Neural Network Language Models

Nicholas Dronen
ndronen@gmail.com

May 29, 2014

Pearson

FT FINANCIAL TIMES

The Economist

PEARSON Addison Wesley

PEARSON Longman

PEARSON Education

PEARSON Assessments

Language is unstructured

Supervised learning

$\mathbb{X}$ Training matrix

$\mathbb{Y}$ Target vector

$\boldsymbol{x}$ Feature vector

$\hat{\boldsymbol{y}}$ Prediction

$$\mathbb{F}(\mathbb{X}, \mathbb{Y}) \to (\boldsymbol{f}(\boldsymbol{x}) \to \hat{\boldsymbol{y}})$$

Language models (LMs)

- Language as a sequence
- Tasks
    - Training
        - Learn joint probability of word given context
    - Predicting
        - Compute probability of word and context
        - Generate sequences

LMs are unsupervised

$\mathbb{X}$ Sequences of words

$\boldsymbol{x}$ A sequence of words $\boldsymbol{w}\ \boldsymbol{w}_{\boldsymbol{n}-1}\ldots$

$\hat{\boldsymbol{P}}(\boldsymbol{x})$ Probability of $\boldsymbol{w}$ given context

$$\mathbb{F}(\mathbb{X}) \to (\boldsymbol{f}(\boldsymbol{x}) \to \hat{\boldsymbol{P}}(\boldsymbol{x}))$$

N-grams

Unigrams: N-grams with N=1.

$$P(w_n) = \frac{C(w_n)}{\sum_w C(w)}$$

N-grams

Bigrams: N-grams with N=2.

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

Trigrams, quadrigrams, ...

N-gram example

this is the way the world ends

this is the way the **world ends**
this is the way the **ladies ride**
this is the way the **world ends dexter**
this is the way the **bunny hops song**

About 376,000,000 results (0.23 seconds)

not with a bang but a whimper

not with a **bang but a whimper**
not with a **club the heart is broken**
not with a **whisper but with a bang**
not with a **bang damon knight**

About 963,000 results (0.31 seconds)

N-gram example

Phrase	Counts (Google)
This is the way the world	2,750,000
This is the way the world ***ends***	1,920,000
This is the way the world ***is***	537,000
This is the way the world ***turns***	319,000
This is the way the world ***works***	211,000
This is the way the world ***bling***	0

N-gram example (cont.)

$$\begin{aligned} P(w|w_{n-1}\ldots) &= P(\textit{ends} \mid \textit{this is the way the world}) \\ &= 1,920,000/2,750,000 \\ &= 0.69 \end{aligned}$$

Evaluating LMs

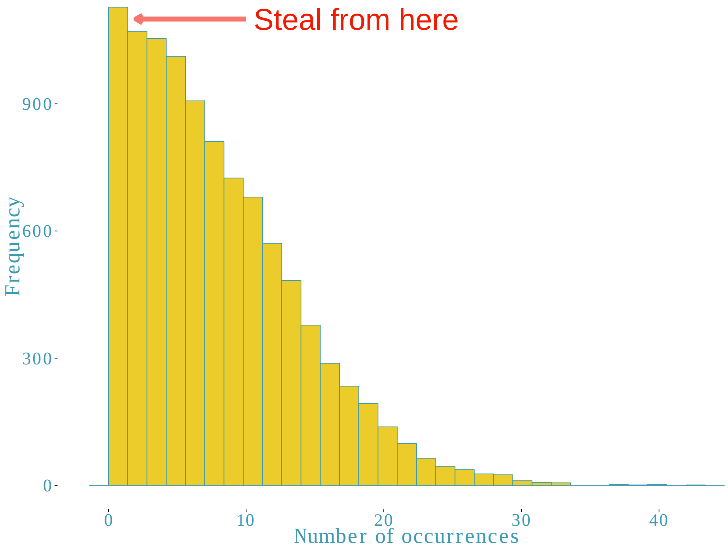
Type	Loss function
Regressor	$\frac{1}{N}\sum(\boldsymbol{y}-\hat{\boldsymbol{y}})^2$
Classifier	$\frac{1}{N}\sum \boldsymbol{I}(\boldsymbol{y},\hat{\boldsymbol{y}})$
Language model	$\sqrt[N]{\sum_{i=1}^{N}\frac{1}{P(w_i \mid w_{i-1})}}$

A problem

N-grams that don't occur in the training set cause N-gram probability to go to 0.

$$
\begin{aligned}
\boldsymbol{P}(\boldsymbol{w}|\boldsymbol{w}_{\boldsymbol{n}-1}\ldots) &= \boldsymbol{P}(\textbf{\textit{bling}} \mid \textbf{\textit{this is the way the world}}) \\
&= 0/2,750,000 \\
&= 0
\end{aligned}
$$

Solution - Good-Turing



Solution - Backoff

3-gram ? $\rightarrow$ 2-gram ? $\rightarrow$ 1-gram ?

Solution - Interpolation

$$\boldsymbol{LM} = \alpha_1 \boldsymbol{lm}_1(\boldsymbol{w_n} \boldsymbol{w_{n-1}}) + \cdots + \alpha_{\boldsymbol{n}} \boldsymbol{lm}_2(\boldsymbol{w_n} \boldsymbol{w_{n-1}})$$

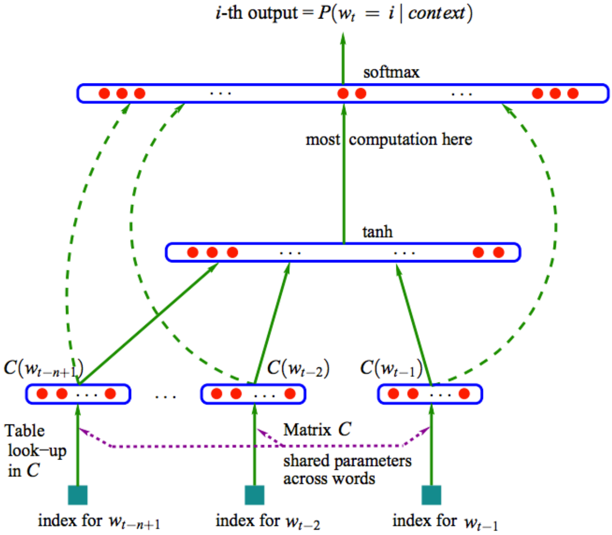
Questions before moving to neural network (NN) LMs?

NN LMs are different

$\mathbb{X}$	Sequences of text
$\boldsymbol{x}$	Sequence of text $\boldsymbol{w}\ \boldsymbol{w}_{\boldsymbol{n}-1}\ \ldots$
$\hat{\boldsymbol{P}}(\boldsymbol{x})$	Probability of $\boldsymbol{w}$ given context
$\boldsymbol{W}$	Matrix of word representations

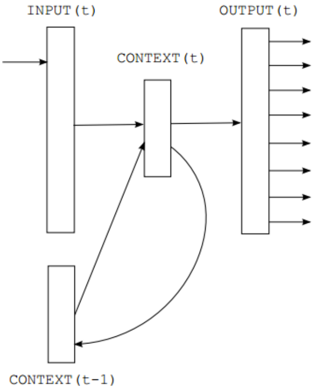
$$\mathbb{F}(\mathbb{X}) \rightarrow (\boldsymbol{f}(\boldsymbol{x}) \rightarrow \hat{\boldsymbol{P}}(\boldsymbol{x}), \boldsymbol{W})$$

Early neural network LM



A Neural Probabilistic Language Model, Bengio et al, 2003

Recurrent neural network LM



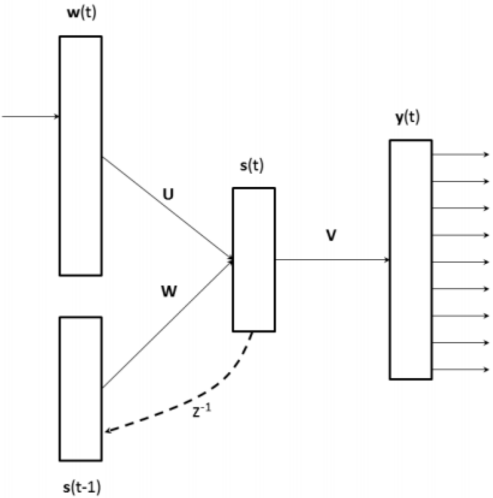
Recurrent neural network based language model, Mikolov et al, 2010

Recurrent neural network LM

Model	# words	Perplexity
KN5	200K	336
KN5 + RNN	200K	271
KN5	1M	287
KN5 + RNN	1M	225
KN5	6.4M	221
KN5 + RNN	6.4M	156

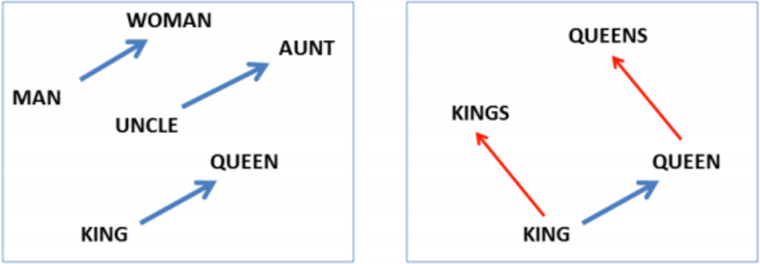
Performance on WSJ dev set.

Learning representations



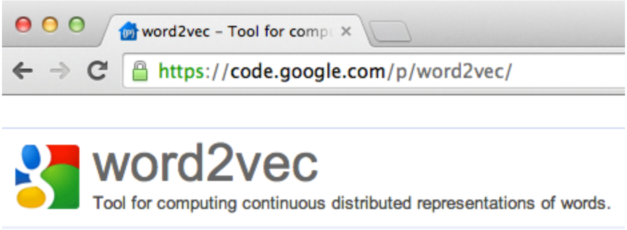
Linguistic Regularities in Continuous Space Word Representations, Mikolov et al, 2013

Properties of learned representations



Linguistic Regularities in Continuous Space Word Representations, Mikolov et al, 2013

A fast neural network LM



Efficient Estimation of Word Representations in Vector Space, Mikolov et al, 2013

Also try this online demo.

Questions?

Feel free to contact me at
ndronen@gmail.com