University of Szeged

Department of Informatics

Implementing a self-sizing convolutional neural network

Bachelor's thesis

Author:

Jenei Bendegúz

student

Supervisor:

Varga László Gábor

lecturer

Supervisor:

Berend Gábor

lecturer

Szeged

2018

Contents

Task proposal

The task of the candidate is to create a framework in which convolutional neural networks can be trained, and the size of these networks are determined automatically. The framework should be able to perform the following:

- Training of a fully convolutional neural network to solve a simple image processing task (ex.: edge detection, automatic thresholding, etc.).

- Determining the size of the ideal neural network. An ideal sized neural network solves the problem with a given layer count and structure with the maximum possible accuracy, while keeping the neuron count minimal among layers.

Content summary

This thesis falls into the topics *machine learning* and *digital image processing*, since it uses a machine learning technique of training a *fully convolutional network* to perform *edge detection*, which is an image processing task. The focus is not on the training or task itself, instead it is about optimizing the parameters of the network, specifically minimizing the number of layers and the number of masks in the layers, while not sacrificing the accuracy of the network.

The optimal network size is determined by a training process which is capable of saving its state, and comparing or reverting to previously saved states. It is also capable of altering the inner workings of the network temporarily, to test and compare different configurations. The previous techniques are combined with different metrics, and the training stops when an exiting criterion is met, resulting in a network configuration that is determined optimal by the current strategy.

In the end, a naive strategy seemed the most efficient, which randomly selects masks in the network and disables them temporarily to see if the training can continue without them without a drop in accuracy. A more intelligent method also seemed feasible in some situations, while being more conservative in reducing the count of masks, but it was overperformed by the random method.

The framework was coded in the *Python* language, utilizing the *tensorflow* libraries. To aid performance, the training was hardware-accelerated by using *NVIDIA* graphics cards with CUDA cores, which are operated by tensorflow through the *CUDA* and *cuDNN* libraries.

Introduction

TODO: rizsa

Chapter 1

Technical overview

1.1 Minimal fully convolutional networks

The goal of this thesis is to create a framework which is capable of training a deep convolutional neural network for simple two-dimensional image processing tasks, to provide ways to make this network as small as possible without sacrificing a considerable amount of its accuracy.

TODO: nn & dnn First is a short description of *artificial neural networks*, or *NNs*, on which convolutional neural networks are based. Artificial neural networks are inspired by the biological neural networks in the brain. An artificial neuron is a simple model of a biological neuron. They can have any number of inputs, which originate from other neurons, or directly from the input, or sensory organs in case of real neurons, and they have exactly one output, which can be connected to an input of another neuron, or channeled to the output. When a neuron gets its inputs, it produces a weighted sum of them, adds a bias, applies a function called activation function, and finally puts the result on its output. The *weights* and *bias* within a neuron are its unique parameters that define is characteristics. The activation function is used to introduce some nonlinearity, it is usually a simple linear or exponential function, often cutting negative values, or values falling outside of a range, substituting them with zeros. When a neuron passes it's output to another neuron instead of the output, we are speaking of a *deep or multi-layered neural network.*

To mimic the learning capability of intelligent species, artificial neural networks perform feedback based learning by altering the unique weights and biases of the neurons. An objective function is defined for every neural network, which produces a value which is proportional to how well the network performs. For example, in case of classification, it can be proportional to how confident the network is about an incorrect class, where a

greater values mean poorer performances. Using this value, which is often called error or score, and derivatives, the weight and bias parameters of the neurons are altered based on their contribution to the error or score, improving its value and thus the performance of the network. This process is called *error back-propagation* or *training*.

The size of a deep neural network is defined by two parameters: *number of layers*, and *number of neurons* in each layer, which can be unique among layers. These parameters must be specified when building a neural network for training. This thesis will try to provide automatic strategies to come up with beneficial values for these two parameters, as opposed to defining them with trial & error, repeated manual testing, or based on experience.

Fully convolutional networks, or *FCNs* are derived from convolutional neural networks, but they are not exactly neural due to the lack of artificial neurons in the network. The neurons were used at the last layers, called fully connected layers, right after the convolutional ones. The fully connected layers were used to produce numerical values from the pictures the convolutional layers provide, for classification or regression. FCNs on the other hand let the images be images, making a network which gets images as input and throws back images as output. This makes them able to mark interesting parts of an image, or perform numerous other image processing tasks.

The network needs a task which gets an image as input and gives back another image. This leads us to simple, but non-trivial two-dimensional image processing task, which can be learned in feasible times, but the quality of the edge maps are somewhat subjective, with each method having its strong and weak areas. Below are the the two chosen image processing tasks. Other, very different tasks could have been chosen, which is a possible direction for further experimentation.

1.2 Edge detection

An *edge* in image processing is a sudden change in pixel intensity through an image. This is common on the edges of objects, or at the intersection of different colors in a pattern. There is no criteria for the actual required value of the change intensity-change per pixel, so there is no single right solution for edge detecting an image. There are multiple techniques available for detecting edges, each having its own strengths and weaknesses.

There are two classes of edge detection, based on their output format: *continuous* and *discrete*.

Continuous detectors produce an image of the same dimensions as the source image, where a pixels intensity value corresponds to how strong of an edge is present at that

position in the original image.

Discrete detection uses boolean values to tell if an edge is present or not. Continuous detections can be converted to discrete with thresholding, or using adaptive thresholding on parts of a continuous edge map.

Edges can occur in any direction on an image. The detectors often use derivatives and gradients to determine the sudden drops and rises in intensity, among both axes. Since most detectors use multiple steps until they produce the final edge map, it makes sense to use a deep, multi-layered network, where a layer can more-or-less represent a step in the process. *Convolving* an image with an appropriate mask is also used in edge detection. A *convolutional network* is capable of learning and applying combinations of masks automatically.

1.2.1 Conventional algorithmic edge detectors

The following two edge detectors are important, as they are widely used algorithms, their output will serve as examples of edge maps.

The *Sobel–Feldman operator*, or *Sobel filter* consists of two discrete matrices to be convolved with the source image. This will result in two edge map images, or gradients, one for the horizontal and one for the vertical axis. Then the two matrices can be combined by calculating the geometric mean pixel-wise. The result is a continuous edge map.

The *Canny edge detector* is a more complex one, consisting of five steps, including filtering, gradient computing, which can be learned by the network using appropriate masks, and including non-maximal edge suppression, double thresholding, and hysteresis to suppress weak edges, which is not done using convolution, so it cannot directly be learned by a convolutional network. This will force the network to find its way around the problem, come up with alternative methods using convolutional masks in order to optimize the objective function. The Canny edge detector as opposed to the Sobel operator will result in a discrete edge map.

The network will be trained with raw images, and images ran trough the Canny and Sobel edge detectors. The objective of the network will be to mimic the outcome of the detectors. First, we will teach with the Sobel detector. This should be the easier task for the network, since everything can be done with convolutions with the correct masks. Then we will use the Canny detector, which is considered a harder task due to the multiple steps

including different algorithms.

So the input will be a raw image, encoded and compressed in png format, this is what the network will get to work on, and the sample output will be an edge map, made with either the Sobel operator or the Canny operator. The objective function will ensure that the actual output is as close to the sample output as possible, by producing a high penalty if they differ.

1.3 Tools

1.3.1 Software

There are a variety of machine learning toolkits and frameworks to choose from. Some of the major criteria is the ability to freely customize as much aspects as possible, while also being able to quickly produce a working implementation for fast experimentation. These seemingly conflicting points ask for a framework which is not too high or too low-level. The physical tools, the hardware, must be chosen as well, based heavily on availability and possession.

Python is an open source interpreted programming language, which is widely used in machine learning, or for writing simple tools, thanks to its very fluent syntax, which can feel similar to writing in plain English. Most major machine learning frameworks (tensorflow, Caffe, Caffe2, CNTK and so on) support Python, often as the main language. It is easy and fast to implement and test a new idea, due to it being a high level language, but the performance cost of it is negligible, because the most demanding part of machine learning is not the network building part, which is the part Python is used for, but the training itself, repeatedly evaluating the network, calculating the error and propagating it backwards, and this part is usually made to be hardware accelerated by most frameworks.

Tensorflow, being developed by Google, is a framework for making and using computational graph, supporting the creation of deep neural networks. It is used both for research and as a backend for PC and mobile applications, or the core on which higher level frameworks build on. Tensorflow is based on Python, and it has limited support for C++. There are two ways to make graphs with the toolkit: high level and low level. In the higher level, there are several predefined layers available, these can be parametrized stacked on top of each other to produce a graph. The most common machine learning network layouts are supported by this method. In the lower level, exact operations (addition, convolution, transposing, etc.) and the flow of the data must be coded. This method is used to develop tools for higher level tensorflow, and for research, because of the freedom

of configuration it provides. While being lower level, there are still numerous tools and helper functions available, even image file decoders and preprocessors for example. Another strength of tensorflow is hardware acceleration. With a supported NVIDIA graphics card, computational graphs can be executed in GPU accelerated mode, which is often faster than CPU mode by several magnitudes.

CUDA by *NVIDIA* is required to utilize the computing capacity of a supported GPU, and run tensorflow in GPU accelerated mode. CUDA is crossplatform, available for Windows, macOS and Linux, and it is downloadable freely.

cuDNN, the *CUDA Deep Neural Network* library provides GPU-accelerated routines for common machine learning tasks. It is also available for free, but a registration is required, specifying the plan of usage and associated institutes as well. Tensorflow, when used in GPU mode, relies on this library. One negative effect of this is the non-deterministic nature of some routines, and tensorflow provides no way around it. This way, getting the same exact outcome within two different runs are not possible. Using seeding to achieve this would produce more comparable result, without this technique more testing is required. The exact operation which is affected will be described at section 2.2.2.

1.3.2 Hardware

Initially a relatively low performance notebook was used to perform all the training. The main benefit of using this notebook was its built-in dedicated GPU, the NVIDIA GeForce 840M. It has 384 CUDA cores with compute capability version 5.0, which means it is able to run tensorflow in GPU-mode. Tensorflow demands compute capability to be 3.0 or higher to run in GPU-mode, 3.0 and 3.5 are common among machine learning toolkits. The notebook had 4GB RAM, and 2GB VRAM.

Later, several tests were run remotely at a higher performance desktop computer, while the notebook was used to develop the models and check if they work as intended. These runs were often longer than a week, depending on the sample size. TODO: specs This PC was an Intel Core i7-6700 with 16GB RAM, and it had an NVIDIA Tesla K40 GPU with 12GB VRAM and 2880 CUDA cores.

Chapter 2

Implementation

2.1 Preparing the input images

The Visual Object Classes Challenge 2012 (VOC2012) dataset was used as input images. The VOC dataset contains colored pictures of 20 classes in different settings. Classes include aeroplane, people, bicycle, horse and so on. These are fine for simple image processing tasks, there are a wide variety of objects in various distances and quantities. The classes are not needed for us, since we are outputting an image the same dimensions as the input, instead of performing classification. Without the classes, the expected output for teaching had to be generated from the input images, this was done with matlab scripts and built-in image processing functions.

The images and edge maps are fed to the network during training. The images must be preprocessed to be in an efficient format for training. The goal is to have matrices which can run trough the FCN and produce an output. First the center of the images are cropped, because the input pipeline uses a fixed structure, and the dataset has pictures of varying size. The size of the area to be cropped is 256×256, meaning this is the dimension of the images the network will work on. Then the images are decoded to pure matrices containing floating point values. After this we will have three dimensional images due to the RGB color palette, this is reduced to two dimensions by converting the image matrices to grayscale. This way the color information is lost, but color edge detection is a whole different area. Finally, a number of images are batched together to enable bulk processing of inputs, leaving us with $BATCHSIZE \times WIDTH \times HEIGHT$ dimensions for a single input object.

2.2 Network structure

This section explains the parameters of the FCN, including the types and sizes of the layers, the objective functions and the metrics used.

In tensorflow, the network we build is a computational graph. The vertices are operation, like computations and the edges are the data flowing in one direction, joined to one or more vertices. The edges are also called tensors, they can be simple numbers or multi-dimensional matrices, like images or groups of images. When defining multi-dimensional variables, the tensorflow-recommended *NHWC* format is used where possible:
N: batch, *H*: height, *W*: width, *C*: channels. This forces some consistency into the code and structure, most built-in tensorflow functions expect this format, so no reshaping is required, and there is also a benefit in performance, since this is the optimal format for CUDA operations as well.

Our FCN is made up by three layers, the first layer gets the reprocessed image from the previous section, while the third layer produces the output image. The layers are convolutional with an added bias, initialized randomly with a predefined distribution.

Each *convolutional layer* has its own set of five masks sized 5×5. Strides and dilations are set to one in all dimensions, this means no pixels are skipped when convolving the whole image. A 5x5 mask allows for more complexity than a 3×3 mask, while being able to produce the same results if the weights on the edges are small enough. After a convolution is performed with a mask, a scalar bias is added elementwise to the resulting pixel.

Truncated normal initialization is used before the first training. The values are derived from a general normal distribution parametrized by the given mean and deviation. Truncated means that values with more than two standard deviations distance from the mean are redrawn, this will chop off the edges of the histogram, producing a more even distribution with fewer outliers.

TODO: activation The *activation function* is *ReLu* (rectified linear unit) across the whole network. ReLu is a simple linear rectifier, returning zero for negative values and the original value otherwise. It is a convenient choice because doe to its linearity it does not deform the images. It is also fast to compute and widely used for convolutional networks. Since ReLu has no upper limit, the final output must be scaled down and possibly rounded to fit an image format standard, usually $0 - 255$ or $0 - 1$, in our case $0 - 255$ png.

2.2.1 The objective function and output

The objective function must return a single number which the optimizer will try to minimize or maximize. In our case this will be minimization, the tensorflow optimizers default to this. The framework being made for this thesis has some options to specify the objective function before training, so testing different losses can be automatized. A few fitting objective functions were tested, and the ones which produce high accuracy the faster were kept.

The error-calculating operation lives between the output and training operation in the graph.

For teaching with the Sobel edge map dataset, we have a continuous output and a set of continuous example images, so the objective function must take two batches of continuous matrices to calculate the error. L1 and L2 losses work with the difference of the matrices. L1 loss or mean absolute error (MAE) calculates the mean of the absolute difference matrix, L2 loss or mean squared error (MSE) calculates the average of the squared values of the difference matrix. These losses are common with regression. The reason behind squaring the values is to achieve shorter training times, because bigger errors are penalized exponentially more.

For teaching with the Canny edge map dataset, we have a continuous output, but a discrete expected output, matrices that containing either ones or zeros. L1 and L2 loss can be also used here. To account for false positives and negatives, and true positives and negatives, a modification of confusion metrics are also used, namely F1 score operates on purely discrete (positive and negative) values, and here only the expected output is discrete, this will not be true f1 loss, but it will follow the same principles. The result must also be negated, meaning subtracted from one, to get a loss instead of a score, which is proportional to the accuracy.

2.2.2 Training and error back-propagation

The training operation comes last after all the layers, right after the operation that calculates the error from the output. It computes the gradients of all the operations in the graph, and then changes the trainable variables weighted by their contribution to the error.

One of the tensorflow built-in optimizers, *Adam* is used. It features momentum-based dynamic learning rate, which promises improved convergence time. This optimizer has two built-in running variables, which change over time, resulting in smaller likelihood as time progresses during training. Because of cuDNN, as mentioned in section 1.3.1, Adam is slightly non-deterministic even when used with the same parameters and random-seed.

Training and *evaluation* steps are alternating during a training session. Training steps take up the majority of the time, while evaluation or validation steps are around ten or more times less frequent during a run.

A *step* during training consist of producing one error value from the input and output, and performing back-propagation once to modify the trainable variables, or weights. This does not mean that one step takes exactly one image, since batching is used, so one step deals with a batch of images. The processing time a resources needed for a step depends on the batch size.

With each *training step*, the error is calculated and back-propagation is performed. A training step only gets data fed from the training dataset, which is the largest dataset out of the two. For both the Sobel and Canny tasks, we use 20000 images for training and roughly 14000 images for evaluation.

An *evaluation step* stops right before calculating the error, and only produces the output image, so no back-propagation is happening here. This leaves the network untouched. The purpose of this is to measure the accuracy on the evaluation set, which is invisible to the network during training. Evaluation is also used to produce more and prettier metrics for the human observer. Evaluation runs less rarely than training, so heavier computation on metrics are possible without slowing down the training process.

2.3 Minimization

Minimization refers to reduction in the size of the FCN. This section is about the methods and tricks which make it possible to change the network mid-training, step back to an earlier state, and compare states. This way determining the minimal size for the FCN can be automatized.

After each change in the network, including all of the methods described below, additional steps are necessary. Adam, the optimizer used uses parameters that change over time. These must be reset to give the modified network a fresh start in training. Adam has accumulators, and also a variable storing the step number, which represents the passing of time. The greater these variables, the less likely that the trainable variables receive big updates, when they are reset, we are giving each network configuration equal chances.

These variables are not easily accessible, there are no helper methods provided. We must manually reach into the graph and reinitialize them.

Neglecting neurons

To simulate a graph with fewer masks, or „neurons”, the masks can be disabled any time during training. A mask is multidimensional matrix, basically a set of mask, one mask for each edge, or input that the vertex, or operation receives. Since the graph cannot be modified, we include switches into it, which tell exactly which neurons are disabled. The disabled masks always produce an output of zeros, whatever their input might be. Disabling masks provides no performance advantage, since they are still in the graph, taking up memory, but if we find that the results are still accurate with a few masks disabled, it means that the problem is solvable with fewer masks.

Freezing neurons

Freezing a mask means that the frozen mask will not receive updates via back-propagation, but will still produce an output. This is implemented similarly to disabled masks, providing switches that toggle between giving the mask zero-gradient, or retaining its value.

Reinitializing neurons

Reinitializing the masks instead of giving them zero values is also considered. The reason is that before the first run, the masks were given initial values based on normal distribution, instead of zeros, which should provide accelerated learning speed.

Stepping back

After a couple of training steps, we evaluate the accuracy, and save a snapshot of the graph to disk. This allows for resumed training later. This way we can compare thecurrent metric with earlier ones, and revert to them if we wish, for example when the accuracy had decreased since. Tensorflow provides functions to save and resume the graph structure, called meta-graph, and a checkpoint, which will store the current values of all free variables.

This is not hassle free. Tensorflow stores the absolute paths of the meta-graph and checkpoints into a plain text file name ’checkpoint’. In order to be able to move a saved model around, this file is altered by the training script to reflect the actual folders when resuming the network.

Another problem is that by default, tensorflow automatically saves every variable it can find, and constant values are incorporated into the graph. The graph cannot be touched after it was reloaded, so everything what needs changing later must be a variable. In order

to prevent the saver from incorporating the input file list into te graph, and bloating it with thousands of file paths on each resume, the input pipeline must be set up separately from the rest of the graph, which needs to be feedable with the filenames. This way the saved model can be resumed and initialized with a new set of inputs, without storing them in the graph.

2.3.1 Testing the worth of a neuron

There are a few ways to utilize freezing and disabling to determine how much a single mask contributes to the overall accuracy, and see how the network performs without it.

One disabled

One selected mask is disabled. This is useful to test the accuracy without that mask, and also observe what the optimizer will teach this mask the next time.

One disabled and frozen

The same as the above, plus the mask is frozen as well. This would provide insight to how the rest of the network will change when this mask is removed, whether it will learn to take its role, or remain unchanged.

One disabled, others frozen

The rest of the net is frozen, this will prevent the net from compensating the absence of the mask, but lets the mask relearn.

2.3.2 The minimization process

This process encapsulates multiple training sessions, and uses the tools and methods described above between them in hope of determining a minimal configuration.

First, a training session is invoked, this is the longer than the later ones, to produce a working network that is close to convergence and is capable of inference. When this happens can be measured by checking when the error stagnates. The network is then evaluated and saved to disk. Then one or more masks are *selected* and altered. Altering includes freezing and disabling masks, and there are multiple strategies for *selection*, they are described below. After this, another training session and evaluation is performed with the altered network. This is usually shorter, because the network is already semi-converged. The results of the new evaluation and the old one are compared, and if the

new one is significantly weaker, the old network is reloaded from disk. Then comes another selection, which makes sure to select different masks if the network was restored from disk. The sequence of training, evaluation, comparison and *selection* is repeated until a stopping criteria is met. We are left with the final network, and a one-hot matrix which describes which masks are to be removed to get the minimal structure.

2.3.3 Selection

Masks have to be selected, for freezing or disabling, the choice should reflect the contribution of the selected mask to the overall performance, or in other words, the least important mask should be the one that gets selected, so the network could possibly perform with one less mask.

Cosine-similarity

Cosine similarity tries to represent how similar two n-dimensional vectors are. A mask is a set of two-dimensional matrices plus scalar biases, so they are reshaped to a vector format, with the bias appended to the end of the vector. Then all the vectors are compared to each other using cosine similarity, and the end result is stored in a matrix, where the i,j element represents the cosine similarity between the the mask in the i. place and the mask in the j. place. These value range from -1 to 1, 1 is most similar (same), while negative numbers mean opposites, vectors in the opposite direction. This is done for each layer, producing a matrix for each layer.

Using the matrices, selection could be done by selecting the mask which is most similar to the others, so it has the highest average similarity, or the one which has the maximal single similarity. Absolute value could be used on the similarity values, sine negative similarity could still mean similarity in the FCN.

An extension to this method is to use the second, or n-th maximal mask after the previous one failed and produced a significant decrease in performance.

Naive selection strategies

Sequential This method simply goes through every mask in a sequential order. This is mostly useful for comparison for smarter strategies.

Random Randomly picks a mask. Also useful for comparison, but it also could be a viable strategy.

Chapter 3

Results

3.1 Metrics

In addition to the objective function, or error, which is a metric itself the following metrics were used to judge the performance. These were used in the validation steps, or separately from training to benchmark the network.

Universal Image Quality Index

This metric is used to rank the similarity of two images. It is produced from three components, measuring the correlation, luminance difference and contrast difference between the two input images. It works by splitting the images into sections with a chosen size, and calculating the metric on those. We calculate an average of these to get a single value. When used alongside the simpler MAE and f1 score, this could provide an alternative metric and view on accuracy. A matlab implementation was provided by the authors, which was then converted to python so it can be integrated into the training process.

Visualization

Visual representation of the masks, and the input images it travels through the stages of the FCN are printed as pictures to disk. This way the inner workings of the net are somewhat observable. Masks, and the image apart from the final output can include negative values. The negative values are red, and the positive ones are green, and their intensity represents the absolute value.

Chapter 4

Conclusion

4.1 Successful strategies

Chapter 5

Appendix

Statement

Alulírott ..... szakos hallgató, kijelentem, hogy a dolgozatomat a Szegedi Tudományegyetem, Informatikai Intézet ..... Tanszékén készítettem, ... diploma megszerzése érdekében.

Kijelentem, hogy a dolgozatot más szakon korábban nem védtem meg, saját munkám eredménye, és csak a hivatkozott forrásokat (szakirodalom, eszközök, stb.) használtam fel.

Tudomásul veszem, hogy szakdolgozatomat / diplomamunkámat a Szegedi Tudományegyetem Informatikai Intézet könyvtárában, a helyben olvasható könyvek között helyezik el.

Szeged, October 7, 2018 .....

aláírás

Acknowledgments

Ezúton szeretnék köszönetet mondani *X. Y-nak* ezért és ezért …