Filter-Invariant Image Classification on Social Media Photos

Yu-Hsiu Chen, Ting-Hsuan Chao, Sheng-Yi Bai, Yen-Liang Lin, Wen-Chin Chen, Winston H. Hsu

National Taiwan University, Taipei, Taiwan

ABSTRACT

With the popularity of social media nowadays, tons of photos are uploaded everyday. To understand the image content, image classification becomes a very essential technique for plenty of applications (e.g., object detection, image caption generation). Convolutional Neural Network (CNN) has been shown as the state-of-the-art approach for image classification. However, one of the characteristics in social media photos is that they are often applied with photo filters, especially on Instagram. We find that prior works do not aware of this trend in social media photos and fail on filtered images. Thus, we propose a novel CNN architecture that utilizes the power of pairwise constraint by combining Siamese network and the proposed adaptive margin contrastive loss with our discriminative pair sampling method to solve the problem of filter bias. To the best of our knowledge, this is the first work to tackle filter bias on CNN and achieve state-of-the-art performance on a filtered subset of ILSVRC2012.

Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: PATTERN RECOGNITION—*Applications*

Keywords

Convolutional Neural Network (CNN); Siamese Network; Image Classification; Photo Filter; Filter Bias

1. INTRODUCTION

With the emergence of social media in recent years, lots of photos are uploaded everyday (e.g., 70M photos per day on instagram[1]). Understanding the stories behind the photos is becoming a strong need for analyzing the user behavior on social media. With widespread photo editing and sharing services like Instagram and Facebook, people can easily change the lighting and the color tone by applying filters (e.g., Hudson available in Instagram) with just a few clicks. Based on the statistics of Instagram images, around 54% of photos have been enhanced by filters.[2] Thus, considering the side effects brought by the filtered images has become an issue that we cannot neglect.

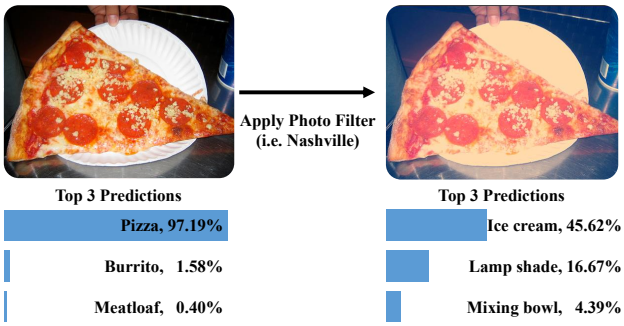

Figure 1: An example shows that a correctly predicted image with high confidence will be mispredicted after applying photo filters. In this case, the photo is applied with "Nashville" filter on Instagram. We call this phenomenon filter bias.

Recently, deep learning is playing an important role in image classification. Convolutional Neural Networks (CNN), like AlexNet [6], VGG-Net [10] and GoogLeNet [11], achieve state-of-the-art performance. However, most of the traditional models ignore the effects brought by photo filters in image classification. For instance, as Figure 1 shows, a normal image which is predicted correctly with high confidence would be messed up after applying filters while humans can still recognize it correctly. Besides, Table 1 shows that the performance of most of the well-known models would degrade drastically after simply applying filters to validation images (e.g., top-5 accuracy from 80% to 54% on AlexNet [6]). These results suggest the previous work in machine learning, even with the state-of-the-art performance, are vulnerable to the filtered images. We call this phenomenon **filter bias**.

A lot of work has been published on image classification with CNN, but few of them take care of the filter bias. Domain adaptation tackles the problem of domain shift by adapting one model from a source domain to another target domain. In this scenario, we can treat one filter as one domain and use domain adaptation works, such as [9, 14] to deal with the filter bias. However, it might be hard to get filter information of a photo on social media. In addition, if we treat one filter as one domain, there would be infinitely

[1]https://instagram.com/press/

[2]http://filterfakers.com/filterguide

MM'15, October 26–30, 2015, Brisbane, Australia.

ACM 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806348 .

Table 1: The top-1/top-5 accuracy on the original and filtered images. The performance among most of the state-of-the-art CNN models degrades severely after applying filters to the same validation set (ILSVRC2012 Val). Here, we use filter "Valencia" available on Instagram.

ILSVRC2012 Val	Without filter	With filter
AlexNet [6]	56.87%/80.30%	30.14%/53.70%
VGG-Net [10]	68.27%/92.50%	50.95%/74.58%
GoogLeNet [11]	68.70%/88.90%	47.85%/73.02%

many domains that we need to adapt to since we can easily change the parameters of a filter (e.g., the lighting and the color tone) to generate a brand new filter. Moreover, the focus of domain adaptation works is limited to one-to-one, or many-to-one domain adaptation, and it might be problematic to adapt to the domains of all filters.

Guo and Wang [3] discover the problem that a filtered image would harm the descriptor of SIFT and try to learn a more robust domain-invariant descriptor. However, it is limited in a shallow model and test on a small scale of data and filter types. Recent works [12, 2] explore the adversarial examples that can easily "hack" the deep CNN models. We can take any arbitrary image and fool a state-of-the-art CNN model to classify it as whatever class we want while the noises we added are almost imperceptible to the human eye. These works give us hints about the root of filter bias and imply that it is not dependent on the CNN architecture, i.e., it might be suffered from the linear nature of CNN.

Thus, we propose a pairwise regularization method and a discriminative pair sampling technique to relieve the filter bias and learn representations that are more robust and filter-invariant. Meanwhile, we also create a challenging filtered image dataset—Filter100—that contains around 1 million filtered photos. To the best of our knowledge, this is the first work to address the problem of filter bias on CNN-based image classification on social media photos.

2. PROPOSED METHOD

As we stated in the previous section, we attempt to learn filter-invariant representations that are less vulnerable to filtered images. The intuitive way to solve this problem is data augmentation. It is easy to generate filtered images from original images; however, it is hard to generate and include all types of filtered images for training since the size of training data would grow excessively and training models on this scale of images is unrealistic. Hence, we introduce a novel CNN architecture which learns robust representations by integrating a Siamese network with our proposed adaptive margin contrastive loss. Moreover, the discriminative pair sampling prevents the size of training data from overgrowth and improves the performance. We start by generating affordable types of filtered images and use a pairwise constraint to regularize the classification loss which helps filter-invariant representation learning.

2.1 CNN Structure

We propose a novel CNN architecture, which is illustrated in Figure 2. It is a Siamese network with both columns composed by AlexNet [6] and two additional adaptive margin contrastive losses between them to force these layers to learn filter-invariant representations. A Siamese network is a

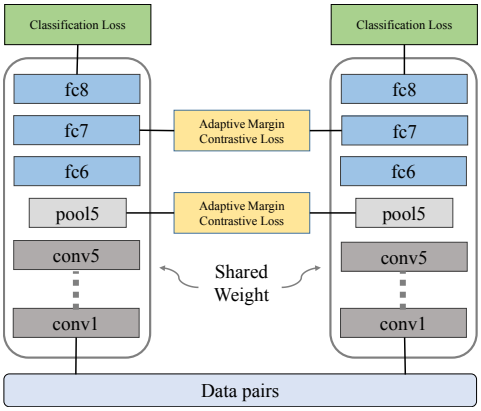

Figure 2: Our proposed CNN architecture for filter-invariant representations learning. It is a Siamese network composed of two AlexNet [6] and the additional adaptive margin contrastive loss. Combining the classification loss with two adaptive margin contrastive losses helps us to learn both semantic meaningful and filter-invariant representations.

weakly-supervised learning structure first proposed by [1]. It learns the similarity relationship from labeled pairs of data as similar or dissimilar. Though Siamese is a two-column architecture, the parameters of the layers on both columns are identical. It has been shown that Siamese networks are very effective for learning invariant representations [4, 13].

2.2 Pairwise Definition For Siamese Network

For Siamese network, it requires pairwise data for the training process. First, we define the similar and dissimilar pairs in our work. We treat the image pairs that are generated from the same image but with different types of filters (we treat original images as one kind of filtered images) as similar pair. Since the two filtered images are exactly generated from the same original image, we expect their representations to be more similar and try to make these two embeddings closer (e.g., dog images in Figure 3). For the dissimilar pairs, we sample those image pairs that have the same filter type but are drawn from different classes.

2.3 Adaptive Margin Contrastive Loss

After selecting pairwise training data, it is essential to measure the quality of learned representations. Hence, we propose a novel loss function based on the contrastive loss in [4] and it is defined as the following:

$$\mathcal{L}_r(x_i, x_j, s_{i,j}) = \begin{cases} \mathcal{D}(x_i, x_j) & \text{if } s_{i,j} = 1 \\ max(0, m_{adapt} - \mathcal{D}(x_i, x_j)) & \text{if } s_{i,j} = 0 \end{cases},$$

where $\mathcal{D}(x_i, x_i) = \|x_i - x_j\|_2^2$. We denote x_i and x_j as the representations from image i and j respectively. $s_{i,j}$ equals to 1 if x_i and x_j are regarded as a similar pair, otherwise as a dissimilar pair, and m_{adapt} is our proposed adaptive margin. The physical meaning of adaptive margin contrastive loss is to make the representations of similar pairs closer and try to keep that of dissimilar pairs at least m_{adapt} away from each other. The illustration of adaptive margin contrastive loss function is shown in Figure 3. The idea of adaptive margin m_{adapt} is inspired by [7]. For each iteration, we calculate the pairwise distance $\mathcal{D}$ of all the similar and dissimilar pairs in

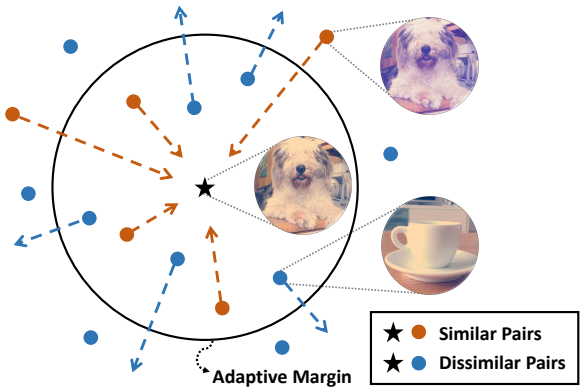

Figure 3: Illustration of adaptive margin contrastive loss. We visualize the relationship between sample point (star symbol), its corresponding similar pairs (orange points) and dissimilar pairs (blue points) on a 2D-plane. The idea is to encourage the similar pairs (generated from the same image but with different filter types) to be closer, and keep dissimilar pairs (drawn from different class but with same filter type) at least m_{adapt} apart from each other.

one batch. Then, we calculate the average pair distance between similar pairs as $mean_{sim}$, as well as $mean_{dis}$ for dissimilar pairs. After that, we set the adaptive margin m_{adapt} for each iteration as the following:

$$m_{adapt} = \frac{1}{2}(mean_{sim} + mean_{dis}).$$

The idea is to separate the representations of similar and dissimilar pairs to avoid confusing. By using adaptive margin rather than a fixed margin, we can get rid of the effort for tuning the margin parameter and adapt to datasets without any pre-processing. We apply two adaptive margin contrastive losses on both *pool5* and *fc7* layer since they are the last layer of the convolutional part and the fully-connected part respectively. Therefore, the final objective loss function of our proposed method is

$$\mathcal{L} = \mathcal{L}_c(X, Y) + \lambda \mathcal{L}_r(P, S).$$

We denote X as the training data, Y as the labels, $\mathcal{L}_c$ as classification loss and $\mathcal{L}_r$ means the adaptive margin contrastive loss we propose. P are the training pairs we sampled from X, and S are the collections of indicators for similar and dissimilar pairs. Hence, combining with the classification loss and the adaptive margin contrastive loss we added, we can learn a representation that is both semantic meaningful and filter-invariant.

2.4 Discriminative Pair Sampling

In the training phase, we may want to learn as many pairs as possible to avoid overfitting. However, the number of possible pairs increases quadratically with the number of selected filters, and makes it unrealistic to include all the possible pairs in training set since it will be computationally prohibitive. For instance, as described in Section 2.2, 50,000 original images with 5 types of filtered images in training set will result in more than 10 billions of possible pairs. Intuitively, we can use uniform sampling to limit the number of pairs in training data, but it will give a sub-optimal solution. Therefore, we propose a discriminative pair sampling method which tries to include as many informative pairs as possible in a given number of selection. The idea is to select similar pairs whose changes of the color and lighting are more drastic, which might be more informative for training. For every original image, we sample similar pairs with the probabilities proportional to the pairwise distance based on *fc7* layer representations. For instance, if we include original images and filtered images with N types of filter, we have all C_2^{N+1} possible similar pair combinations for every single image g, which is denoted as U. If we want to select k similar pairs from all possible combinations U with the representation on *fc7*, which is denoted as x. We sample similar pairs with the probabilities of each pair combination as the following:

$$Prob(g_{i,j}) = \frac{\mathcal{D}(x_i, x_j)}{\sum_{u,v \subset U} \mathcal{D}(x_u, x_v)},$$

where $g_{i,j}$ means the similar pair of image g with filter type i and j. By using this discriminative pair sampling method, we can learn in a more efficient and effective way.

3. EXPERIMENTS

Since we want to learn a both filter-invariant and semantic meaningful representation, we use zero-shot testing, in other words, test models on the validation images applied with filters which are unseen in training. Therefore, we can ensure we learned robust representations that are less vulnerable to different types of filters. In this section, we describe the details of the experiments we conducted.

3.1 Filter100

To generate a large scale dataset with filters, we randomly sample 100 classes from ILSVRC2012 [8] with 500 images per class. After that, we apply 18 types of popular filters available on Instagram to generate filtered images and finally form a dataset that contains 0.95M images (original images included) in total. We call this dataset **Filter100**.

For one single training set, we randomly select 5 types of filters and include the corresponding filtered images along with the 50,000 original images as training set, which contains 300,000 images in total. In validation, we use zero-shot testing, i.e., we use the filtered validation set only contains 13 unseen filters to evaluate the robustness of learned representations. There are total 65,000 images in a single validation set. We repeat the procedure 5 times on 5 different train/validation sets and report the average accuracy of the 5 sets to reduce the bias caused by the selection of filters.

3.2 Experiment Settings

For the implementation, we use Caffe [5] to implement our adaptive margin contrastive loss layer and the Siamese network structure as we illustrated in Figure 2. We sample 6 similar pairs and 6 dissimilar pairs from every original images, which results in 0.6M training pairs for each training set by using our discriminative pair sampling method. After that, we use the pre-trained Caffenet model as the initial weights in our fine-tuning process by using the CNN architecture we designed.

3.3 Experiment Results

Table 2 shows the result of pure AlexNet fine-tuned with only 50,000 original images in Filter100 and the performance degrades from around 80% (test on only validation images

Table 2: Our method outperforms all the other works, like pure AlexNet [6] fine-tuned on 100 classes subset of ILSVRC2012 without any filtered images and the one fine-tuned on Filter100, on all 5 train/test sets with different filter selections. Also, our work surpasses the state-of-the-art in domain adaptation [14] and implies that domain adaptation cannot adapt to such a "filter domain" and cannot relieve the filter bias problem.

	Set 1	Set 2	Set 3	Set4	Set5	Average
AlexNet [6] (fine-tuned without filtered images)	65.93%	66.96%	67.33%	67.93%	65.96%	66.82%
AlexNet [6] (fine-tuned on Filter100)	70.53%	70.69%	73.93%	73.22%	69.60%	71.59%
DDC [14]	70.42%	70.67%	73.55%	72.86%	68.93%	71.29%
Adaptive Margin Contrastive Loss + Uniform Sampling **(ours)**	71.55%	71.24%	74.70%	73.87%	70.01%	72.34%
Adaptive Margin Contrastive Loss + Discriminative Pair Sampling **(ours)**	**71.76%**	**71.65%**	**74.93%**	**74.19%**	**70.29%**	**72.56%**

with no filter) to around 67% (with filter) in all validation set, which is concurrent with the result in Table 1. Then, we compare with AlexNet fine-tuned with Filter100, and the result shows that the traditional data augmentation techniques reduce the performance degradation caused by filter bias. However, our work surpasses it by utilizing pairwise constraint. Furthermore, compare with uniform sampling, discriminative pair sampling effectively limits the pair number in training and further improves performance. In addition, we observe that the mean distance of similar pairs on the learned *fc7* representations in our work is much smaller than the one in other work, which is about 9 times smaller than the one in AlexNet fine-tuned on Filter100. That is to say, an image and its filtered images are more similar after considering pairwise constraint in the learning process, which is coincident with our intuition. We think it is one of the key points to learn filter-invariant representations.

Moreover, our work outperforms the state-of-the-art in domain adaptation, DDC [14]. We treat original images as source domain and both original and filtered images as target domain to do supervised domain adaptation. The results indicate that it might be hard to solve filter bias by simply adapting to such a big "filter domain." To sum up, we outperforms all of the literature aforementioned in all 5 train/validation set and achieves state-of-the-art performance in image classification on social media photos.

4. CONCLUSIONS AND FUTURE WORKS

We point out the problem of filter bias in image classification on social media photos and propose a novel CNN architecture leveraging the pairwise constraint to relieve this phenomenon. We propose an adaptive margin loss function which dynamically selects the most suitable parameter for different batches in the training process. Furthermore, the discriminative pair sampling method we propose effectively limits the pair number needed in training and improves the performance. To the best of our knowledge, this is the first work to deal with the problem of filter bias in CNN-based image classification and our work surpasses other works and achieve state-of-the-art performance.

Determining the optimal selection of filters included in training is still an issue. We will exploit more cues about how different types of filters affect the CNN model in our future work. Moreover, though the implementation of our architecture is based on AlexNet [6], the adaptive margin contrastive loss we propose is not limited to AlexNet. We will work on other state-of-the-art CNN architectures, like GoogLeNet [11] and VGG-Net [10] to find out a generic method that can learn a more filter-invariant and semantic meaningful representation to improve the performance of image classification on social media photos.

5. ACKNOWLEDGEMENTS

This work was supported in part by HTC and MediaTek.

6. REFERENCES

[1] 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. *IJPRAI*, 1993.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

[3] Z. Guo and Z. J. Wang. An adaptive descriptor design for object recognition in the wild. In *ICCV, 2013*.

[4] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[7] J. Lin, O. Morère, V. Chandrasekhar, A. Veillard, and H. Goh. Deephash: Getting regularization, depth and fine-tuning right. *CoRR*, abs/1501.04711, 2015.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[9] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*. 2010.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CVPR*, 2015.

[12] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *ICLR*, 2014.

[13] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[14] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.