Inverse classification using generative models

Exposé Bachelorarbeit

Marius Hobbhahn
4003731

A major problem in current research on neural networks, especially in classification and recognition tasks, is the size of training data sets needed, to yield accurate results. This is primarily due to way in which the network is trained to represent real world concepts. A neural network trains on many variations of the same concept, i.e 1000 pictures of a cat from different angles with separate light intensities. After this training process it can distinguish different categories with very high accuracy in many applications. Reducing the number of variations would be beneficial to cut costs and training time. In the optimal case a reduction would go as far as one variant per concept. The network would then be able to see one representation of a concept and be able to identify others since it learns the essential features of this idea. Human learning indicates that this is possible since it is very unlikely that children have huge data sets stored in their memory after birth but are still able to generalize concepts from very few examples.

The way in which we try to achieve this goal is by combining a generative model with a inverse classification approach. The process is structured in two steps. The first part is the training of a generative recurrent neural network (RNN) on a number of sequences representing letters. This is likely done by feeding the network a one-hot vector, representing a given class until it is able to produce this letter on its own. This has already been done experimentally to great success in [MMS17]. The second part of the model is the inverse classification. Here, an undefined class, i.e. a vector full of zeros is presented as input and the picture to be classified given as the corresponding label. Likely, the difference between generated output and label will be rather high. Gradients are then back-propagated through the network and a prediction for a class is given inversely. This prediction is put into the network

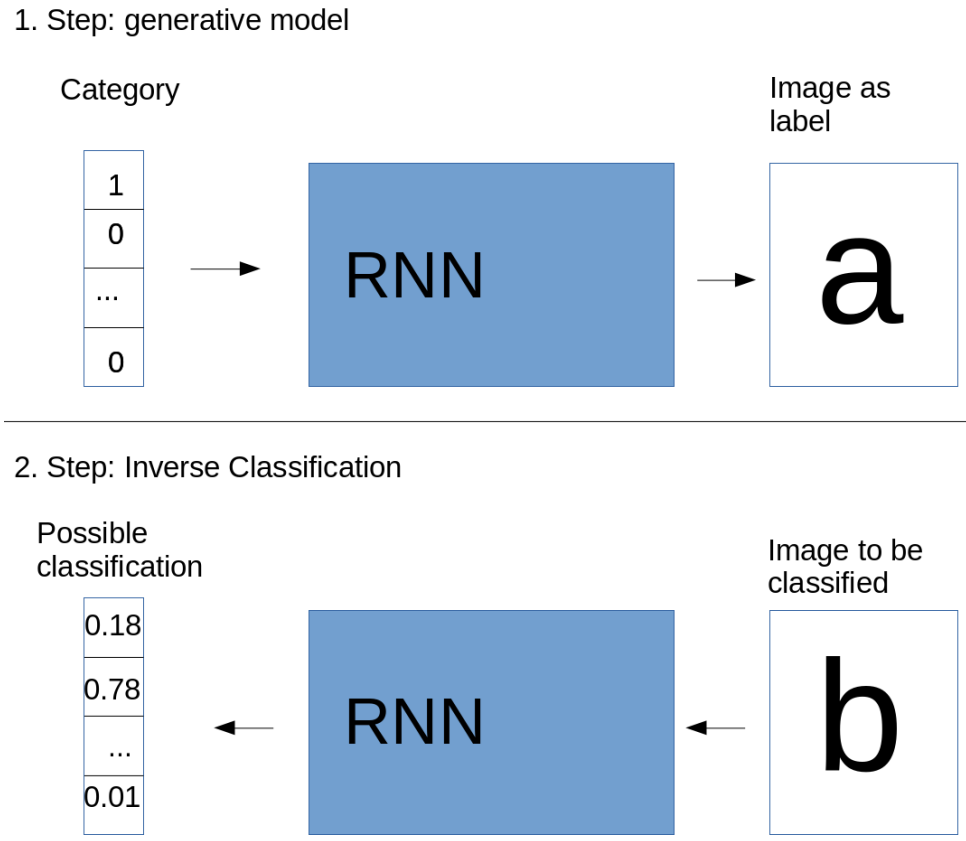

Figure 1: Top: A generative model is trained by input of a one-hot vector and showing sequences representing handwritten letters as labels.
Bottom: Inverse classification process: class is guessed, compared to actual sequence and then adapted. After some iterations class should be converging to the real letter.

again and the process is repeated until the class vector converges. In the optimal case it again is a one-hot vector representing the letter of the image that we wanted to classify. Importantly, the networks weights are not changed in the second step but only the current representation of the class vector. The concept of inverse action inference [OSFB17] and, more importantly, context input inference [MDAS18] has been shown successfully. The whole model can be seen in figure 1.

A similar approach, combining a generative model with a classification network, is taken in variational autoencoders, published in [KW13]. This work tries to extent autoencoders in two ways. Firstly, the generative and the classifying model are the same network, while in autoencoders two different models are used. Having the same models for both tasks might lead to more meaningful internal representations of concepts such as features of letters. Secondly, the latent space for autoencoders is generated during the training process and therefore not naturally interpretable by

humans. In our case the latent space is essentially the class distribution for the letters, hopefully increasing interpretability of the classification results.

If the RNN is unable to fulfill this task, a second approach with variational autoencoders might be tried.

Time-table:

- 1. month: Find suitable datasets, implement RNN read related work
- 2. month: Implement inverse classification, optimize RNN and run tests
- 3. month: If possible implement a variational autoencoder and combine with inverse classification
- 4. month: Write down results of the experiments and compare the results to expectations and similar work

References

[KW13] D. P Kingma and M. Welling. Auto-Encoding Variational Bayes. *ArXiv e-prints*, December 2013.

[MDAS18] Martin V. Butz, David Bilkey, Alistair Knott, and Sebastian Otte. REPRISE: A Retrospective and Prospective Inference Scheme. 2018.

[MMS17] Mitja Nikolaus, Martin V. Butz, and Sebastian Otte. Investigating the Generative Capabilities of Recurrent Neural Networks. October 2017.

[OSFB17] Sebastian Otte, Theresa Schmitt, Karl Friston, and Martin V. Butz. Inferring adaptive goal-directed behavior within recurrent neural networks. In Alessandra Lintas, Stefano Rovetta, Paul F.M.J. Verschure, and Alessandro E.P. Villa, editors, *Artificial Neural Networks and Machine Learning – ICANN 2017*, pages 227–235, Cham, 2017. Springer International Publishing.