Advancements in Semi-Supervised Learning: Data Distillation

Chris Mitsopoulos

COMP 135 - Project 03

Traditionally, research in Machine Learning has focused on two types of learning for building classification, regression and other types of models: supervised and unsupervised learning. In supervised learning, the model learns a mapping from input to output through experience, by relying on labeled example input data [1]. Assume we are classifying images of apples and oranges. Each example in our dataset (that consists of pixel values) will contain a label that states whether the current example is an apple or an orange. This is the label of our input. We then hope that the model will then be able to generalize well to unseen data after learning from example inputs. In other words, since we want the model to generalize to any new data point in our current domain, we are making the implicit assumption that our dataset (sample) is representative of the distribution of our total population of examples. In unsupervised learning, inferences are drawn from input data that **do not** have labeled outputs. This is done through searching for similarities and groupings in the dataset (in the form of clustering into groups for example) [2]. Semi-supervised learning acts as a middle ground between the two learning techniques mentioned, with the main idea being that a mixture of labeled and unlabeled data can be used to train models. By drawing from the robustness and proven success of supervised learning as well as the raw amount of unsupervised data available from internet sources, more data is ultimately used for training which in turn leads to better accuracy in unseen data and a more robust model. Researchers in semi-supervised learning believe there is value in being able to readily use raw unlabeled data in conjunction with clean labeled datasets. In 2017, Facebook AI research (FAIR) [3] published a paper called "*Data Distillation: Towards Omni-Supervised Learning"* [4] in which they explain the development of an omni-supervised learning technique, which they claim is lower-bounded by the accuracy of training on all annotated data [4] (the accuracy of supervised learning).

Due to the scale and globalization of the internet today, heaps of unlabeled images, textual data and videos are more prevalent and readily available than ever. On the other hand, labelling datasets is an arduous task that has to be performed by humans. Generating clean and consistently labeled datasets is time-consuming and acts as a bottleneck for supervised learning. FAIR proposed the idea of using "data distillation" [4], to generate labels for unlabeled data and using those generated labels as extra data to retrain and improve the model in a supervised learning manner. This idea of Pseudo-labelling is not new and has been explored in semi-supervised learning literature. For example, in an image classification problem, pseudo-labels could be created for unlabeled images, by training the model with labeled and unlabeled images simultaneously and using the majority classification result for unlabelled images to assign them labels [5]. However, the general risk of training a model using its own prediction stems from the model misclassifying unlabeled inputs and then enforcing its mistake by falsely updating weights in a Neural Network, which could lead to an increase in error and reduction in accuracy. Also, typically these semi-supervised experiments are simulated using a fully labeled dataset (even though a percentage of these labels will be discarded to evaluate the ability to pseudo-label), which means that the models trained are upper bounded by a model trained using the whole labeled dataset in a supervised learning fashion [6]. However, with the advancements of State-of-the-art supervised learning models that can classify unseen data with a very high degree of confidence, such feedback loops can be reconsidered.

In Machine Learning research, transformations on input data, such as horizontal flips or pixel translations in image data, are used to increase learning data for many processes like image classification and object detection [7], in order to increase the accuracy and robustness of the models trained. The process of data distillation uses this idea as part of the pipeline, by performing transformations on an input image to create several input images, running these examples through the same model and ultimately creating a label from the ensemble of runs on each transformation image. The result is “a single model run on multiple transformed copies of unlabeled data”[4].

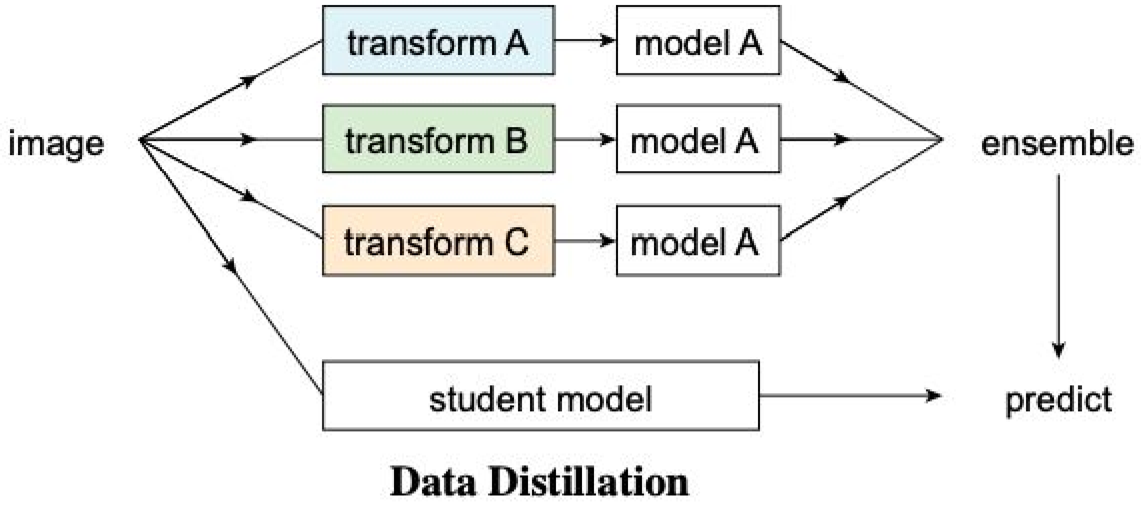

Figure 1: Model A outputs of each image transformation are combined to generate a label for the image which is then used to train the student model [4]

In principle, the process of data distillation is very intuitive. The model is trained on labeled data, like any other model that employs fully-supervised learning. The hope is that labeled data will be able to create a robust model that is close to state-of-the-art and will be able to classify correctly to a high degree of confidence. The same trained model is fed with multiple transformations of examples in the unlabelled dataset [4], leading to a number of different predictions for each data point (as there are many transformations of a single input). All predictions are then aggregated into a single ensemble prediction and a label is appended to the current unlabeled data point being inspected. This labeled data point is then added to the training set (which only has labeled data) and the model as a whole is retrained. The idea of feeding multiple transformations of a single example to the same model is referred to as multi-transform inference [4] in the paper. How are all these transformations then aggregated to generate a single class label? Depending on the task in hand, be it classification or human pose estimation for example, "hard" labels (labels that are categorical by nature) are preferred so that no other changes need to be made to the structure of the network. On the other hand, if the labels generated were probability vectors (ie a vector of the same length as the number of output classes, with each index storing the probability of input being in that class), the techniques used for gradient descent and backpropagation would have to be restructured to work with these "soft" values, whereas by using hard labels no changes have to be made to accompany label generation. Therefore, additional logic is required for label generation depending on the task at hand, a potential drawback of data distillation. Once generated, the newly automatically labeled data are indistinguishable from the labelled data used in the supervised learning section of the procedure. To ensure that the model follows a trajectory of improvement in accuracy and reduction in error, and that the model is not re-trained with a high amount of newly labelled data, the ratio of automatically created labels to manually created labels already in the dataset at each training batch is predetermined and constant, to ensure that erroneous gradient updates stemming from falsely automatically labeled data are minimized [4]. Experiments results of using data distillation on keypoint detection on the COCO dataset [8] along with an unlabeled dataset from a different source (and so supposedly a different distribution), were favorable, with a higher resulting average precision than that obtained from using the same model on solely labeled dataset [4].

Advancements in semi-supervised learning are exciting as they build on top of a well-grounded part of ML literature that has shown state-of-the-art capabilities (supervised learning with large datasets), while also harnessing the scale of raw data that is readily available to researchers today but needs time and preparation to prove useful.

References

1. Chapelle, O., Scholkopf, B., & Zien, A. *Semi-Supervised Learning*. Retrieved from http://www.acad.bg/ebook/ml/MITPress- SemiSupervised Learning.pdf

2. Unsupervised Learning. (n.d.). Retrieved December 16, 2019, from https://www.mathworks.com/discovery/unsupervised-learning.html.

3. Facebook AI. (n.d.). Retrieved December 16, 2019, from https://ai.facebook.com/.

4. Radosavovic, Ilija, Dollár, Piotr, Girshick, Ross, & Kaiming. (2017, December 12). Data Distillation: Towards Omni-Supervised Learning. Retrieved from https://arxiv.org/abs/1712.04440.

5. Lee, Dong-Hyun. (2013). Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. ICML 2013 Workshop : Challenges in Representation Learning (WREPL).

6. Rodriguez, J. (2018, September 26). Facebook Believes in Omni-Supervised Learning. Retrieved December 16, 2019, from https://towardsdatascience.com/facebook-believes-in-omni-supervised-learning-78f37253f4f4.

7. Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T., Shlens, J., & Le, Q.V. (2019). Learning Data Augmentation Strategies for Object Detection. *ArXiv, abs/1906.11172*.

8. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L. (2014). Microsoft COCO: Common Objects in Context. *ECCV*.