Deep Label Distribution Learning with Label Ambiguity

Literature Review

Ⅰ. Introduction.

Currently, Convolutional Neural Networks (ConvNets) have achieved excellent performances in various visual recognition tasks. However, for some domains including apparent age estimation, head pose estimation, multi-label classification and semantic segmentation, it is difficult for us to collect enough training samples with precise labels. Fortunately, Jianxin Wu and his team found that there exists ambiguous information among labels. In this paper, they proposed DLDL method to utilize the label ambiguity in both feature learning and classifier learning. Experiments showed that this approach achieves better performance compared to other state-of-the-art methods.

Ⅱ. Necessity

Due to the inevitability of ambiguity and the uncertainty of labeling pictures, it is difficult to collect a large and accurately labeled training images in the domains mentioned above. Therefore, exploiting deep learning methods with few samples and fuzzy labels has become an attractive and challenging topic. Although there already exists several methods to handle label ambiguity, they may either fall into the local optimal solution, or require a large amount of data. The authors reckon that the reason for these defects is that they didn't make good use of label ambiguity. If we make good use of it, a more robust method with better performance will occur.

Ⅲ. Formalization, Simplification and Verification

Jianxin Wu and his team recognized that there is high correlation among input images with similar outputs. To utilize this correlation information in both feature learning and classifier learning, deep learning method is a good choice because of its end-to-end characteristic. To formalize correlation as well as other types of ambiguity, the authors quantize the range of possible y values to estimate an entire label distribution. Therefore, the deep learning machine is able to take care of the ambiguity among labels. In other words, this ambiguity can be formalized as learning a conditional probability mass function $\hat{y} = p(y|X;\theta)$ from D, in which θ is the parameters in the framework, such that $\hat{y}$ is similar to y.

To simplify this idea, the authors first propose that the label set defined for the specific task is complete. I think that this simplification is natural because if some possible ys don't have their corresponding members in the label set, we will have to find out the parameters related to these abnormal outputs. In this case, we will be deviated from looking for label ambiguity to predicting the label. Second, they assume that we will only focus on the last fully connected layer in a deep ConvNet. This is also a natural simplification because this layer is directly related to the output labels and our goal, label ambiguity. On the contrary, the problem will become more complicated if we consider all fully connected layers.

Based on the simplified idea mentioned above, the author gave the details of the fully connected layer in DLDL. First, DLDL choose KL-divergence to measure the similarity between distribution $\hat{y}$ and y. Thus, the loss function T under the KL-divergence measurement can be given. By applying the chain rule, we will have the derivative of T with respect to θ. After updating $\hat{y}$ until convergence, we use this network to predict labels. Second, the paper argues that the output of DLDL may be different according to the expected labels. If we expect that the class

label is a single one, DLDL outputs the most probable label. In contrast, DLDL outputs expectation of $\hat{y}$ as label if the expected output is a real number. In general, DLDL is suitable for both classification and regression tasks.

Now that we have a broad view of DLDL, it's time to apply the DLDL method to the four recognition tasks. In these tasks, the paper further simplifies its assumptions.

For age estimation, it assumes that the probabilities should concentrate around the ground-truth age. Personally, I think this is too vague because its difficult to define *concentrate around*. What's more, the article quantizes y using a normal distribution, but Morph dataset provides the $\text{date\_of\_arrest} - \text{date\_of\_birth}$ as its ground-truth [1], which isn't a normal distribution but a determined number. It is unreasonable to set $\sigma = 2$ for the distribution in this case.

For head pose estimation, it assumes that the covariance matrix Σ is diagonal, which means that pitch and yaw angles are independent to each other. But, in reality, because it is hard to pitch $90°$ and yaw $90°$ at the same time and it's easy to either pitch $90°$ or yaw $90°$ respectively, pitch and yaw are not independent. However, this simplification is still a reasonable one because pitch and yaw can be treated as independent random variables apart from few severe cases. Besides, this simplification further simplifies the calculation.

For multi-label classification, the author defines probabilities for different types of labels to use label ambiguity. I think this is convincing since treating `Difficult` as `Positive` or `Negative` are both adding unnecessary inductive preferences to the dataset.

For semantic segmentation, the article focuses on the ambiguity in the boundaries. It proposes a Gaussian kernel matrix to modify the original label distribution. I think that this kernel helps smooth the boundaries and reduce the noise. As a result, these inherently ambiguous pixels are smoothed and the result may be better.

As for the training details, training from scratch and fine-tuning are proper candidates if we try to keep the features of a deep learning method while integrating DLDL method. The author proves that DLDL achieves higher accuracy with more robustness through several experiments. Moreover, they claimed that DLDL have clear semantic clustering in low dimensional space. It can reduce over-fitting, accelerate convergence and perform well in small datasets or sparse labels.

Ⅳ. Discussions

Finally, I'd like to discuss the ideas in the paper because there are several interesting topics to think deeper. Firstly, what if the label set is not complete? This leads the problem to a Transfer Learning one. Secondly, what if we change the measurement of the similarity between the ground-truth and the label distribution? There are other measurements such as Hellinger distance and Bhattacharyya distance. Is it more convincing to replace *divergence* with *distance*? Thirdly, what if we apply more datasets on age estimation problem? As we can see, DLDL isn't the best method for ChaLearn dataset, and the author only used one dataset to verify that DLDL method performs well when μ and σ is known or when σ is unknown. I believe that this is far from enough. Fourthly, what if we add the DLDL to a new fully connected layer following the last fully connected layer? Will we get better performances for the model? Last but not least, what if we apply the DLDL method to other fields in computer vision, even those fields other than computer vision? As far as I'm concerned, the above questions are worth thinking in the future.

References

[1] G. B. &. B. Yip, “MORPH-II Dataset, Summary and Cleaning,” 16 June 2017. [联机]. Available: https://garrettbingham.com/files/Cleaning%20MORPH-II%20Presentation.pdf.