Imperial College London

MEng Individual Project

A Comparison of Features Learned on Natural Vs Drawn Image Datasets

WANG BIYUAN 00861345

Department of Bioengineering
Dr. Anil Bharath
Kai Arulkumaran

Submitted in partial fulfilment of the requirements for the Award of the MEng in Biomedical Engineering and the Diploma of Imperial College.

January 9, 2017 Word Count: 2442

Contents

1 Project Specification

This project aims to make a comparison between the image features learned on natural and drawn image datasets qualitatively. In the past few years, there are a lot of researches have shown that convolutional neural network can deliver prominent performance on image understanding tasks especially visual classification [1]. However, our understanding of image representations learned by these networks has been limited to natural images so far. There is still not enough insight into the internal operations of these convolutional neural networks (CNN) models as well as how and what they learn from the datasets in order to perform related tasks.

One of the explicit ways to interpret image representations is the implementation of visualisation techniques. Human-comprehensible interpretation of the behavior of a deep CNN model can be achieved by inverting them back to pixel space. This project is to study and visualise the features learned from different layers of a deep CNN model, and to compare the features learned by networks of similar architecture but trained on two types of data of great contrasts in style, natural images and illustrations. The results obtained from the visualisation are expected to help people explore more details about the deep CNN model in computer vision and therefore encourage efficient designs of network to perform more complex tasks.

This project will be carried in Torch7 in Ubuntu operating system, which is a scientific computing framework with wide support for machine learning algorithms [2].

2 Ethical Analysis

In the long term the project aims to enhance the understanding of the CNN networks learning on non-natural images. This would have a positive impact on advanced researches of either hand written or drawn work as the more we understand about the computer vision representations learned by the network, the more likely to design effective networks with less dependent on an empirical point of view.

The main aspect in this project that may turn into a subject of ethical debate is the potential misuse of the data collected for the model training.

To facilitate deep learning, large amount of natural images and illustrations were collected for the purpose of training CNN models. The database could include images that are not suitable for work. Inappropriate manipulation of the database can cause harm or offence.

All the images selected as database should receive consent for natural images if the images are referred to identifiable individuals. The consent is also expected to receive from authors if illustrations are taken due to the copyright and intellectual reasons.

Researchers can easily hide from public view while using the image database [3]. It is important for researchers to assure the information are only for academic use. Finally, as the results of this project will involve a series of images, which are expected obtained by visualising the image representations, Their authenticity should be guaranteed.

3 Preparation Work and Literature Review

Basic background knowledge of machine learning and deep learning is essential for this project. Two open courses and several requisite supplementary reading material have been mainly gone through as preparation work.

Stanford University Open Course: Machine Learning/ Lecturer: Andrew Ng [4] Professor Andrew Ng from Stanford University gave a series of lectures to help novice perceive the basic concepts and ideas from the field of machine learning. It mainly introduced the general forms of linear and logistic regression problems under supervised learning. Moreover, loss evaluation and optimisation algorithms such as cost function and gradient descent as well as different

regularisers were also mentioned. These terms are supposed to improve the learning quality so that the model could give prediction with high accuracy. In the latter lectures, Professor Ng also drew the topic into neural networks. This model was inspired by human brains. It simulates biological neurons by setting up unit modules in a model which response to different simulations.

Oxford University Course: Machine Learning/ Lecturer: Nando de Freitas [5] Neural network has improved the prediction results a lot. However, for complex polynomial problems, a more advanced network is required. Lectures given Professor Nando De Freitas talked about machine learning from the field of deep neural networks. More detailed optimisers and the architecture of a model were elaborated. It also came along with a series of practicals which allows to implement a complete model and check the performance on classification task.

Several papers related to visualisation of networks trained by natural images and a network trained by illustrations were reviewed as an indication to start my project. Further more, two papers related to style transfer were also reviewed as the background of my further work.

Mahendran A, Vedaldi A. Understanding deep image representations by inverting them[C]//2015 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, 2015: 5188-5196. [6] It is known that two totally different images could share the same representations. An encoding of an image should not be uniquely invertible and it is possible to have reconstructions of large difference from these representations. Pre-images are the notable representations in images, which are the crucial elements used in reconstruction. When two irrelevant images have common representations, not all pre-images are useful for image reconstruction. They proposed to restrict the pre-images used for reconstruction to the set of natural images. In this paper from Vedaldi's group, they used different regulariser terms as natural image priors to make sure the reconstructions have the same statistics as target images. Then direct analysis of representations was carried out by charaterising the image information they retained compared with the target image. More details about invariances captured by the representation can be evaluated by reconstructing a number of possible samples and making the comparison as before. They also evaluated on different regulariSation penalties as natural image priors in order to obtain more accurate reconstruction.

Mahendran A, Vedaldi A. Visualizing deep convolutional neural networks using natural pre-images[J]. International Journal of Computer Vision, 2016: 1-23. [1] Vedaldi's group pointed out that due to most of the image processing methods are based on suitable image representations rather than on image itself, there is still lack of understanding of image representations. They proposed three types of visualisations under the natural pre-images framework to explore the interpretation of representations. Natural pre-images is the notable representations in images from real world. This is used to constraint the visualisation to sensible natural images. In their experiments, they tuned regularisers and optimisation algorithms to investigate the effects on each layer induced by different parameters.
The first visualisation method is **Inversion**. In this method, the pre-images of a target image are found by passing through a deep CNN model. New images are reconstructed according to those obtained pre-images and evaluated by loss function to access reconstruction quality. The second and third methods **Activation maximization** and **Caricaturization** regard the output of a scoring function as a weight to select the representations which should be maximally activated. **Activation maximization** sets the one-hot indicator vector to show the representation being visualised. Then the same objective function and loss function are applied. The maximally activated neuron output is visualised. **Caricaturization** are similar to **Activation maximization** in principle but have different normalisation factors and optimisation methods
.

Saito M, Matsui Y. Illustration2Vec: a semantic vector representation of illustrations[C]//SIGGRAPH Asia 2015 Technical Briefs. ACM, 2015: 5. [7] This is a paper about networks trained by illustrations. The authors trained a deep CNN model to extract attributes of a single illustration, and mapped them to a semantic vector space so that people are able to search images with similar features. This CNN model is different from others as it is supposed to predict multiple tags in a single image, which means more capable of precisely estimating local information. It was modified from a VGG model A in [8] and combined with a NIN model, more concretely as shown in Fig 1 to replace the fully connected layers by convolutional layers in order to get more local details.

VGG (model A, 11 layers)	VGG + NIN model
input ($3 \times 224 \times 224$)	
conv + max-pooling layers	
FC-4096 FC-4096 FC-1539	conv3-1024 conv3-1024 conv3-1539
sigmoid layer	average-pooling layer sigmoid layer

Figure 1: Modification on VGG model A by replacing the fully connected layers by convolutional layers to extract more local information. The final modified architecture is a combination of VGG model A [8] and NIN model.

Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]// European Conference on Computer Vision. Springer International Publishing, 2014: 818-833. [9] The configuration of large convolutional neural network models and general visualisation techniques are studied in this paper. The techniques of **deconvnet** was highlighted. The feature visualisation was obtained by means of backpropagation of the output from each layer and then projected them back to pixel space to reveal the activated part on feature map.

Gatys L A, Ecker A S, Bethge M. A neural algorithm of artistic style[J]. arXiv preprint arXiv:1508.06576, 2015. [10] It was found that deep neural networks can artificially create artistic effects on images by using neural representations to separate and recombine the contents and styles of arbitrary images. This paper offers an algorithmic understanding of how this artistic effects transfer between images. They also implemented the VGG-Network and built a style representation to find the feature correlations and used to generate the new texture.

Gatys L A, Ecker A S, Bethge M, et al. Controlling Perceptual Factors in Neural Style Transfer[J]. arXiv preprint arXiv:1611.07865, 2016. [11] As the two Vedaldi's papers have mentioned that a pre-image is found simultaneously reproducing features and statics of content and style images. More advanced algorithms used in style transfer were introduced in this paper. They introduced control over spatial location, colour information and spatial scale and decomposed style into these three factors. Such techniques can be used to generate different images that share common deep features, which could help the understanding of the invariance properties of the representations.

4 Implementation Plan

My implementation plan and timeline of this project is shown in details in Fig 3.

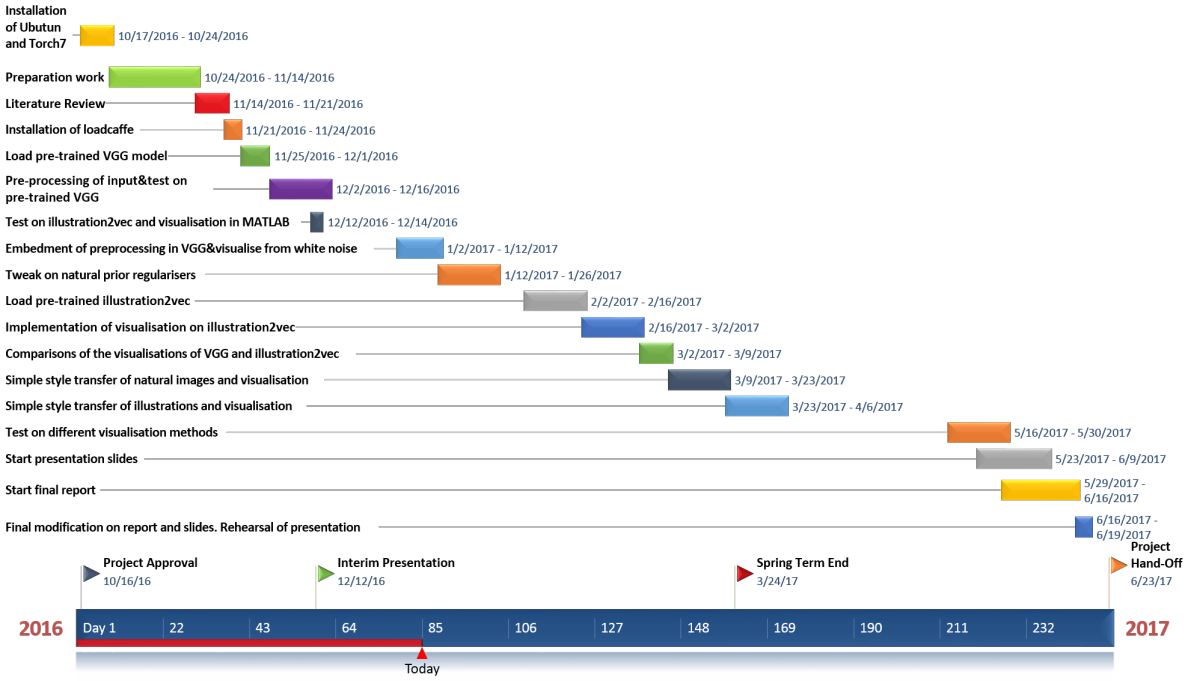

Figure 2: Gantt chart

5 Preliminary Results

5.1 Caffe Model Load

In this project, I would take advantage of the VGG models pretrained on ImageNet datasets to realise the visualisation in Torch. Due to the VGG models are constructed in Caffe framework, it takes several steps to install protocol buffer so that allows Caffe networks to load in Torch7. considering the computational speed, at this early stage of the project I selected a relatively small model, *VGG_CNN_S*, which is a network of 8 convolutional and fully connected layers. The network architecture is outlined in Table 1.

Name	Size	Stride
conv1	$96 \times 7 \times 7$	2
$\times 3$ pool	$96 \times 7 \times 7$	2
conv2	$256 \times 5 \times 5$	1
$\times 2$ pool	$256 \times 5 \times 5$	1
conv3	$512 \times 3 \times 3$	1
conv4	$512 \times 3 \times 3$	1
conv5	$512 \times 3 \times 3$	1
$\times 3$ pool	$512 \times 3 \times 3$	1
fc6	4069	dropout
fc7	4096	dropout
fc8	1000	softmax

Table 1: Layer configuration of VGG_CNN_S model [12]

5.2 Preprocessing of Input Images

Images are firstly rescaled so that the smallest side is 256 while preserving the aspect ratio. As specified in the prototxt file of *VGG_CNN_S*, input images to the model must satisfy the dimensions of $3 \times 224 \times 224$ (Colour channel×Length×Width) as the model is trained on 224x224 centre crops sampled from images. The released mean BGR image should be also subtracted from the 224x224 crops to normalise color channel values to center around the means, so that it will be in the range of $[-125, 130]$ rather than the range of $[0, 255]$. The pixels loaded in Torch is ranged from 0 to 1, these numbers are supposed to be rescaled to 0 to 255 which is required by the VGG model. Finally, due to the OpvenCV convention RGB colour channels should be swapped to BGR channels.

In general, the preprocessing is concluded in five steps:

- Rescale the smallest side of an input image to 256 while preserving the aspect ratio
- Crop image to 3×224×224
- Rescale pixels from 0-1 to 0-255
- Subtract image mean
- Change colour channel from RGB to BGR

5.3 Classification

Preprocessed images were passed through the network. The output of the model is a tensor that contains predicted log probabilities of 1000 classes. The higher the value is, the more likely that input image belongs to the corresponding class. To access the performance on classification task of this model, I tested with five images including three living creatures (dogs and cats) and two non-living objects (bucket and truck) and found the indices of the highest three probabilities and mapped them back to the class label file. The prediction results are shown in Table 2 and Fig 3.

Picture	Predicted Class	Log Probability
1	tub, vat	0.294
	cup	0.103
	Siamese cat, Siamese	0.093
2	Egyptian cat	0.543
	tabby, tabby cat	0.087
	tiger cat	0.069
3	golden retriever	0.957
	Labrador retriever	0.031
	kuvasz	0.004
4	trailer truck	0.938
	moving van	0.061
	paasenger car, coach, carriage	0.001
5	bucket, pail	0.725
	measuring cup	0.127
	cocktail shaker	0.063

Table 2: Image classification test on VGG_CNN_S model [13]. Five pictures were tested and the highest three possible labels were listed with corresponding probabilities.

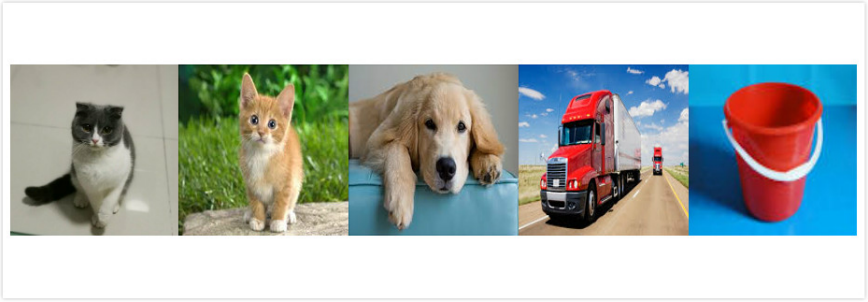

Figure 3: Tested images: from left to right corresponding to picture 1 to 5.

The results have been proved the outstanding performance of a deep CNN model in image recognition. The prediction can be made not only on the categories of input images, but also very specific species of that object even with a background interference. However there is still error in the prediction as shown in Table 2, the tag with the highest probability of picture 1 is '*tub*' rather than a cat category.

5.4 Test on *illustration2vec model* Demo

This is a test of the *illu2vec demo* [7], which involves the implementation of the model trained on illustrations. I tested with one of my illustrations on this *illustration2vec* model demo to access what it can capture from hand-drawn art. The tested illustration and tag prediction was shown in Fig 4 and Table 3. The result has shown that this model predicted fairly accurate tags on the sketch. It is notable that the model gave a precise prediction tag of school uniforms, which could be a widely various element in illustrations. Moreover, as this character was sketched in her common look, the model even predicted its copyrights with a very high confidence.

Tag Type	Tag	Confidence
General	rown hair	91.1%
	1 girl	88.4%
	brown eyes	76.2%
	short hair	51.5%
	school uniform	33.0%
	short	27.9%
Character	Misaka Mikoto	75.0%
Copyright	To Aru Majutsu no Index	97.4%
	To Aru Kagaku no Railgun	76.5%
Rating	Safe	98.7%
	Questionable	1.05%

Table 3: Test on *illustration2vec* model demo [14]

Figure 4: Tested illustration

5.5 Retrieve of Visualisation of Layers in Caffe-AlexNet Model

In Vedaldis first paper ***Understanding deep image representations by inverting them*** [6], reconstruction from representations captured by a Caffe-AlexNet model was realised. The whole project was implemented in MATLAB and relied on the *matconvnet* and *vlfeat* packages. I retrieved their code in MATLAB and tried to reproduce the visualisation of each layer's output. Fig 5 shows a complete reconstruction of a natural image. From the picture it can be observed that each layer of CNN captured different types of structure in the image, from instance-specific information to large variations in layouts with the increase of depth sequentially.

Figure 5: Visualisation of Layers in Caffe-Alex Model

5.6 Embedment of Preprocessing

To further improve the implemented code at natural image reconstruction stage, I planned to embed the preprocessing of input image in the model architecture. As a result, the configuration of a modified model is shown in Table 4

input		
Mul Constant 255		
Rescale smallest side to 256 with aspect ratio preserved		
Crop 224×224		
Select colour channel		
R submodel	G submodel	B submodel
Subtract R mean	Subtract G mean	Subtract B mean
Join submodels in BGR order		
VGG_CNN_S		
Output		

Table 4: Preprocessing embedded model

5.7 Preliminary Visualisation from White Noise

To have a generic understanding in visualisation, I started it with a white noise image passing through the *VGG_CNN_S* model and visualised the output of one randomly selected filter in the network. After several iterations, it can be observed that an arbitrary feature of the white noise image itself was extracted by the selected filter and had been gradually visualised as the iteration increased. Fig 6 shows the visualisation in gradual change.

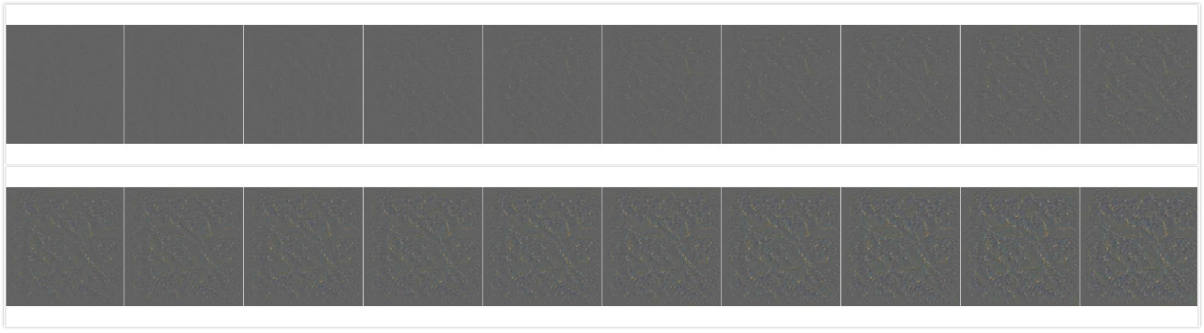

Figure 6: Visualisation of a white noise image. Filter Index=110. Iterations=20.

6 Evaluation

At the end of the project I expect to have the visualisations of image representations captured by natural-image-trained and illustration-trained network respectively. The point of evaluation is to appraise what are the features learned by a network trained on non-natural datasets and how they are different from those of natural images, by the means of making comparison of them qualitatively. For the same type of objects depicted in the image, features learned by two networks with similar architectures but utterly different training data will be assessed to investigate their distinct properties as well as invariance.

For more advanced evaluation on the invariance properties of these very distinct image features, style transfer would be implemented to obtain even more different image styles and compare with the original ones.

References

[1] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, pages 1–23, 2016.

[2] Torch — scientific computing for luajit. `http://torch.ch/`. (Accessed on 01/07/2017).

[3] Rose Wiles, Jon Prosser, Anna Bagnoli, Andrew Clark, Katherine Davies, Sally Holland, and Emma Renold. Visual ethics: Ethical issues in visual research. 2008.

[4] Machine learning - stanford university — coursera. `https://www.coursera.org/learn/machine-learning`. (Accessed on 01/07/2017).

[5] Machine learning. `https://www.cs.ox.ac.uk/people/nando.defreitas/machinelearning/`. (Accessed on 01/07/2017).

[6] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5188–5196. IEEE, 2015.

[7] Masaki Saito and Yusuke Matsui. Illustration2vec: A semantic vector representation of illustrations. In *SIGGRAPH Asia 2015 Technical Briefs*, SA '15, pages 5:1–5:4, New York, NY, USA, 2015. ACM.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[10] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[11] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. *arXiv preprint arXiv:1611.07865*, 2016.

[12] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *CoRR*, abs/1405.3531, 2014.

[13] Cnn_s model from the bmvc-2014 paper "return of the devil in the details: Delving deep into convolutional nets". `https://gist.github.com/ksimonyan/fd8800eeb36e276cd6f9`. (Accessed on 01/07/2017).

[14] Illustration2vec demo. `http://demo.illustration2vec.net/`. (Accessed on 01/07/2017).