Automated Label Noise Identification for Facial Attribute Recognition

Jeremy Speth and Emily M. Hand
University of Nevada, Reno
jeremyspeth@nevada.unr.edu, emhand@unr.edu

Abstract

Current state-of-the-art facial attribute recognition techniques use exceedingly deep convolutional neural networks (CNNs), which require large human-annotated datasets that are costly and time-consuming to collect. In most domains, there are several large-scale datasets for researchers to work with. In facial attribute recognition, there is only one large-scale dataset available – CelebA – causing researchers to rely too heavily on this one set of data. While CelebA provides the scale necessary for training deep networks, there are several types of noise present in the dataset. We address the problem of label noise by introducing a novel multi-label verification framework to identify mislabeled samples. Our work is applicable to data collection, cleaning, and multi-label verification. Our method is used to analyze label noise in CelebA and perform extensive experiments with additive noise to show the efficacy of the proposed approach.

1. Introduction

Attributes provide an intuitive and compact way to describe real-world objects using natural language. Recently, the rapid performance gains on facial attribute recognition (e.g. *hair color*, *gender*, etc.) utilizing CelebA [5] have begun to plateau for several reasons including label noise. We aim to assess the levels of noise present in CelebA with a flexible approach that may be extended to any multi-label dataset. This work addresses two types of noise present in CelebA: 1) incorrect labels and 2) ambiguous labels. With the prevalence of label noise and label ambiguity detailed in [4], an automated process for the identification of such noise in CelebA will have a significant impact on the research community. We propose a method to identify mislabeled samples by performing attribute verification between a candidate sample and a set of representative samples. We design a multi-label siamese CNN for embedding samples in a lower-dimensional space where distance metrics correspond with attribute similarity. The distances between two embedded samples are used for attribute verification – a binary prediction of the pair's semantic similarity for all attributes. This siamese network is then used to identify noisy samples by comparing many representative samples to a candidate sample.

2. Related Work

The proposed approach is influenced by *classificaton filtering* originally formulated in [3]. Classification filtering uses an initial model for classification and removes all samples that are misclassified [7]. Removing samples based on a single misclassification places a great deal of confidence in the initial model and removes many correctly labeled samples. Additionally, classification filtering does not address the problem of outlier versus noise classification, and simply removes outliers as well as noise. A natural improvement over single model classification filtering is to use an ensemble of models [1]. Ensembles are beneficial since difficult samples are less likely to be misclassified by the entire ensemble, and less confidence is placed on any single classifier. A downside to using ensembles is the need for designing and training multiple models. Cluster-based approaches have been used to improve classification accuracy when training with noisy data [6]. Evaluating several clustering algorithms on a dataset multiple times with different parameters allows for noisy samples to be identified based on the clustering results. The approach has not been tested on any image datasets, making it difficult to discern whether extracted image features could be clustered to provide comparable results. The proposed approach uses a single model with multiple sample comparisons to combine the advantages of the above approaches. Previous methods struggle to achieve high precision which is necessary for classification filtering without human oversight. For a thorough survey of label noise correction literature, see [2].

3. Approach

We utilize a verification framework to identify noisy samples by comparing candidate samples to a set of representative exemplars. Given a sample that is predicted to be more similar to samples of the opposite class than to those

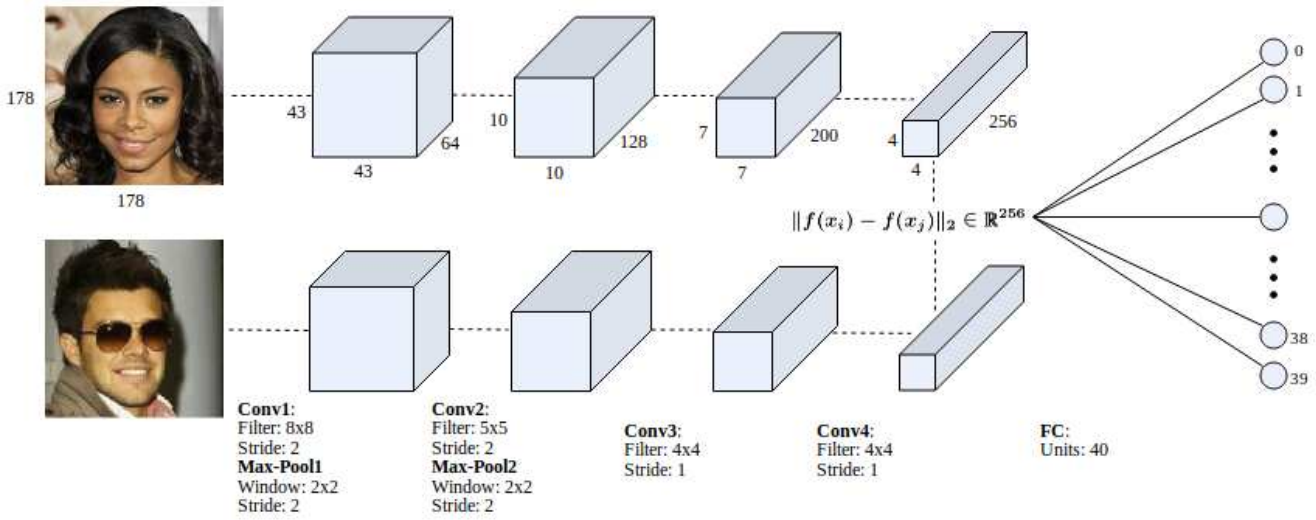

Figure 1: Siamese network taking two input images from CelebA. Pair-wise euclidean distance is taken between each feature map in the final convolutional layer to produce a distance vector. The distance vector is then fed into a fully connected layer which produces an output of the similarity for all 40 labeled facial attributes between the images.

of the same, we can characterize that sample as noisy.

3.1. Attribute Verification

We define the attribute verification process as mapping two input samples (x_i, y_i) and (x_j, y_j) to the joint label vector $Y \in \{0, 1\}^k$, where k is the number of attributes. The joint label is given as $Y_a = 0$ if attribute a is the same for the input samples and $Y_a = 1$ if they are different. We use a CNN as the base for the siamese verification network, which embeds an image into a lower-dimensional space. The symmetry of siamese networks then allows for comparison between high-level features of a pair of input images, typically with a distance metric. Figure 1 shows the layout of the proposed network architecture in detail.

We use euclidean distance as our fixed metric when comparing the feature maps of two input images from the siamese network. Specifically, each path of the siamese network propagates its input image through the CNN. Following the final convolution, each path produces several 2D feature maps. Each of the feature maps is paired with its corresponding feature map in the other network stream, and the euclidean distance is taken between them to yield a distance vector.

We consider the output for each attribute to be some weighted combination of the distances between feature maps. This weighted combination is learned by a fully-connected layer between the distance vector and the verification outputs. The model is trained end-to-end using binary cross entropy loss rather than learning the CNN embedding and similarity prediction separately.

3.2. Label Noise Detection

Once the siamese network has been trained for the task of attribute verification, it can be used to identify noisy samples. We use pairwise similarity between a candidate sample (x_c, y_c) and a representative exemplar (x_r, y_r) to find contradictions. A contradiction occurs when verifying attributes if the joint label differs from the predicted similarity for that attribute. A sample is confidently identified as noisy if verifications with a set of multiple representative samples produces some number of contradictions.

We define a threshold, $t \in [0, 1]$, for tagging a sample as noisy based on the ratio of the number of contradictions to the size of the representative set. The size of the representative set corresponds to the number of pairwise comparisons to be made. Representative sets are constructed by randomly sampling from the training set. Large representative sets provide more coverage of the general distribution of an attribute at the cost of increased computation time. Increasing t corresponds to emphasizing precision and avoiding falsely tagging samples. Adjusting t greatly influences the results as shown in figure 2.

In practice it is necessary to balance the distribution of positive and negative labels in the representative sets for each attribute. By enforcing the same number of positive and negative samples, candidates are only tagged by verifiers that have learned both similarity and dissimilarity in comparisons. This acts as a filtering mechanism for weak verifiers that may falsely tag candidates. For simplicity in balancing the label distribution we generate representative sets for a single attribute at a time. The process is parallelizable if necessary, since attributes are tagged independently.

4. Evaluation

We evaluate our approach by analyzing tagging results on CelebA with simulated noise. We evaluate the ability of our model based on precision – the number of correct tagged samples over the total number of tagged samples. Previous methods yield poor precision and we find it the best metric to emphasize minimal human supervision when identifying label noise. *A model with high precision and low recall may miss noisy samples, but removes the need for human oversight.*

4.1. Data

We evaluate our method on the CelebA dataset. CelebA has over 200,000 images each labeled with 40 binary attributes (e.g. *blond hair*, *smiling*, *goatee*, etc.). The attributes cover a wide range of visual characteristics such as gender, facial hair, hair color, facial structure, age, accessories, and other describable features.

4.2. Implementation

We implement the siamese CNN model depicted in Figure 1 and train the model by minimizing end-to-end binary cross-entropy loss on the training set. We use batches of size 64 which are dynamically balanced with an approach similar to [4]. Each batch from CelebA is randomly cropped to form 178×178 images. We select the model which achieves the highest accuracy for attribute verification on sample pairs from the validation set. Given the trained siamese network, we follow the approach described in section 3.2 to inspect every attribute label for every sample in the test set. We create balanced representative sets from the training data and tag samples that exceed the threshold t defined by the number of contradictions over the size of the representative set.

4.3. Additive Noise Results

To evaluate the proposed approach, various levels of label noise were added to the test set in CelebA. With a test set of size N_t, we add label noise with level $z \in [0, 1]$ by randomly flipping $z \times (40 \times N_t)$ original labels. We maintain the original labels in the training set and randomly sample from it to construct our representative sets. This experiment simulates accumulating data from a potentially noisy source after training a verification model with a relatively noise-free dataset. We find this experiment particularly relevant to quality assurance during the data labeling process or cleaning a previously labeled large-scale dataset.

The proposed approach is capable of achieving high precision as the threshold t is increased. Figure 2 shows the precision and recall with different thresholds and noise levels. For a threshold of 90% contradictions between a candidate sample and the representative set nearly all tested noise levels achieve precision over 0.9. The precision improves at each threshold as the noise level increases, while the recall remains nearly constant. The threshold greatly influences both the recall and precision, with lower thresholds allowing for a greater number of incorrect tags and identification of more noisy samples. All experiments use a representative set of size 1000 for each attribute.

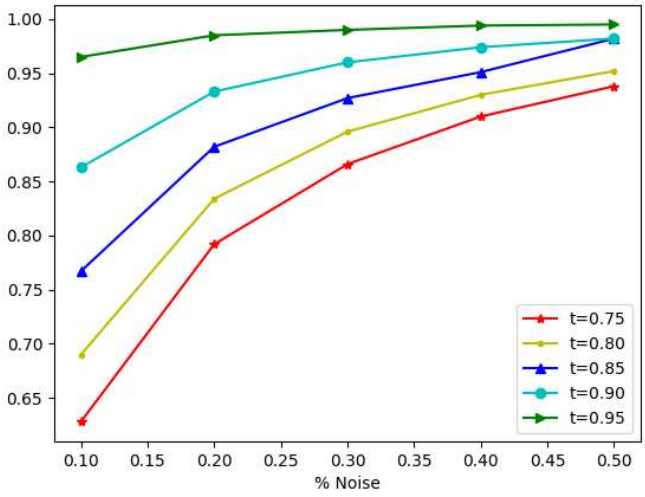

Figure 2: The precision on the CelebA test set with different thresholds and levels of additive noise. The precision increases as both the threshold and level of noise increase. The recall for each t remains constant relative to the noise level. Recall for $t = \{0.75, 0.80, 0.85, 0.90, 0.95\}$ are $\{0.604, 0.525, 0.413, 0.276, 0.058\}$ respectively. Correctly labeled samples are not tagged when the precision is high, which allows for the process to be automated.

5. Conclusion

We presented a technique for identifying label noise in large-scale multi-label datasets. The presented approach improves upon classification filtering methods which place overconfidence in a single predictor and ensemble methods which require building multiple models. Leveraging the verification framework gives the benefits of both approaches by using multiple predictions from a single model. Pairwise comparisons between candidate samples and a set of representative exemplars allows us to approximate whether a candidate's label contradicts the general distribution of an attribute. Adjusting the set of samples for comparisons gives our approach the flexibility necessary to handle any dataset. Our results on CelebA with additive noise show our approach is capable of achieving high precision in identifying mislabeled samples without human oversight. As supervised methods increasingly rely on benchmark datasets, we believe the proposed approach will help ensure future models are trained and evaluated on reliable data.

References

[1] Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167, 1999.

[2] B. Frenay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, May 2014.

[3] Dragan Gamberger, Rudjer Boskovic, Nada Lavrac, and Ciril Groselj. Experiments with noise filtering in a medical domain. In *Proceedings of International Conference on Machine Learning*, pages 143–151, 1999.

[4] Emily Hand, Carlos Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *AAAI Conference on Artificial Intelligence*, 2018.

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[6] B. Nicholson, J. Zhang, V. S. Sheng, and Z. Wang. Label noise correction methods. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9, Oct 2015.

[7] Jaree Thongkam, Guandong Xu, Yanchun Zhang, and Fuchun Huang. Support vector machine for outlier detection in breast cancer survivability prediction. In *Advanced Web and Network Technologies, and Applications*, pages 99–109, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.