Streszczenie

Głębokie Uczenie (ang. *Deep Learning*, DL) pozwala osiągnąć rewolucyjne wyniki w takich dziedzinach jak wizja komputerowa, czy przetwarzanie języka naturalnego. Niestety, teoretyczne własności sieci neuronowych (modele rozważane w DL) które stoją za tym sukcesem są dalej niejasne. Na przykład sieci neuronowe mają często o rzęd wielkości więcej parametrów niż rozmiar zbioru trenującego. Jest to sprzeczne z intuicją, że proste modele powinny generalizować lepiej (tj. działać lepiej na nowych przykładach). Ta pozorna sprzeczność motywuje badanie teoretycznych podstaw sieci neuronowych.

Naszym punktem startowym jest obserwacja, że większość głębokich sieci neuronowych jest trenowana z użyciem prostego algorytmu Stochastic Gradient Descent (SGD). Wcześniejsze prace pokazują, że wybór metod optymalizacji wpływa na generalizację modelu. W ramach tej pracy rozwiniemy ten nurt badań.

Na pracę doktorską składa się sześć publikacji. Głównym rezultatem jest zestaw teoretycznych i empirycznych wyników pogłębiających nasze zrozumienie trenowania sieci z użyciem SGD. Podsumowywując najważniejsze z nich to:

- Na krzywiznę funkcji kosztu ma duży wpływ krok uczenia i rozmiar próbki (hiperparametry SGD). Nie tylko w końcowym minimum, ale też w czasie całego trenowania.
- Skala szumu stochastycznego w SGD jest kontrolowana przez iloraz kroku uczenia do rozmiaru próbki. Ten iloraz jest kluczowym czynnikiem wpywającym na dynamikę uczenia i końcową generalizację.
- Po początkowej fazie uczenia poziom skomplikowania funkcji (w przestrzeni wejścia) i krzywizna funkcji kosztu są w dużej mierze ustalone.

Z praktycznego punktu widzenia wyniki w tej rozprawie mogą pomóc praktykom trenować sieci neuronowe, przykładowo upraszczając dobór hiperparametrów SGD. Mamy także nadzieję, że nasza praca przyczyni się w przyszłości do rozwoju optymalizatorów dostosowanych do sieci neuronowych.

Stanisław Jastrzębski

Summary

Deep Learning (DL) has led to numerous breakthroughs in fields such as computer vision or natural language processing. However, theoretical reasons behind this success remain still unclear. For example, the number of parameters of a deep network routinely exceeds by orders of magnitude the number of training examples. This is somewhat in contrast with the intuition that simpler models should generalize better (i.e. work better on unseen examples); as John von Neumann said: *"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk"*. This apparent tension between the success of DL and lack of understanding of its theoretical underpinnings motivates this work.

The approach we take starts from the observation that *the common denominator of most modern deep networks is that they are trained using stochastic gradient descent (SGD)*, a simple optimization algorithm. Previous work has argued that optimization acts as an implicit regularizer. Depending on the optimization algorithm, or its hyperparameters, the final solution can have different generalization properties. We build on this line of thought.

This thesis is composed of six papers. To summarize, we consider as the key contribution the set of empirical and theoretical results about SGD–based training of deep networks. We highlight here three main take-aways:

- Curvature of the loss surface is highly influenced by the learning rate and the batch size (hyperparameters of SGD); both during training and at the final minima.
- The scale of stochastic noise in SGD is controlled by the learning rate to batch size ratio, and is a crucial factor for the learning dynamics and the final generalization performance.
- After the early phase of training both the complexity of the function in the input space and curvature of the loss surface are largely determined.

Practically speaking, results in this thesis can help practitioners train neural networks; for instance by simplifying finding the optimal hyperparameters in SGD. Besides, we hope our work will help develop in the future new optimizers tailor-fit to deep neural networks.

Stanisław Jastrzębski