An Empirical Study of Example Forgetting During Deep Neural Network Learning

Björn Bebensee (2019–21343)
Topics in Artificial Intelligence

November 26, 2019

Toneva et al. [1] define "forgetting events" as the transition of a training example during a classification task from being classified correctly to incorrectly during the training process. In this paper they investigate and try to characterize such forgetting events which lead to *catastrophic forgetting* (i.e. forgetting what was previously learned when trained across multiple tasks). Conversely the authors define *unforgettable* examples as training examples that are never forgotten across subsequent training tasks.

In order to characterize example forgetting Toneva et al. train a classification model for the MNIST, permuted MNIST and CIFAR-10 datasets. Out of the tested datasets, MNIST has the biggest portion of unforgettable examples (as images are less complex). During training, many forgettable examples need to be presented more often than unforgettable ones until they are first learned (correctly classified). Upon analysis of the unforgettable examples, they find that most forgettable training examples exhibit atypical class characteristics and are thus harder for the model to learn correctly. Based on their findings, Toneva et al. hypothesize that unforgettable examples contribute less and are less informative than forgettable examples. To test this they remove unforgettable examples from the dataset. Surprisingly up to 30% of the CIFAR-10 dataset can be removed while still maintaining competitive performance.

Finally, the authors confirm that this behaviour persists across different model architectures and that there is a large intersection of unforgettable training examples in different architectures (i.e. these examples are unforgettable in all architectures).

References

[1] Toneva, Mariya, et al. "An empirical study of example forgetting during deep neural network learning." *arXiv preprint arXiv:1812.05159* (2018).