Social Similar Image Detection and Ranking Based on Subreddit Topics

Gendong Zhang
Stanford University
zgdsh29@stanford.edu

Zixuan Zhou
Stanford University
zixuan95@stanford.edu

Liuming Zhao
Stanford University
lxz299@stanford.edu

Abstract

Recently extensive researches have been done on semantic analysis to Reddit contextual post but only few on visual learning of Reddit image contents. In this project, we investigated both image classifications and rankings based on Reddit topics. Three deep learning models have been fully analyzed: Siamese Network, Triplet Network and VisNet. Triplet Network has the best performance among all 3 models. Triplet Network can achieve the highest of 90.7% top3 and 0.43 MRR when detecting and ranking duplicate images, which achieves much higher accuracy compared with Siamese Network and VisNet. In addition to fake image detection, we also extended our Triplet model into NSFW image detections, in which the Triplet Network can reach over 70% accuracy in detecting and ranking NSFW images. This Triplet model can also outperform VGG-16 and ResNet-18 with 91.7% accuracy in classifying Reddit image based on subreddit topics.

1. Introduction

1.1. Background

As an online social media and forum, Reddit contains more than 70 million submissions, 700 million comments and over 540 million monthly visitors. The site is divided into several sub-communities, referred as "subreddit" based on different topics. Users can post comments along with some images under the corresponding subreddit topics. Thus, Reddit offers larger textual and visual information for various analysis. A lot of studies have already been done with the semantic analysis to Reddit contextual information, such as classification and generation of Reddit Post titles, but only few focus on applying machine learning models to Reddit post images. In this project, we exploded several deep learning techniques for Reddit images based on their subreddit topics.

1.2. Objective

Image Ranking, Fake Image Detection, NSFW Image Detection, Image Classification

1.3. Related Work

Near duplicate image detection has been an important task in image search and ranking field. Chum [3] proposed to use a visual vocabulary of vector quantized local feature descriptors (SIFT) and enhanced min-Hash techniques to retrieve image in the context of near duplicate image detection. Liu [6] used a image ranking approach to web image search, in which he used social data from social media platform jointly with visual data to improve the relevance between returned images and user intentions. Wang [9] utilized the Convolution Neural network to detect and find the fine grained image similarity.

2. Dataset

For this project, we will use Reddit and CASIA v2 [4] dataset to train and test our model.

2.1. Reddit Images

The Reddit dataset consists of 30479 images scrawled from Reddit site, each of which is labeled by its corresponding subreddit topic. In total, we got 15 different subreddit topics for all the images scrawled from Reddit site: beard, cycling, cats, dog pictures, fishing, golf, guns, lego, minecraft, NBA, pizza, pokemon, starwars, watches, xray. We used 28979 images for training and 1500 images for testing. For the test set, apart from the 1500 images from 15 subreddit topics listed above, we also added additional 500 images from extra 5 NSFW subreddits for NSFW Image Detection task. The distribution of each subreddit in the dataset is uneven, and within one subreddit, images do not look the same, since users can post any image of their interest to any subreddit. Figure 2 is an example of fishing subreddit. As can be seen that man with fish, fish or even no man no fish images can be included in the fishing subreddit, therefore, this dataset is very challenging for training.

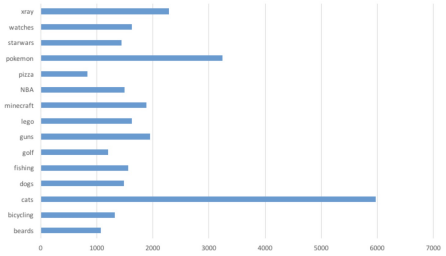

Figure 1: List of the 15 subreddits of the training images from Reddit site

Figure 2: One Set of Example of Fishing Subreddit

2.2. CASIA dataset

In addition to the Reddit images, we also used CASIA dataset specifically for fake image detection task. The CASIA dataset consists of 1935 authentic images with 5123 fake images. The fake images were generated by using image slicing and post-processing operations on randomly chosen authentic images. Every fake image generated from the same real image was labeled in the same group alone with the real image. Examples of CASIA dataset from two image groups are in Figure 3.

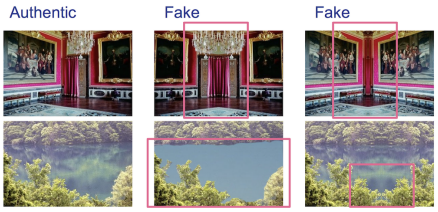

Figure 3: Examples of CASIA dataset from two image groups

Since the fake image detection task requires multiple dataset partition, we further specified the split of training and test set in the later section.

2.3. Data Preprocessing

For the data preprocessing, we resized all the images into size of 100x100 RGB format for both Reddit and CASIA dataset. For the Reddit dataset, we also detected and eliminated all the blank or lost images as we performed web scrawling.

3. Architectures

3.1. Siamese Network

The first architecture we used for duplicate image detection and image ranking tasks is Siamese Network. Siamese networks were first introduced in the early 1990s by Bromley and LeCun to solve signature verification as an image matching problem (Bromley et al., 1993)[2]. A Siamese Network consists of one pair of twin CNN architectures. One of CNN model receives anchor image and another one receives either a positive image that is similar to the anchor image or a negative image that is dissimilar to the anchor image. The twin CNN architecture has shared weights, each of which computes the feature vectors of the input image. The overall architecture of Siamese Network is shown in Figure 4. Contrastive loss function was used to train the

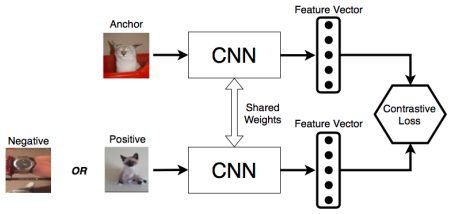

Figure 4: Architecture of Siamese Network

Siamese Network by computing the L2 distance of the output feature vectors. The formula of Contrastive loss function is in Equation (1):

$$Loss = (1-Y)\frac{1}{2}(Dw)^2+(Y)\frac{1}{2}max(0,m-(Dw)^2) \quad (1)$$

In the Equation (1), D_w is the L2 distance of the output feature vector pairs, Y is the label to indicate whether the input image pair is from the same class (1 indicates the same class and 0 not the same class) and m is the margin.

3.2. Triplet Network

The first architecture we used was Triplet Network. Triplet Network was proposed by Wang et al. (2014)[1], which aims for learning a ranking for image information retrieval. We also applied this model for our duplicate image detection and image ranking tasks based on Reddit topics in our project. A Triplet Network consists of three parallel CNN architectures with shared weights. The Triplet network receives one anchor image, one positive image and one negative image each time. Each CNN architecture computes the feature vectors of the input image and the network will output three feature vectors. The overall architecture of Triplet Network is shown in Figure 5. Triplet loss func-

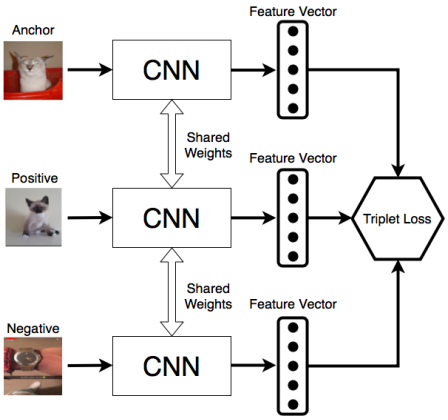

Figure 5: Architecture of Triplet Network

tion was used to train the Triplet Network by computing the L2 distance of the output feature vectors. The Equation (2) shows the Triplet loss function:

$$Loss = ||f_a - f_p||_2^2 - ||f_a - f_n||_2^2 + m \qquad (2)$$

In the Equation (2),f_a is the output feature vector of anchor image, f_p is the output feature vector of positive image, f_n is the output feature vector of negative image and m is the margin.

3.3. VisNet

In our project, we also explored VisNet for duplicate image detection and image ranking tasks. Similar to Triplet Network, VisNet also consists of three parallel deep neural networks with shared weights. Unlike Triplet Network, each deep neural network of VisNet contains three CNN models as shown below[10]:

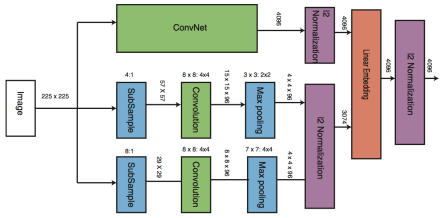

Figure 6: Architecture of VisNet[10]

According to the figure above, each image was feed into three paths[10]. The first path is basically a pre-trained VGG-16 model as shown in the top green box, which is used to extract complex features of the input image. The other two paths consist of two shallower convNets that mainly extract low-resolution visual features of the input image. In the final layer of deep CNN model, we used linear embedding layer to combine feature vectors coming from three paths. Each time, we trained triple images simultaneously: one is the query image, one is the positive image and another one is negative image. In the final step, we evaluated the similarities among these three images using contrastive loss function as shown below:

$$Loss = (1-Y)\frac{1}{2}(Dw)^2 + (Y)\frac{1}{2}max(0, m-(Dw)^2) \quad (3)$$

D_w is the L2 distance of the output feature vector pairs, Y is the image label and m is the margin.

4. Experiments

4.1. Image Ranking

For Image ranking task, we experimented with three network architectures as mentioned in the Architectures.

Dataset We used the Reddit dataset 28979 images from 15 subreddits for training, and another 1500 images from these subreddits for testing.

Training Strategy For the VisNet, we babysit the training process, with learning rate ranging from 0.01 to 0.00001 for different epochs, based on the rise and fail of the loss, we manually changed the learning rate. We also tried different optimizer such as Adam and RMSprop. For both Siamese Network and Triplet Network we used 0.0005 and Adam optimizer.

Discussion First we tried VisNet with our Reddit dataset. We randomly sampled from the dataset to create triplets, one image to be the query image, one image from the same subreddit to be the positive image and another from arbitrary subreddit except the query subreddit to be the negative image. For our dataset which is not clean and arranged, we was only able to train the loss to 0.56, not as the original paper [9] that is able to reduce the loss to a small number. We speculate the cause of this is because the their dataset was cleaner and well-organized, and this could lead to a easy-tuning model. For Siamese Network and Triplet Network we basically followed the same procedure to create pairs and triplets, and we managed to decrease the loss, as can be seen in Figure 7. Note that we only trained these two networks 60 epochs, since they had already converged. Test results for cat subreddit and golf reddit are showed in Figure 8, 9. The first row is the ranking outcome of VisNet, the second is for Triplet Network and the last one is for Siamese Network. For each row the first image is the query image. As from these two results, we can see our models were able to detect the similar images to query image at a high accuracy. For simple subreddit class cat, our Triplet Network could even retrieve the ten cat images without mistakes for top ten similar images. Even for the hard subreddit class golf, our model was able to achieve a fairly high accuracy

for the top ten similar images. However, the VisNet performed poorly, these could due to our Reddit dataset is not friendly to this model. According to [10], a huge dataset (more than one million) is required to train VisNet in this case. In addition, the VGG-16 convNet in the VisNet might capture too much detail of each image, causing overfitting problem, since within one subreddit, images can be very different from each other. Over-learning each image could give rise to low capacity of recognizing new images. While the shallower convNet architecture in our models can generalize the images in one subreddit.

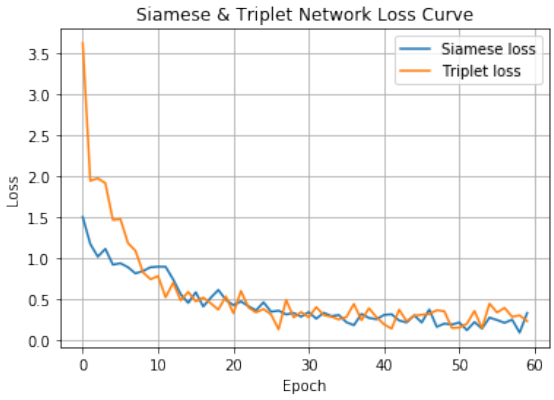

Figure 7: Loss curves of Siamese Network and Triplet Network

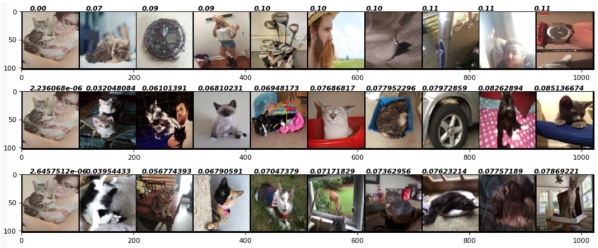

Figure 8: One example of cat image ranking

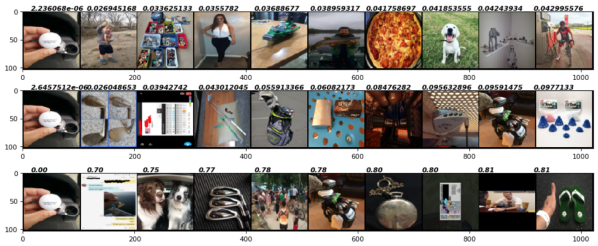

Figure 9: One example of golf image ranking

4.2. Fake Image Detection

We performed multiple evaluations for fake image detection task by using different datasets. Siamese Network, Triplet Network and VisNet are the three deep learning models we mainly explore and compare for fake image detection.

Datasets. We experimented our deep learning models using 3 different partitioned datasets to evaluate the performance as shown in the following. **(1) Pre-training on Reddit:** the training set were 28979 images from Reddit and the test set were all the 7058 images from CASIA. Since we already have the pre-trained model on Reddit data for image ranking task, we directly used the pre-trained model to test on the entire CASIA dataset. **(2) CASIA dataset:** We only used CASIA dataset to train and validate our models. The training set contained 5917 CASIA images with 1623 groups of authentic and fake images. The test set contained 1141 CASIA images with 312 groups of authentic and fake images. **(3) CASIA dataset with Reddit background images:** The partition of CASIA dataset was the same as the partition in (2). In addition, we added 28979 images from Reddit as background images to the test set.

Training Strategy. During the training process, we created customized data loader to generate triplet training sample (anchor, positive and negative images) for both Triplet Network and VisNet. The positive image was randomly chosen from the same group of the anchor image and the negative image was chosen from a different group. For Siamese Network, two images were randomly chosen each time. They could either from the same group (anchor and positive) or from the different group (anchor and negative). We used batch size of 64 with the learning rate of 0.005 when training Siamese Network, Triplet Network and VisNet. Adam optimizer has been used during the training.

Testing Strategy. During the testing process, we extracted 5-dimensional output feature vector for each query image by feeding the query image into well-trained deep learning models, and then compared with the feature vector of each test image by computing the L2 distance. We retrieved top 3 closest images using KNN. We evaluated the model's performance using top3 accuracy and MRR(Mean Reciprocal Rank).

Discussion. The retrieval accuracy of fake images is shown in table 1. Triplet Network worked the best among all three deep neural networks we experimented. We were able to achieve highest top3 accuracy of 90.7 percent by only using CASIA dataset to train and test. Even with additional Reddit images added to the test set as background images, 70.4 percent top3 accuracy is still a promising result considering the fact that we need to retrieve fake images among total of 30120 images. Another interesting result is that even without using CASIA as training set, our triplet model still achieved 51.1 percent top3 accuracy in predicting fake im-

	Pretrained (Reddit)		CASIA 2.0		CASIA 2.0 & Reddit Background	
	Top3	MRR	Top3	MRR	Top3	MRR
VisNet	47.2%	0.13	81.2%	0.41	63.5%	0.31
Siamese Net	41.2%	0.11	72.2%	0.35	54.2%	0.26
Triplet Net	51.1%	0.16	90.7%	0.43	70.4%	0.34

Table 1: Results for three Networks

ages of new groups of CASIA dataset.

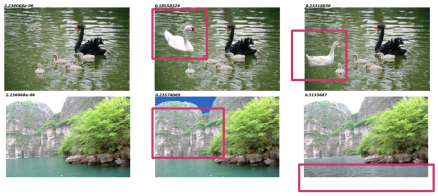

Figure 10: One example of retrieved fake images

4.3. NSFW Image Retrieval

Retrieving NSFW (Not Safe For Work) images can be another application of our models, and it can be quite useful if we can retrieve these images and identify their urls in time for a safe and cleaner online community.

The Triplet network model we used are pre-trained on 15 subreddits, and these NSFW subreddits are hidden for the training process.

Test Strategy. In this section, rather than MRR and top3, we applied another way to evaluate the performance, since the task is relatively simple, and original two methods will cause overestimation on results. In our evaluation, for a query image, we calculated score of correct labeled images out of top 10 similar images, and within same NSFW subreddit, we averaged the score and obtained the final accuracy.

Discussion. The retrieval accuracy is shown in Figure 11, which shows that for each NSFW subreddit, pre-trained Triplet network can reach more than 70% accuracy. In Figure 12, Triplet network can correctly label nine out of ten images, and one error is a cat images, since there exists uneven distribution of our dataset, and the number of images in Cat is almost twice as the number within other subreddit, as shown in Figure 1.

4.4. Image Classification

In the testing parts of sections 4.1-4.3, we compared query image with entire test set, which is unreasonable for a huge dataset. One improved method is to classify the query image first, and then search within certain class. So far,

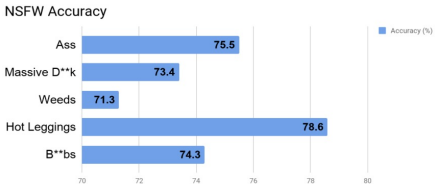

Figure 11: One example of Pre-training model on Reddit

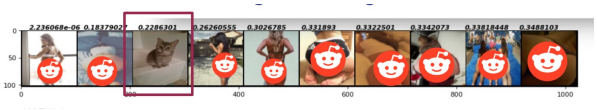

Figure 12: One example of image retrieval on NSFW subreddits with Triplet network

both of two models performed well on image ranking related task on Reddit dataset. We tried to extend our application and applied Siamese Network and Triplet Network to solve classification problem.

Transformation method. As shown in Figure 13, we first calculated the mean feature vector of samples within a certain subreddit in advance, representing that class. Then the query image was embedded with our models. In the end, we compared its feature vector with each mean feature vector, and labeled it with a subreddit name. Besides, another way we tried was K-mean algorithms for those feature vector, but it decreased the performance.

t-SNE results. T-Distributed Stochastic Neighbor Embedding (t-SNE) [8] is a technique for dimensionality reduction that is applied for the visualization of high-dimensional datasets. We used t-SNE to visualize embedded training set with Triplet network, as shown in Figure 14 while each color stands for a corresponding subreddit. Some clusters can be seen in the Figure, and there exist overlaps between subreddits, which is helpful to explain why mean feature vector method works better than K-mean.

Discussion. In our early experiments, a small dataset (around 2000) was used to train our model, and the results we obtained are below 30%, but when we increased the size to entire 15 sub-reddit (around 30000), the performance jumped to a very high level, which proves that large dataset can improve the algorithm significantly. We compared our models with VGG-16 [7] and ResNet-18 [5]. For VGGNet and ResNet, we trained them on 15 subreddit and fine-tuned the hyper-parameters. According to the result in Table 2, Siamese network and Triplet network outperformed rest two models, and Triplet network can achieve more than 91% accuracy. As far as we are concerned, users uploaded images to websites to their wish, and the dataset is too messy for VGGNet and ResNet to capture the gen-

VGG-16	ResNet-18	Siamese Network	Triplet Network
70.2%	78.5%	80.8%	91.7%

Table 2: Results of classification of Networks

eral features, while CNN layers in our models are relatively simple, and thus networks are trained more efficiently on this task.

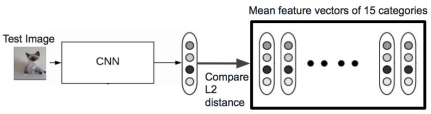

Figure 13: One example of Pre-training model on Reddit

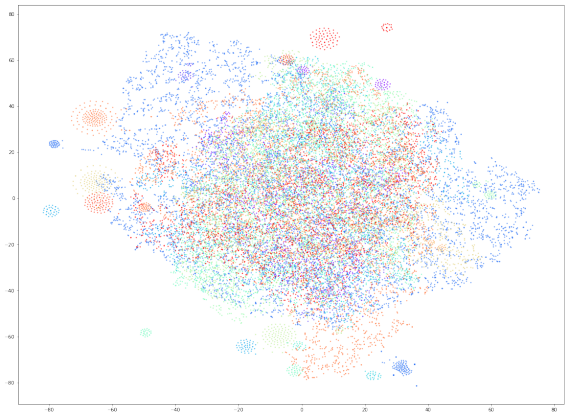

Figure 14: T-SNE on training set embedded with Triplet Network

5. Conclusion

For this messy dataset, we explored several ways to target image ranking and classification tasks. The Triplet Network of ours was able to achieve 51.1% for top3 accuracy and 0.16 MRR with the pre-trained model of Reddit dataset, and 90.7% top3 and 0.43 MRR when identifying fake images. This Triplet model was also able to outperform VGG-16 and ResNet-18 with 91.7% accuracy for correctly classifying the distance among the query image, positive image and negative image. Our model could also be extended to NSFW image retrieval and fake image detection. In the future work, there are two ways to better the performance, in which one is to make the dataset more clean, and another is to use a huge dataset. Besides, to continue this work, LSH algorithms can be used to deal with the scalable similar images detection task for a large-scale dataset.

References

[1] *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015. 2

[2] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. *Internat. Journ. of Pattern Recog. and Artific. Intell.*, 7(04):669–688, 1993. 2

[3] O. Chum, J. Philbin, and A. Zisserman. Near duplicate image detection: min-hash and tf-idf weighting, 01 2008. 1

[4] J. Dong, W. Wang, and T. Tan. Casia image tampering detection evaluation database, 07 2013. 1

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 5

[6] S. Liu, P. Cui, H. Luan, W. Zhu, S. Yang, and Q. Tian. Social visual image ranking for web image search. In S. Li, A. El Saddik, M. Wang, T. Mei, N. Sebe, S. Yan, R. Hong, and C. Gurrin, editors, *Advances in Multimedia Modeling*, pages 239–249, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. 1

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 5

[8] L. van der Maaten and G. Hinton. Visualizing data using t-sne, 2008. 5

[9] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. *CoRR*, abs/1404.4661, 2014. 1, 3

[10] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, pages 1386–1393. IEEE Computer Society, 2014. 3, 4