Learning Semantics of Classes in Image Classification

The task of classifying objects from images has been a hot research topic in the last years, and the advent of deep learning has improved performance drastically. Typically, deep networks for image classification contain several convolutional layers, followed by at least one fully-connected layer to produce the logits, which are later going through a SoftMax classifier. This Softmax classifier trains the network such that it predicts the correct class, and confidences of all other classes are reduced in the same way, independently on the actual semantic similarity of classes. For example, when the correct class is a Granny Smith apple, for the softmax classifier it is as bad to predict a Gala apple as to predict a baseball, a hammer or an airplane. Thus, there is nothing that tells the classifier about semantic similarities between classes.

The ImageNet dataset [Russakovsky et al., 2015] is composed of 1000 classes, which are semantically aligned in the WordNet tree structure [Miller, 1998]. Hence, through some similarity metrics such as in [Kaushik, 2022], this tree structure provides information on how far (semantically) two different classes are. The goal of this Master thesis is to investigate the errors made by a classifier trained using standard SoftMax training. When a sample is classified incorrectly, how semantically far is the correct class from the predicted one?

After this metric is developed, the semantic similarity of classes should be incorporated into the training of a classifier. One possible solution is to assure that deep features of related classes have low distances. Another solution could be to have multiple tasks in one network: First predict the category and then the final class. Further approaches include the modification of the SoftMax activation or the categorical cross-entropy loss to incorporate semantic information.

Requirements:

- Participation in my Deep Learning course.
- A reasonable understanding of deep neural networks and how they learn.
- Programming experience in python and a deep learning framework, optimally, pytorch.
- Decent understanding of written English.

References

[Kaushik, 2022] Kaushik, R. (2022). Portability of targeted adversarial attacks. Master's thesis, University of Zurich.

[Miller, 1998] Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.