M.Sc. Artificial Intelligence (Extended) Research Project Agreement

PERSONAL INFORMATION STUDENT

Family name : Kimman

First name : Nils

Student number : s1007368

E-mail address : nilskimman@xs4all.nl

Course code / ECs : MKI94 (45 EC) (please select one)

Note: unless otherwise decided by the exam board, students that started the programme in the study year 2017-2018 take MKI92 or MKI94; students started before take MKI92 or MKI93.

AI specialisation: (please select one)
(2019-later) Cognitive Computing

Supervision

Project type : internal (please select one)

Note: internal projects are supervised and assessed by AI staff members. Affiliated projects are supervised by an affiliated staff member (e.g., DCC, CS, MPI, CLS etc) and assessed by an AI staff member. External (non-affiliated) internships are supervised by both an external supervisor and an AI staff member (mostly within a thesis circle).

Details of Supervisor 1 (Internal supervisor and/or assessor – must be AI staff member)

Name :Umut Güçlü

Details of Supervisor 2 (Internal / Affiliated / External supervisors)

Name :Thirza Dado?_____

Institute :Radboud/Donders?_____

Telephone :_____

E-mail address :_____

(*Add if there are more supervisors involved*)

Details of hosting institute

Name :Radboud University + Donders

Address :_____

Contact person :_____

Telephone :_____

E-mail address :_____

PROJECT DESCRIPTION

1. TITLE OF PROJECT: SPATIAL ENCODING MODEL AFRT TO VISUALISE NEURON RECEPTIVE FIELDS

2. ABSTRACT:

3. PROJECT DESCRIPTION: (MAX. 2000 WORDS) *1359*

BACKGROUND OF PROJECT (THEORY, STATE OF THE ART)

Spatial transformer networks [1] are a type of network that learn spatial transformations of the input data. As the name suggests, they learn spatial (affine) transformations efficiently, something classical convolutional networks struggle with. The obvious benefit of this is in real world applications, where image classification often fails due to inputs where the viewpoint will not always be perfect. Being able to still recognise an object if it is viewed from a slightly different angle is essential for any realistic application. A spatial transformer is actually more like a module that can be included in a typical neural network architecture. The way they work is by transforming the input into a canonical, or expected input, which the neural network *can* recognise.
This all sounds very similar to how humans perceive objects: spatially manipulating an object typically has little effect on how we perceive it. This means our brains must have a similar mechanism, and this is where the notion of receptive fields comes in. Different neurons activate for different stimuli, which means information about the spatial orientation of objects is encapsulated in the brain [4]. This turns into invariance under spatial manipulations, and is thus the direct thing we are modelling. This also means there is a direct parallel between the artificial and the natural, which can be beneficial in either direction.
The type of network we use is an Affine Feature Response Transform (AFRT), which is based on spatial transformation networks. It consists of three components: an affine layer, a feature model, and a linear response layer. The affine layer provides the spatial flexibility required to model the receptive fields. The features are then extracted using Alexnet [3], and the brain response is predicted using a single learnt linear layer.

Spatial transformers typically also consist of three components, although we only use two of them. Firstly, a localisation network learns the spatial transformation parameters based on the feature map. In our case this is an affine transformation for a 2D image, meaning we have 6 parameters. Then the image is transformed onto a grid using a grid generator, according to the given transformation parameters. And finally, the grid points are sampled to create the output map; important to note is that this operation is (sub)differentiable, and very efficient on a GPU. For our purposes, we only use the grid generator and sampler, because we have labelled data we can train the network on. However, the affine parameters only have an indirect effect on the end result (the individual voxel responses), meaning training may be difficult.

The feature model we use is Alexnet, which is a deep convolutional neural network used for image classification. It consists of multiple convolutional layers, some of which are followed by pooling layers. The final two layers are fully connected layers where dropout is applied ($p = 0.5$). The only activation function used is the rectified linear unit, on all the convolutional and fully connected layers. The important part is that this model performs very well and the parameters are available online, meaning training this network is already done, which saves a lot of time. We use this model to extract features from the images,

and determine the brain response via a single fully connected layer from there. Note that this model is pre-trained and that AFRT only learns parameters for the affine layer and the final response layer. This is important because it drastically reduces the number of parameters that need to be learned. The affine layer has 6 parameters, XY translation, scale, and rotation, and the response layer has a different number of parameters based on what layer we want to analyse; this ranges from 65 to 385.

There is a similarity between biological neurons and the artificial neurons found in neural networks. It has been shown that the receptive fields of both gradually increase in complexity for e.g. visual processing [2]. This means we can theoretically analyse the receptive fields of the neurons in an artificial neural network to learn something about the receptive fields of biological neurons. The receptive field of a visual neuron specifies the part of the visual field as well as possible transformations for which that neuron activates. For example, some neurons only look at the bottom left corner of an image, or others only look at straight lines. Since this coincides with the function of our Affine layer, the neurons in this layer (or more specifically the affine parameters in this layer) can provide us with insights about the receptive fields of biological neurons

Aim of the project (research question, motivation, impact, importance)

Our main goal is to analyse the receptive fields of different neurons in Alexnet, and generalise that to receptive fields in different visual areas in the brain. Because there is such a parallel between the brain and deep (convolutional) neural networks, it is likely that this generalisation is valid to at least some degree. And even if such a generalisation is incorrect, this will give us more information on how biological and artificial neural networks differ. On top of that, learning the receptive fields in neurons in Alexnet can already help develop new CNN architectures which can be more effective.
However, if we learn the receptive fields of biological neurons, this can tell us much more about specific visual disorders and how to treat them effectively. The most obvious example is visual prostheses and how to improve stimulation in the visual cortex.

Project plan (approach, methods, design, analyses)

The basic premise is training the AFRT model parameters for a specific Alexnet layer. The parameters of interest then are of course the affine transformations for specific samples. However, these have to be learned from data that is put through a feature model and a dense layer (which has to be learned synchronously), meaning learning is not as straightforward. This means training is likely very strict and figuring out how to train these components requires going in steps, increasing the complexity gradually.

4. Schedule: (max. 1 page)

Specifically state the start and end dates of the various phases of the project

01/09/2022 - 01/11/2022 - training AFRT on simulation data on a single layer
01/11/2022 - 01/01/2022 - implement real data and train again
01/01/2022 - 01/02/2022 - generalise to different Alexnet layers

01/02/2022 - 01/04/2022 - analyse results and write report

5. Scientific, societal and/or technological relevance: (About 250 words)

Describe here the broader context and relEvance of your project

Alexnet is an often used network for image classification tasks, so learning more about how it works is always beneficial. On top of that, learning the receptive fields in Alexnet can help develop new CNN architectures. These architectures can perhaps achieve better performance in certain situations, or provide further insight into neural networks.
Learning more about biological neuron receptive fields is integral to improving certain visual prostheses. Stimulating the visual area in the case of damaged eyes (the most common visual impairment) can generate phosphenes on the visual field, which can be used to create low resolution binary pictures that detect edges and contours. Knowledge of how visual neurons function can aid this type of research majorly.

6. References:

[1]: Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. Advances in neural information processing systems, 28.
[2]: Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience, 35(27), 10005-10014.
[3]: Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.
[4]: Duhamel, J. R., Bremmer, F., Ben Hamed, S., & Graf, W. (1997). Spatial invariance of visual receptive fields in parietal cortex neurons. Nature, 389(6653), 845-848.

7. APPENDIX: INTERNSHIP CONTRACT (if applicable)

aDD HERE ALL FORMAL AGREEMENTS MADE WITH THE host instiute WITH RESPECT TO ALLOWANCES (IF ANY), FACILITIES, CONFIDENTIALLY ETC.

Checks

- € The research proposal was discussed with, and agreed upon by, the master thesis coordinator and the internal supervisor/assessor (AI staff member).

- € At the start of the research project, I (will) have obtained at least 48 EC of course credits in the M.Sc. AI programme.

- € If appropriate, I have discussed the *Checklist for External Projects* with my external and internal supervisor and an internship contract has been agreed upon, checked by the master thesis coordinator, and added as an appendix to the proposal.

Signatures

These signatures confirm the accuracy of all statements made on this form and agree to all principles and articles as stated in the "Rules and Regulations MSc Internship / Research project in Artificial Intelligence" (the most recent document can be found on Brightspace)

Student

Name Date Signature

Internal supervisor (formal assessor)

Name Date Signature

Second internal/affiliated/external supervisor (add more if applicable)

Name Date Signature