Bias in Image and Caption Datasets

[redacted]
Department of Computer Science[redacted]
[redacted]

Abstract

Datasets are playing a more and more important role in object recognition and caption generation projects. Currently, there are numerous datasets available and each of these datasets has their main focus. However, many researchers merely focus on the result of their methods based on one dataset, which makes it hard to decide the real-world performance of the method. Furthermore, considering the bias of datasets, focusing on the result based on one dataset will prevent the method from generally applicable. In this paper, we focus on the bias of some commonly used datasets, COCO and IAPR-TC-12. We train a model using same amount of inputs from both datasets. If these two datasets have no bias, the model will have a hard time distinguishing samples from one dataset to another. We hope our research can make researchers realize that doing testing on merely one dataset can reach bias results and researchers should calculate their results based on more datasets.

1. Introduction

Datasets are one of the most important component of object recognition research. Researchers are always familiar with numerous datasets, because they are reading papers using difference datasets all day. These datasets are expected to be unbiased, which means when researchers are exploring samples for these datasets, they are trying to present the real world in all aspects. However, this is almost an impossible task for the author of datasets. Hard as they tried, bias in these datasets are always an inevitable problem, which means these datasets failed to become the true representation of the world. It is universally acknowledged that the ultimate goal for object recognition research is to building a model that can be used to the world.

In addition to theoretical problems, the bias in datasets leads to many unnoticed realistic problems. On the one hand, researchers always train and test their model based on merely one dataset and they compare their result with other models using the same dataset. However, since there are bias in datasets, results presented by researchers are not representing the result of the world. On the other hand, sometimes researchers pay too much attention on their accuracy rather than the value of their research. When researchers are trying to write a paper, everyone want to "beat" other papers in accuracy, which makes their paper to be better or more meaningful. However, in order to maximize the accuracy, some special optimization methods are taken, regardless of whether these methods will work in other datasets. Although their methods ranked top in this dataset, the same method may not work for the same problem using other dataset. Instead of generating a include-all dataset, maybe training and testing over another dataset is a fair way to calculate the performance of a method. Furthermore, focusing on the best accuracy leads to another problems. New innovative methods are hard to beat old well-optimized methods, so they are likely to be ignored. These new methods will lose their chance to be noticed and optimized.

To find out whether there are bias between these datasets, we train a model that anticipates the dataset to which the samples belong. We take images and captions from two popular datasets as the input and trained two models respectively. If these datasets are unbiased, our model will fail to learn useful knowledge from training samples, and have trouble distinguish test samples.

Our first goal for this paper is to bring this problem to the table. The bias in dataset may be hard to solve, but researchers can take some actions to alleviate the effect of datasets to their research. Second, we think that it is a good idea not focusing on the accuracy and result of the paper, but focusing on the idea and universality. Finally, we provide some suggestion in reducing the bias in datasets. We sincerely hope that the bias in datasets can be settled in the future.

2. Related Work

We gained the idea of this paper from [5]. They conduct two different experiment to prove the bias between datasets. First, they perform a toy experiment, which trains a model to decide which dataset an image belongs to. 12-way support vector machine classifier is trained and tested

and reaches 39 percent accuracy. Second, as the main experiment, they perform "car" and "person" detection and classification on six popular datasets. As for object detection, they use Dalal Triggs [2], which contains a histograms of oriented gradients detector followed by a linear support vector machine. As for classification, they use bag-of-words with a non-linear gaussian kernel support vector machine. Their research mainly focuses on the bias in images from datasets, and they even provide a value of these datasets. Since there are bias in datasets, [4] focuses on proposing a model that shows bias in datasets in training stage. Their model learns from both bias vectors associated with datasets and the weights in visual world. To calculate the performance of the model, they compare its performance to a classic support vector machine using a new dataset. According to their research, it is helpful to reduce the bias by merging several datasets. In this paper, we focus on proving the bias in datasets and provide possible solution to this problem.

3. Models

We train two different models for images and captions respectively to achieve better performance. In part one of this section, we discuss our convolution neural network for image distinguishing task. In part two of this section, we introduce the pre-trained BERT model which we used for caption distinguished task.

3.1. Model for Images

We design a convolution neural network for images distinguishing. As shown in the Figure 1, our convolution neural network consist of three convolution layers, three max pooling layers and two fully connected layers in our model. After each convolution layer, there is a max-pooling layer. The final decision is made based on the value of the final fully connected layer. The reason why we design this network is that convolution neural networks has great performance in image related tasks. Convolution layers can extract regional features effectively, while max pooling layers are able to avoid over fitting and increase fault tolerant.

3.2. Model for Captions

As for captions distinguishing model, we take pre-trained BERT [3] as the model. Pre-trained BERT is first semi-supervised [1] trained on large amount of text. Then it is supervised trained on a specific task with labeled datasets. BERT consistes of a large amount of encoders and fits natural language understanding tasks. In our experiment, we take captions labeled with their dataset name as input. The output vector of BERT is used as the input for a classifier of caption distinguishing decision. Our model chooses a single-layer neural network as the final classifier. Using pre-trained BERT model will save the time and resources for training a new model from scratch.

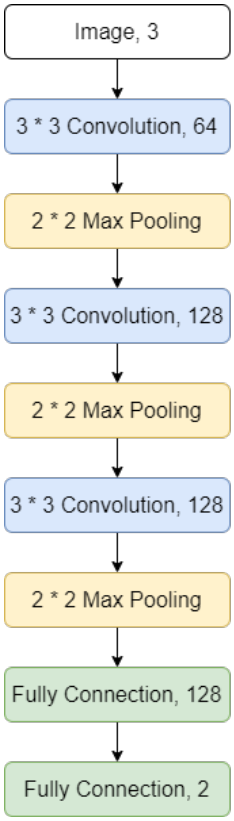

Figure 1. Overview of our image distinguishing model. This model takes an image as input and decides which dataset it belongs to. The input image will go through three convolution layers and three max pooling layers. The final decision is made based on the value of final fully connected layer.

4. Experiments and Results

Intuitively, if the overall accuracy is above 0.5, there is bias between these two datasets. In order to reach unbiased experiment result, we run the training and testing process for three times and calculated the average result. In part one of this section, we discuss our experiment and result for image distinguishing tasks. In part two of this section, we discuss our experiment and result for caption distinguishing tasks.

First, we split samples from two datasets into three categories: 7000 for training, 2000 for validation and 1000 for testing in total from both datasets. We label the samples based on which dataset these samples belong to. After training, we calculate the confusion matrix and the accuracy based on the testing samples.

4.1. Experimental Results for Images

In image distinguishing experiment, the number of epochs is 50, and batch size is 10. We choose the learning rate as 0.001. The accuracy and loss trend is shown in Figure 2. At the end of the training process, the accuracy and loss become stable in Figure 2, which means training process is adequate. Since our validation set does not overlap with our testing set, generally speaking, the validation accuracy of our model is a little less than the testing accuracy of our model.

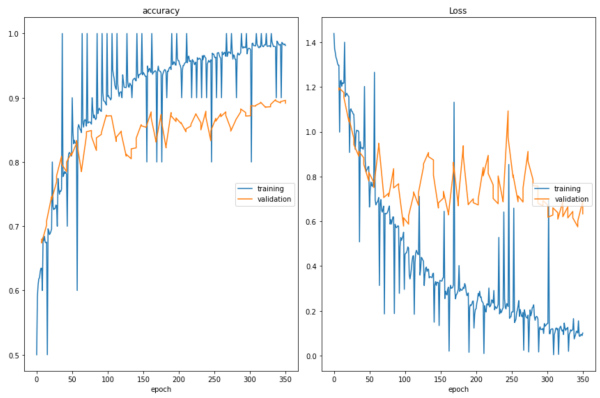

Figure 2. Accuracy and loss in image model training and validation. This model shows the accuracy and loss trend throughout the training process. As we can see, the accuracy and loss become stable and reach a great result at the end of training process.

As shown in the Table 1, our model predicts that 509 images are from COCO and 491 images are from IAPR-TC-12. Furthermore, the overall accuracy is 91.1 percent, which means our model had great performance in distinguishing images from these two datasets.

	Predicted COCO	Predicted IAPR-TC-12	
Actually COCO	460	40	500
Actually IAPR-TC-12	49	451	500
	509	491	

Table 1. Confusion matrix for image distinguishing. There are 500 images from COCO and 500 images from IAPR-TC-12. Our model predicts that 509 images are from COCO and 491 images are from IAPR-TC-12. This table indicates that our model successfully distinguish images from these two datasets.

4.2. Experimental Results for Captions

In caption distinguishing experiment, the number of epochs is 30, and the batch size is 10. We choose the learning rate as 0.001. The accuracy and loss trend is shown in Figure 3. As we can see from Figure 3, the validation accuracy and loss is stays in a bad performance level, and the training accuracy and loss changes rapidly throughout the training process. This figure indicates that our model fails to extract useful features.

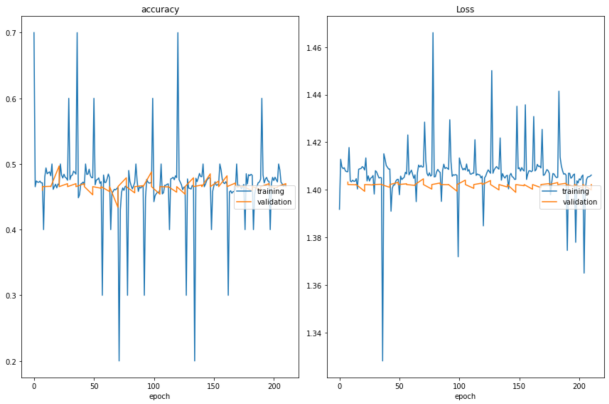

Figure 3. Accuracy and loss in caption model training and validation. This model shows the accuracy and loss trend throughout the training process. As we can see, the accuracy and loss of validation set stays in a bad performance level, and the accuracy and loss of training set changes rapidly.

As shown in the Table 2, our model predicts that 56 captions are from COCO and 944 captions are from IAPR-TC-12. The accuracy of testing set is 47.8 percent, which is less than 50 percent. The result shows that our model is unreliable in this task because it simply predicts most of the captions to be in IAPR-TC-12.

	Predicted COCO	Predicted IAPR-TC-12	
Actually COCO	17	483	500
Actually IAPR-TC-12	39	461	500
	56	944	

Table 2. Confusion matrix for caption distinguishing. There are 500 captions from COCO and 500 captions from IAPR-TC-12. Our model predicts that 56 captions are from COCO and 944 captions are from IAPR-TC-12. This table indicates that our model fails to distinguish captions from these two datasets.

5. Discussion

Based on our previous experiment and result, there are obvious bias between images from COCO and IAPR-TC-12, but our model fails to find significant bias between captions. Some people may argue that instead of blaming the dataset or their authors, using machine learning methods to find minor difference between datasets are the root cause

for significant bias in datasets. In fact, our models are not complex or fine-tuned for images or caption distinguishing tasks, but the bias in images from these two datasets is still undeniable. Instead of blaming the authors of these datasets, we believe that the root cause of the bias in images from datasets is the difference of background. For instance, "house" only refers to "single-unit residential building" in the United States, while it means "general residential building" in China. This is the bias from selection. Furthermore, bias in location is also an important factor. "House" may be bigger and moderner in developed country or region instead of undeveloped country or region. Different regions have different preferred type of residence. In contrast to the significant difference in images, our model fails to find significant bias between captions. On the one hand, our model may be not powerful enough to solve this problem. On the other hand, when authors of datasets are generating these captions, they are using the clearest and simplest way to illustrate the image, which leads to the failure of our model. Regardless of the background of caption generators, they are capable of generating the simplest caption about the image, so there are little bias in captions.

There are two meaningful questions based the result of our research. The first question is how to create an unbiased dataset. There is no doubt that this was a challenging task, since it is hard to become unbiased in real world. However, as the caption datasets are unbiased, diversity may be a great solution to mitigate this issue. For example, hiring people with different background in sample labelling and selection will be a great idea. Furthermore, researchers can randomly select samples from more than one existing datasets to train there model more comprehensively. Bias in datasets will be reduced significantly with these two methods. Another question is that: What should we do regarding the bias of datasets. In our opinion, the ultimate goal for object recognition and caption generation research is to better understand the world. If there is specific limitation of research area, researchers should chooose dataset that matches the actual world in that criterion. If the researchers are trying to find a general method, choosing different datasets with different focus might be a great idea. To sum up, researchers should choose their datasets based on the need of their project.

6. Future Work

Although we have reached some conclusion in this paper, there are still many possible improvements to reach better results.

Regarding the images distinguishing tasks, we think training the model with same type of images can be a better idea. For example, we can take images of the "cars" from both datasets and train a "cars" model to find out the bias in "cars" images. This can eliminate the impact of a certain dataset lacking a certain data type on the results. Furthermore, calculating the overall accuracy based on the frequency of each image type will provide a result that is closer to real world.

Unfortunately, our model for caption distinguishing task does not perform as we expected. It fails to distinguish captions from these two datasets. There are two possible reasons that may lead to this result. First, there is no significant bias between captions in these two datasets at all. We need to compare captions from other datasets to reach more reliable decision. Second, our model does not work in this task, which means we need to try new models.

7. Acknowledgement

[redacted]

References

[1] A. M. Dai and Q. V. Le. Semi-supervised sequence learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 3079–3087. Curran Associates, Inc., 2015.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba. Undoing the damage of dataset bias. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, pages 158–171, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[5] A. Torralba and A. Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.