Notes on “ORDER–EMBEDDINGS OF IMAGES AND LANGUAGE (2016)”

1.Motivations

- In a word, they adapt partial order to Embedding explicitly.
- More specifically, hypernymy, textual entailment, and image captioning can be seen as special cases of a single visual-semantic hierarchy over words, sentences, and images. In this paper, they advocate for explicitly modeling the partial order structure of this hierarchy.

2. Contributions

- They firstly proposed a general approach of learning Order-Embedding.
- In different applications, the main difference is how to sample negative examples.

2.1 Definition

- They tackle this problem by learning a mapping from X into a partially ordered embedding space $(Y, \preceq_Y)$ and to predict the ordering of an unseen pair in X based on its ordering in the embedding space.
- We define $\preceq$ as $x \preceq y$ *if and only if* $\bigwedge_{i=1}^{N} x_i \geq y_i$
- Instead of viewing our embeddings as single points $x \in \mathbb{R}_+^N$, we can also view them as sets $\{y : x \preceq y\}$.

2.2 Loss Function

- We define $E(x, y) = \|\max(0, y - x)\|^2$.
- Then the loss function is $Loss = \sum_{(u,v)\in P} E(f(u), f(v)) + \sum_{(u',v')\in N} \max\{0, \alpha - E(f(u'), f(v'))\}$

3. Experiments and Results

- They tested 3 applications of Order-Embedding. Those 3 applications are in the fields of hypernym prediction, caption image retrieval, and textual entailment.
- In a word, this idea works. And in some cases, it works great.
- Noticeable, in caption image retrieval, they use two-level partial order with captions above the images they describe. Shallow partial orders like this work because: **Symmetric similarity should fail when an image has captions with very different levels of detail, because the captions are so dissimilar that it is impossible to map both their embeddings close to the same image embedding. Order-embeddings don’t have this problem: the less detailed**

caption can be embedded very far away from the image while remaining above it in the partial order.

4. Insights and Questions

- Since the visual-semantic hierarchy is an antisymmetric relation, they expect this approach to introduce systematic model error. This *order-preserving* approach, instead of the conventional *distance-preserving* approach, directly impose the transitivity and antisymmetry of the partial order. In this way, models don't need to learn this prior by themselves.
- As it should be extremely sparse in $\mathbb{R}_+^N$, if we randomly embed an instance, it's very likely that it can't be compared with other instances. This is especially dangerous as $N \geq 1000$ in latter the two applications. But why this is not a problem in practice? I'm not so sure yet.