Word Embeddings 2 - Class Discussions

Jalaj
February 18, 2016

Opening Remarks - Word embeddings as a concept are intriguing. The approaches are mostly adhoc but show good empirical performance.

Paper 1 - Skip Gram Model (Mikolov et.al.)

1. How do word representations encode linguistic regularities and patterns? Consider this example: vec("Madrid") - vec("Spain") + vec("France") is closer to vec("Paris") than to any other word. This points to a "linear structure" in these language patterns.

2. Its like subtracting and adding contexts. From vec("Madrid") we removed its context by subtracting vec("Spain") then added the context of vec("France") to get to vec("Paris").

3. Question: Why do these patterns (which can be represented by simple algebraic operations) occur?

4. Skip-gram model: Objective is to obtain word representations that are useful in predicting surrounding words in a document. Given training words $w_1, w_2, \dots w_T$ and training context c, we want to maximize:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

where

$$p(w_{t+j}|w_t) = p(w_O|w_I) = \frac{\exp(\chi_{w_O}^T \rho_{w_I})}{\sum_{w+1}^{W}\exp(\chi_w^T \rho_{w_I})} \tag{1}$$

The parameters of this model are $\theta = \{\chi_{1:W}, \rho_{1:W}\}$. Problem: Need to compute derivatives for optimization which is very expensive.

5. View $p(w_O|w_I)$ as potential function. Can we do variational inference or stochastic variational inference here?

6. Skip gram model vs Bengio's (NPL) model, CBOW (Continuous Bag of Words Representation): In skip-gram we are predicting contextual words given the current word $\sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t)$. In NPL, CBOW: given the contextual words, we are predicting the current word $p(w_t|w_{t-1}, \ldots w_{t-c})$. We are still encoding contextual information in both; does this flip matter?

7. Didn't really discuss Hierarchical Softmax which is a computationally efficient approximation to the full Softmax.

8. NCE (Noise Contrastive Estimation) is introduced as an alternative to Hierarchical Softmax and Negative sampling is introduced as a simplification to NCE.

9. NCE - connections to Generative Adversarial networks? Possible relations to model checking and goodness of fit? We didn't really comment much here.

10. Negative Sampling: Has the following objective function which replaces $p(w_O|w_I)$ as in (1)

$$p(w_O|w_I) = \chi_{w_O}^T \rho_{w_I} - \log(1 + \exp(\chi_{w_O}^T \rho_{w_I})) \tag{2}$$

$$+ \sum_{k}^{K} -(\chi_{w_k}^T \rho_{w_I} - \log(1 + \exp(\chi_{w_k}^T \rho_{w_I}))) \tag{3}$$

11. Intuition: For each word w_I consider 2 classes - words w_O that co-occur with w_I and words w_k that don't co-occur with w_I (let's refer to them as noise). Maximizing $p(w_O|w_I)$ as given in (2) is maximizing the probability of being able to distinguish between w_O and the noise words w_k.

12. Can we add some regularization term to (2)? Since we are simultaneously learning $\{\chi_{1:W}, \rho_{1:W}\}$ this makes sense.

13. There was no consensus whether (2) is some sort of Taylor approximation or Monte Carlo estimate of (1). Or whether its a new objective function itself which gives high-quality word representations. Can variational inference or stochastic variational inference give something like Negative Sampling?

14. I go with the first view. Negative Sampling is some sort of approximation to NCE which in itself approximates (1).

Paper 2 - GloVe (Manning et.al.)

1. Two broad class of methods:
   (a) "global" matrix factorization (of the word-word co-occurrence matrix) type: which perform poorly on word analogy tasks
   (b) methods which take into account "local" contexts like the skip-gram model which does better on analogy tasks (indicating its learned a finer structure). But the model is trained on separate local context windows instead of leveraging information from the entire corpus.
2. Model: X - matrix of word-word co-occurrences, X_{ij} - the number of times word j occurs in the context of word j, $X_i = \sum_k X_{ik}$ - the number of times any word appears in the context of word k.
3. The contextual information is indeed captured through this big matrix. This combines "global" matrix factorization type ideas and local contextual type ideas.
4. $p_{ij} = p(w_O = j | w_I = i) \approx \frac{X_{ij}}{X_i}$. By assuming a linear structure, they make a log-linear model:
$$\log(p_{ij}) = \chi_j^T \rho_i \quad \Leftrightarrow \quad \log(X_{ij}) \approx \chi_j^T \rho_i + b_i + b_j$$
for some bias terms b_i and b_j. I have continued the notation of χ_j being the word representation and as in (1)
5. Criticism about the authors assuming a linear relationship in the model rather than coming up with a model that exhibits linear relationships (contrast this with Arora et. al. who come up with an underlying generative model and derive the "linear structure").
6. They then go on to solve a weighted least squares problem with the loss function as:
$$\sum_{i,j} f(X_{i,j})(\chi_j^T \rho_i + b_i + b_j - log(X_{ij}))^2$$

7. $f(\cdot)$ is just a weight function to account for "contagion" (a feature of certain words occurring together in many articles like the frequent co-occurrence of "in", "the"). The choice of $f(\cdot)$ seems ad-hoc. Better ways to deal with it? Integrated/Robust models?

8. A comment was made in class as to why the squared loss was chosen in the objective function. Objective function of the skip-gram model can be reduced to minimizing the weighted sum of cross entropy error (with long tails) and also requires computing the normalizing constant, both of which are undesirable. As an alternative, they choose the squared loss function; a very nice discussion on this is given in Section 3.1

9. There was discussion on some model selection questions that can be explored.

   (a) Choice of dimension of embeddings? Size of the context window?

   (b) Can we have different context lengths?? Can we weight along context lengths?

   (c) Bayesian Non-Parametric models for choosing context length? Cross-validation can be expensive.

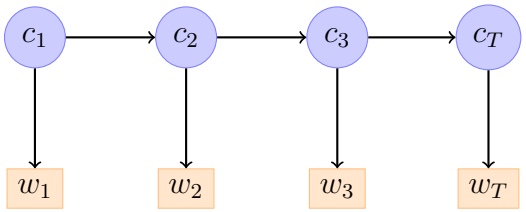
Paper 3 - RAND-WALK (Arora et.al.)

1. In the PMI (Pairwise Mutual Information) Model, we have a symmetric matrix whose entries are $P(w, w') = \log \frac{p(w,w')}{p(w)p(w')}$ where $p(w)$ is the marginal probability of word w and $p(w, w')$ is the probability of words w, w' appearing together in a given context size.

2. Low rank SVD on the PMI matrix gives the word vectors. Since the approximation $v_w^T v_{w'} \approx PMI(w, w')$ is good in practice, it points to the linear structure. Arora et. al. provide a generative model and show that

$$PMI(w, w') = \frac{v_w^T v_{w'}}{d} + \mathcal{O}(\epsilon) \tag{4}$$

   for some constants d and ϵ. Thus their generative model implies linear structure (as against Pennington et. al. who sort of assume it).



3. Their model is a latent variable random walk model as shown above.

4. Here $c_t \in R^d$ represents the discourse vector at time t, representing the topics currently being talked about. Every latent word vector $v_w \in R^d$ captures the correlation of word w with the discourse as:

$$p(\text{word } w \text{ emitted at time t}|c_t) \propto \exp(c_t^T v_w) \tag{5}$$

5. Using (5) and integrating over the discourse vectors c, they are able to derive (4).

6. We briefly tried understanding the relationship between analogies and likelihood ratios as given in Section 3 i.e. justification of the following statement:"woman" is the solution to the analogy "king:queen::man:y" because

$$\frac{p(\chi|king)}{p(\chi|queen)} \approx \frac{p(\chi|man)}{p(\chi|woman)}$$

   where χ is any word

   Below are some of my thoughts (which the readers may or may not find useful) on this issue; inspired by Maja and Adi's comments on piazza. They mention a nice thing about "interchangeability".

7. One observation is to look at $p(\chi|king)$ as the probability of χ occurring in the context of the word "king". Now how does ratio help? Lets view χ as "probe words". For probe words which are neither related to "king" or "queen" like "tree" or related to both like "marriage", the ratio is close to 1. For probe words that favor "king" more, like "fight" the ratio will be large and for those which favor "queen" like "beauty" it will be small.

This is the same intuition as given by Pennington et. al. in Table 1; one can see the magnitudes to get a sense of how ratios help in nicely classifying probe words into 3 buckets. Now we want a word y, whose context classifies probe words into the same buckets (when the ratio is taken with the context of "man").

8. We want to look for "y" such that the probe words which capture similar dimensions of "king" and "queen" would also capture similar dimensions of "man" and "y". Probe words which favor the king capture some dimension w.r.t. queen - say "masculinity" and those which favor the "queen" capture "femininity"

   (Pardon any poor choice of examples in the above paragraphs)

9. In some sense these probe words define a direction in the space of word representations - to go from "king" to "queen". Following the same direction from "man" leads to "woman"

10. This is more formally put in the random walk model. There's an underlying slow random walk of the discourse vector c_t and the probe words capture similar features between "king" and "man" by their correlation to c_t.

$$\frac{p(\chi|king)}{p(\chi|man)} = \sum_t \frac{\frac{\exp(c_t^T v_w)}{\exp(c_t^T v_{king})}}{\frac{\exp(c_t^T v_w)}{\exp(c_t^T v_{man})}} = \sum_t \exp(c_t^T (v_{king} - v_{man}))$$

$$\frac{p(\chi|queen)}{p(\chi|y)} = \sum_t \exp(c_t^T (v_{queen} - v_y))$$

   So essentially $(v_{queen} - v_y)$ must be aligned along $(v_{king} - v_{man})$

11. Probe words are used since c_t are unobserved latent variables that are marginalized out. In equations 3.4, 3.5 and 3.6, they show that $v_{king} - v_{man} = v_{queen} - v_{woman} = \mu_R +$ small noise. Proof of small noise follows from the isotropic assumption.