Ensemble Based Image Retrieval for Textual Descriptions

Abstract

This papers aims to find, amidst a collection of images, the one that is most related to a given set of descriptions. To aid in this, we are additionally provided with tags and extracted ResNet features for the images. We propose an ensemble architecture that involves Bag of Words (BoW) representations of the descriptions and tags that are used to train a pair of regression models. The paper compares and contrasts the approaches taken along the way to our final configuration, with results shown for each. Finally, we discuss further improvements and recommendations to achieve better results.

1 Introduction

This papers approaches the problem of image retrieval, i.e., finding amidst a set of images the one that is best described by the given sets of descriptions.

To aid us in this, we are provided features extracted from the pool5 and fc1000 layers of ResNet, a deep-learned convolutional neural network. These features provide a class prediction for an image that has been learnt through successive layers of convolutions and pooling with the aid of big amounts of labeled images, e.g., ImageNet.

Traditionally in such scenarios, the image is represented by its extracted features and the text descriptions are represented by a bag of words. Following this, transformations are performed on either side that move them to a common subspace where similarities can be computed.

Our work follows such an approach. But, our novelty is that, in order to provide an added layer of confidence and deterrence from feature noise, we predict the most likely tags for a given description and add its weight in considering similarities.

2 Baseline Configuration

The task was to create a mapping between the description and the images in the high-dimensional subspace. The following is a list of steps we followed in the pursuit of finding an optimal architecture and algorithm that produce the best results.

2.1 Pre-processing

For the baseline solution we started by pre-processing the descriptions. We used python methods and nltk models to achieve this. The following are the steps performed:

- Stopword Removal
- Punctuations
- Lowercasing
- Lemmatization

2.2 Pool5 vs FC1000

We had two types of features from the ResNet, pool5 and fc1000, the latter being the final classification layer which ties the images to the concepts. We ran the regression models using both individually and got the results as given in table 1. The pool5 tend to perform better in all the cases. This maybe due to the fact that, being a penultimate layer, it captures more underlying features and consequently provides more richness than fc1000.

2.3 Word2Vec and Dimensionality Reduction

We tried using **Word2Vec** to convert each word in the description into an embedding. The output of this processing was a 300 dimensional vector which we had to map to 2048 vector of ResNet pool5 features.

The initial approach to reducing the pool5 features involved selecting columns randomly and by order of variance.

This consequently led to the idea of applying PCA on the image features to achieve optimum selection. Through trial and error, we discovered that using 80 principal components yielded the optimum accuracy on our training-test split.

We also trained the models using a BoW model for the descriptions and got mildly better performance under some regressions. The results of these approaches are shown in Table 1.

2.4 Regression Models

We trained models using various Regression models. PLSR gave good results but at the expense of a lot of computation time and resources. Finally, Kernel Ridge Regression was chosen as it gave the best results along with shortest execution time, which allowed us to iterate.

The scores were obtained based on calculating the Cosine Distance from the predicted image features for a given description to the actual image. These scores for the various approaches are displayed in table 1 below. The MAP 20 output for the best approach came out to be **0.301**

Configurations	MAP@20 Score
WORD2VEC-POOL5-TOP400VAR	0.122
BOW-POOL5-RAND100	0.158
BOW-POOL5-TOP400VAR	0.222
BOW-FC1000-PCA80	0.26
BOW-POOL5-PCA80	0.301

Table 1: Performance of different configurations on baseline approach

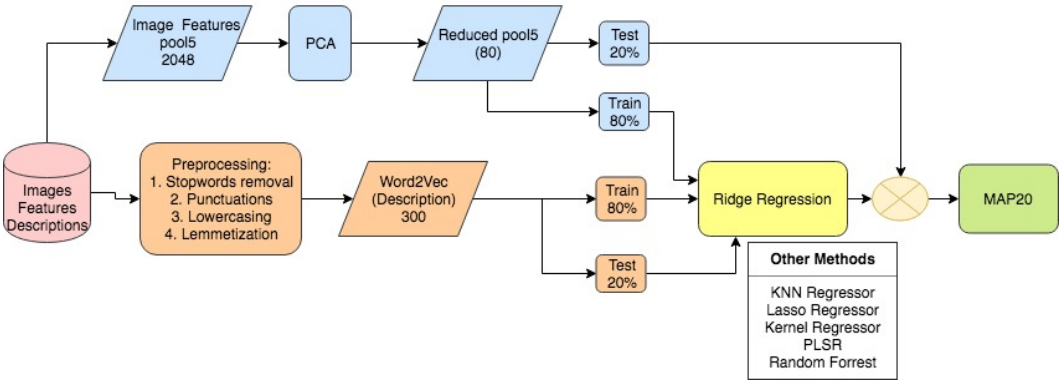

Figure 1: Baseline Approach

3 Motivations for Final Approach

In deciding where to focus our efforts to further improve the accuracy, we arrived on the below areas.

3.1 Combined Bag Of Words and Word2Vec Representation

A drawback of Bow is that its vocabulary is bounded by the words used in the multi-modal training data, which is at a relatively small scale compared to a text corpus containing millions of words. To compensate for such a loss, we further leveraged word2vec.[1]

By learning from a large-scale text corpus, the vocabulary of word2vec is much larger than its BoW counterpart. We obtain the embedding vector of the sentence by mean pooling over its words, i.e.,

$$s_{word2vec} := \frac{1}{|q|} \sum_{n=1}^{\infty} v(w) \tag{1}$$

Multi-scale sentence vectorization is obtained by concatenating the three representations, that is

$$s(q) = [s_{bow}(q), s_{word2vec}(q)] \tag{2}$$

3.2 Tags Prediction

On inspecting the images chosen in the top 20, we noticed that the most likely image was being pushed further back in the rankings due to the appearance of images that had seemingly no relation with the descriptions.

We attributed this to noise in the learnt CNN model that was making it have a similar distance d_i to the description vector. We realised that one effective way to weed out the unrelated images was to leverage the tags information.

$$d_i = \alpha_1 . d_f + \alpha_2 . d_{tag} \tag{3}$$

In the above equation d_{tag} refers to the cosine distance between the vectorized representation of the predicted tags and the actual tags (if they exist). The weight coefficients were estimated based on an iterative approach that maximized the accuracy on the training data.

4 Final Configuration

The figure shown depicts our final approach taken for the problem, which is an ensemble based method trained on descriptions to predict both tags and the pool5 ResNet features. The predicted features are then compared with the actual ones to output the top 20 images.

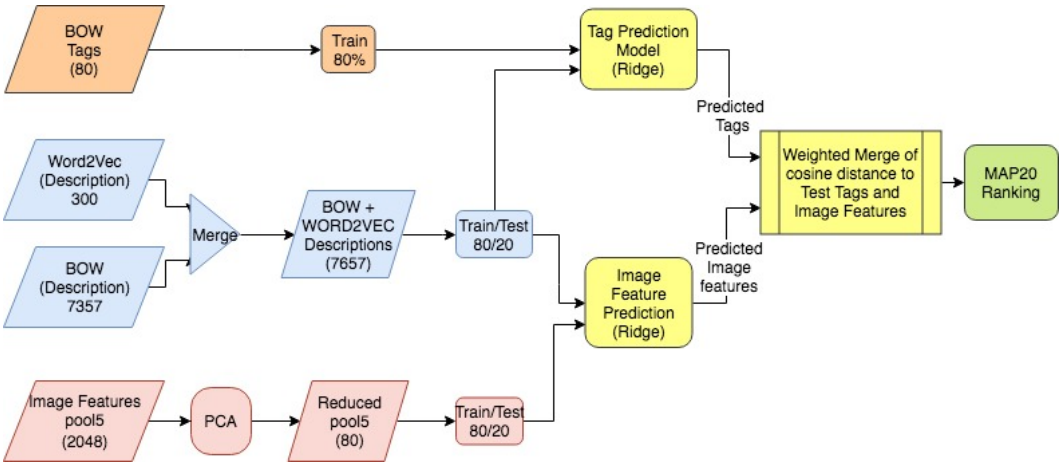

Figure 2: Ensemble based final approach

5 Experimentation and Evaluation

The evaluation metrics used for this exercise is MAP@20(Mean Average Precision at 20). This is given in the formula below wher i refers to the index of the one true image in the list of 20.

$$score = \frac{20 + 1 - i}{20} \tag{4}$$

Below are the results of running our final configuration with different regression models.

Model	MAP@20 Score
Random Forest	0.202
Lasso	0.324
Ridge	0.385
Kernel Ridge	0.4008

Table 2: Performance of different models in Final approach

6 Kaggle Score

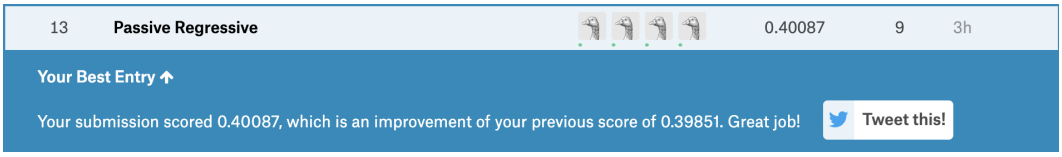

Figure 3: Final score on Kaggle

7 Conclusion

The challenge of accurately matching images to descriptions is a complex problem that can be approached from a multitude of angles. Through this experiment, the accuracy of multiple algorithms and models were compared to each other. We tried to use different architectures and combine multiple models together to get better results. It was found that the best model is one that requires extensive data pre-processing and makes use of a variety of algorithms whose predictions can be combined in order to get the most accurate predictions. We achieved accuracy of **0.40** using basic machine learning techniques.

The future work would be to get a better mapping from description by using n-grams and features which preserve contextual information in the data. Trying out Deep Learning models is also an option but would require more data and computation. Another approach could be to use the features to predict the descriptions for a test image and rank it based on the similarity with the given description. We also found PLSR to be promising in terms of accuracy but could not iterate enough to get to a conclusion. It remains as another viable option in the presence of additional computational resources.

References

[1] Jianfeng Dong, Xirong Li, and Cees G. M. Snoek. *"Predicting Visual Features from Text for Image and Video Caption Retrieval"*. IEEE TRANSACTIONS ON MULTIMEDIA, 2018.