

CS224n

Thomas Lu

Note: statements that I am unsure about are enclosed in double square brackets, e.g. $[[\text{Every even integer greater than 4 can be expressed as the sum of two primes.}]]$

1 Lecture 1: Introduction

This lecture basically just gives a high level overview of the fields of Natural Language Processing (NLP) and Deep Learning (DL). Some highlights from the lecture:

- NLP is a study that aims for computers to understand natural (human) language well enough to perform useful tasks.
- NLP has a lot of useful applications (virtual assistant, customer support automation, etc.) and has taken off commercially in the past few years.
- What is special about NLP (vs. other subfields of ML)?
 - Language is specifically constructed to convey information. It's not an environmental signal (e.g. an Amazon purchase, which I probably made for a reason other than conveying information).
 - Language is (mostly) a discrete/symbolic/categorical system, but our brains, which process language, are continuous systems.
- What is special about DL (vs. other ML techniques)?
 - Traditionally, learning an ML model requires an engineer/researcher to design and build relatively human-interpretable features. The machine then basically does some numerical optimization on these features to produce something useful. Note that this still requires a lot of intuitive human work to interpret data and figure out what parts of it might be useful.
 - In contrast, deep learning just feeds “raw data” or very simplistic representations of data to a deep neural net, which then learns its own effective representations of the data, obviating the need for feature engineering.
 - DL has performed very well recently, e.g. speech recognition, ImageNet.
- Why is NLP hard?
 - Human language contains lots of contextual information. The meaning of a sentence may depend on something that was said a few sentences ago, a piece of information from another document, current events, or something observable in the environment (“Did you see that dog?” refers presumably to a dog in the speaker’s field of vision).
 - Human language is also ambiguous by design; since speech is pretty slow at transmitting information, people will almost always leave “relatively obvious” things unsaid in order to improve communication efficiency. Listeners are expected to infer these things. This is very different from computer languages, which must specify every possibility.

2 Lecture 2: word2vec

One part of NLP is word meaning representation. How do we represent the definition of a word so that a computer can do something useful with that representation?

One method is WordNet, which is a database of words that groups words with similar meaning together into synsets, which are also arranged in a kind of hierarchy. But this has some problems:

- Synonym/hypernym relationships miss a lot of nuance (e.g. “expert”, “proficient”, “good” don’t really mean the same thing).
- It’s hard to keep the database up-to-date when new words are added or existing words gain/lose usages and meanings.
- Creating the database is subjective and labor-intensive.
- Given two words, there isn’t a good way to get a measure of how similar they are.

This is an example of a discrete/categorical/symbolic representation of words, where each word represents a single concept. Words might be related, but separate words are generally regarded as separate entities. If these words were vectors, each one would be a V -dimensional vector of zeroes with a single 1, where V is the vocabulary size. This is sometimes referred to as a “localist” or “one-hot” representation.

A problem with these representations is that they lack similarity measures; we could try to manually encode some (as with WordNet), but they don’t exist in the representations themselves (e.g. any two distinct word vectors are orthogonal). We can instead try to encode words so that they include similarity information, where a dot product of two words can give you a measure of their similarity. This gives us representations of words as shorter vectors, and this is known as a “distributed” representation: instead of the meaning of a word being “localized” to a single component of a vector, it’s “distributed” across multiple components.

One idea for generating distributed representations is the idea of distributional similarity, where we figure out the meaning of a word based on the contexts it appears in, i.e. its distribution in text. (The word “distributional” in “distributional similarity” is not related to the word “distributed” in “distributed representation”.) This idea is summed up in a well-known quote: “You shall know a word by the company it keeps.” (John Rupert Firth)

word2vec is a well-known distributed representation model. Its key idea is to predict **outer words** that will appear in the context of a **center word**. More specifically, given a center word w_c , it attempts to predict the probabilities $p(w_o|w_c)$ that various outer words w_o will appear in the context of w_c . Supposing we have a training corpus of T words, the training process will attempt to maximize

$$\prod_{t=1}^T \prod_{\substack{-m \leq j \leq m, \\ j \neq 0}} p(w_{t+j}|w_t; \theta),$$

where θ is the parametrization of the model, the context of a word is defined as the m words before and after it, w_i is the i -th word in the corpus, and $p(w_i|w_j; \theta)$ denotes the model’s predicted probability of w_i appearing in the context of w_j given its parametrization θ . Equivalently, the training process attempts to minimize the objective function

$$J(\theta) = - \sum_{t=1}^T \sum_{\substack{-m \leq j \leq m, \\ j \neq 0}} \log p(w_{t+j}|w_t; \theta).$$

More concretely, the model is parametrized by $2|V|$ word vectors, where V is the vocabulary. Each word $w \in V$ has two vector representations: a center word representation v_w and an outer word representation

u_w . The model then predicts

$$p(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}.$$

More intuitively, this means that the model predicts that o is likely to appear in the context of c if the outer word representation of o is similar to the center word representation of c .

Because the training corpus for a word2vec model is typically very large, the model is usually trained using stochastic gradient descent, regarding each context window as a single training example. Now the gradient of J with respect to θ is a little tricky to notate, but since J is just a sum of terms of the form $\log p(w_o|w_c)$, the partial derivatives of that term will be quite informative. We have:

$$\begin{aligned} \frac{\partial}{\partial u_o} \log p(o|c) &= \frac{\partial}{\partial u_o} \log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} \\ &= \frac{\partial}{\partial u_o} \left[\log \exp(u_o^T v_c) - \log \left(\sum_{w \in V} \exp(u_w^T v_c) \right) \right] \\ &= v_c - \frac{1}{\sum_{w \in V} \exp(u_w^T v_c)} \frac{\partial}{\partial u_o} \sum_{w \in V} \exp(u_w^T v_c) \\ &= v_c - \frac{1}{\sum_{w \in V} \exp(u_w^T v_c)} \frac{\partial}{\partial u_o} \exp(u_o^T v_c) \\ &= v_c - \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} v_c \\ &= v_c (1 - p(o|c)) . \end{aligned}$$

Similarly, we can calculate

$$\begin{aligned} \frac{\partial}{\partial v_c} \log p(o|c) &= \frac{\partial}{\partial v_c} \log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} \\ &= \frac{\partial}{\partial v_c} \left[\log \exp(u_o^T v_c) - \log \left(\sum_{w \in V} \exp(u_w^T v_c) \right) \right] \\ &= u_o - \frac{1}{\sum_{w \in V} \exp(u_w^T v_c)} \frac{\partial}{\partial v_c} \sum_{w \in V} \exp(u_w^T v_c) \\ &= u_o - \sum_{w \in V} \frac{\exp(u_w^T v_c)}{\sum_{w' \in V} \exp(u_{w'}^T v_c)} u_w \\ &= u_o - \sum_{w \in V} p(w|c) u_w \\ &= u_o (1 - p(o|c)) - \sum_{\substack{w \in V \\ w \neq o}} p(w|c) u_w . \end{aligned}$$

Intuitively, this means that when we encounter a word o in the context of another word c , we should:

- Move the outer word representation u_o of o towards the center word representation v_c of c by an amount negatively correlated with the prior prediction $p(o|c)$. This means that we should make a larger update for more unexpected events.
- Move the center word representation v_c of c :
 - towards the outer word representation u_o of o by an amount negatively correlated with the prior prediction $p(o|c)$, meaning that we should make a larger update for more unexpected events; and

- away from the outer word representations u_w for each w_o by an amount positively correlated with the prior prediction $p(w|c)$, meaning that we should make a larger update for more unexpected events (since observing o in the context of c means that we did not observe w in the location of o).

This matches our intuition that we expect to see o in the context of c if u_o is similar to v_c .

3 Lecture 3: GloVe

3.1 word2vec revisited

Recall that the skip-gram (word2vec) model is typically trained using stochastic gradient descent. Because the parametrization of the model is very large ($2Vd$ terms, where V is the vocabulary size and d is the dimensionality of the vector representations of the words) and the calculated gradient at each iteration is comparatively sparse, we should avoid computing and passing around complete gradient vectors when performing SGD.

Previously, we used the following probability in the skip-gram model:

$$p(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}.$$

The numerator of this is easy to compute, but the denominator is pretty expensive, as the sum has to be taken across the entire vocabulary. Instead of computing the full sum, the actual word2vec model uses a slightly different objective function that allows it to randomly sample from the vocabulary instead of iterating over the whole vocabulary:

$$J(\theta) = \sum_{t=1}^T J_t(\theta) = \sum_{t=1}^T \left[\log \sigma(u_o^T v_c) + \sum_{\substack{j \sim P(V) \\ k \text{ samples}}} \log \sigma(-u_j^T v_c) \right],$$

where T represents our corpus, σ is the sigmoid function, $\sigma(z) = (1 + e^{-z})^{-1}$, $P(V)$ is a probability distribution over the vocabulary V , \sim indicates random sampling, and k is a constant, usually around 10. Furthermore, instead of minimizing this objective function, the model actually seeks to maximize it. Intuitively, this means that the model attempts to maximize $u_o^T v_c$ when o occurs in the context of c and minimize $u_w^T v_c$ for w that do not appear in the context of c , which is pretty similar to what we were doing with the previous objective function.

One detail on the objective function: the probability distribution $P(V)$ that is usually used is $P(w) = U(w)^{3/4}/Z$, where $U(w)$ is the unigram frequency of $w \in V$ and Z is a normalizing factor to make $\sum_{w \in V} P(w) = 1$. The exponent of $3/4$ “smoothes out” the distribution a bit, so that more frequent items are less likely to be selected than less frequent items in proportion to their frequencies.

An alternative to the skip-gram model is the similar Continuous Bag-Of-Words (CBOW) model. While the skip-gram model attempts to predict the probability of an outside word appearing given a center word, the CBOW model attempts to predict the probability of the center word occurring given the sum of the (vector representation of the) outside words.

3.2 Co-occurrence matrices

Co-occurrence matrices are another way to construct vector representations of words. To start, we can create a $|V| \times |V|$ matrix X and iterate through our corpus, incrementing X_{ij} every time we see $i, j \in V$

occurring together. We can define “occurring together”, or co-occurrence, in various ways. Some common ones are:

- Window-based co-occurrence. This is similar to what the skip-gram model does: we say that two words co-occur if one appears in a window around the other (e.g. the window from 8 words before to 8 words after).
- [[Full-document co-occurrence. Here we say that two words co-occur if they both appear in some document in our corpus.]]
- [[Word-document co-occurrence. In this definition, we need to change our matrix X . We will still have a row for each word, but instead of having a column for each word as well, we have a column for each document in our corpus, and we say that a word and a document co-occur if the word appears in that document.]]

It turns out that window-based co-occurrence allows us to represent both syntactic and semantic information in the vector representations, while [[full-document co-occurrence and word-document co-occurrence allow us to capture topics and perform Latent Semantic Analysis.]] Unless otherwise specified, further discussion in this section will be about window-based co-occurrence matrices.

These “raw” co-occurrence matrices, however, are extremely storage intensive; a 100,000 word vocabulary would yield a matrix with 10 billion entries when using window-based co-occurrence. These matrices are also rather sparse, which negatively impacts the robustness of models trained on them. We can solve both of these problems by reducing the column space using singular value decomposition (SVD). Other improvements that we can make to this model include:

- To prevent stopwords from creating very spiky counts, we can cap frequency counts when assembling the pre-SVD matrix or ignore stopwords altogether.
- We can weight co-occurrences; e.g. maybe we should weight the co-occurrence of two words more if they are adjacent rather than 5 words apart (e.g. by adding a larger or smaller increment to the corresponding entry in the pre-SVD co-occurrence matrix).

However, co-occurrence matrices still have some issues:

- Adding words to the vocabulary is difficult - we have to grow both the row space and the column space.
- SVD is an expensive operation; on an $m \times n$ matrix with $n < m$, it has runtime $O(mn^2)$.
- The model only captures word similarity and doesn’t really capture any additional patterns.
- It’s hard to make the model not give excessive significance to large counts.

On the other hand, co-occurrence matrices have some advantages as well:

- It’s relatively fast to train.
- It uses the statistics of the corpus efficiently - after a single training pass, we can perform whatever further transformation/analysis we want on the co-occurrence counts.

As for the skip-gram model:

- The representations it produces are often better suited for downstream tasks.
- It captures patterns beyond just word similarity - it also gets things like syntactical relationships.
- The training time scales with the corpus size. It can take a very long time to train on a large corpus.
- It doesn’t really let us use corpus statistics well, as there isn’t really any intermediate representation of the data besides the trained model itself.

3.3 GloVe

Global Vectors, or GloVe, is another method for training vector representations of words. It compiles a co-occurrence matrix P , and then seeks to minimize the objective function

$$J(\theta) = \frac{1}{2} \sum_{i,j \in V} f(P_{ij}) (u_i^T v_j - \log P_{ij})^2.$$

f here is a function that is increasing until a cutoff point, after which it becomes constant. The parameters of the model here are the word vectors u_i and v_j , similar to the skip-gram model.

Intuitively, we weight common co-occurrences more heavily in this objective function, but we have a cap on how heavily any one can be weighted. We also want $u_i^T v_j$ to be larger, i.e. u_i and v_j to be more similar, when P_{ij} is larger (i.e. when i and j co-occur more frequently). At the end when we've trained our u_i and v_j , the final representation x_w that we choose for each word $w \in V$ is simply $x_w = u_w + v_w$.

This method scales to large corpora but has been shown to also work well for small corpora. It's like a best-of-both-worlds between word2vec and co-occurrence matrices: it's fast to train and uses statistics efficiently, but also captures information beyond word similarity and caps the influence of large co-occurrence counts. Furthermore, it avoids the computational cost of performing SVD on a large matrix.

3.4 Evaluation of vector representations of words

Once we have a model, we'd like to evaluate how well it performs. In general, there are two categories of evaluation metrics for any model: intrinsic and extrinsic.

- Intrinsic metrics measure how well the model performs some specific or intermediate subtask, such as how well vector similarities of words map to human judgments of similarities of words. They:
 - are typically easy to compute and measure and
 - often can give you more insight into your own system by showing you which hyperparameters affect your system, but
 - oftentimes don't clearly show whether the model is actually good at some more externally useful task.
- Extrinsic metrics measure how well the model performs some more useful downstream task, such as machine translation. Extrinsic metrics:
 - are often slow to compute (e.g. it might take a while to train the machine translation model from your word representations) and
 - can be unclear on whether your component (e.g. word representations) of the larger system (e.g. machine translation system) specifically improved performance on its own if other parts of the system have also changed.

One intrinsic metric is word vector analogy accuracy. For example, if we give the analogy task

Paris:France::Rome:??

to the model, the model should be able to determine that the correct answer is Italy. Typically word representation models do this by finding the word w that maximizes

$$\frac{(x_b - x_a + x_c)^T x_w}{\|x_b - x_a + x_c\| \|x_w\|},$$

where the analogy is given as **a:b::c:???**. We can then measure how well the model works by feeding it many such analogies and examining how many it gets correct.

4 Lecture 4: Word Window Classification and Neural Networks

Previously we have discussed unsupervised models for learning word representations, but eventually we want to use these representations to train supervised predictive models.

In a typical classification problem, we'll have a training dataset $(x_i, y_i), 1 \leq i \leq N$, where the x_i are our inputs and the y_i are labels that we want to train our model to predict. A legacy technique for training these models is logistic regression, a.k.a. softmax. In this method, we train a weight matrix W with one row per class, and then attempt to adjust it so that our predictions

$$p(y|x) = \frac{\exp(W_y x)}{\sum_c \exp(W_c x)}$$

match the training examples as closely as possible. Here W_c represents the row of weights in W corresponding to the specific class c . A common loss function for softmax is the cross-entropy loss:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N -\log \frac{\exp(W_{y_i} x_i)}{\sum_c \exp(W_c x_i)},$$

where θ represents all the parameters of the model (in this case W). Oftentimes in practice we also add a regularization term $\lambda \theta^T \theta$, which helps prevent overfitting. Note that this method treats the inputs $X = [x_1^T, x_2^T, \dots, x_N^T]^T$ as constant. In contrast, when we use deep learning models, we usually update our word vectors (i.e. X) as well. However, because this vastly increases our parameter space, we need to be careful not to overfit. Often if we have a small training set, we shouldn't update word vectors, as overfitting would be a big problem.

4.1 Tips for vector calculus

There will be a lot of vector calculus involved in calculating gradients for neural network (NN) models. Some tips for when it gets tricky:

- Carefully define all variables and keep track of their dimensionality. Checking dimensionality is a great sanity check for whether you've made any obvious mistakes.
- When using the chain rule, keep track of variable dependencies to make sure you haven't forgotten a necessary step of differentiation.
- Casework can sometimes make a calculation more intuitive (e.g. gradients for the correct class vs. the incorrect classes).
- It can sometimes also be easier to calculate derivatives element-by-element instead of trying to differentiate the whole vector function at once.
- After doing casework and elementwise differentiation, try to vectorize your notation again. It's important to vectorize your implementations in code, so you'll need to vectorize the notation at some point.

4.2 Window classification

A common prediction task is to predict whether a center word in a window of words is a named person/place/organization or not a named entity. Typically the input to the model is the concatenation of the word vectors in the window (so if we had d -dimensional word vectors and used a window with m words before and after the center word, our input would be a vector of dimensionality $(2m+1)d$).

We could approach this problem with the softmax technique, but softmax is only capable of learning linear decision boundaries. NNs, on the other hand, can learn more complex nonlinear boundaries.

4.3 Neural networks

A neural network is made up of many neurons, arranged in sequential layers. We call the first layer the input layer, since it just contains the input values, and the final layer the output layer. The intermediate layers between them are commonly referred to as hidden layers. Each neuron has a weight vector w , an input vector x , a bias unit b , and an activation function f , and produces output $f(wx + b)$. A common activation function is the sigmoid function $f(z) = (1 + e^{-z})^{-1}$.

The reason that a neural network can learn more complex decision boundaries is that it can feed the outputs of neurons into another layer of neurons: instead of being constrained to immediately predict the output, it can learn an intermediate representation that might be more helpful for predicting the output. Note however that this requires a nonlinear activation function like the sigmoid; if we had a linear activation function, like the identity function, our neural network would still just be a linear classifier at the end of the day.

Back to our example problem - imagine that we're using a simple one-hidden-layer NN to predict whether the center word in a window is a named entity location or not. We have an input layer x_j which together with a weight matrix W produces a hidden layer a_i , which finally produces a score s . For simplicity, we will use a simple model where s is a linear score: $s = U^T a$ for some weight vector U . We will use the max-margin loss function:

$$J = \max(0, 1 - s + s_c),$$

where s is predicted score of a positive example (where the center word is a named location) and s_c is the predicted score of a negative example (where the center word is not a named location). In practice, for a single positive example, we usually randomly sample some s_c rather than computing the max over the whole training corpus. Intuitively, this loss function tries to find a decision boundary that separates positive and negative samples by a "sufficiently large" margin ("sufficiently large" is defined here as 1, but it can be tuned as a hyperparameter).

Note that it is very common to write $a = f(z)$, so that $a = f(Wx + b)$ and $z = Wx + b$. (Here f is evaluated element-wise.)

We now compute the gradient of J in parts. When $J = 0$ we have $\vec{\nabla} J = 0$, so assume $J > 0$. Then $\vec{\nabla} J = -\vec{\nabla} s + \vec{\nabla} s_c$. We will just compute $\vec{\nabla} s$ for illustration. For the partial derivatives with respect to U , we have:

$$\frac{\partial s}{\partial U} = \frac{\partial}{\partial U} U^T a = a.$$

For the partial derivatives with respect to W , we have:

$$\frac{\partial s}{\partial W_{ij}} = \frac{\partial}{\partial W_{ij}} U^T a = \frac{\partial}{\partial W_{ij}} U^T f(z) = \frac{\partial}{\partial W_{ij}} U^T f(Wx + b).$$

Note that only the i -th entry of $f(Wx + b)$ is nonconstant with respect to W_{ij} , so we have

$$\begin{aligned} \frac{\partial}{\partial W_{ij}} U^T f(Wx + b) &= \frac{\partial}{\partial W_{ij}} U_i f(W_i x + b_i) \\ &= U_i f'(z_i) x_j \\ &= \delta_i x_j, \end{aligned}$$

where we have written $\delta_i = U_i f'(z_i)$. Now note that by combining all of these entries, we can write:

$$\frac{\partial s}{\partial W} = \delta x^T.$$

To calculate the partials with respect to b , we have:

$$\begin{aligned}
\frac{\partial s}{\partial b_i} &= \frac{\partial}{\partial b_i} U^T f(Wx + b) = \frac{\partial}{\partial b_i} U_i f(W_i x + b_i) \\
&= U_i f'(W_i x + b_i) \\
&= \delta_i \\
\Rightarrow \frac{\partial s}{\partial b} &= \delta.
\end{aligned}$$

Finally it remains to find the partials with respect to x . We have:

$$\begin{aligned}
\frac{\partial s}{\partial x_j} &= \frac{\partial}{\partial x_j} U^T a \\
&= \sum_i \frac{\partial}{\partial x_j} U_i a_i \\
&= \sum_i U_i \frac{\partial}{\partial x_j} f(W_i x + b_i) \\
&= \sum_i U_i f'(W_i x + b_i) \frac{\partial}{\partial x_j} W_i x \\
&= \sum_i \delta_i W_{ij} \\
\Rightarrow \frac{\partial s}{\partial x} &= W^T \delta.
\end{aligned}$$