Learning Deep Structure-Preserving Image-Text Embeddings

2019.2.22

Paper

- Wang, Liwei, Yin Li, and Svetlana Lazebnik. "Learning deep structure-preserving image-text embeddings." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Instroduction

- Computer vision is moving from predicting discrete, categorical labels to generating rich descriptions of visual data.

- A core problem for these applications is **how to measure the semantic similarity** between visual data (e.g., an input image or region) and text data (a sentence or phrase).

- A common solution is to **learn a joint embedding for images and text into a shared latent space** where vectors from the two different modalities can be compared directly. This space is usually of low dimension and is very convenient for cross-view tasks such as image-to-text and text-to-image retrieval.

Instroduction

- Several recent embedding methods are based on Canonical Correlation Analysis (CCA), which finds **linear projections that maximize the correlation** between projected vectors from the two views.

- However**, CCA is hard to scale to large amounts of data**. In particular, stochastic gradient descent (SGD) techniques cannot guarantee a good solution to the original generalized eigenvalue problem, since covariance estimated in **each small batch (due to the GPU memory limit) is extremely unstable**.

Instroduction

- An alternative to CCA is to learn a joint embedding space using SGDwith a ranking loss. WSABIE and DeVISE learn linear transformations of visual and textual features to the shared space using a single-directional ranking loss that applies a margin-based penalty to incorrect annotationsthat get ranked higher than correct ones for each training image.

- Compared to CCA-based methods, this ranking loss easily scales to large amounts of data with stochastic optimization in training.

- As a more powerful objective function, a few other works have proposed a bi-directional ranking loss that, in addition to ensuring that correct sentences for each training image get ranked above incorrect ones, also ensures thatfor each sentence, the image described by that sentence gets ranked above images described by other sentences.

- However, to date, it has proven frustratingly difficult to beat CCA with an SGD-trained embedding (properly normalized CCA)

- Another strand of research on multi-modal embeddings is based on deep learning. By making it possible learn nonlinear mappings, deep methods can provide greater representational power than methods based on linear projections.

Instroduction

- In this work, we propose to learn an image-text embedding using a two-view neural network with two layers of nonlinearities on top of any representations of the image and text views.

- These representations can be given by the outputs of two pre-trained networks, off-the-shelf feature extractors, or trained jointly end-to-end with the embedding.

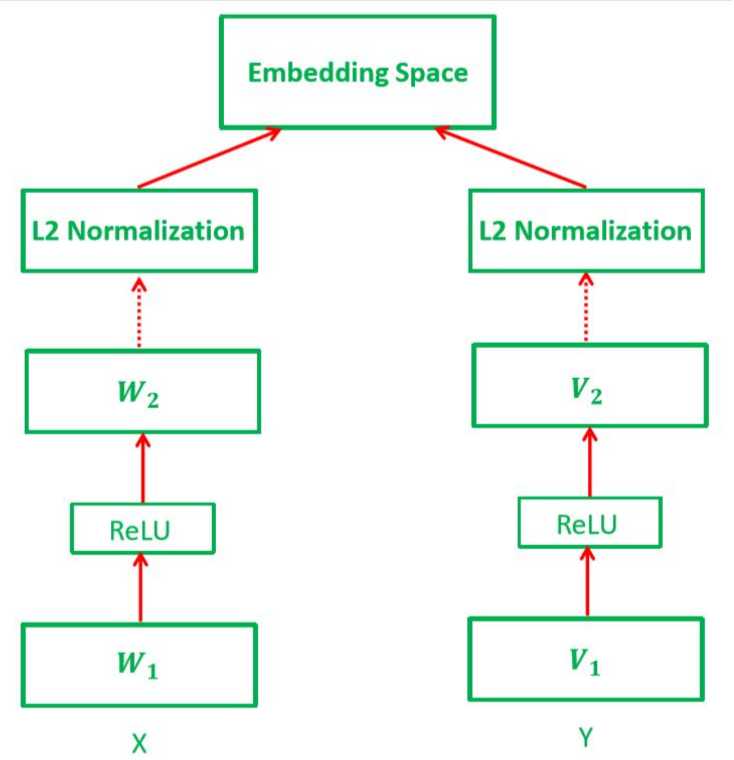


Figure 1. Our model structure: there are two branches in the network, one for images (X) and the other for text (Y). Each branch consists of fully connected layers with ReLU nonlinearities between them, followed by L2 normalization at the end.

Instroduction

- To train this network, we use a bi-directional loss function, combined with constraints that preserve neighborhood structure within each individual view.

- Our method can avoids Deep CCA's training-time difficulties associated with covariance matrix estimation.

- Our network also gains in accuracy by performing feature normalization (L2 and batch normalization) before the embedding loss layer.

- Finally, two branches share weights.

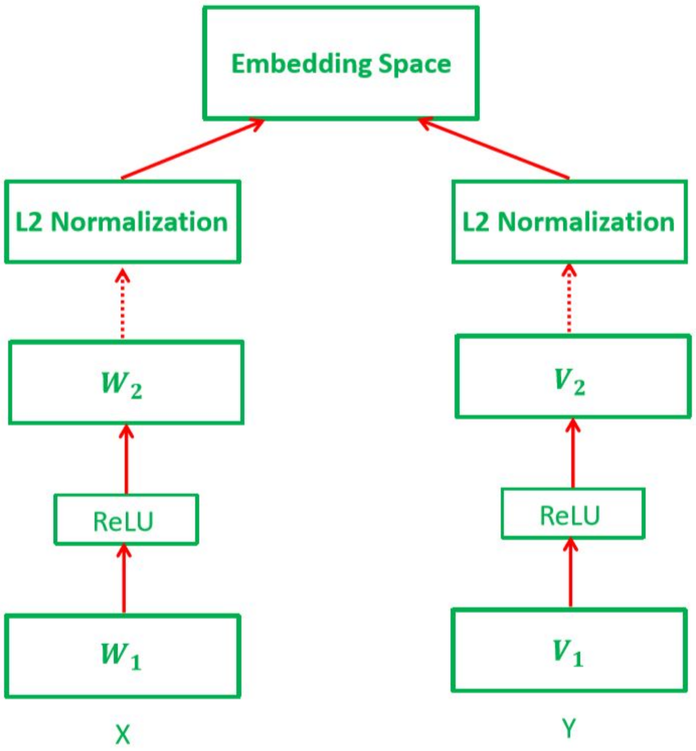


Figure 1. Our model structure: there are two branches in the network, one for images (X) and the other for text (Y). Each branch consists of fully connected layers with ReLU nonlinearities between them, followed by L2 normalization at the end.

Deep Structure-Preserving Embedding

Network Structure

- X,Y : collections of training images and sentences, each encoded according to their own feature vector representation.

- d(x, y) : Euclidean distance between image and sentence vectors in the embedded space.

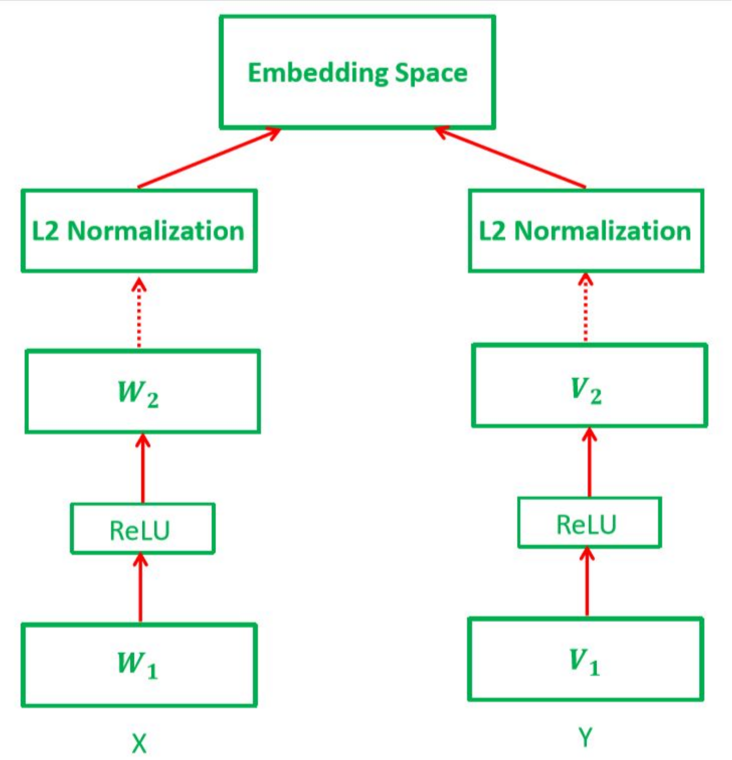


Figure 1. Our model structure: there are two branches in the network, one for images (X) and the other for text (Y). Each branch consists of fully connected layers with ReLU nonlinearities between them, followed by L2 normalization at the end.

Deep Structure-Preserving Embedding

Training Objective

$$d(x_j, y_i) + m < d(x_{\acute{k}}, y_i) \;\; \forall x_j \in X_i^+, \forall x_{\acute{k}} \in X_i^- \;\; (2)$$

Deep Structure-Preserving Embedding

Training Objective

- **Structure-preserving constraints**
  - $N(x_i)$: neighborhood of x_i containing images that share the same meaning. set of images described by the same sentence as x_i
  - we want to enforce a margin of m between $N(x_i)$ and any point outside of the neighborhood:

  - Analogously to (3), we define the constraints for the sentence side as :

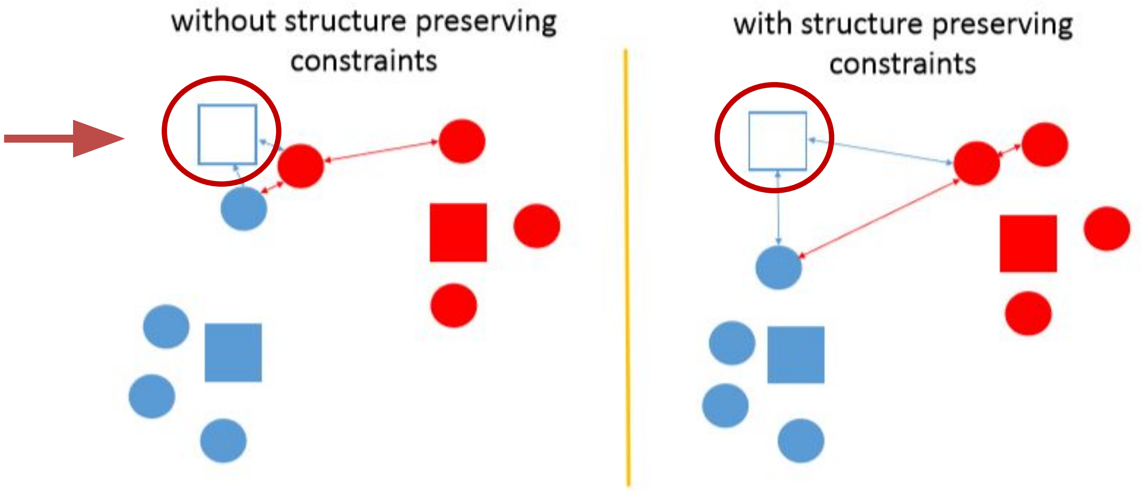
$$d(y_i, y_j) + m < d(y_i, y_{\acute{k}}) \ \ \forall y_j \in N(y_i), N(y_i) \nexists \forall y_{\acute{k}} \ (4)$$

  - $N(y_i)$ contains sentences describing the same image.

Deep Structure-Preserving Embedding

Training Objective

- Each square (representing an image) is closer to all circles of the same color (representing its corresponding sentences) than to any circles of the other color. Similarly, for any circle (sentence), the closest square (image) has the same color.


without structure preserving constraints
with structure preserving constraints

New image

Figure 2. Illustration of the proposed structure-preserving constraints for joint embedding learning (see text). Rectangles represent images and circles represent sentences. Same color indicates matching images and sentences.

Deep Structure-Preserving Embedding

Training Objective

Deep Structure-Preserving Embedding

Training Objective

- **Triplet sampling**
  - we sample triplets within each minibatch and optimize our loss function using SGD.
  - instead of choosing the most violating negative match in all instance space, **we select top K most violated matches in each mini-batch.**

  - This is done by computing pairwise similarities between all (xi , yj), (xi , xj) and (yi , yj) within the mini-batch.
  - For each positive pair (i.e., a ground truth image-sentence pair, two neighboring images, or two neighboring sentences), we then find at most top K violations of each relevant constraint (we use K = 50 in the implementation, although most pairs have many fewer violations).

  - we randomly sample 1500 pairs (xi , yi) to form our minibatches.
  - for each xi in a given minibatch, we add one more positive sentence distinct from the ones that may already be included among the sampled pairs, resulting in mini-batches of variable size.

Experiments

- we analyze the contributions of different components of our method

- we evaluate our method on image to-sentence and sentence-to-image retrieval
    - dataset : popular Flickr30K and MSCOCO datasets

- we evaluate our method phrase localization on
    - dataset: new Flickr30K Entities dataset.

Experiments

Features and Network Settings

- Given an image, we extract the 4096-dimensional activations from the 19-layer VGG model.

- To represent sentences and phrases, we primarily use the Fisher vector (FV) representation.
    - 300(word2vec) * 30(a codebook with 30 centers) * 2(first and second-order information) = 18000
    - using PCA, 18000 -> 6000 (To save memory and training time)

- we are also interested in exploring the effectiveness of our approach on top of simpler text representations.
    - To this end, we include results on 300-dimensional means of word2vec vectors of words in each sentence/phrase, and on tf-idf-weighted bagof-words vectors.
    - For the Flickr30K dataset, our dictionary size (and descriptor dimensionality) is 3000, and for MSCOCO, it is 5600.

Experiments

Features and Network Settings

- For our experiments using tf-idf or FV text features,
    - set the embedding dimension to be 512

- On the image (X) side, when using 4096-dimensional visual features,
    - W1 is a 4096 × 2048 matrix.
    - W2 is a 2048 × 512 matrix.
    - the output dimensions of the two layers are [2048, 512].

- On the text (Y) side, the output dimensions of the V1 and V2 layers are [2048, 512].
- For the experiments using 300-D word2vec features, we use a lower dimension (256) for the embedding space and the intermediate layers output are accordingly changed to [1024, 256].

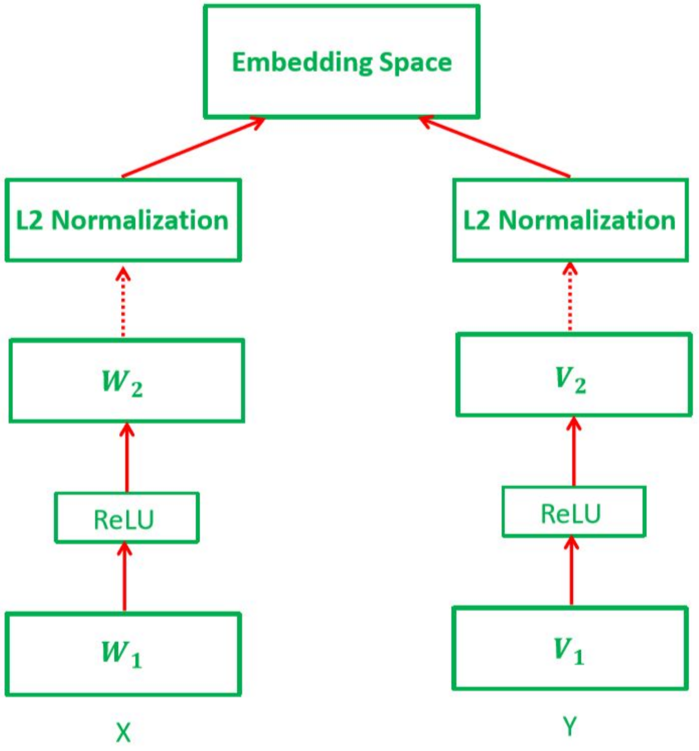


Figure 1. Our model structure: there are two branches in the network, one for images (X) and the other for text (Y). Each branch consists of fully connected layers with ReLU nonlinearities between them, followed by L2 normalization at the end.

Experiments

Image-sentence retrieval

- image-to-sentence and sentence-to-image retrieval on the standard Flickr30K and MSCOCO datasets.
    - Flickr30K : 31783 images accompanied by five descriptive sentences each.
    - MSCOCO : 123000 images, also with five sentences each.

- For Flickr30K, given a test set of 1000 images and 5000 corresponding sentences, we use the images to retrieve sentences and vice versa, and report performance as Recall@K (K = 1, 5, 10), or the percentage of queries for which at least one correct ground truth match was ranked among the top K matches.

- For MSCOCO, we also report results on 1000 test images and their corresponding sentences.

Experiments

Image-sentence retrieval

Experiments

Image-sentence retrieval

- For MSCOCO, results on 1000 test images are listed in Table 2. The trends are the same as in Table 1.
- We have also tried fine-tuning the VGG network by backpropagating our loss function through all the VGG layers, and obtained about 0.5% additional improvement.

Experiments

Phrase Localization on Ficker30K Entities

- The recently published Flickr30K Entities dataset allows us to learn correspondences between phrases and image regions.
  - Specifically, the annotations in this dataset provide links from 244K mentions of distinct entities in sentences to 276K ground truth bounding boxes (some entities consist of multiple instances, such as “group of people”).

- . For text, in this section we use only the FV feature. Thus, the input dimension of X is 4096 and the input dimension of Y is 6000 as before (reduced by PCA from the original 18000-D FV). We use the two-layer network structure with [8192, 4096] as the intermediate layer dimensions on both the X and Y sides (note that on the X side, the intermediate layer actually doubles the feature dimension).

- it is resampled with at most ten regions per phrase

- we augment the mini-batches by sampling not only additional positive phrases for regions, but also additional positive regions for phrases, to make sure that we have as many triplets as possible for structure-preserving constraints on the region side (eq. 3) and the phrase side (eq. 4).

Experiments

Phrase Localization on Ficker30K Entities

- In order to further improve the accuracy of our embedding, we need to refine it using negative data from background and poorly localized regions.

- To do this, we take the embedding trained without negative mining, and for each unique phrase in the training set, calculate the distance between this phrase and the ground truth boxes as well as all our proposal boxes.

- Then we record those "hard negative" boxes that are closer to the phrase than the ground truth boxes. For efficiency, we only sample at most 50 hard negative regions for each unique phrase.

- Next, we continue training our region-phrase model on a training set augmented with these hard negative boxes, using only the bi-directional ranking constraints (eqs. 1 and 2).

- We exclude the structure-preserving constraints because they would now be even more severely outnumbered by the bidirectional ranking constraints

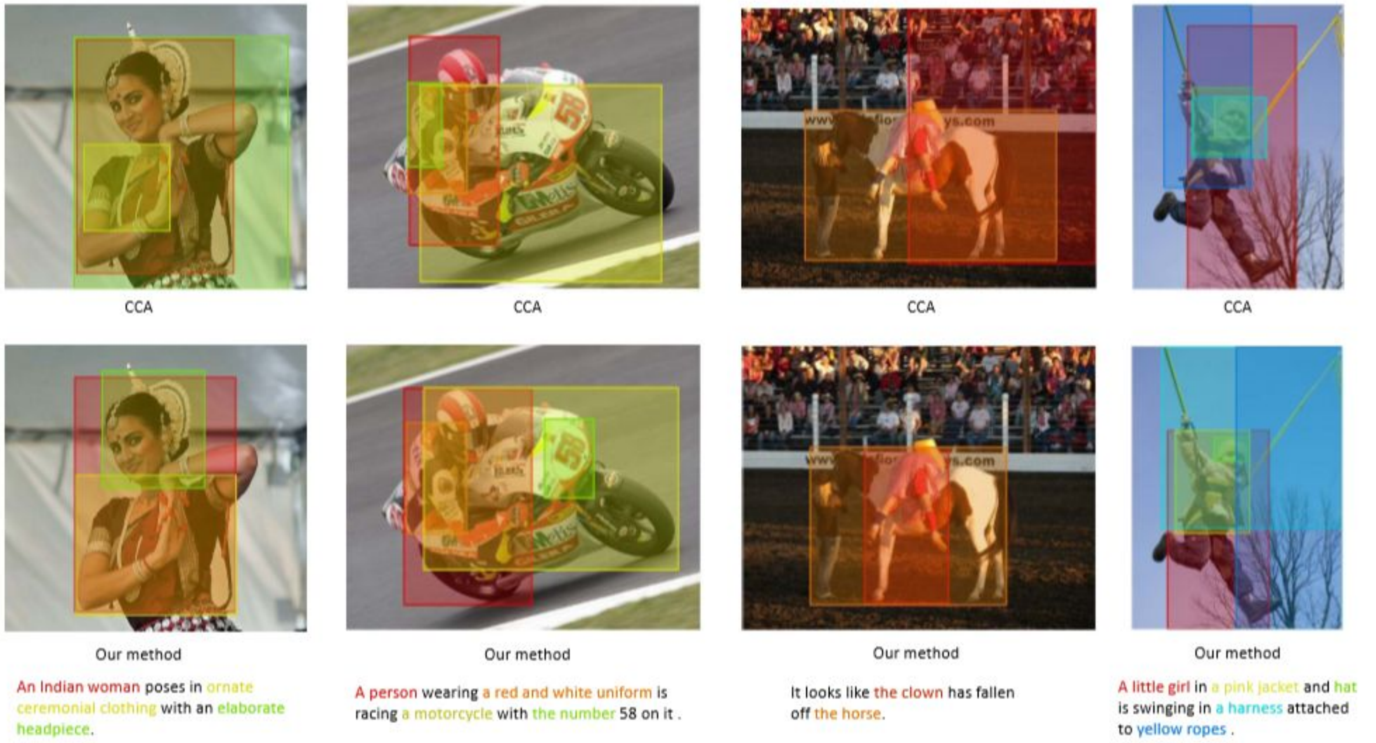
Experiments

Phrase Localization on Ficker30K Entities

- Flickr30K Entities training set, for all 130K pairs, there are around 70K unique phrases and 80K regions described by a single phrase

Experiments

Phrase Localization on Ficker30K Entities



An Indian woman poses in ornate ceremonial clothing with an elaborate headpiece.

A person wearing a red and white uniform is racing a motorcycle with the number 58 on it .

It looks like the clown has fallen off the horse.

A little girl in a pink jacket and hat is swinging in a harness attached to yellow ropes .

Figure 3. Example phrase localization results. For each image and reference sentence, phrases and best-scoring corresponding regions are shown in the same color. The first row shows the output of the CCA method [37] and the second row shows the output of our best model (fine-tuned model (d) in Table 3 with negative mining). For the first (left) example, our method gives more accurate bounding boxes for the clothing and headpiece. For the second example, our method finds the correct bounding box for the number 58 while CCA completely misses it; for the third column, our method gives much tighter boxes for the horse and clown; and for the last example, our method accurately locates the hat and jacket.

Conclusion

- This paper has proposed an image-text embedding method in which a **two-branch network** with multiple layers is trained **using a margin-based objective function consisting of bi-directional ranking terms and structure-preserving terms** inspired by metric learning.

- Our architecture is simple and flexible, and **can be applied to various kinds of visual and textual features**. Extensive experiments demonstrate that the components of our system are well chosen and all the terms in our objective function are justified.

- To the best of our knowledge, our retrieval results on Flickr30K and MSCOCO datasets considerably exceed the state of the art, and we also demonstrate convincing **improvements over CCA on the new problem of phrase localization**on the Flickr30K Entities dataset.

Thank You!

Do you have any question?