Bachelor practical: Report

Marco Unternährer

December 22, 2015

1 Introduction

Training a spiking neural network is a challenging task. Neuromorphic vision researchers have to create spiking datasets since there is a lack of such datasets publicly available . One method for converting static image datasets into neuromorphic vision datasets was proposed by moving a sensor in front of an image [11].

Another approach for training a spiking neural network was developed by researchers at the Institute of Neuroinformatics [6]. A non-spiking neural network is trained by using a regular static image dataset like MNIST or Caltech-101. The resulting weights of the trained neural network can then be used to create a spiking neural network. One constraint for this to work is that units must have zero bias.

The task for this bachelor practical was to train a convolutional neural network on the Caltech-101 dataset.

2 Image Classification on Caltech-101

This section will give a short introduction to the Caltech-101 dataset, the used methods and techniques.

2.1 Caltech-101

Caltech-101 is a dataset of pictures of objects, which was collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato at the California Institute of Technology[7]. Most categories have about 50 images, but the number ranges from 40 to 800 per category [7]. Caltech-101 contains a total of 9'146 images, split between 101 distinct object categories (faces, watches, ants, pianos, etc.) and a background category [1]. The size of each image is roughly 300 x 200 pixels [7].

Unlike other datasets, Caltech-101 doesn't have a predefined split between train and test images. [7] suggest to use a fixed number of images per category for training and testing. Due to the fact that categories have vastly different number of images, the accuracy per category can also vary a

lot. On an additional note, [7] pointed out that if a fixed number of images per category is chosen, then the overall error rate should be reported. However, if all the available images are being used for testing, then the average error rate across categories should be reported.

The state of the art classification on the Caltech-101 dataset has an accuracy of 91.44% [8].

2.2 Convolutional Neural Network

A convolutional neural network (CNN) is a type of feed-forward artificial neural network. CNNs are biologically-inspired variants of multilayer perceptrons [2]. From Hubel and Wiesel's early work on the cat's visual cortex [9], it is known that the visual cortex contains a complex arrangement of cells [2]. These cells are sensitive to small sub-regions of the visual field, called a receptive field [2]. The sub-regions are tiled to cover the entire visual field. These cells act as local filters over the input space and are well-suited to exploit the strong spatially local correlation present in natural images. Additionally, two basic cell types have been identified: Simple cells respond maximally to specific edge-like patterns within their receptive field. Complex cells have larger receptive fields and are locally invariant to the exact position of the pattern [2].

Like their biological relatives, units in a CNN take advantage of the 2D structure of the input data (images or other 2D input such as a speech signal). This is achieved with local connections and tied weights followed by some form of pooling which results in translation invariant features [3].

2.3 Batch Normalization

bn [10]

2.4 Cyclical Learning Rates

In [4], Bengio says that the learning rate is the most important hyper-parameter to optimize and if "there is only time to optimize one hyper-parameter and one uses stochastic gradient descent, then this is the hyper-parameter that is worth tuning". It is well known that too small a learning rate will make a training algorithm converge slowly while too large a learning rate will make the training algorithm diverge [16]. Hence one must experiment with a variety of learning rates and schedules (i.e., the timing of learning rate changes) [13].

The conventional wisdom is that the learning rate should be a single value that monotonically decreases during the training [13]. However, Smith demonstrates the surprising phenomenon that increasing the learning rate is overall beneficial and thus proposes to let the global learning rate vary cyclically within a band of values rather than setting it to a fixed value [13].

3 Experimental Setup

3.1 Data Pre-processing & Data Augmentation

Since the heights and widths of the Caltech-101 images vary, all images were scaled to a common size. Like [11], each image was resized to be as large as possible while maintaining the original aspect ratio and ensuring that width does not exceed 240 pixels and height does not exceed 180 pixels.

After running a few experiments with a fairly deep architecture, it became clear, that the roughly 9k images of the Caltech-101 dataset were not enough to accomplish a decent test accuracy. Starting from the original dataset, data augmentation was pursued by randomly transform each picture ten times, and thus leading to approximately 100k images (including the original dataset).

Using an image data generator [5], we ended up with the following augmentation parameters:

- rotation: random with angles in the range 0° to 20°
- translation: random with shift between 0% and 20% of total width/height
- flipping: horizontal/vertical, yes or no (Bernoulli)
- zooming: random zoom axis by +/- 20%

The created images were then resized again, but with the difference that the images were padded with the edge pixels to fill the needed aspect ratio. The pixel values were normalized to the range of (0, 1) after loading the images. Data standardization, subtracting the mean and dividing by the standard deviation, has sped up the training and was also applied.

The generated images are available online and can be downloaded under [15].

3.2 Architecture

Our architectures were inspired by [12]. However, due to the fact that the neural network will be converted into a spiking neural network, we had two constraints. The first constraint was that units must not have any bias. The second constraint had to do with performance: Since the network should be able to classify pictures near real-time, we limited the depth of our architecture to three convolutional and two fully connected layers. Obviously, when introducing batch normalization, there are more layers. In spite of those, the speed of classification shouldn't fundamentally change, since batch normalization is only a linear transformation.

3.3 Training

The splitting of the generated images into training and test samples was stratified, meaning that each class had 90% in the training set and 10% in the test set.

Stochastic gradient descent was used as the optimizer, with Nesterov momentum of 0.9 and a learning rate decay of 5e-4. Batch size was 64 most of the time, but was decreased for some architecture configurations due to GPU memory overallocation.

Layer type	Parameters
convolution	128x5x5, stride 2x2
relu	
maxpool	2x2
convolution	256x3x3
relu	
maxpool	2x2
convolution	512x3x3
relu	
maxpool	2x2
full	1024
relu	
full	102
softmax	

Table 1: Base Network Architecture

Layer type	Parameters
convolution	128x5x5, stride 2x2
batch normalization	
relu	
maxpool	2x2
convolution	256x3x3
batch normalization	
relu	
maxpool	2x2
convolution	512x3x3
batch normalization	
relu	
maxpool	2x2
full	1024
relu	
full	102
softmax	

Table 2: Network Architecture with Batch Normalization

3.4 Implementation Details

The neural network framework of choice was Keras [5]. Keras is a minimalist, highly modular neural networks library, written in Python and running on top of Theano. Being Python-based and especially developed with a focus on enabling fast experimentation, it was a good candidate for this project. Most importantly, Keras seemed to provide almost all needed functionality out of the box. Another relevant aspect, that lead to the decision to use Keras, was that model architectures

Layer type	Parameters
convolution	128x5x5, stride 2x2
relu	
maxpool	2x2
dropout	0.35
convolution	256x3x3
relu	
maxpool	2x2
dropout	0.35
convolution	512x3x3
relu	
maxpool	2x2
dropout	0.35
full	1024
relu	
dropout	0.5
full	102
softmax	

Table 3: Network Architecture without Batch Normalization

are created with code. That made rapid prototyping and experimentation feasible.

The networks were trained on a system equipped with a single NVIDIA GTX 970 GPU, and training the best performing net took only ???? to converge.

The source code for the entire bachelor practical can be found at [14].

4 Results

w/o batch normalization, w/o data normalization with bn, with data normalization

5 Conclusion

enough data is important bn can speed up training time significantly

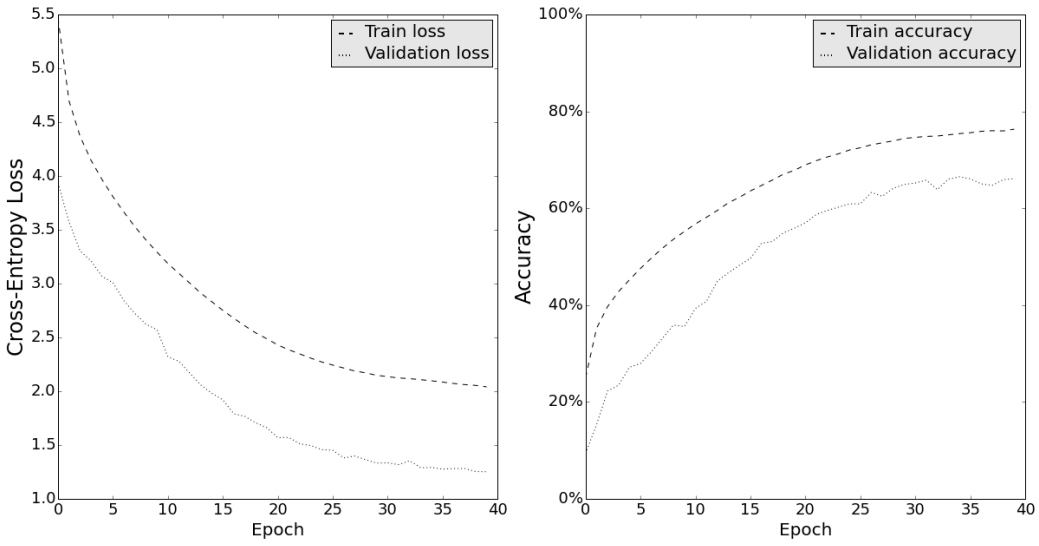

Figure 1: Training/testing with batch normalization

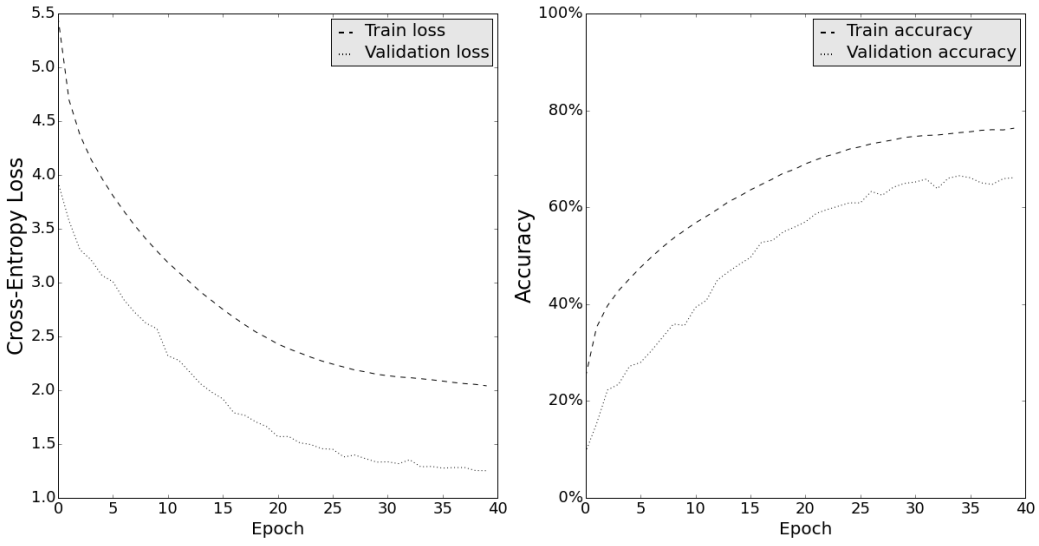

Figure 2: Training/testing without batch normalization

References

[1] Caltech 101 Wikipedia. URL https://en.wikipedia.org/wiki/Caltech_101. Last visited 2015-12-19.

[2] Convolutional Neural Networks (LeNet). URL http://deeplearning.net/tutorial/lenet.html. Last visited 2015-12-13.

[3] Convolutional Neural Network. URL http://ufldl.stanford.edu/tutorial/supervised/ConvolutionalNeuralNetwork/. Last visited 2015-12-13.

[4] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. *arXiv*, jun 2012. URL http://arxiv.org/abs/1206.5533.

[5] François Chollet. Keras. URL http://keras.io.

[6] Peter U Diehl, Daniel Neil, Jonathan Binas, Matthew Cook, Shih-chii Liu, and Michael Pfeiffer. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, jul 2015. ISBN 978-1-4799-1960-4. doi: 10.1109/IJCNN.2015.7280696. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7280696.

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, apr 2007. ISSN 10773142. doi: 10.1016/j.cviu.2005.09.012. URL http://linkinghub.elsevier.com/retrieve/pii/S1077314206001688.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *Eccv*, pages 346–361, 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2015.2389824. URL http://arxiv.org/abs/1406.4729v1.

[9] D H Hubel and T N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, mar 1968. ISSN 0022-3751. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1557912/.

[10] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Arxiv*, 2015. URL http://arxiv.org/abs/1502.03167.

[11] Garrick Orchard, Ajinkya Jayawant, Gregory Cohen, and Nitish Thakor. Converting Static Image Datasets to Spiking Neuromorphic Datasets Using Saccades. pages 1–15, jul 2015. ISSN 1662-453X. doi: 10.3389/fnins.2015.00437. URL http://arxiv.org/abs/1507.07629.

[12] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, pages 1–14, sep 2014. URL http://arxiv.org/abs/1409.1556.

[13] Leslie N. Smith. No More Pesky Learning Rate Guessing Games. *Arxiv*, jun 2015. URL http://arxiv.org/abs/1506.01186.

[14] Marco Unternährer. INI Caltech101 Code, 2015. URL https://github.com/marcuniq/ini_caltech101.

[15] Marco Unternährer. INI Caltech101 Generated Images, 2015. URL https://drive.google.com/folderview?id=0B6t56IB_eb6hVFRGOFp3QVpaR2M&usp=sharing.

[16] Matthew D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. *arXiv*, page 6, 2012. URL http://arxiv.org/abs/1212.5701.