COMPUTER VISION

Toward principled regularization of deep networks—From weight decay to feature contraction

Atsuto Maki

Training deep artificial neural networks for classification problems may benefit from exploiting intrinsic class similarities by way of network regularization that compensates for a drawback in the commonly used target error.

Imagine that you are 2 years old and being quizzed on what you see in a photo of a leopard with your little eyes. You might answer "a cat," and your parents might say, "yeah, not quite but similar," giving partial credit for your answer. When training an artificial neural network that tells you a class, metrics for its output seldom account for class similarities unless some sort of previous knowledge is available. However, a few recent studies on network regularization (*1*, *2*) introduced effective approaches to model generalization, and interestingly, they can be viewed as commonly exploiting class similarities that are intrinsic to the training data.

Computer vision has seen unprecedented progress with deep learning, especially deep convolutional networks (ConvNets), which involve millions of parameters across layers. Next to optimization techniques, regularization techniques are central to the success of learning, being expected to alleviate overfitting. Yet, understanding their mechanism remains a big challenge, while it may shed light on the causalities of the performance of deep learning in general. This article aims to provide insight for developing a principled technique of regularization in supervised learning with ConvNets, in particular, by illuminating a desirable interpretation of target error in classification problems.

A number of methods are known for network regularization and often used in combinations. See (*3*) for a comprehensive review of popular techniques. According to the taxonomy (*4*), the cost function for optimizing ConvNets is twofold: a target error function and a regularization term. The former depends on the consistency between the predictions and the given ground truth, whereas the latter assigns a penalty independent of the target—something that restricts the model to be simple, e.g., the standard L2 norm for the model parameters to be suppressed (also known as weight decay). Those two typically guide the model in opposite directions.

Consider training an artificial neural network that predicts, given an image x, categorical distribution $p(y|x)$ over classes y based on the activations of the output layer. It is broadly taken for granted that the target labels include a single ground-truth label (one-hot vector); the label is 1 for the true class and 0 for all the rest as being false. Then, the target error is normally defined by the cross-entropy between $p(y|x)$ and a one-hot vector. With mutually exclusive target as such, the network tends to be trained to give a high confidence to "leopard" (see Fig. 1, left, where x is an image of a leopard) and regard other classes as equally wrong, whether cat or motorcycle.

Nevertheless, is a hard target ignoring class similarities ideal from the viewpoint of maximizing the information in the training data? This question is related to the intuition underlying so-called dark knowledge: "The relative probabilities of incorrect answers tell us a lot," on which Hinton *et al.* based their use of soft target (*5*) in a smaller network. In my view, the notion of class similarities is crucial for a proper use of target error, although it seems underestimated in the standard framework. In fact, human perception would admit that some incorrect classes are less wrong, and such information may be useful for a broad range of tasks, particularly in robot learning (*6*), where perception is key to many applications. For example, grasping inherently involves a detection problem (*7*) for which classification plays an essential role because a robot often needs to recognize various types of similar objects in the scene, including their precise locations.

Class similarities can be taken into consideration in two ways in network regularization, i.e., on either of the two distributions compared in the aforementioned cross-entropy loss: the ground-truth or prediction distribution. Let us first consider manipulating the ground-truth distribution. One simple form is label-smoothing regularization (*8*), which replaces the target labels with a mixture of the original one-hot representation and a fixed uniform distribution. For more elaborate target class probabilities, a hierarchical model of visual concepts was used for detection purposes (*9*). However, finding reasonable target probabilities representing class similarities is not straightforward.

The second approach is to induce relative probabilities that reflect the intrinsic class similarities to emerge on predictions p (see Fig. 1, right), hence allowing the distribution to have a higher entropy. This can be encouraged by adding a specific penalty on top of the cross-entropy loss. Two types of losses were independently suggested (*1*, *2*), and they can be considered alternatives for penalizing peaky distributions: Confidence penalty (*1*) is a direct addition of negative entropy of p and has been reported with its connection to label-smoothing to improve state-of-the-art models across various benchmarks. Feature contraction (*2*) adds the L2 norm of the feature vector from a certain layer, typically the second to last layer. Formula wise, it looks deceptively similar to weight decay, but it affects the elements of feature vector. It suppresses high values, both positive and negative, which is propagated to the prediction distribution, while keeping lower values on other nodes less affected. Early studies showed its powerful effect in transfer learning without sacrificing the training speed [an available demo code (*2*)].

From the perspective of network optimization, another justification for those losses is the magnitude of gradients (*2*, *8*). That is, they promote a larger gradient owing to the prediction distributions with higher entropies. Maintaining some gradients ought to

KTH Royal Institute of Technology, Stockholm, Sweden.Email: atsuto@kth.se

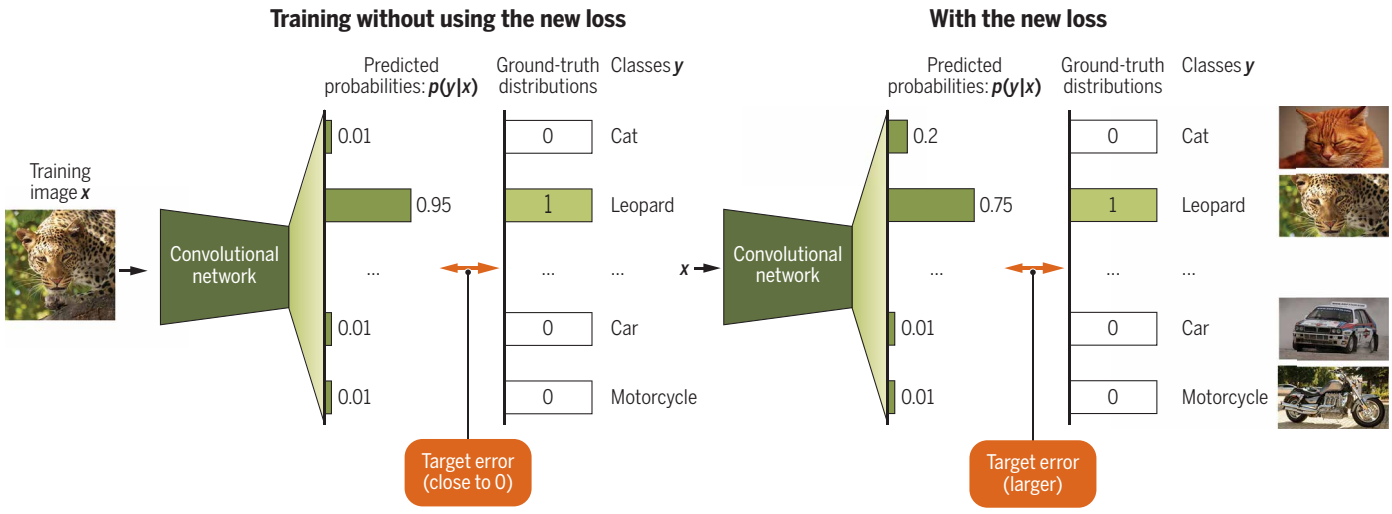

Fig. 1. Schematics of target errors nearer convergence in training a ConvNet for an image classification problem. (**Left**) The network may be overfitting as it places almost all probability on a single class, leopard. (**Right**) Relative probabilities have been induced as {0.2, 0.75, … 0.01, 0.01} reflecting some intrinsic class similarities across {cat, leopard, … car, motorcycle} by a new loss.

help reach a better configuration because overfitting is likely when the network places full probability on a true class (*8*). The new losses (*1*, *2*) counteract it as they exploit the intrinsic class similarities.

Here, explicit regularization techniques were primarily considered in favor of prediction distributions with higher entropy for an overall goal of model generalization. Interesting topics to be explored include (i) how robust the new losses make the network against adversarial samples and (ii) how the increased gradients help optimization on the nonconvex loss landscape. Those also apply widely to other regularization techniques. In this respect, another open question is their relation to implicit regularization (*10*) performed by stochastic gradient descent. It is ascribed to over-parameterization and hidden complexity and offers an important research direction toward principled regularization of deep networks.

REFERENCES AND NOTES

1. G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, G. E. Hinton, Regularizing neural networks by penalizing confident output distributions, in *Proceedings of the International Conference on Learning Representations, Workshop Track* (ICLR-WS, 2017).
2. V. Li, A. Maki, Feature Contraction: New ConvNet regularization in image classification, in *Proceedings of the British Machine Vision Conference, 213* (BMVC, 2018); demo code available at https://github.com/VladimirLi/feature_contraction_example.
3. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning (Chapter 7)* (MIT Press, 2016).
4. J. Kukačka, V. Golkov, D. Cremers, Regularization for deep learning: A taxonomy. arXiv:1710.10686 [cs.LG] (29 October 2017).
5. G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network. arXiv:1503.02531 [stat.ML] (9 March 2015).
6. D. Kragic, M. Björkman, H. I. Christensen, J.-O. Eklundh, Vision for robotic object manipulation in domestic settings. *Robot. Auton. Syst.* **52**, 85–100 (2005).
7. I. Lenz, H. Lee, A. Saxena, Deep learning for detecting robotic grasps. *Int. J. Robot. Res.* **34**, 705–724 (2015).
8. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in *Proceedings of the IEEE International Conference Computer Vision and Pattern Recognition* (CVPR, 2016), pp. 2818–2826.
9. J. Redmon, A. Farhadi, YOLO9000: Better, faster, stronger, in *Proceedings of the IEEE International Conference Computer Vision and Pattern Recognition* (CVPR, 2017), pp. 6517–6525.
10. B. Neyshabur, R. Tomioka, N. Srebro, In search of the real inductive bias: On the role of implicit regularization in deep learning, in *Proceedings of the International Conference on Learning Representations, Workshop Track* (ICLR-WS, 2015).

Acknowledgments: I would like to thank Vladimir Li for fruitful discussions as well as for having provided the demo code of feature contraction regularization referred to in this article. The comments from anonymous reviewers are gratefully acknowledged.

10.1126/scirobotics.aaw1329

Citation: A. Maki, Toward principled regularization of deep networks—From weight decay to feature contraction. *Sci. Robot.* **4**, eaaw1329 (2019).

Science Robotics

Toward principled regularization of deep networks—From weight decay to feature contraction

Atsuto Maki

Sci. Robot. **4** (30), eaaw1329. DOI: 10.1126/scirobotics.aaw1329

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.