Multimodal Image inspired hashtag generator

IRE Team 13

1 Abstract

The goal of this project is to generate hashtags given a multi-modal post on OSM like Instagram that may contain both images and text content. The user might also provide a set of few seed hashtags as input and the generated hashtags should be relevant to these hashtags as well. In the Figure 1, we are given an image, text content and certain hashtags (#dog, #birthday) as the inputs [1]. Our proposed method needs to utilize this data and suggest few other hashtags like #celebration, #party or #puppy for the same post. The project thus requires us to leverage the latest methods in computer vision such as Convolutional Neural Networks (Resnet50, Transfer learning) and in NLP such as glove embeddings, and word2vec to get a high precision/recall on the given task. We propose a two level hierarchical system where the predictions made by an image classifier are subsequently used by the corresponding text based approach.

Figure 1: Predicting the correct hashtag: #dog is difficult without the image [7]

2 Methodologies and Dataset

2.1 Dataset

A few datasets such as YFCC100M [6] and HARRISON dataset [4] deal with the task of hashtag generation using images. However, these datasets do not provide us with text data that is required for the training of our multimodal system. A few other datasets such as [1] deal with the task of personalized image captioning given text and image data along with prior user posts.

[1] There might be cases when any of these inputs is missing. The system should be able to take care of this as well

Thus, we collect our own dataset for the given task by scraping posts from Instagram using Selenium as the official Instagram API has rate limits. Our dataset consists of hashtags collected from a diverse set of 10 topics mentioned below:

1. Pets
2. Jewellery
3. Travel
4. Babies
5. Selfie
6. Art
7. Food
8. Architecture
9. Nature
10. Fashion

For each of the above mentioned topics, we manually chose the top 7 hashtags and scrape first 50 pages of Instagram explore (800-900 posts) for each of the hashtag. So we will end up having a data of 50k to 60k texts and images respectively. [2]. The dataset can be found here: Drive link

2.2 Image Classifier

The image classifier to be used will be a Resnet50 model that gave very high accuracy in the Imagenet challenge. We intend to leverage transfer learning to finetune the model on our dataset. Only the last layer of the model will be trained and we won't perform backpropagation on the remaining layers of the network. The loss function used will be Binary Cross Entropy loss or the normal cross entropy loss depending on whether it is multi-label or single label tasks respectively.

As of now, a Resnet50 [2] model with Binary cross entropy loss was trained to do multilabel predictions. However, the model isn't able to generalize well for the given data and the same model will also be tested for single label topic prediction.

2.3 Text Classifier

We have trained a word embedding approach like Word2Vec [3] and Glove embeddings [5] on our corpus and used distance measures such as cosine similarity to get the most relevant hashtags. As one of the baselines, we also aim to implement a simple topic modelling based approach (LDA) for the task of hashtag generation. The approach is detailed below.

A sample post representing the typical posts on Instagram has been shown in Figure 2.

Figure 2: There is a lack of sufficient text

Typically a post consists of the following:

1. A photo
2. A caption of 2-3 lines
3. A bunch of hashtags

[2]We have collected over 40k posts for Travel and Food already and the method is thus scalable

From above it is clear that the amount of text present is very less and therefore using only the text for training our model won't be enough. We tackle the above problem using below 2 approaches:

- Assuming the hashtags that appear together are semantically related to each other we can also convert them to text and include them in our training corpus. Using the above assumption we prepare our corpus by first preprocessing (removing non-english texts, emoticons, etc.) and then training two different models on the prepared corpus.
  - **Word2Vec**: Using 30 epochs and window size of 5 we trained the model to create 300 dimensional word embeddings.
  - **GloVe**: Using 30 epochs for training we trained the model to create 300 dimensional word embeddings.
- To better capture the semantic relationship between words we use pre-trained word embeddings for a corpus size of 400K words trained on wikipedia data. We train our word2Vec model with our corpus on top of it which helps to capture the context of the instagram corpus. Therefore our vocabulary is able to cover more words and is able to capture both the generic and insta-specific meaning of a word.

2.4 Combining both approaches

We will train the image classifier to predict the category to which a post belongs (10 categories mentioned above), and train 10 different embeddings for each individual topic. We will use glove or ELMO for these word embeddings. During test time, given a new post with an image, the image classifier will classify into one of the 10 categories and then the embedding for that particular category will be used to generate the new hashtags. In case no text is provided, we will simply return the top K hashtags for the topic predicted by the classifier.

3 Implementations and Code

All the codes written by us is available in the given repo: GitHub link. The repo contains all our data collection scripts (Instagram scraper), image classifier code and the code for Word2Vec and Glove embedding training.

The text based approaches perform well for our given data and will be complemented greatly if given additional information by the image classifier as well. Some sample outputs of the text based approach are given in Figure 3 and 4.

Figure 3:
Predicted Hashtags (Approach 1a): #ig #amazing #life #beauty #inspiration
Predicted Hashtags (Approach 1b): #view #mountain #world #amazing #ig
Predicted Hashtags (Approach 2): #mountain #beauty #amazing #view #landscapes

Figure 4:
Predicted Hashtags (Approach 1a): #hiking #mountain #adventure #forest #outdoor
Predicted Hashtags (Approach 1b): #life #city #mountain #lifestyle #trip
Predicted Hashtags (Approach 2): #hiking #mountain #beauty #adventure #landscapes

4 Difference in methodology from scope document

One major difference in methodology is that the initial scope of the project was predicting hashtags only for the travel category. However, in order to introduce more diversity for the models during training, we now intent to train our models on hashtags from multiple topics mentioned above.

Another difference in our methodology was that we intended to present the hashtag generation task using images as a multi-label classification of over 100 hashtags. However, such an approach is slightly ill-posed based on our experiments, as an Instagram post could have hashtags like #Travel, #Travellife, #Travelfun and classifying an image into these categories is an inherently difficult task. We thus, now pose the image classification task as a single category classification of an image into 10 major topics defined earlier by us.

5 New timeline and deliverables

We will train an image classifier and a text based model on the larger dataset separately and then combine them together as mentioned earlier in a hierarchical manner. We will also implement a simple topic modeling and image only baseline to compare them with our proposed methods.

References

[1] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 895–903, 2017.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[4] Minseok Park, Hanxiang Li, and Junmo Kim. Harrison: A benchmark on hashtag recommendation for real-world images in social networks, 2016.

[5] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[6] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[7] Qi Zhang, Jiawen Wang, Haoran Huang, Xuanjing Huang, and Yeyun Gong. Hashtag recommendation for multimodal microblog using co-attention network. In *IJCAI*, pages 3420–3426, 2017.