Carnegie Mellon University
Heinz College

95-865: Self-supervised Learning

Slides by George Chen and Mi Zhou

Even without labels, we can set up a prediction problem!

Hide part of training data and try to predict what you've hid!

Word Embeddings: word2vec

Word2vec model trained by Google on the Google News dataset, on about 100 billion words:

Man is to King as Woman is to _____

Word Embeddings: word2vec

Word2vec model trained by Google on the Google News dataset, on about 100 billion words:

Man is to King as Woman is to Queen

Word Embeddings: word2vec

Word2vec model trained by Google on the Google News dataset, on about 100 billion words:

Man is to King as Woman is to Queen

Which phrase doesn’t fit?
blue, red, green, crimson, transparent

Word Embeddings: word2vec

Word2vec model trained by Google on the Google News dataset, on about 100 billion words:

Man is to King as Woman is to Queen

Which phrase doesn't fit?
blue, red, green, crimson, transparent

Word Embeddings: word2vec

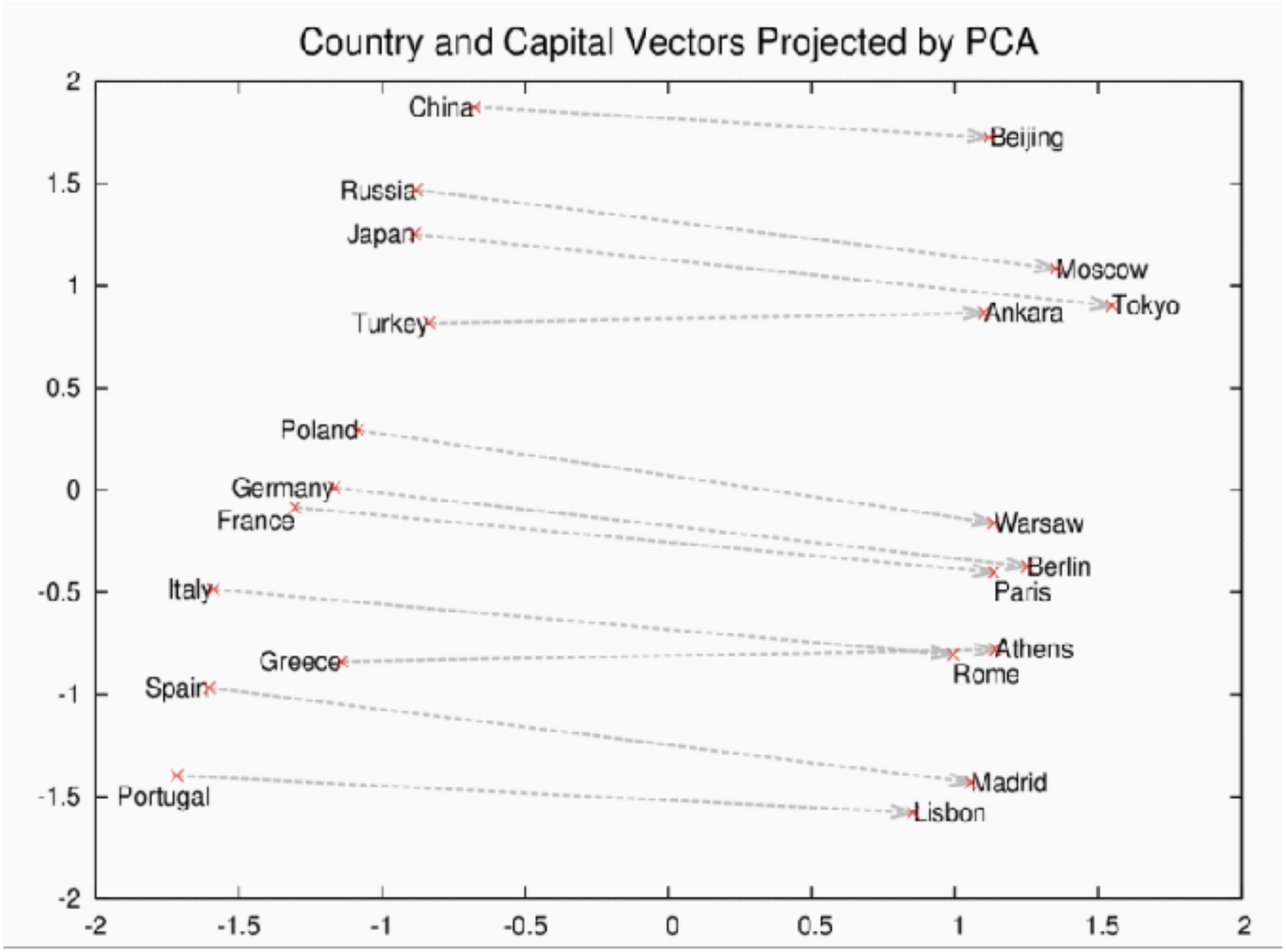


Image source: https://deeplearning4j.org/img/countries_capitals.png

Word Embeddings: word2vec

The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

Predict context of each word!

Training data point: epidemic

“Training label”: the, opioid, or, opioid

Word Embeddings: word2vec

The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

Predict context of each word!

Training data point: or

“Training label”: opioid, epidemic, opioid, crisis

Word Embeddings: word2vec

The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

Predict context of each word!

Training data point: opioid

"Training label": epidemic, or, crisis, is

There are "positive" examples of what context words are for "opioid"

Also provide "negative" examples of words that are *not* likely to be context words (by randomly sampling words elsewhere in document)

Word Embeddings: word2vec

The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

randomly sampled word

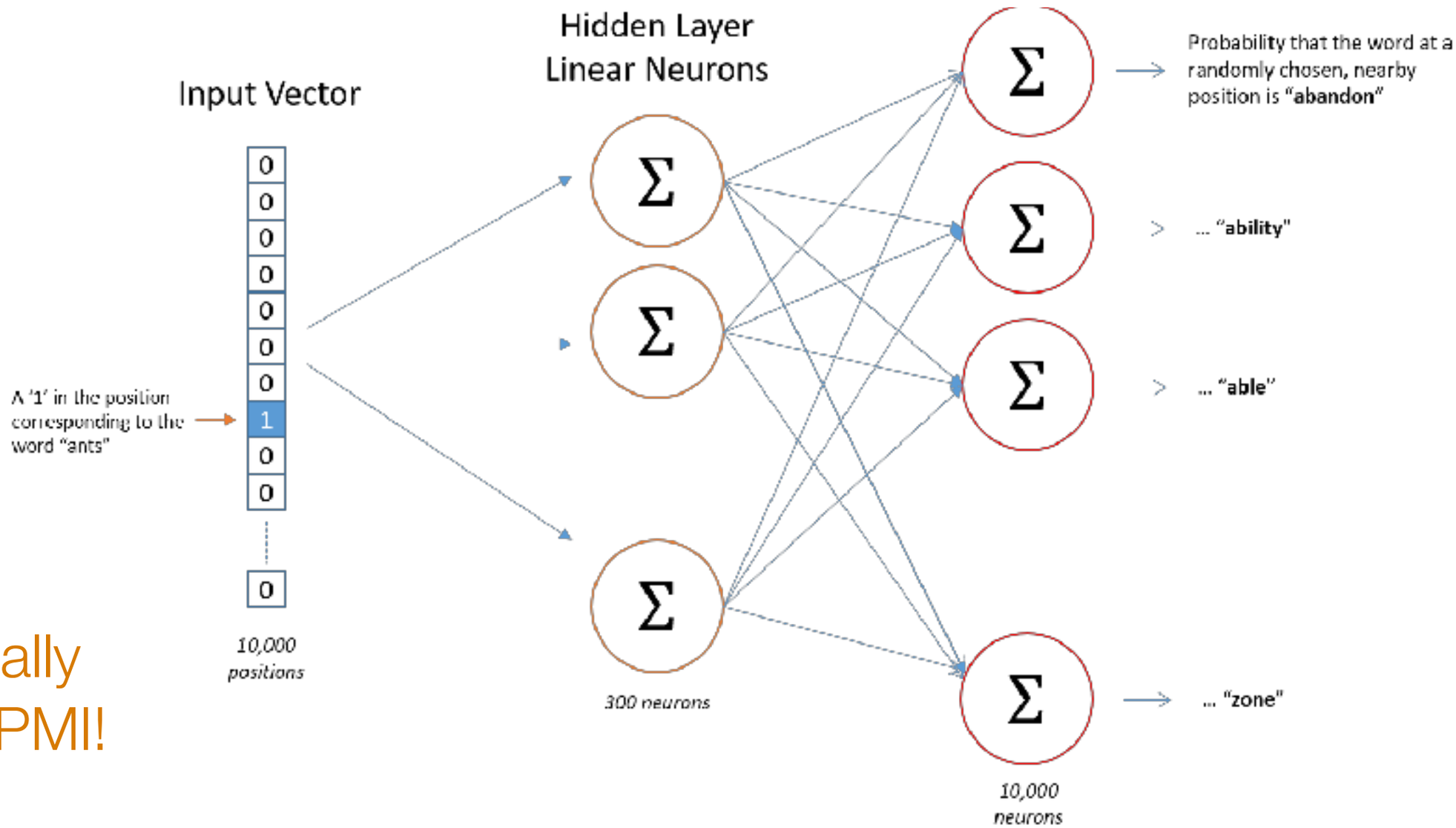
Predict context of each word!

Training data point: opioid

"Negative training label": 2010s

Also provide "negative" examples of words that are *not* likely to be context words (by randomly sampling words elsewhere in document)

Word Embeddings: word2vec



This actually relates to PMI!

Weight matrix: (# words in vocab) by (embedding dim)

Dictionary word *i* has "word embedding" given by row *i* of weight matrix

Image Source: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Self-Supervised Learning

- Key idea: hide part of the training data and try to predict hidden part using other parts of the training data
- No actual training labels required — we are defining what the training labels are just using the unlabeled training data!
- This is an *unsupervised* method that sets up a *supervised prediction* task
- Other word embeddings methods are possible (GLoVe)
  - **Warning:** the default Keras `Embedding` layer does *not* do anything clever like word2vec/GloVe (best to use pre-trained word2vec/GloVe vectors!)