Neural networks

Natural language processing - neural network language model

LANGUAGE MODELING

Topics: n-gram model

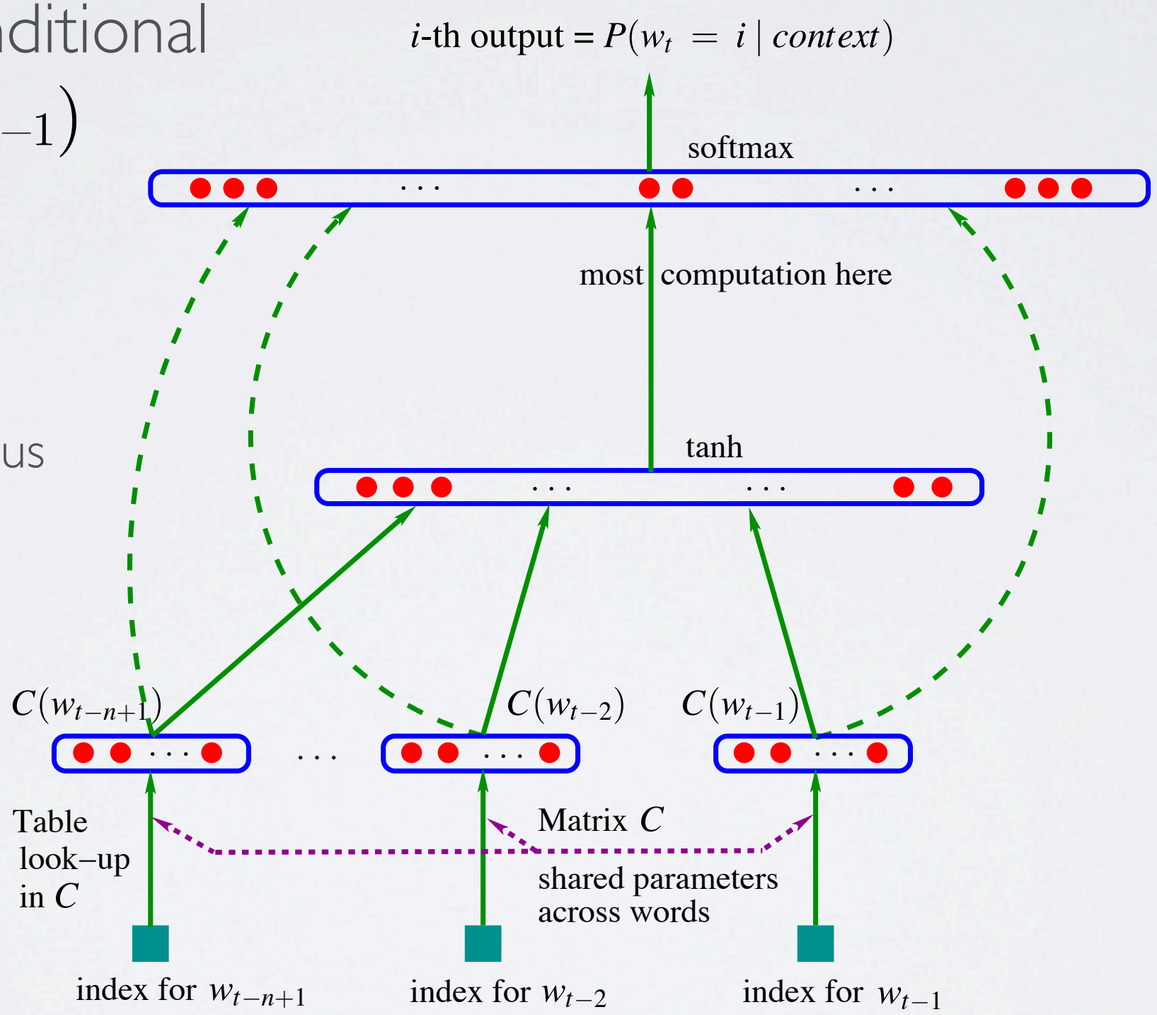
- Issue: data sparsity
  - we want n to be large, for the model to be realistic
  - however, for large values of n, it is likely that a given n-gram will not have been observed in the training corpora
  - smoothing the counts can help
    - combine $\mathbf{count}(w_1\ ,\ w_2\ ,\ w_3\ ,\ w_4)$, $\mathbf{count}(w_2\ ,\ w_3\ ,\ w_4)$, $\mathbf{count}(w_3\ ,\ w_4)$, and $\mathbf{count}(w_4)$ to estimate $p(w_4\,|w_1,\ w_2,\ w_3)$
  - this only partly solves the problem

NEURAL NETWORK LANGUAGE MODEL

Topics: neural network language model

- Solution: model the conditional $p(w_t \mid w_{t-(n-1)}, \ldots, w_{t-1})$ with a neural network
  - learn word representations to allow transfer to n-grams not observed in training corpus

Bengio, Ducharme, Vincent and Jauvin, 2003

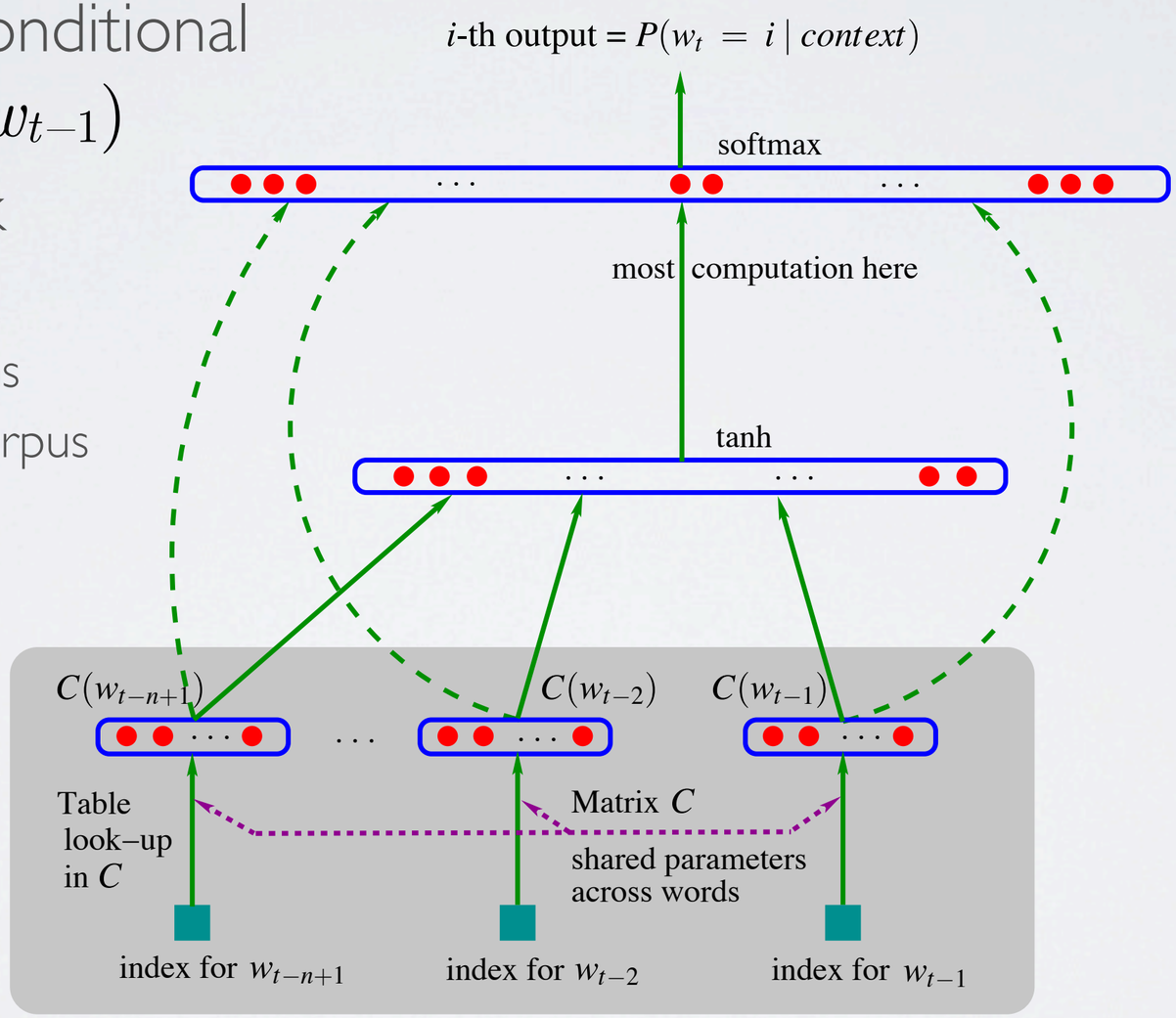


NEURAL NETWORK LANGUAGE MODEL

Topics: neural network language model

- Solution: model the conditional $p(w_t \mid w_{t-(n-1)}, \ldots, w_{t-1})$ with a neural network
  - learn word representations to allow transfer to n-grams not observed in training corpus

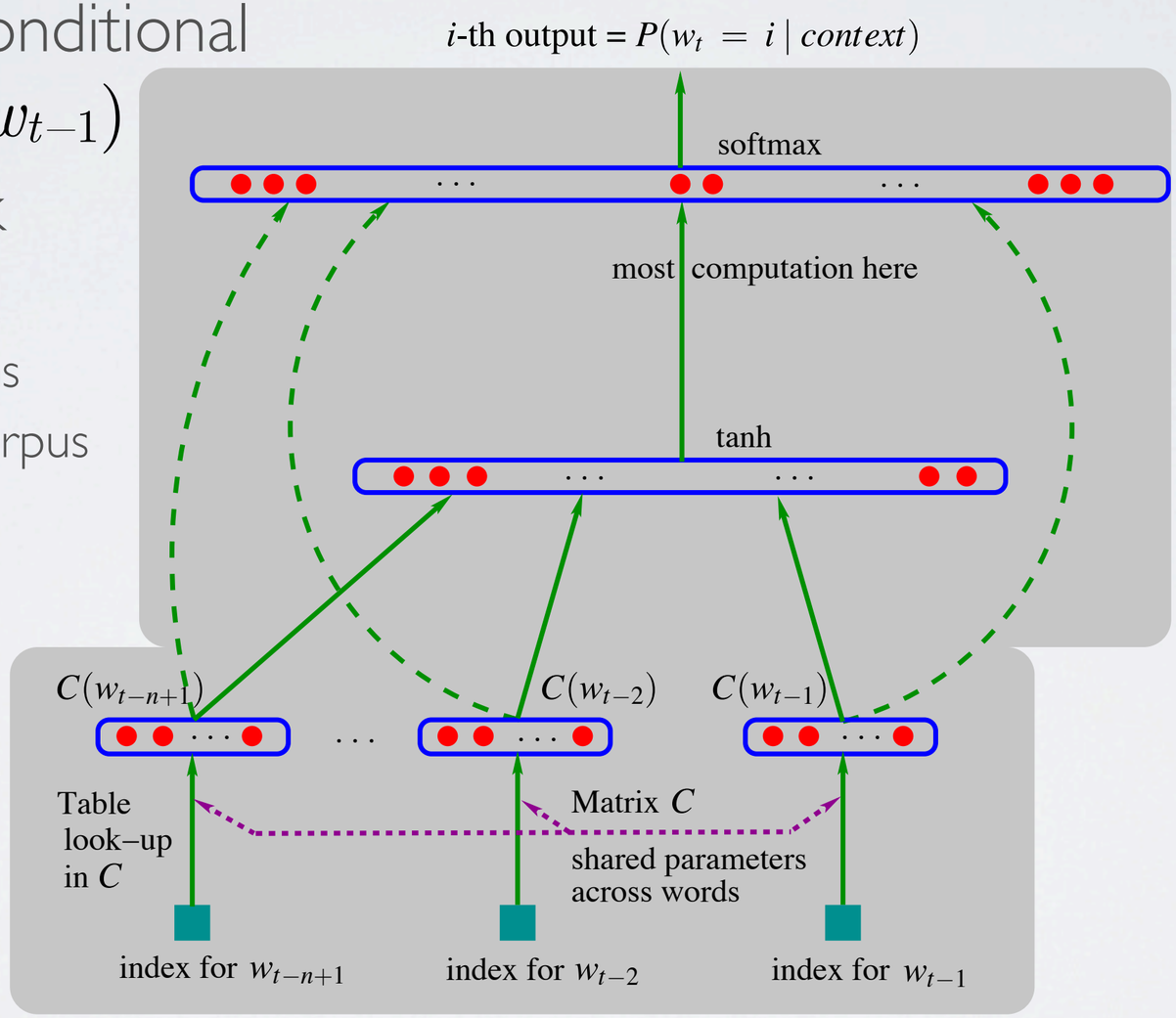
Bengio, Ducharme, Vincent and Jauvin, 2003



NEURAL NETWORK LANGUAGE MODEL

Topics: neural network language model

- Solution: model the conditional $p(w_t \mid w_{t-(n-1)}, \ldots, w_{t-1})$ with a neural network
  - learn word representations to allow transfer to n-grams not observed in training corpus



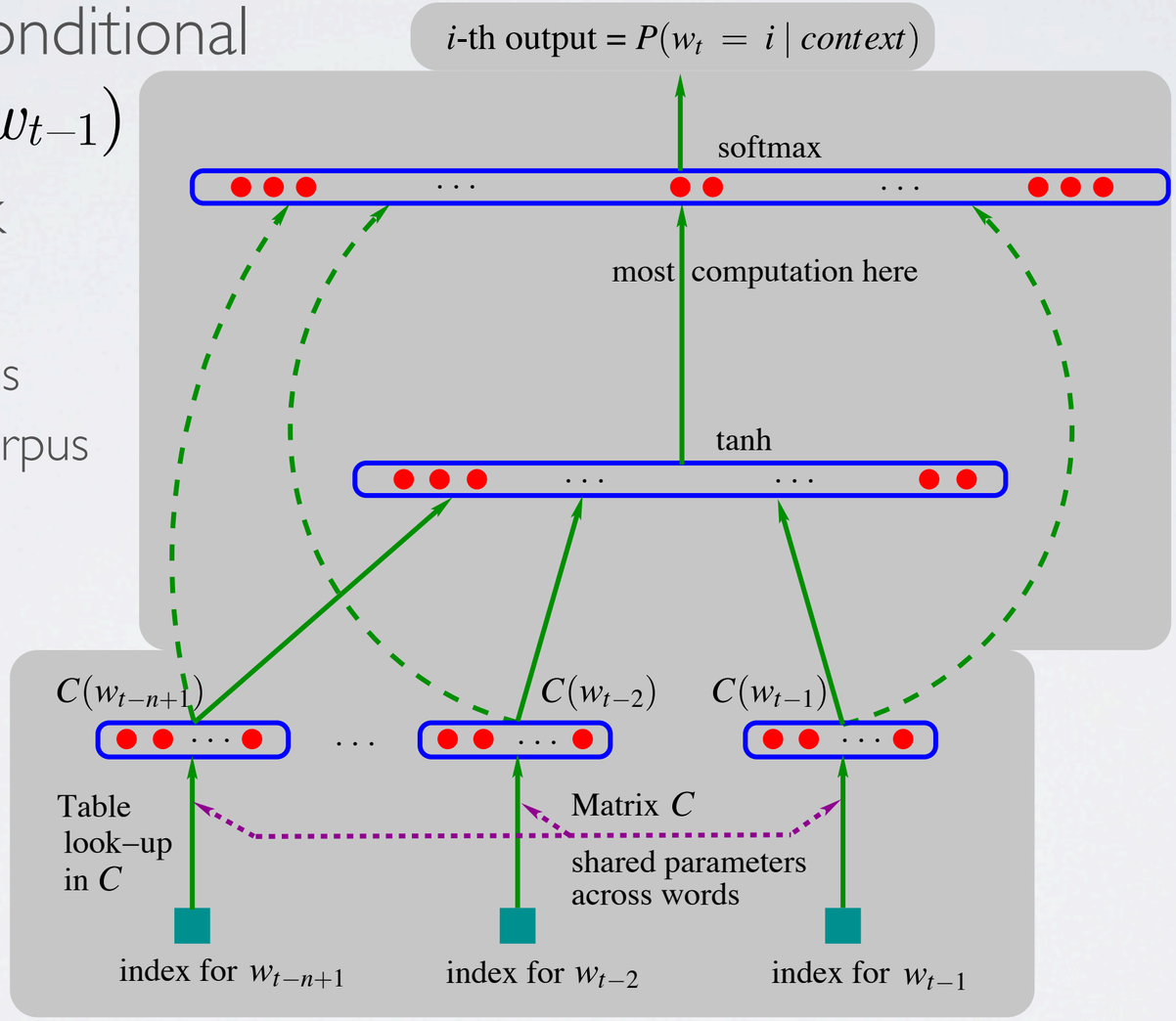
Bengio, Ducharme, Vincent and Jauvin, 2003

NEURAL NETWORK LANGUAGE MODEL

Topics: neural network language model

- Solution: model the conditional $p(w_t \mid w_{t-(n-1)}, \ldots, w_{t-1})$ with a neural network
  - learn word representations to allow transfer to n-grams not observed in training corpus

Bengio, Ducharme, Vincent and Jauvin, 2003



NEURAL NETWORK LANGUAGE MODEL

Topics: neural network language model

- Can potentially generalize to contexts not seen in training set
  - example: $p(\text{“}\texttt{eating}\text{”} \mid \text{“}\texttt{the}\text{”}, \text{“}\texttt{cat}\text{”}, \text{“}\texttt{is}\text{”})$
    - imagine 4-gram [“`the`”,“`cat`”,“`is`”, “`eating`”] is not in training corpus, but [“`the`”,“`dog`”,“`is`”, “`eating`”] is
    - if the word representations of “`cat`” and “`dog`” are similar, then the neural network will be able to generalize to the case of “`cat`”
    - neural network could learn similar word representations for those words based on other 4-grams:

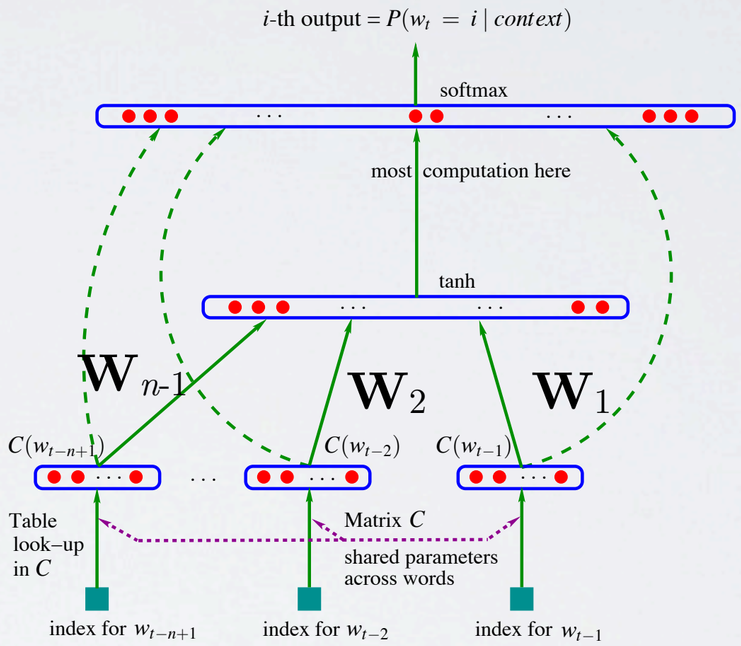
[“`the`”,“`cat`”,“`was`”, “`sleeping`”]
[“`the`”,“`dog`”,“`was`”, “`sleeping`”]

NEURAL NETWORK LANGUAGE MODEL

Topics: word representation gradients

- We know how to propagate gradients in such a network
  - we know how to compute the gradient for the linear activation of the hidden layer $\nabla_{\mathbf{a}(\mathbf{x})} l$
  - let's note the submatrix connecting w_{t-i} and the hidden layer as $\mathbf{W}_i$
- The gradient wrt $C(w)$ for any w is

$$\nabla_{C(w)} l = \sum_{i=1}^{n-1} 1_{(w_{t-i}=w)} \, \mathbf{W}_i^\top \, \nabla_{\mathbf{a}(\mathbf{x})} l$$



Bengio, Ducharme, Vincent and Jauvin, 2003

NEURAL NETWORK LANGUAGE MODEL

Topics: word representation gradients

- Example: [“ `the` ”,“ `dog` ”,“ `and` ”, “ `the` ”, “ `cat` ”]

$$\begin{array}{ccccccccc} w_3 & & w_4 & & w_5 & & w_6 & & w_7 \\ & \searrow & & \searrow & & \searrow & & \searrow & \\ & 21 & & 3 & & 14 & & 21 & \end{array}$$

  - the loss is $l = -\log p(\text{“ } \texttt{cat} \text{ ”} \mid \text{“ } \texttt{the} \text{ ”},\text{“ } \texttt{dog} \text{ ”},\text{“ } \texttt{and} \text{ ”}, \text{“ } \texttt{the} \text{ ”})$
  - $\nabla_{C(3)} l = \mathbf{W}_3^\top \nabla_{\mathbf{a}(\mathbf{x})} l$
  - $\nabla_{C(14)} l = \mathbf{W}_2^\top \nabla_{\mathbf{a}(\mathbf{x})} l$
  - $\nabla_{C(21)} l = \mathbf{W}_1^\top \nabla_{\mathbf{a}(\mathbf{x})} l + \mathbf{W}_4^\top \nabla_{\mathbf{a}(\mathbf{x})} l$
  - $\nabla_{C(w)} l = 0$ for all other words w

- Only need to update the representations $C(3)$, $C(14)$ and $C(21)$,

NEURAL NETWORK LANGUAGE MODEL

Topics: performance evaluation

- In language modeling, a common evaluation metric is the perplexity
  - it is simply the exponential of the average negative log-likelihood
- Evaluation on Brown corpus
  - n-gram model (Kneser-Ney smoothing): **321**
  - neural network language model: **276**
  - neural network + n-gram: **252**

Bengio, Ducharme, Vincent and Jauvin, 2003

NEURAL NETWORK LANGUAGE MODEL

Topics: performance evaluation

- A more interesting (and less straightforward) way of evaluating a language model is within a particular application
  - does a language model improve the performance of a machine translation or speech recognition system

- Later work has shown improvements in both cases
  - Connectionist language modeling for large vocabulary continuous speech recognition
    Schwenk and Gauvain, 2002
  - Continuous-Space Language Models for Statistical Machine Translation
    Schwenk, 2010