Cross-Modal Convolutional Neural Network

Image and Scene Recognition

Marius Marten Kästingschäfer
Maastricht University - Faculty of Psychology and Neuroscience

First Draft - September 2018
A Project supervised by Alexander Kroner and Mario Senden

1 Introduction

Since the AlexNet in 2012 won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) convolutional neural network are state of the art in image processing. Current developments yield in the direction of bigger, deeper and models able to label an rising amount of data. Goal: general model ... take also clip art and .

The project is about training, testing and comparing a single cross-modal network. The network is compared with multiple baselines consisting of a pretrained network that is unadjusted (sequential) and another multi-input model. The pretrained unadjusted network is gonna be the AlexNet due to simplicity of the network [3] or the VGG

The accuracy's of different networks on the Place365 image set (consisting of natural images only) are shown in Figure 1 from [4].

	Validation Set of Places365		Test Set of Places365	
	Top-1 acc.	Top-5 acc.	Top-1 acc.	Top-5 acc.
Places365-AlexNet	53.17%	82.89%	53.31%	82.75%
Places365-GoogLeNet	53.63%	83.88%	53.59%	84.01%
Places365-VGG	**55.24%**	84.91%	**55.19%**	85.01%
Places365-ResNet	54.74%	**85.08%**	54.65%	**85.07%**

Figure 1: Comparison of different networks on Places365 (only natural images)

The multi-input network baseline is coming from [1]. They use ...

Questions: A. How well does an CNN trained on natural images only performance on clip art and drawn images? B. Is it possible to train a sequential model as accurate as a multi-input model with modal specific CNNs? C.

2 Procedure

Data acquiring (http://projects.csail.mit.edu/cmplaces/download.html)

Examples of dataset are shown in 2

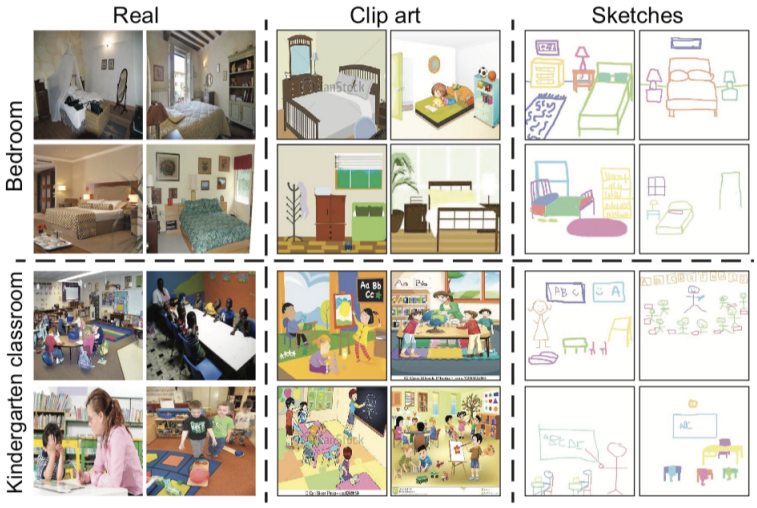

Figure 2: Examples of the cross-modal dataset showing natural images, clip art and sketches or drawings. Source [1]

2.1 Migrate weights into Tensorflow framework

Convert into Tensorflow and Keras The migration from Torch to Tensorflor is done with MMdnn a tool from Microsoft to inter-operate among different deep learning frameworks (https://github.com/Microsoft/MMdnn)

Compute untrained baseline (on the rest of the images) and Split data.

2.2 Using a Pretrained convnet

Why use pretrained networks? (computationally expensive, good feature detectors,..)

Freeze network and adjust layers partly

Possible methods: feature extraction or fine-tuning. The difference is shown in 3.

During **fine-tuning** all weights would be changed during training on the (partly) new task.

During **feature extraction** only the weights of the newly added last layer would change.

(- show AlexNet and adjustments it) (use data augmentation?)

Trying out on local system, Possible to train network and pause it (save adjusted weights after each epoch) https://www.kaggle.com/morenoh149/keras-continue-training

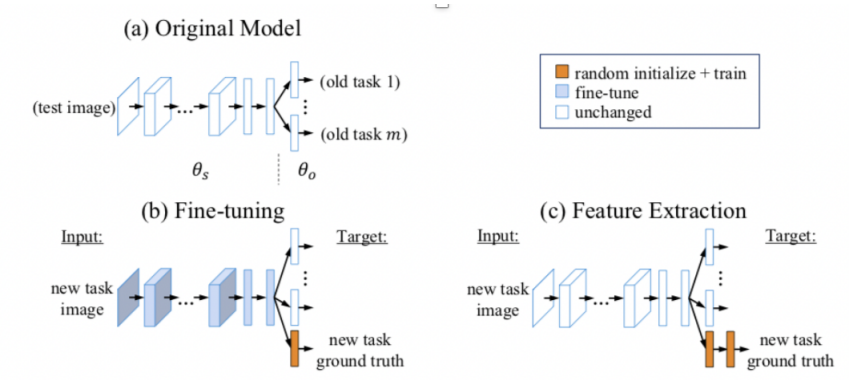

Figure 3: Differences in using pretrained models. Source [2]

Do final calculations on cloud computer (https://console.cloud.google.com/compute/) (Dense layers learn global patterns in their input feature space; whereas convolution layers learn local patterns.)

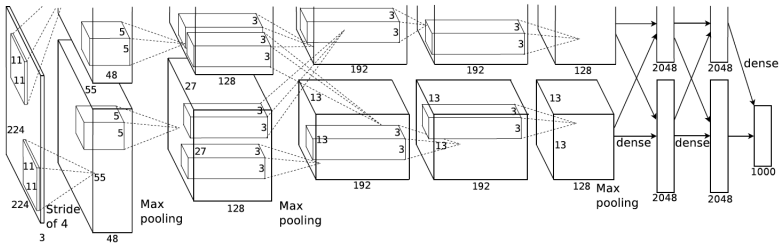

Figure 4: Architecture of AlexNet. Source [3]

2.3 Compare networks

Compare trained network with baseline results (crossentropy more usefuly than accuracy since it takes the distance to goal into account)

3 Optional Procedure

The Marble dates and dealines are the following:

1. Research data collection until January 2019
2. Writing emprirical bachelor thesis until April 2019
3. Deadline submission Intro+Methods Thesis February 2019
4. Deadline submission complete thesis mid April 2019
5. Poster presentations annual Marble conference July 2019

3.1 Visualization

If possible visualizations of the project will be done. How do intermedia actions look like? How does generalization changes with improvement of accruacy? How does abstraction increases witin the network? (show examples -differences in accuracy / headmap) How does differences in arruracy change the representation within the layers.

(https://distill.pub/2017/feature-visualization/)

3.2 Biological processes

Further it could be from additional value to have a look at biological processing underlying vision (object detection, shape and texture differentiation, finding edges, discriminate figure from ground, , color, resolving interference and combining multiple cues especially during motion),

References

[1] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. 2016.

[2] François Chollet. *Deep Learning with Python.* Manning Publications, 2017.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. neural information processing systems. 2012.

[4] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.