Which Inputs of a Neural Net Lead to the Same Output?

Raymond Veldhuis, Zohra Rezgui, Amina Bassit

April 14, 2021

1 Introduction

In this assignment, we want to analyse what type of variations of the input of a neural network lead to the same output. This is important, because sometimes this behaviour is desired, but at other times, it is not because it is unpredictable. An example of a desired mapping of input variations to the same output is robustness to variations of pose or illumination of a facial image in face recognition. This is a desired behaviour, because we don't want the pose or the illumination to affect the recognition result.

2 Background

Neural networks that operate on images, such as autoencoders and classification networks map input images, denoted as vectors $\boldsymbol{x} \in \mathbb{R}^D$ onto a lower dimensional *latent vector*, sometimes also called *embedding*, $\boldsymbol{z} = F(\boldsymbol{x}) \in \mathbb{R}^d$, with $d < D$ and often $d \ll D$. The latent vector is then further used for, for instance, reconstruction or classification. Ideally, the mapping is such that in $\boldsymbol{z}$ the relevant properties of $\boldsymbol{x}$ for the task at hand (e.g. reconstruction of classification) are preserved. The network function $F(\boldsymbol{x})$ is noninvertible and more than one input $\boldsymbol{x}$ can be mapped onto the same latent variable $\boldsymbol{z}$ as shown in Figure 1. Each row of Figure 2 shows the variation of facial images that could lead to the same or very close latent vectors as an artificially generated example.

A bit of math: For a certain $\boldsymbol{z}_0$ we denote the inputs that are mapped onto $\boldsymbol{z}$ by $F^{-1}\{\boldsymbol{z}_0\} = \{\boldsymbol{x}|F(\boldsymbol{x}) = \boldsymbol{z}_0\}$. Note that F^{-1} is not the inverse of F, because F is not invertible. The function F^{-1} operates on sets an is called the *preimage* of a set in mathematics. Formally, $F^{-1}A = \{\boldsymbol{x}|F(\boldsymbol{x}) \in A\}$. Here we are interested in $F^{-1}\{\boldsymbol{z}_0\}$ and maybe, if time permits in $F^{-1}\{\boldsymbol{z}|\|\boldsymbol{z} - \boldsymbol{z}_0\| \leq \epsilon\}$: the set of inputs mapped onto a small neighbourhood of $\boldsymbol{z}_0$.

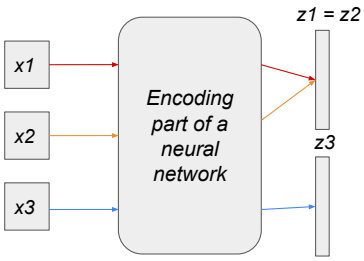

Figure 1: Illustration of samples x_1 and x_2 mapped to the same latent vector by a neural network.

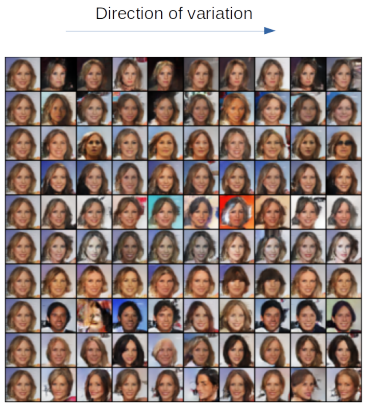

Figure 2: An artificially genereated example of facial images that could be mapped on to the same latent variable. Every row corresponds to one latent variable.

3 The assignment

The goal of this assignment is to study preimages of elements of the latent space of a classifier, and, if time permits, preimages of small neighbourhoods of elements of the latent space of a classifier, by visualisation of these variations in the input image. So, in human language: *We want to see which images are mapped onto the same latent variable.*

More specifically, for a relatively simple classifier, i.e. with a low-dimensional latent space, and for a manageable dataset, in this case the MNIST set of digits, the task is to analyse the preimages of the latent vectors corresponding to the input digits by visualisation.

We suggest to do this by linearisation of the network function F. If we have $F(\boldsymbol{x}_0) = \boldsymbol{z}_0$, then for small deviations $\boldsymbol{\Delta}$ of $\boldsymbol{x}_0$ we may write

$$F(\boldsymbol{x}_0 + \boldsymbol{\Delta}) \simeq \boldsymbol{z}_0 + \frac{\mathrm{d}F(\boldsymbol{x}_0)}{\mathrm{d}\boldsymbol{x}}\boldsymbol{\Delta}.$$

The first derivative $d \times D$-matrix $\frac{\mathrm{d}F(\boldsymbol{x}_0)}{\mathrm{d}\boldsymbol{x}}$ can be found analytically by applying back propagation to the network. We look for vectors Δ that are in the null-space of $\frac{\mathrm{d}F(\boldsymbol{x}_0)}{\mathrm{d}\boldsymbol{x}}$, because for those $\boldsymbol{\Delta}$ we have, by approximation, that $\boldsymbol{x}_0 + \boldsymbol{\Delta} \in F^{-1}\{\boldsymbol{z}_0\}$.

More details can be provided when the research proposal is written.