Deep Learning

Big Data & Machine Learning Bootcamp - Keep Coding

Outline

1. Word representation
2. Using word embeddings
3. Properties of word embeddings
4. Learning word embeddings
5. Word2Vec & Skip-gram
6. GloVe word vectors
7. Sentiment classification
8. Debiasing word embeddings

Word representation

Word embedding: Featurized representation

Feature ***Word and the number in the dictionary***

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97
⋮	⋮	⋮	⋮	⋮	⋮	⋮

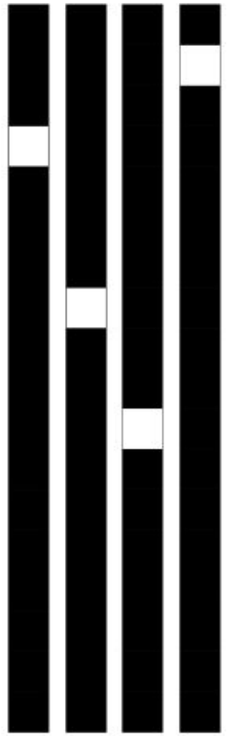
We represent the words by using vectors that are close when the words are similar or they have some semantic similarity.

This table is just to show an example. Usually the vectors are automatically learned and they don't have obvious distinctions as gender, Royal, Age, etc

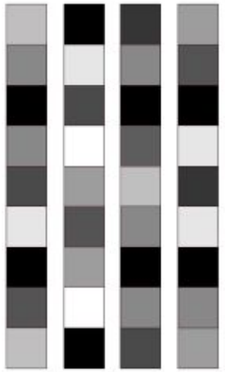
Sources:
- Coursera

Word representation

Word embedding: Featurized representation



One-hot word vectors:
- Sparse
- High-dimensional
- Hard-coded



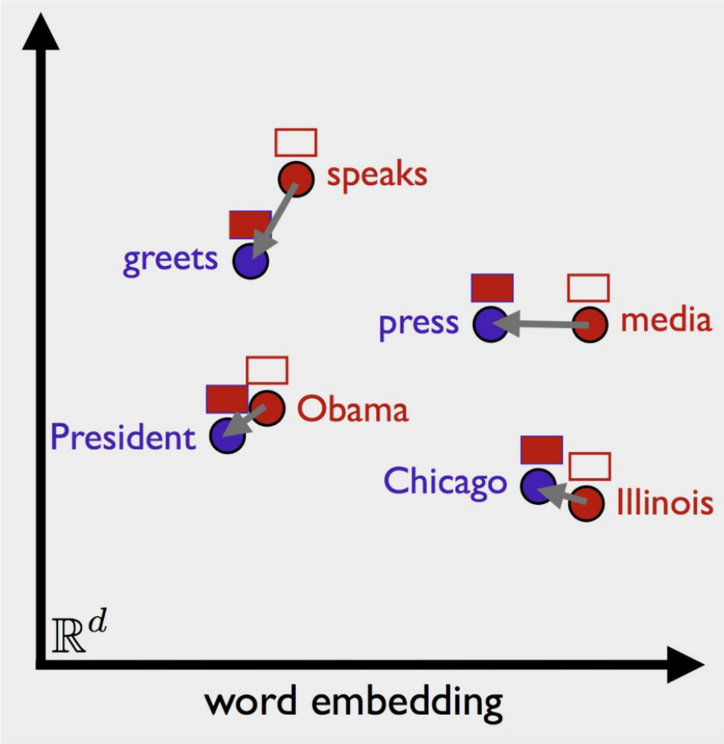
Word embeddings:
- Dense
- Lower-dimensional
- Learned from data

Sources:
- Coursera
- https://jjallaire.github.io/deep-learning-with-r-notebooks/notebooks/6.1-using-word-embeddings.nb.html

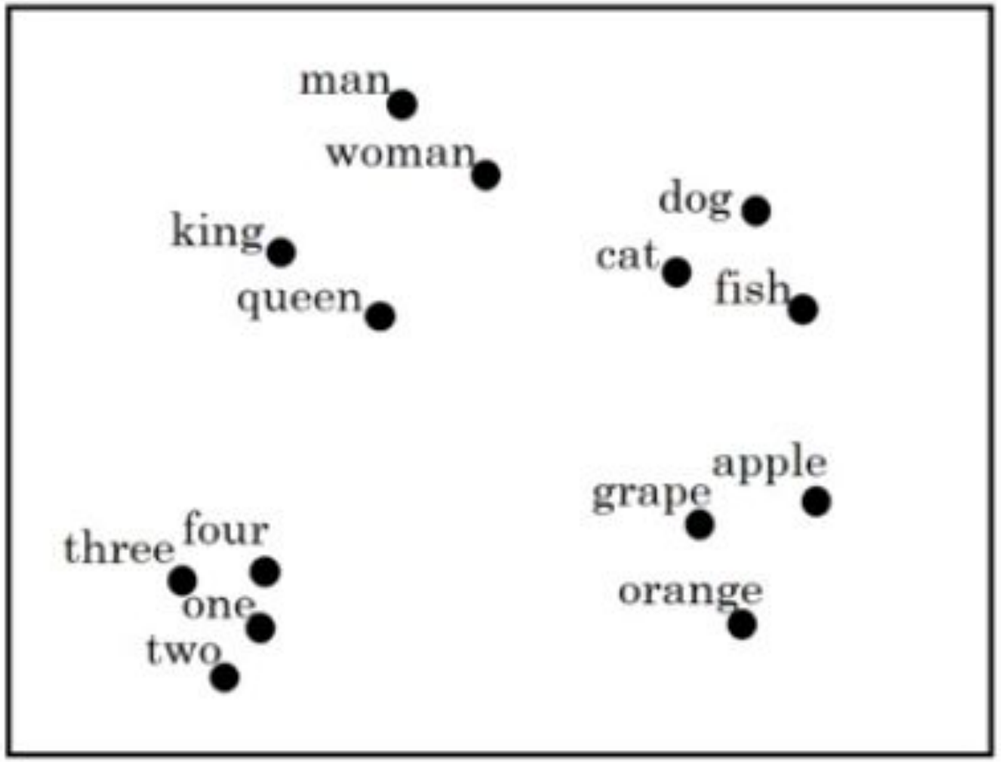
Word representation

Word embedding: Featurized representation

Example 1



Example 2



They are called embeddings because these vectors are represented by a number of variables that can be embedded in a space with the same number of dimensions.

These vectors are embedded in a space. This space can be a 300 dimensional space

Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9, no. Nov (2008): 2579-2605.

Sources:
- Coursera
- https://mlwhiz.com/blog/2017/04/09/word_vec_embeddings_examples_understanding/

Using word embeddings

Word embeddings has many advantages when compared to one hot representation of words. Probably the most important one is transfer learning.

1. You can learn embeddings from a large text corpus (1 to 100 billion words) or download pre-trained embedding online

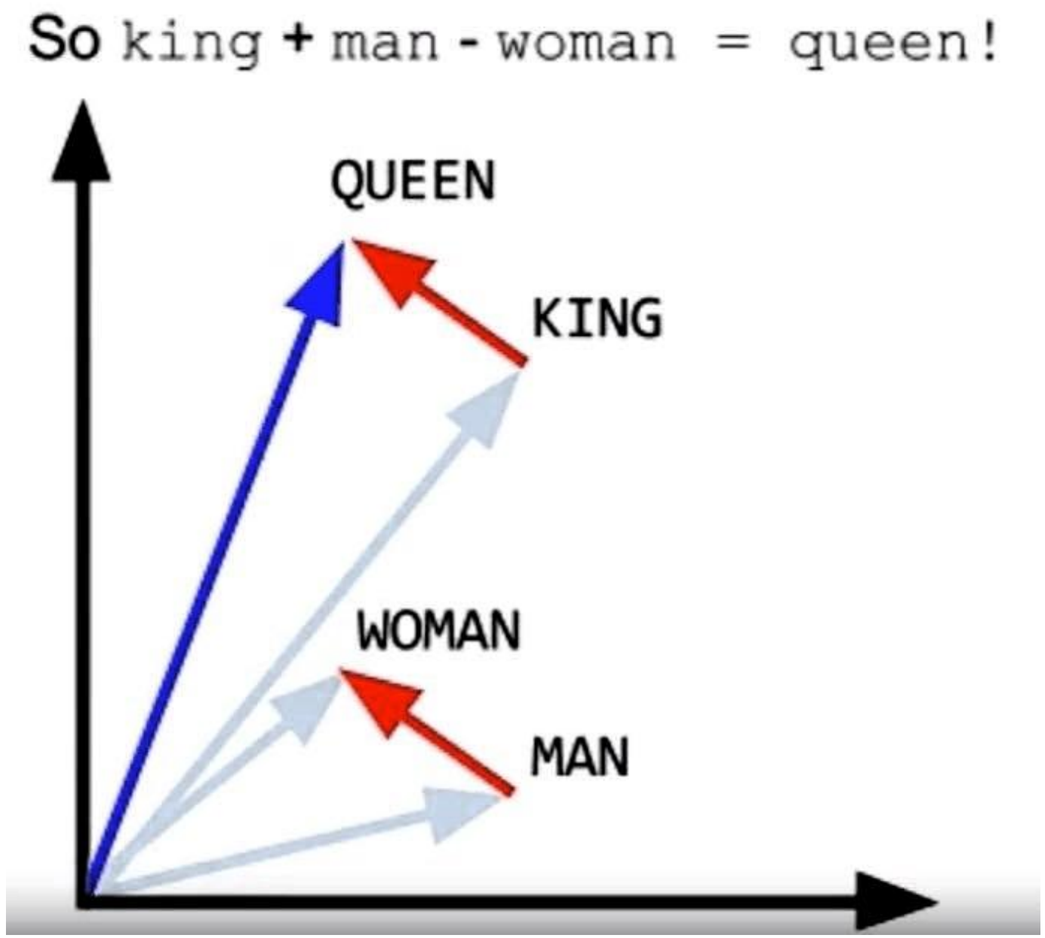
2. Transfer embedding to new task with smaller training set. (Say 100.000 words)
   Really useful for most Natural Language Processing (NLP) tasks but less useful for language modelling and machine translation settings

3. If you want, fine tune the word embedding with new data.

Sources:
- Coursera

Properties of word embeddings

Analogy reasoning.



Red vectors are the vector differences between word pairs ***Man-Woman*** *and* ***King-Queen****.*

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations." In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746-751. 2013.

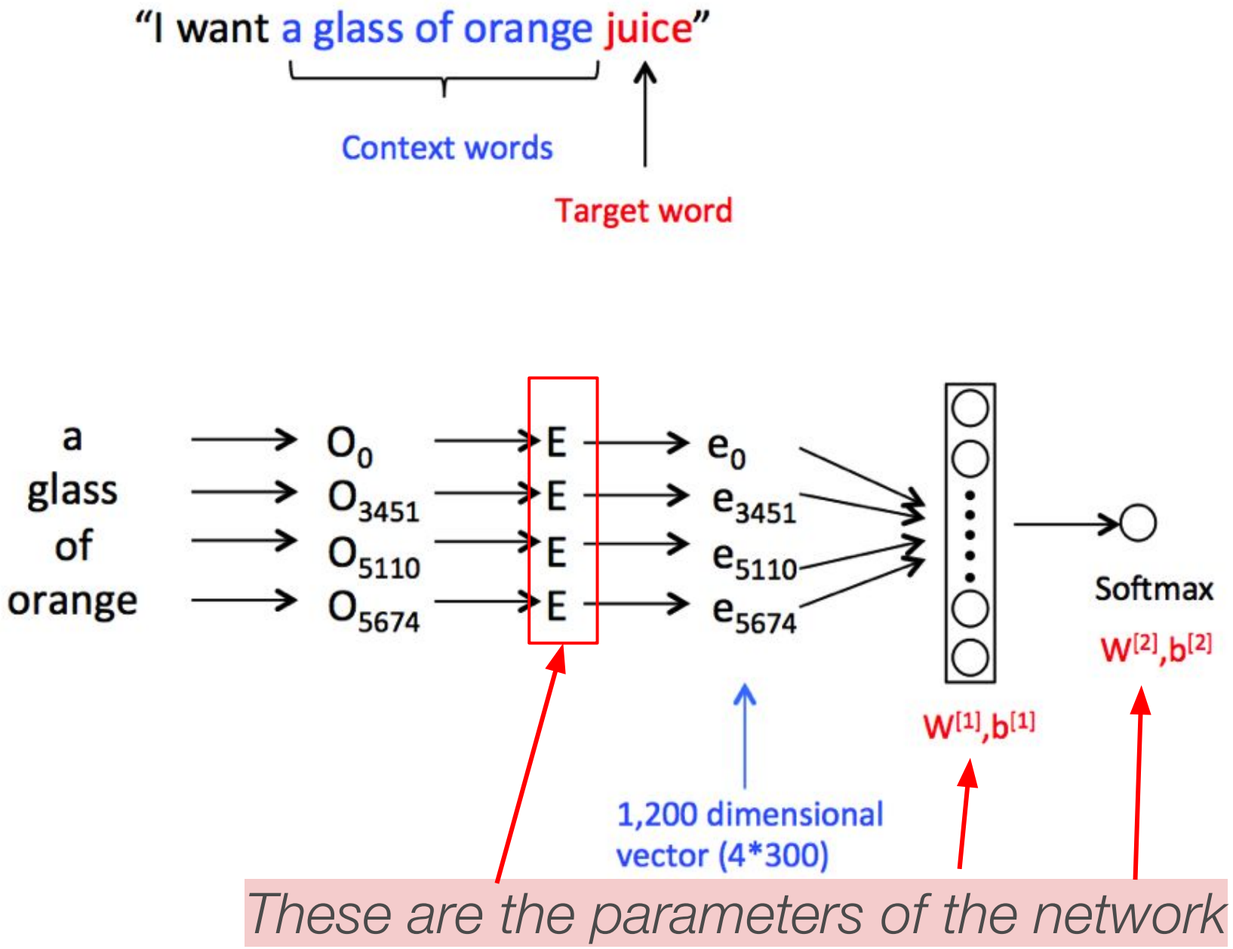
Sources:
- Coursera
- https://twitter.com/kirkdborne/status/1080500437520924672

Learning word embeddings

Now, let's talk about how we can build word embeddings.

Building a neural language model is a reasonable way to learn a set of word embeddings

*We used 4 words history (**context words**) to predict the **target word**.*
By learning those network parameters, this algorithm learns pretty decent word embeddings (Matrix E).



Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. "A neural probabilistic language model." *Journal of machine learning research* 3, no. Feb (2003): 1137-1155.

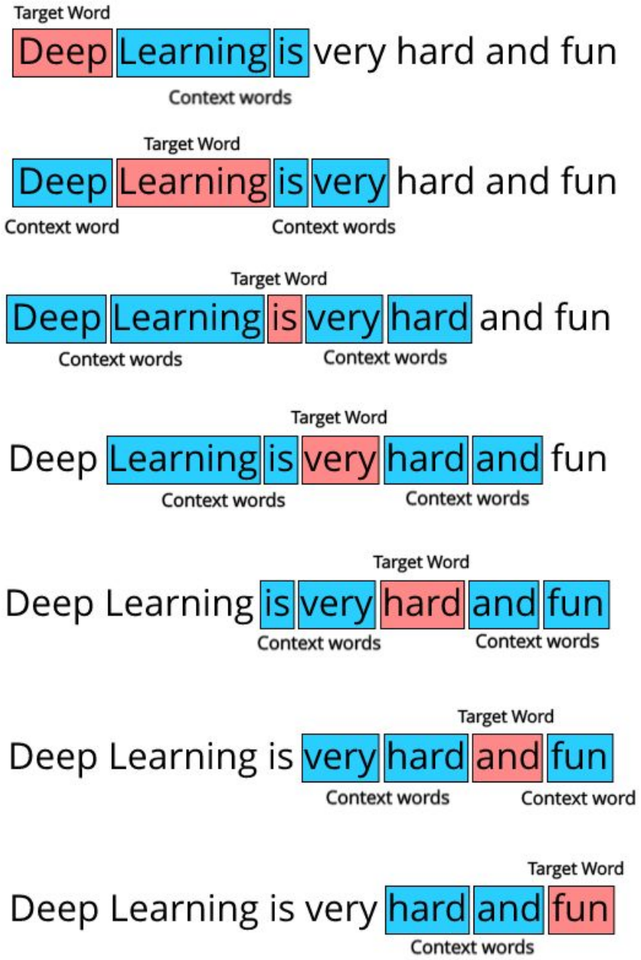
Sources:
- Coursera

Learning word embeddings

Using other context/target pairs.

We can also use other context words instead of using only the previous 4 words. These are the options:

- **Last 4 words**
- **4 words on left and right**
- **Last 1 word**
- **Nearby 1 word (or Skip-gram). More simple but it works really well.**



Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. "A neural probabilistic language model." *Journal of machine learning research* 3, no. Feb (2003): 1137-1155.

Sources:
- Coursera
- https://deeplearningdemystified.com/article/nlp-1

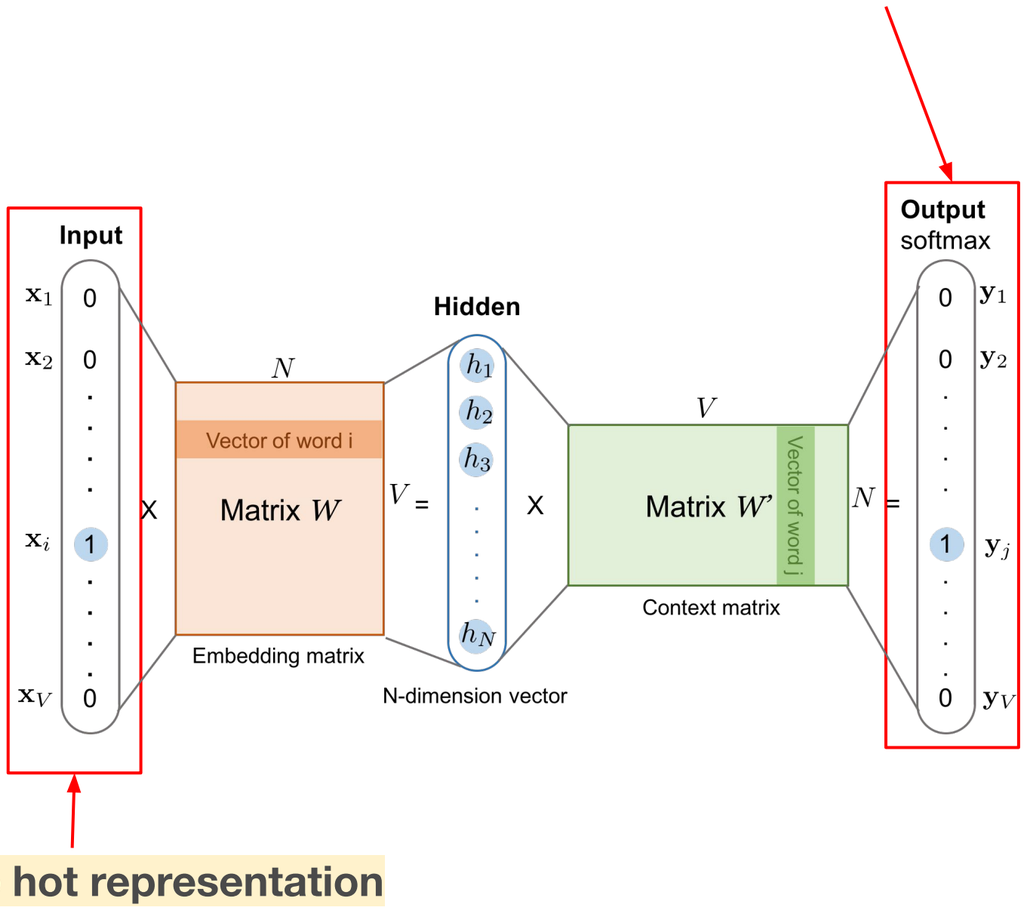
Word2Vec & Skip-gram

Skip-gram: The simpler and more efficient way to learn word embeddings.

Context and target words are selected randomly

Instead of using the 4 previous words as the context, we randomly choose a word from the phrase as ***the context word*** *and randomly choose* ***the target word within a window*** *of 3 or 5 words close to the context word.*

In this way, we can have more samples (pairs target context) to train the supervised problem.



Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

Sources:
- Coursera

GloVe word vectors

GloVe (global vectors for word representation) is not as used as Skip-gram but it has some momentum.

This algorithm works a bit different than previous ones as the loss function is different. It first defines the term:

X_{ij} = ***It is the number of times word j occurs in the context of word i.***

Loss function to get the word embeddings

Weighting term

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

$$f(x) = \begin{cases} (x/x_{\max})^{\alpha} & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$

100 $x_{\max}$, 3/4 α

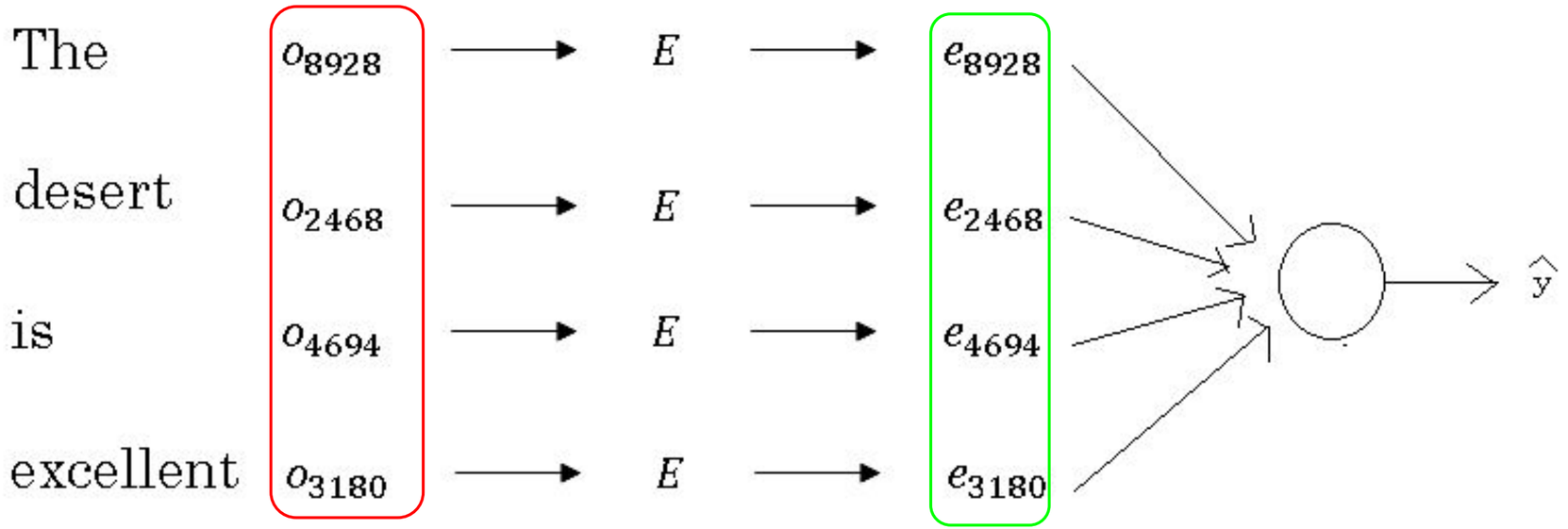
Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

Sources:
- Coursera
- https://jonathan-hui.medium.com/nlp-word-embedding-glove-5e7f523999f6

Sentiment classification

Word embeddings in sentiment classification

Using a simple model to classify sentiment in a phrase



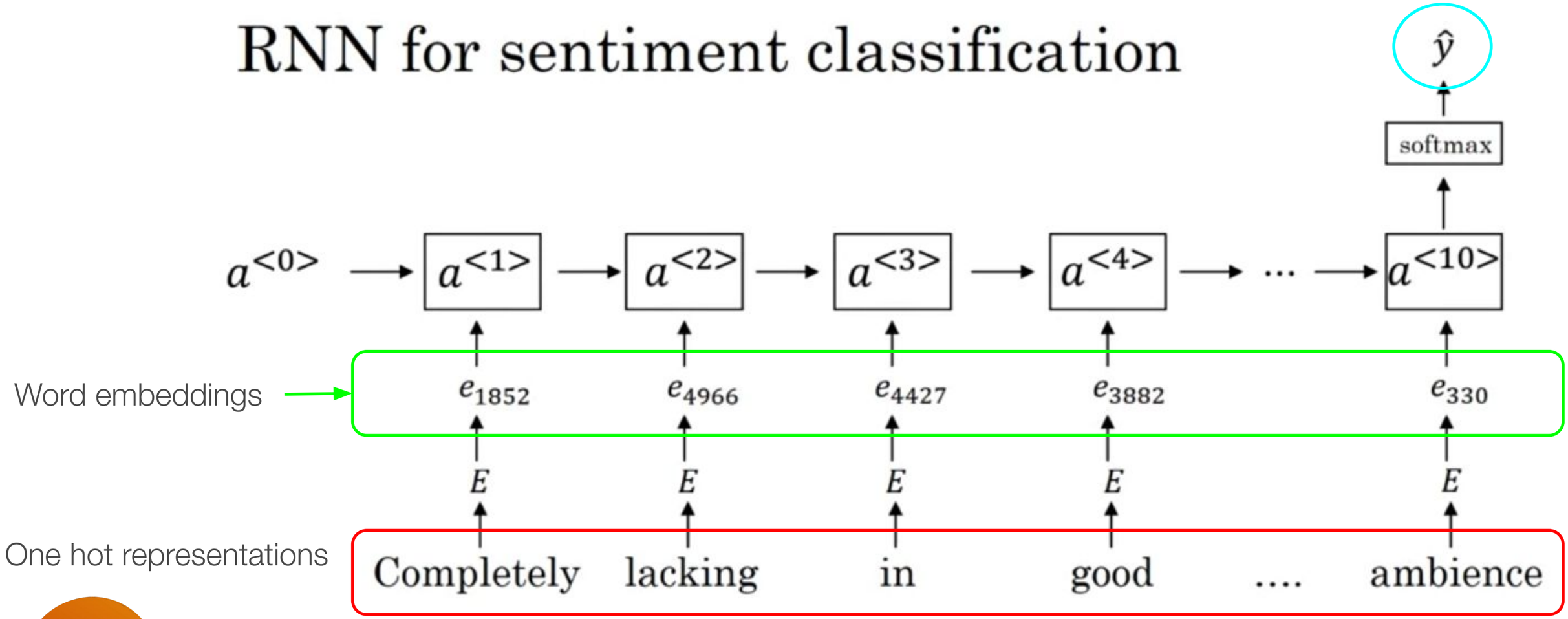
1. Very bad
2. Bad
3. Neutro
4. Good
5. Very good

However, this model ignores word order.
For instance, the phrase **“Completely lacking in good taste, good service, and good ambience”** will be classified as a “good” comment because of the number of “good” words in the phrase

Sources:
- Coursera

Sentiment classification

Many-to-one architecture



Sources:
- Coursera

Debiasing word embeddings

Language models and word embeddings reflect gender, ethnicity, age, sexual orientation, and other biases of the text used to train the model!

Because machine learning tools are being used to make decisions such as loan applications and criminal justice system (sentencing guidelines), **it is important to remove these negative, undesirable and toxic biases.**

Authors of the paper **“Man is to computer programmer as woman is to homemaker? debiasing word embeddings”** proposed a way to do this.

IT IS STILL AN AREA OF RESEARCH

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in neural information processing systems* 29 (2016): 4349-4357.

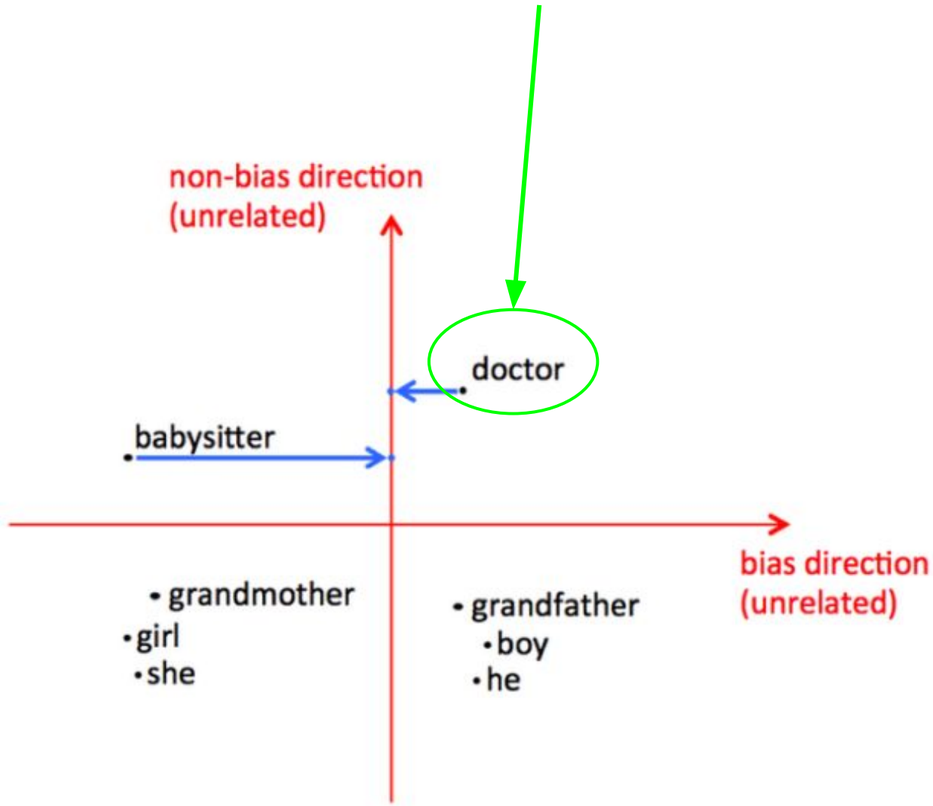
Sources:
- Coursera

Debiasing word embeddings

The word doctor is biased to man

Steps to address bias:

1. Identify bias direction
2. Neutralize: For every word that is not definitional, project to get rid of bias.
   A definitional word doesn't have, by definition, a gender. For instance, doctor or receptionist are non definitional words as they are roles that can be occupied by any gender.
3. Equalize pairs



Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in neural information processing systems* 29 (2016): 4349-4357.

Sources:
- Coursera
- https://vagdevik.wordpress.com/2018/07/08/debiasing-word-embeddings/