Brief History of Deep Learning Training 20200510

Nicola Bernini ,

May 2020

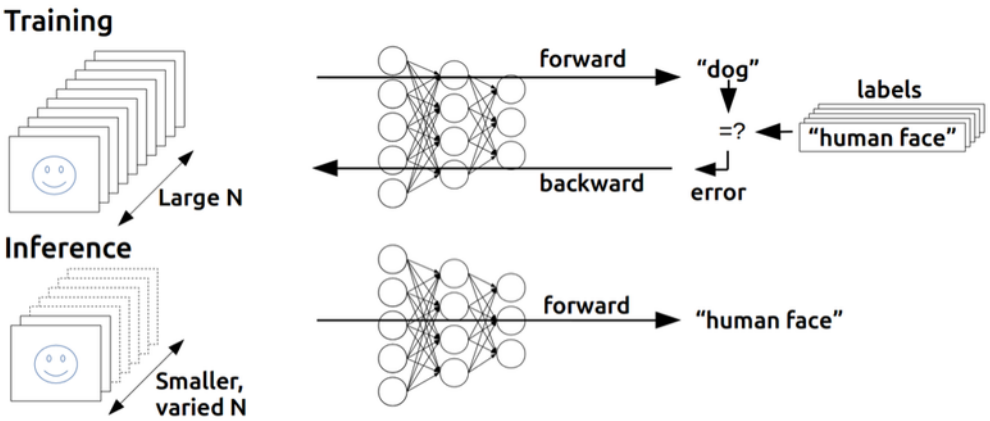

Figure 1: Training

1 Introduction

- It is well known that the Loss Function for DNN Training is highly non convex
- In 1986 Murty, Katta G, Kabadi, Santosh showed that **finding the global optimum** of a **generic non convex function** is NP Complete
- Generic means with no **specific structure** to exploit to make the problem easier
- Unfortunately in 1992 Blum and Rivest showed (quoting the paper)

  TRAINING A 3-NODE NEURAL NETWORK IS NP-COMPLETE
- NOTE: It is important to observe here it is meant to find the global optimum

- so training a Deep NN has been classified as an **intractable problem** at least with the technology so it all depends on the available computational resources: technology plays a big role
- But we all know DNNs are now very commonly used in practice to solve problem, so **how is it possible if they are so hard to train** ?
- Certainly the technology has evolved a lot since then: we now have powerful GPUs allowing us to train way faster than in the past, but is this enough to explore the super-huge weights space of a the DNN models we used today? It is not (yet).
- So what's the secret behind the recent success of DNN Training?
- They key is: **apparently, there are quite a lot of low hanging fruits** meaning that there is an **abundance of local minima which make DNN work well**
- It means **there is no need to find the global minimum** to make the DNN work well in practice, in fact a local minimum is typically associated to good performance which also means **good enough generalization for task-specific applications**
- This is anyway an empirical evidence: practically, when we train a DNN and it works well of course we can't assume we have found the actual global minimum (as it is NP Complete) so we assume we have found a local minimum and which is **good enough** and with **good enough** we mean we have measured the performance of the trained model (very much likely to be in a local minimum) with some metric on a set of data not used for the training (Test Set) and this is OK for us
- NOTE: here there are no theoretical guarantees, just an ex post empirical measure on a certain dataset (Test Set) and a subjective judice