Report HW7
Robotics Multimodal Deep Learning for Object Recognition

Giorgio Giannone

August 1, 2018

Introduction

Our goal is to implement a Network to detect 15 class objects, fusing RGB and Depth information.

Architecture

We build a simple Convolutional Neural Network to solve this problem. In particular the structure is:

$$Conv(64,(5,5)) \rightarrow MaxPool((3,3),(2,2)) \rightarrow Conv(64,(5,5)) \rightarrow MaxPool((3,3),(2,2)) \rightarrow$$

$$Flatten() \rightarrow Dense(384) \rightarrow Dense(192) \rightarrow Dense(15)$$

where $Conv(f,(k,k))$ is a convolutional layer with f filters and kernel size k. $MaxPool((r,r),(s,s))$ is a downsampling layer with downsampling rate r and stride s. $Dense(g)$ is a fully connected layer with g number of units.

Result

We report the results using only RGB (Fig.1), only Depth (Fig.2) and both modalities (Fig.3). We selected an image resolution of (48, 48) and we trained for 100 epochs to speed-up the training procedure.

Discussion

In a multimodal setting we expect to improve the metric using both modalities. In this case we improve over the RGB modality. But we obtain an unexpected result for the depth modality: this can be explained considering that the network is not able to fuse completely the two modalities.

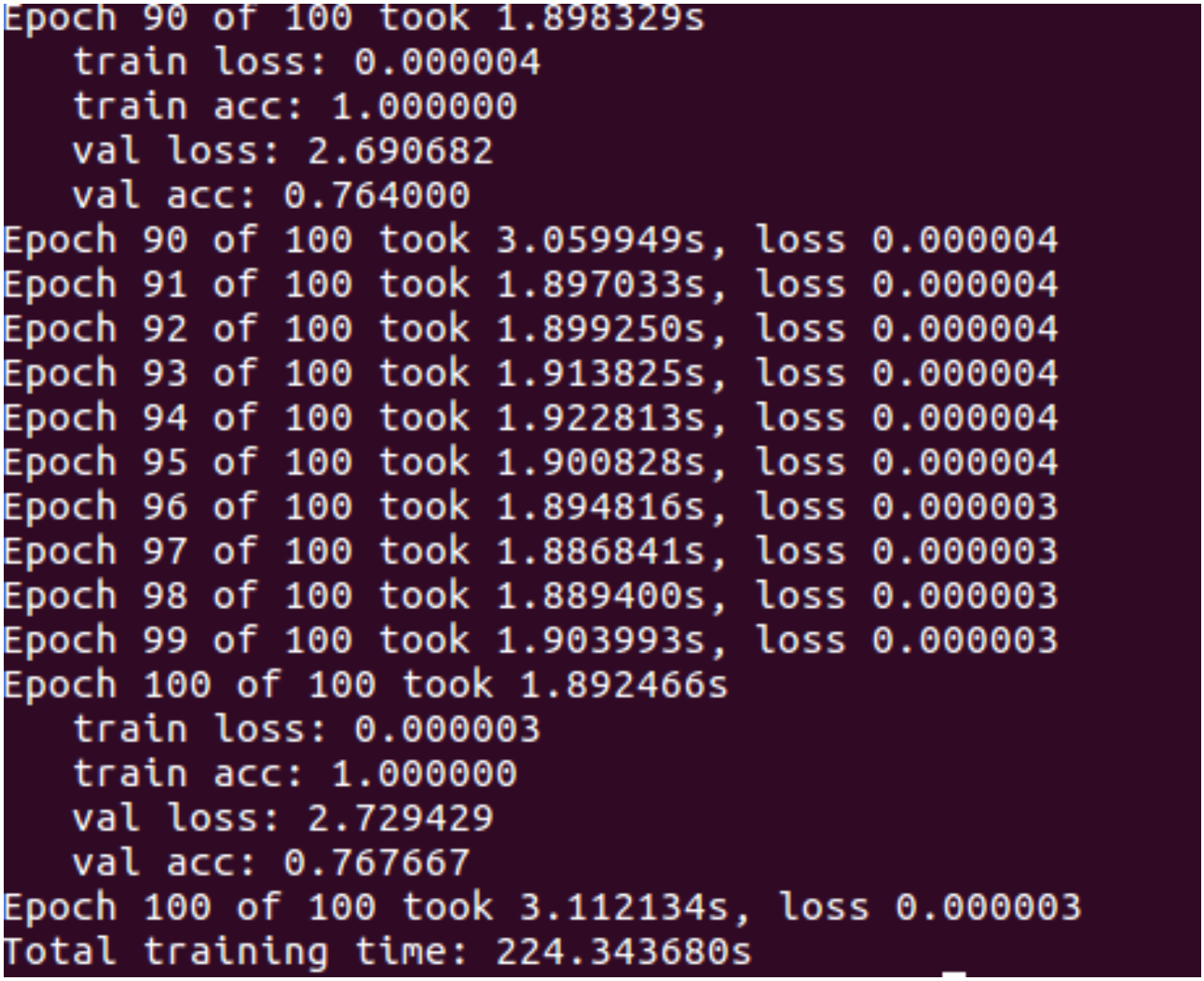

Figure 1: result using only RGB

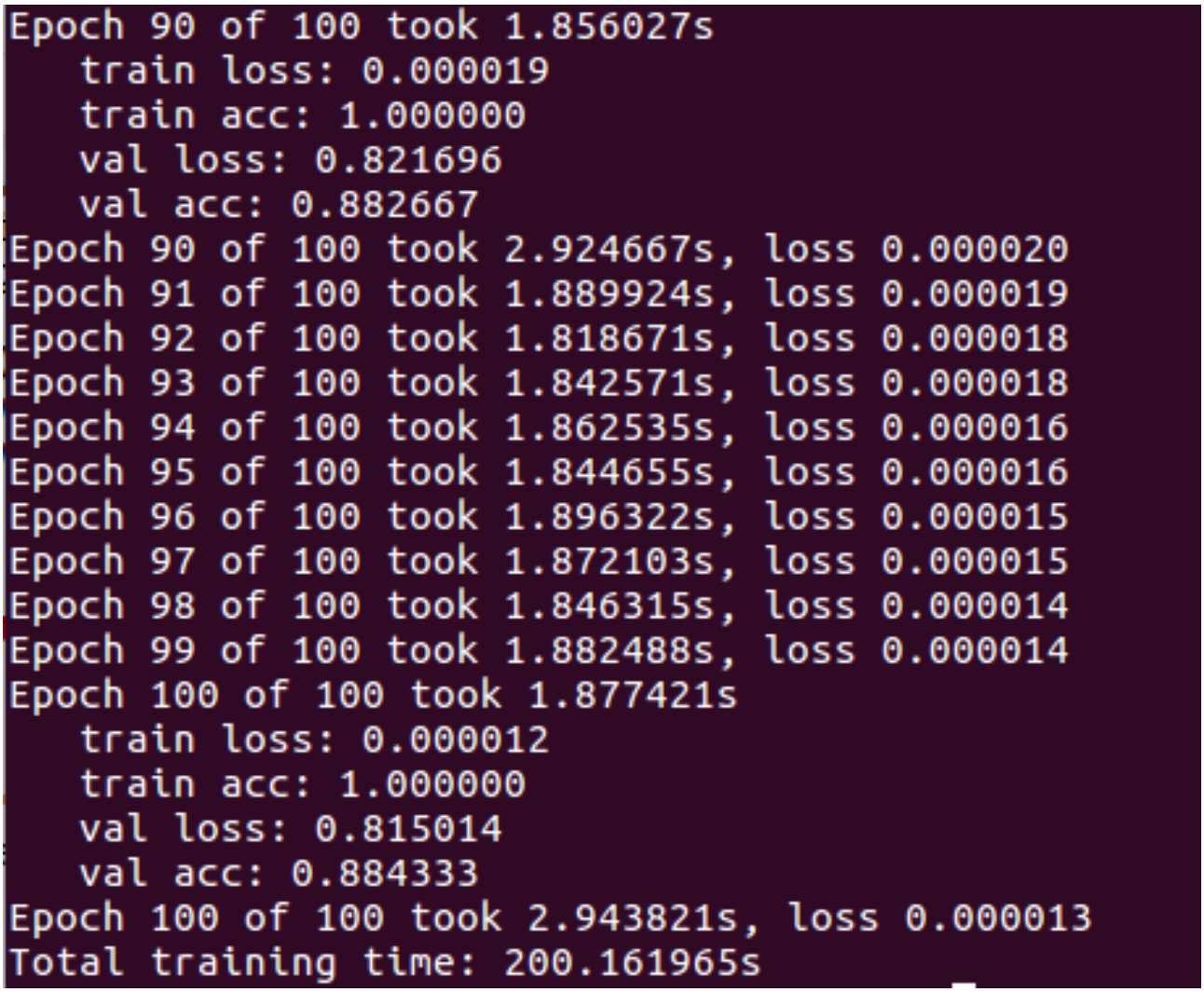

Figure 2: result using only Depth

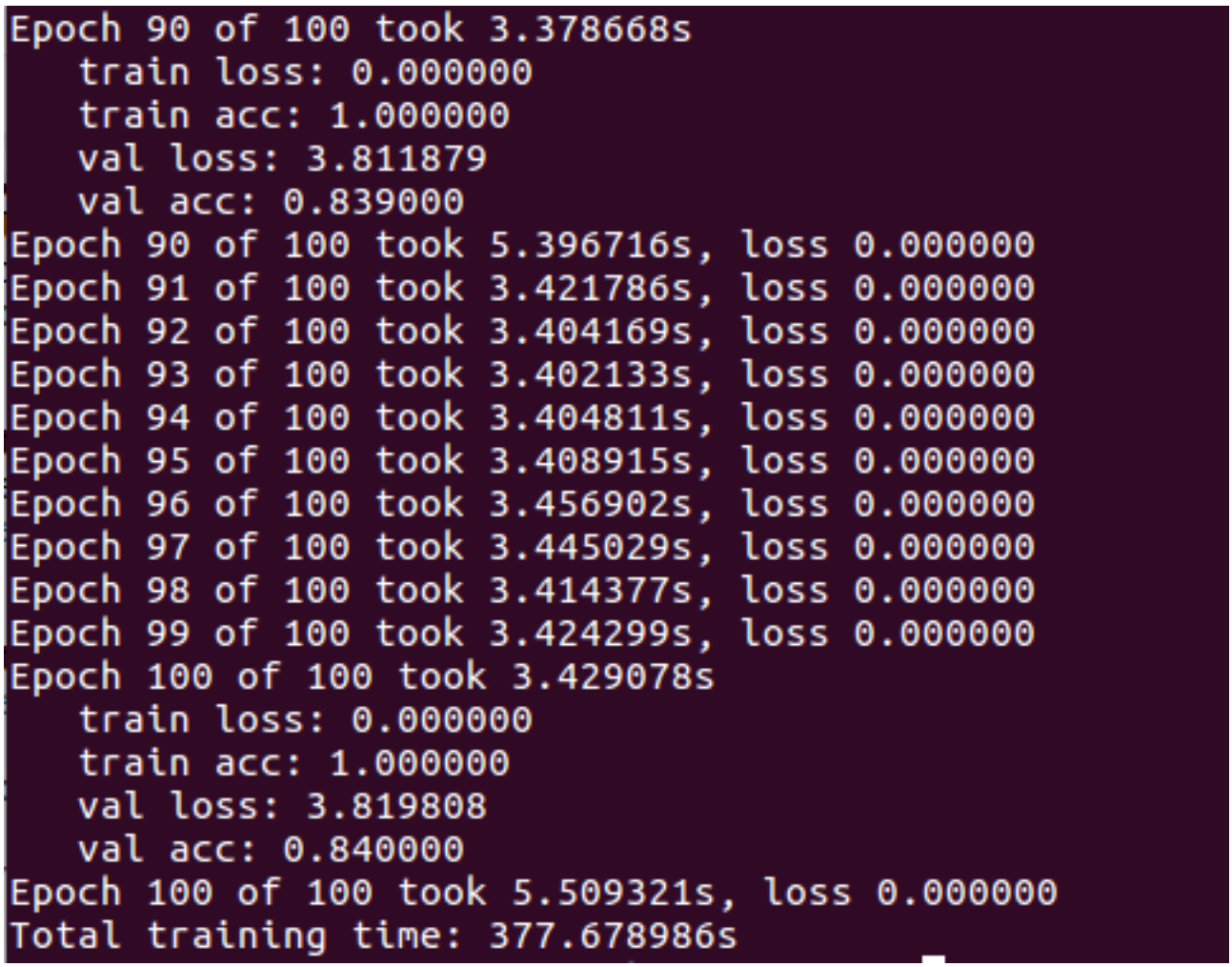

Figure 3: result using RGB and Depth