebi*
University of Gothenburg

Lars Borin**
University of Gothenburg

Adam Jatowt†
Kyoto University

Our languages are in constant flux driven by external factors such as cultural, societal and technological changes, as well as by only partially understood internal motivations. Words acquire new meanings and lose old senses, new words are coined or borrowed from other languages and obsolete words slide into obscurity. Understanding the characteristics of shifts in the meaning and in the use of words is useful for those who work with the content of historical texts, the interested general public, but also in and of itself.

The findings from automatic lexical semantic change detection, and the models of diachronic conceptual change are also currently being incorporated in approaches for measuring document across-time similarity, information retrieval from long-term document archives, the design of OCR algorithms, and so on. In recent years we have seen a surge in interest in the academic community in computational methods and tools supporting inquiry into diachronic conceptual change and lexical replacement. This article provides a comprehensive survey of recent computational techniques to tackle both.

Replication

- Objective: Learn by replicating a paper
- Bonus: replicate and then extend
- Example: Replicate gender bias paper from Garg et. al. on 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 the proposal with as many details as possible
- 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.
- Language model is 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

Evaluation

- Perplexity: inverse probability of test data, averaged over words
- $2^{-\frac{l(w)}{M}}$, where M is the number of words and

$$l(w) = \sum_{m=1}^{M} \log P(w_m \mid w_{m-1}, \ldots, w_1)$$

- Lower perplexity means model is less surprised

Can we do better than N-gram language models?

Language modeling

- Instead of modeling P(x), why not model P(w|c)?
- Cast language modeling as a self-supervised learning task

Language modeling

Context | Word

May the Force be with | ____

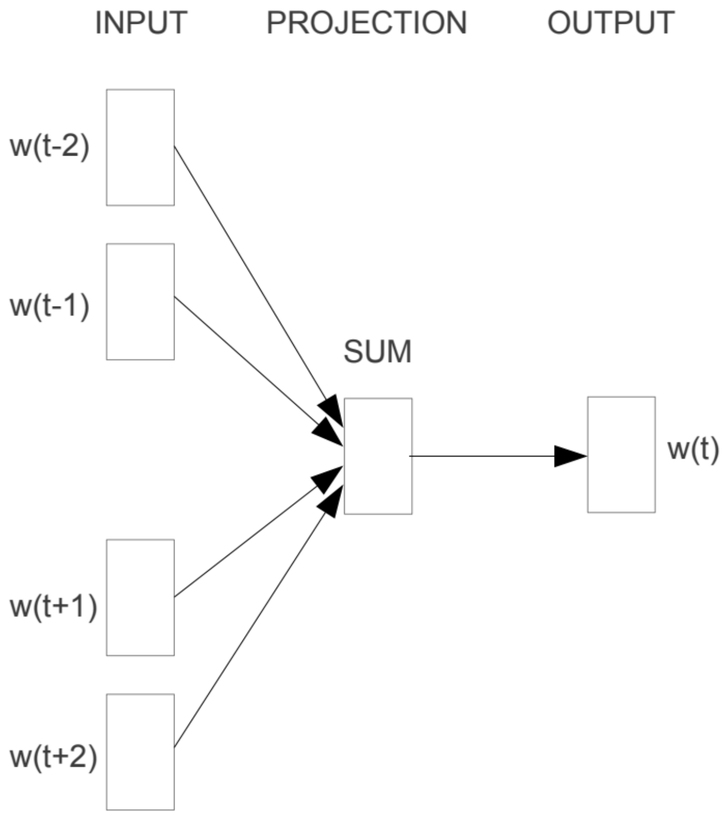
Parametrize $P(w|c) = \frac{\exp(\beta_w v_c)}{\sum_{w'} \exp(\beta_{w'} v_c)}$, where β_w is a dense vector for the word and v_c is the dense vector for context

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



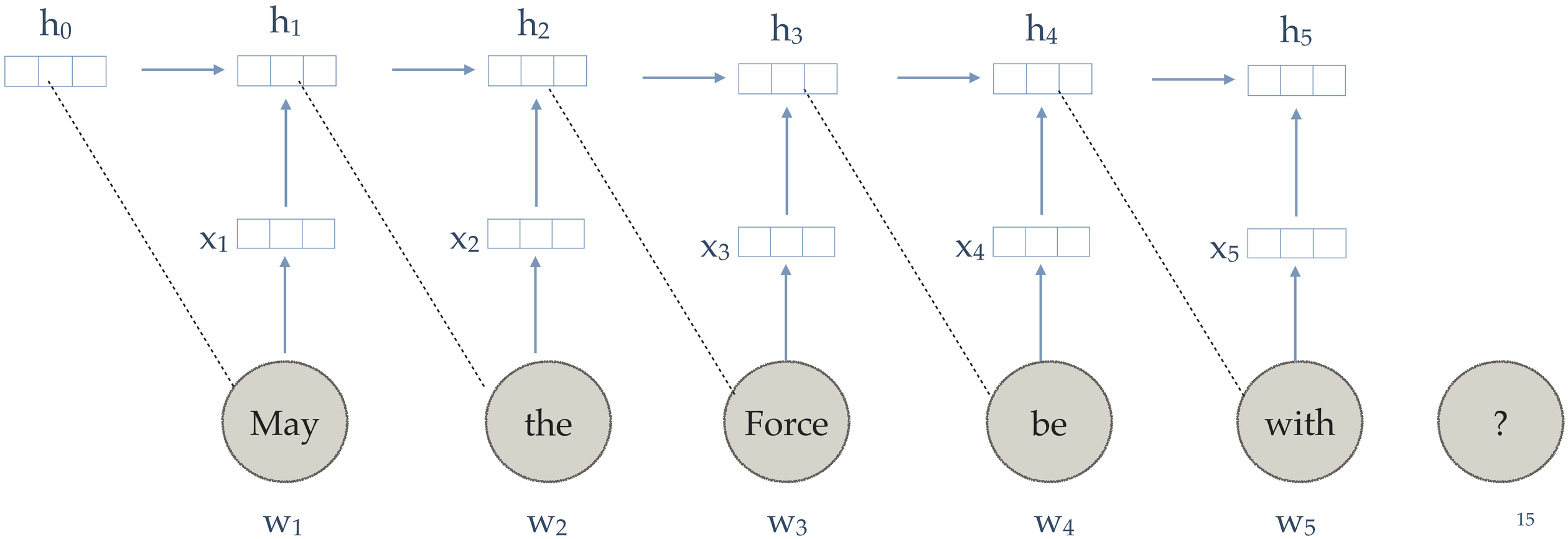
Mikolov et. al. 2013

What can be better ways of coming up with a vector representation of the context?

Context

Word

May the Force be with ____



Recurrent Neural Network LM

At every position m:

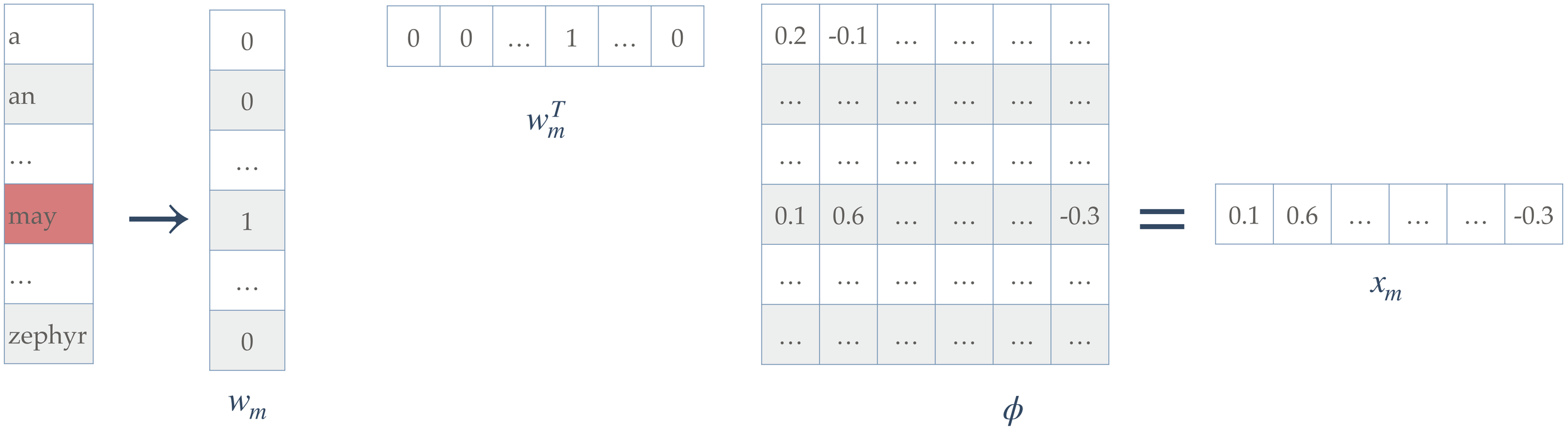
$$x_m = \text{Lookup}(\phi, w_m)$$

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

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

Lookup

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



Recurrent Neural Network LM

At every position m:

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

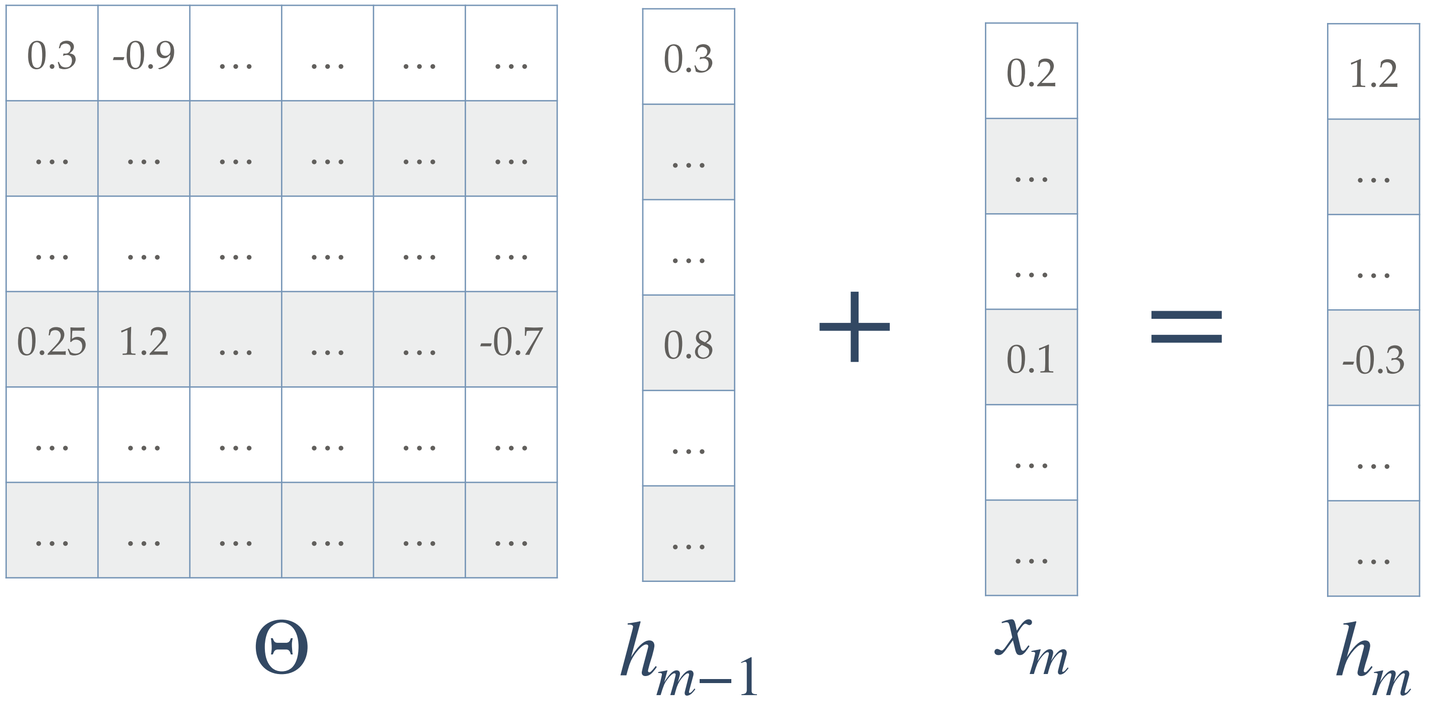
$$h_m = \text{RNN}(x_m, h_{m-1})$$

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

Recurrent Neural Network

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

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



Recurrent Neural Network LM

At every position m:

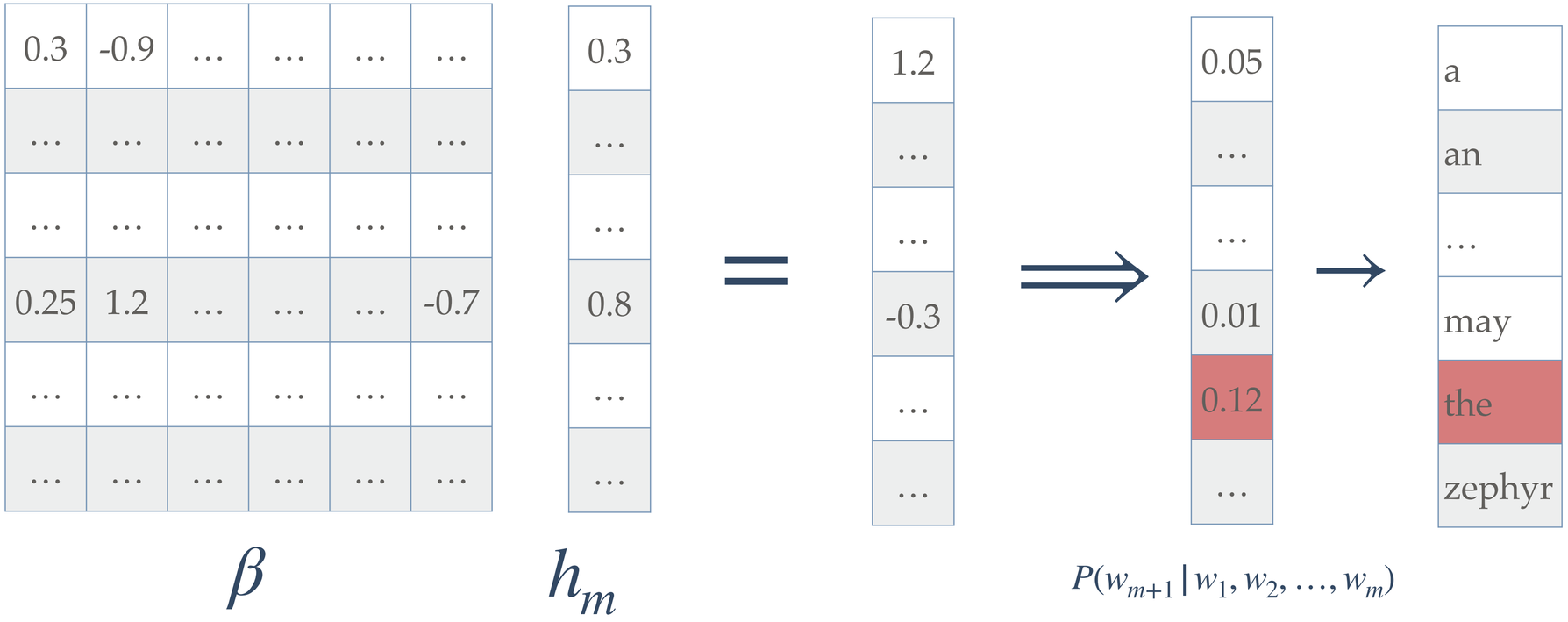
$$x_m = \text{Lookup}(\phi, w_m)$$

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

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

Softmax

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



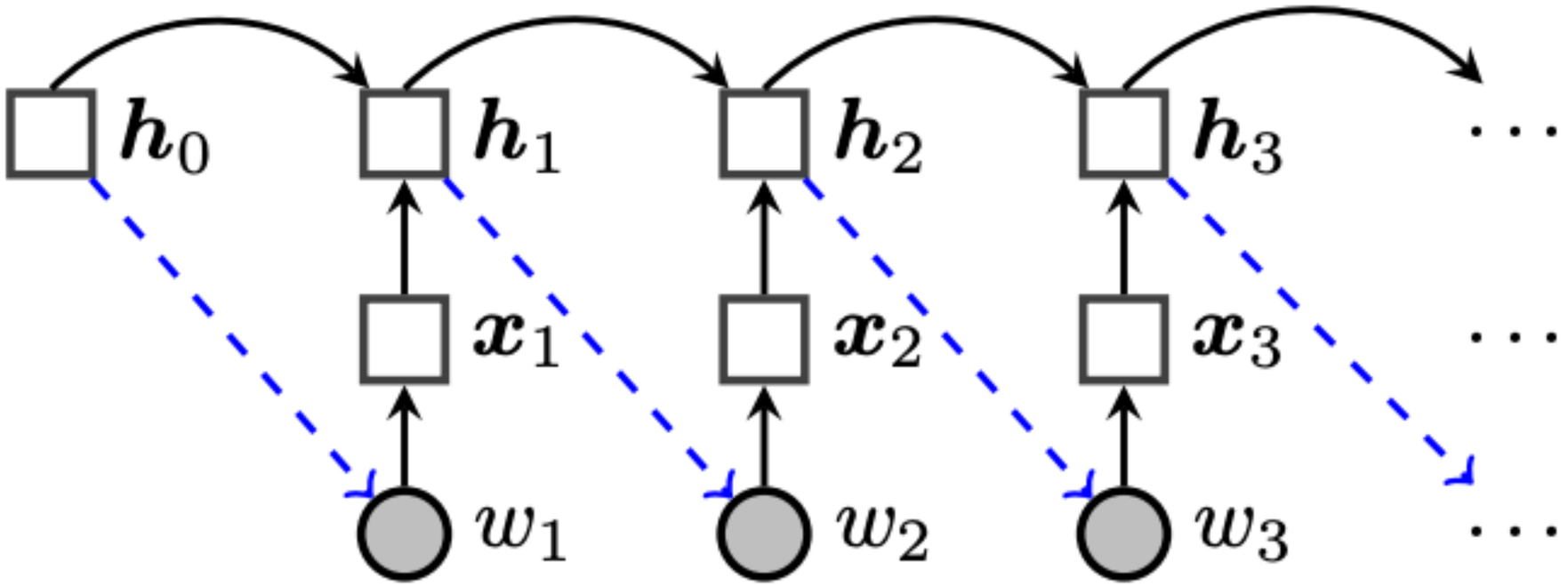
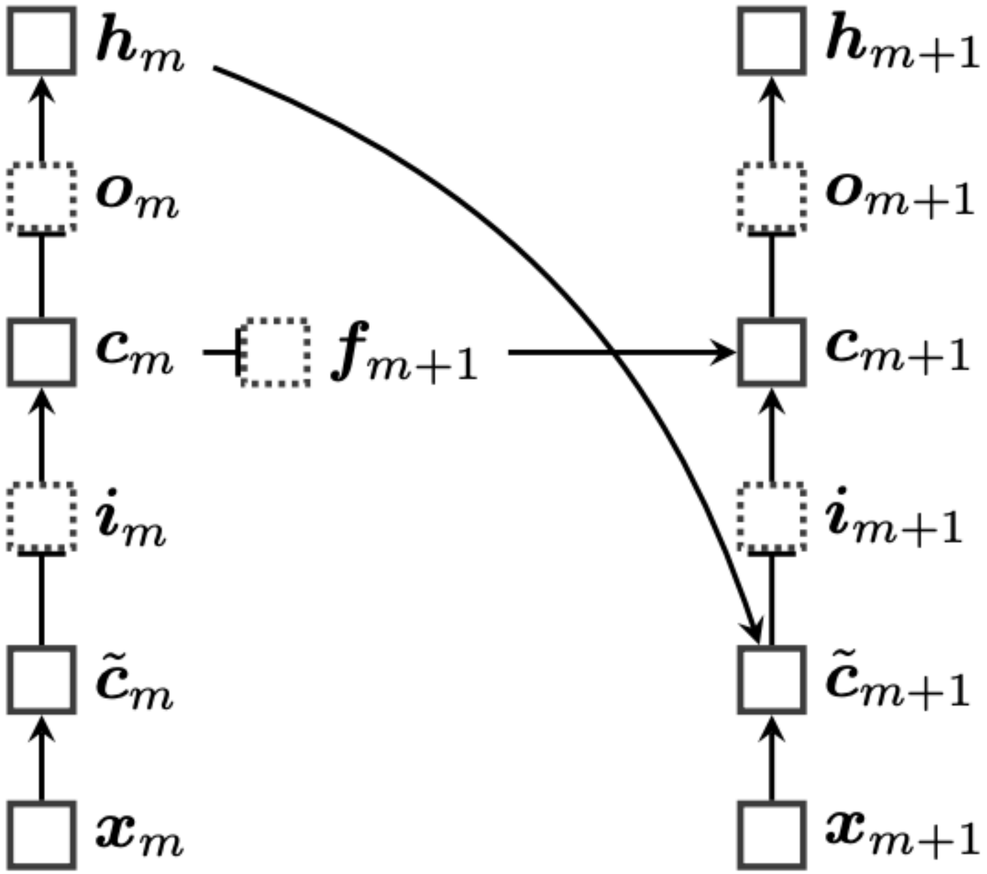


Figure taken from Eisenstein 2018

Long short-term Memories (LSTM)



- Transform x to h by passing x through gating units
- Preserves information propagation over long distances and downweights unimportant contexts in the past

Summary

- Language Modeling is a foundational task in text processing
- Count based language models are easy to interpret but not very powerful for longer sequences
- Neural LMs (RNNs, LSTMs) are extremely powerful

Can we do better?

- Which part of the context is more important or one we should attend?

Attention

- We can calculate a score to each part of a sequence by learning some parameters

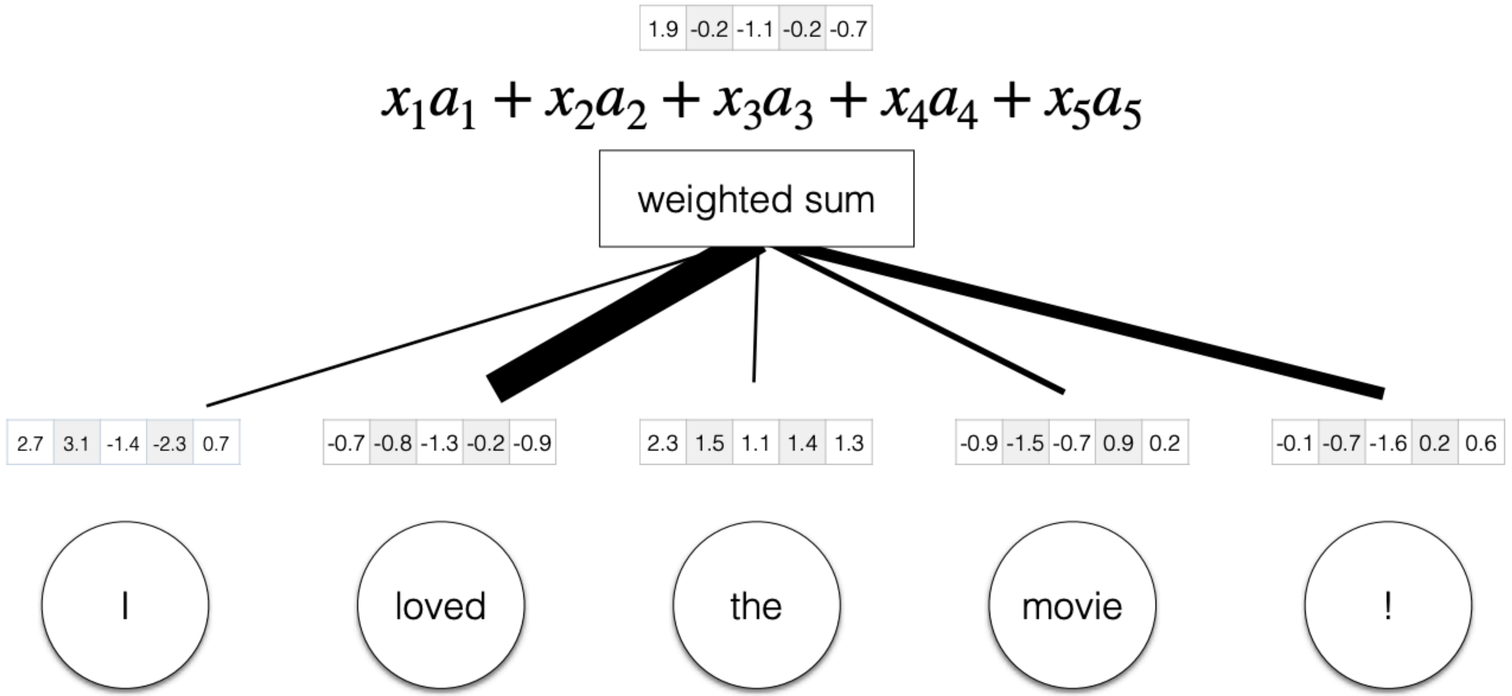


Figure taken from David Bamman's class slides

$$r_i = v^\top x_i$$

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

Self Attention

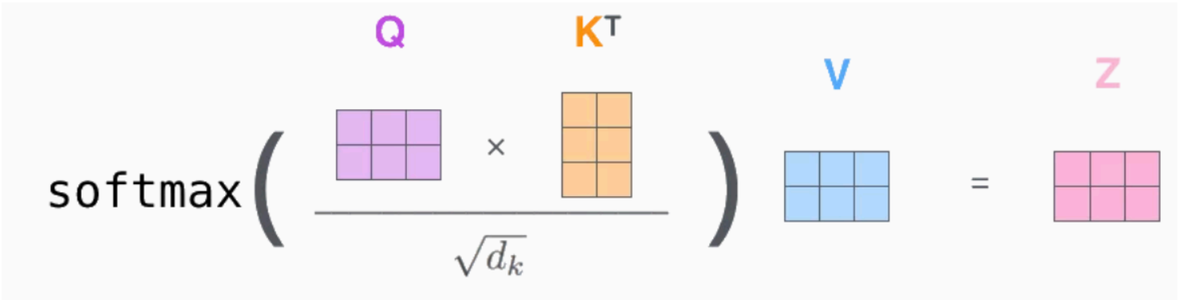
- Given a sequence find importance of every word over every other word in the sequence

Self Attention

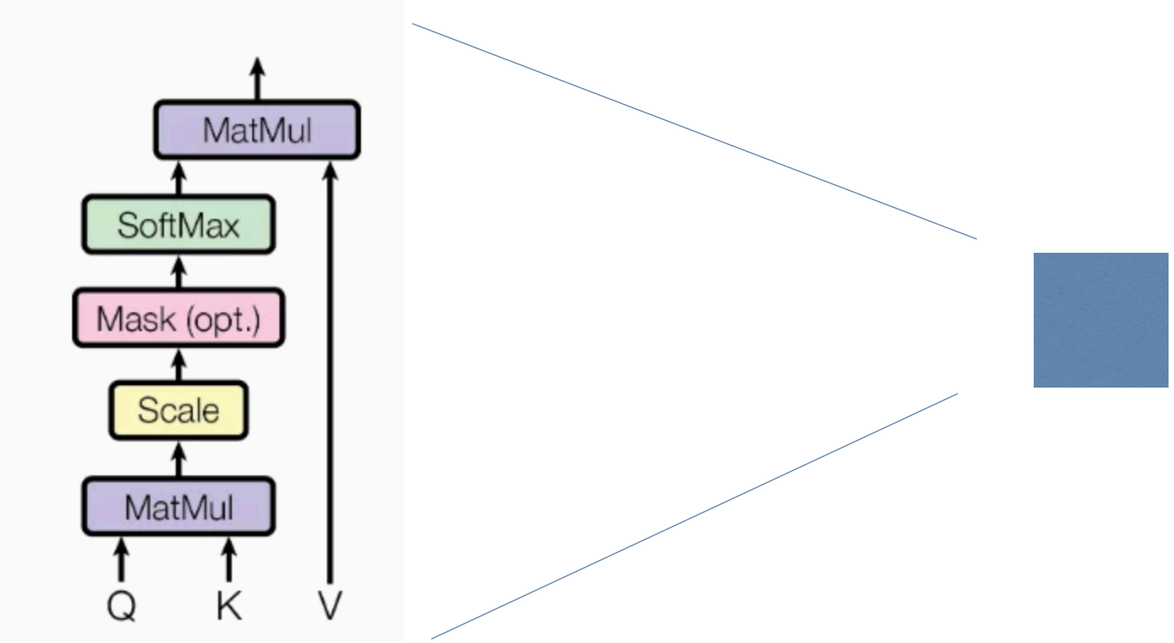
query $Q = XW^Q$

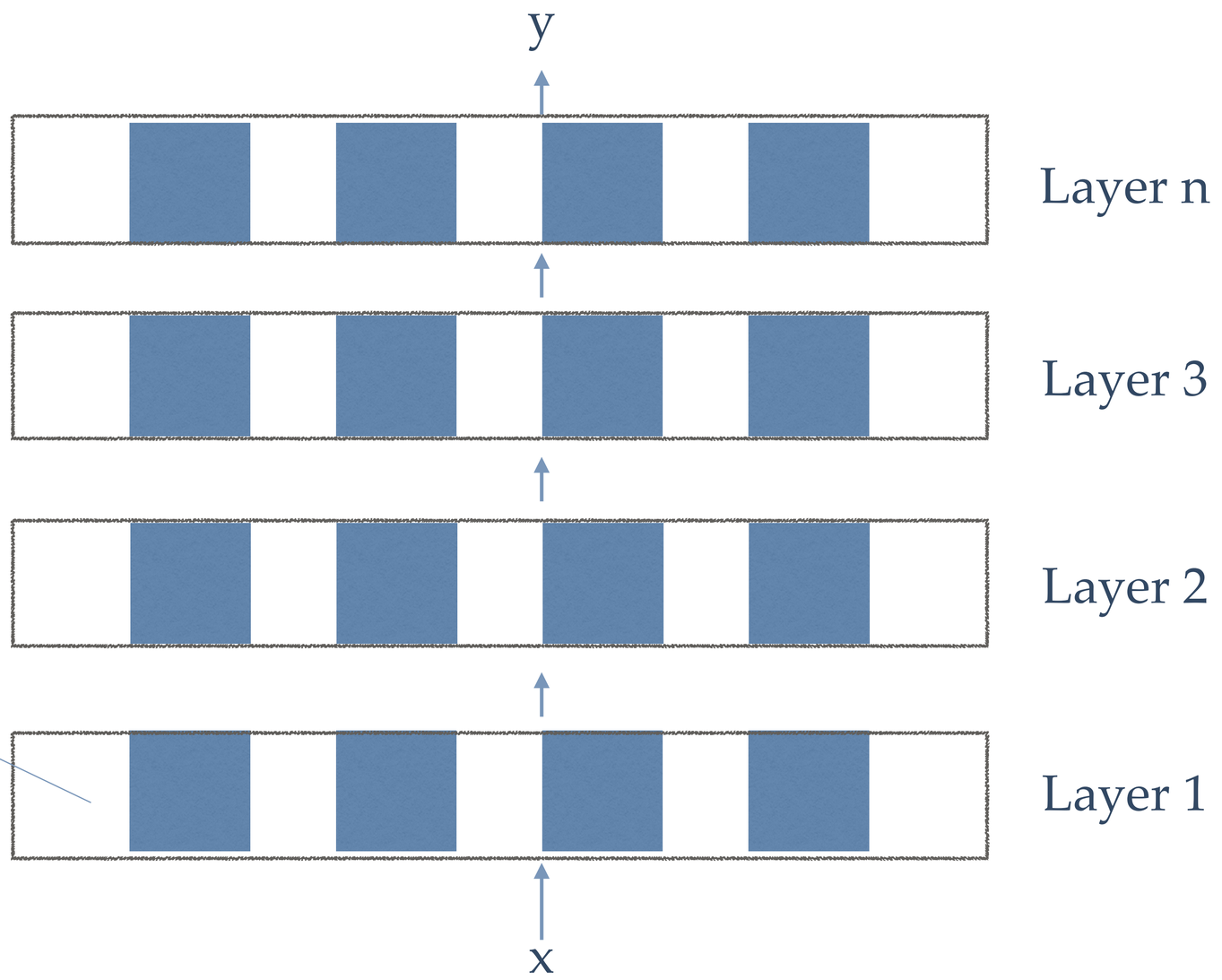
key $K = XW^K$

value $V = XW^V$



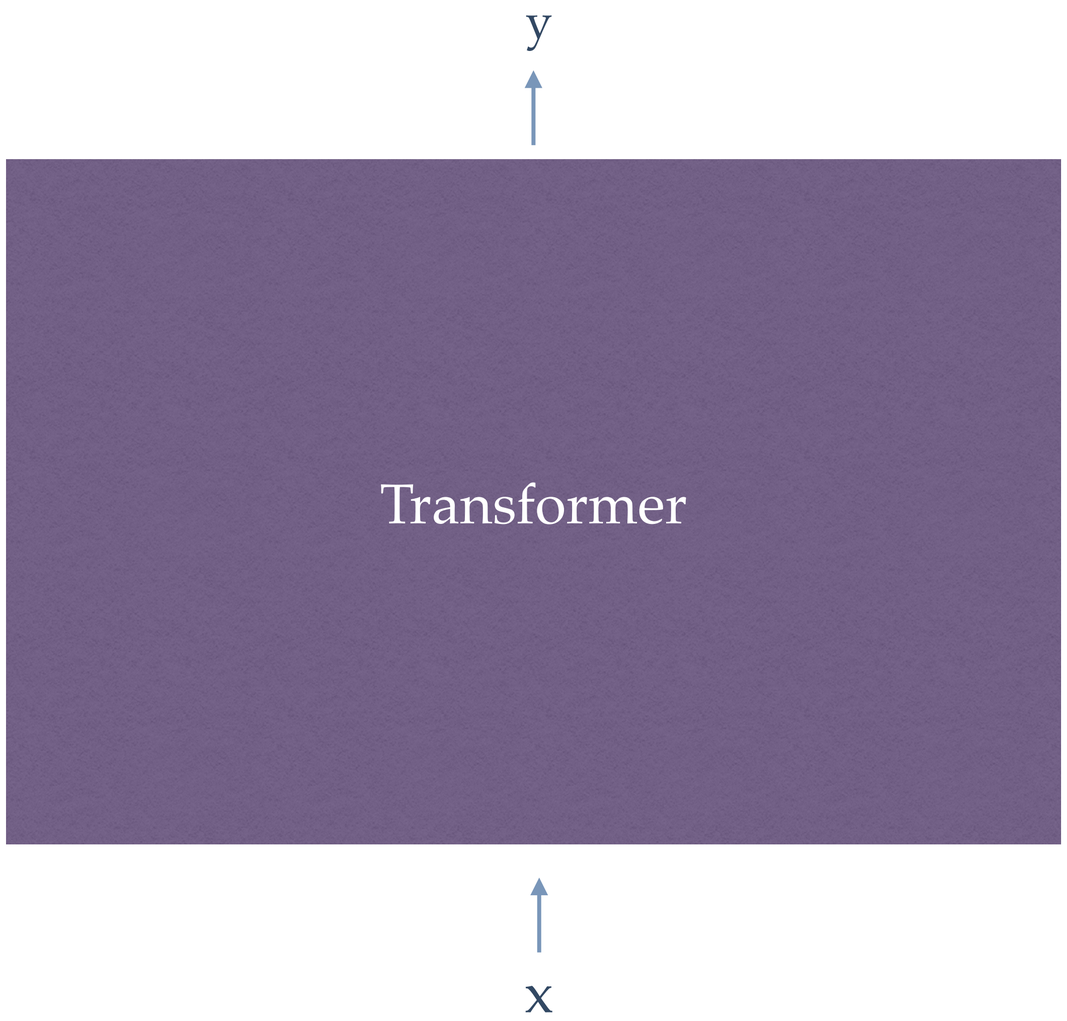
Single head attention





Self attention units
within an attention block

Transformers



LARGE LANGUAGE MODELS

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