Part III Physics Research Project Proposal

Probing the Variance of the Information Phase Transition of Deep Neural Networks for Interpretable System Architectures

James Bernardi
Candidate Number:
Signature:

Supervisor: Professor Pietro Lio'
Helena Andres Terra, Ben Day
Signature:

Department of Computer Science
University of Cambridge
2018

Summary

The Physics of the flow of Information through a Deep Neural Network (DNN) closely mirrors that found in systems from Ideal Gas statistics to the field of Quantum Information. This project seeks to develop a more complete, theoretical understanding of DNNs' learning and data-compression phases when DNNs have the potential to revolutionise fields such as Physics, Medicine, and indeed wider society.

Specifically, this project will apply the Physics of Information Theory to a series of interpretable DNNs to seek insight into the dynamics of the training regime, which otherwise exists as a 'black box'. Several system architectures specifically designed to inject physicality into the networks' decision making processes will be examined to extract meaning from the comparison of the Information Flows during their training regime.

A particular mechanism of interest is the phase transition point between phases in the training regime known as the *drift* and *information compression* phases. The speed and nature by which the system progresses through each phase will allow us to extract meaning from how Networks respond to adjustments made to the representations of the data that the network is training on.

Results from this project may lead to informing future system architectures, for example if one representation emerges as preferential to others in terms of having a faster training process that retains more Information than other representations.

1 Introduction

1.1 Deep Learning

Deep Neural Networks (DNNs) are transforming the world in countless fields. Their ability to operate with higher volumes of data and inputs on much shorter timescales than human computation is already seeing them to outperform human experts in a range of classification tasks such as medical referral decisions [1] and aircraft maneuvers [2].

In general, DNNs are trained by specifying an output $\boldsymbol{y}$ for a given input $\boldsymbol{x}$ (e.g. supplying training data labelled with its outputs). Once the network is trained, it will be able to classify (e.g. provide a predicted output $\hat{\boldsymbol{y}}$ for) previously unseen examples, often performing better than humans. The output vector for each of n 'hidden layers', $\boldsymbol{h}_j$, is given by:

$$\boldsymbol{h}_1 = activation\left(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1\right) \tag{1}$$

$$\boldsymbol{h}_i = activation\left(\boldsymbol{W}_i\boldsymbol{h}_{i-1} + \boldsymbol{b}_i\right) \tag{2}$$

$$\hat{\boldsymbol{y}} = softmax\left(\boldsymbol{W}_n\boldsymbol{h}_{n-1} + \boldsymbol{b}_n\right) \tag{3}$$

For $1 < i < n$. The weight matrices $\boldsymbol{W}_i$ give the weights on the edges between node vectors $\boldsymbol{h}_i$, with bias vectors $\boldsymbol{b}_i$ added. The parameters $\boldsymbol{W}_i$ and $\boldsymbol{h}_i$ are learned through the process of minimising a cost function C by gradient decent, where one example (mean squared error function) is defined as:

$$C = \frac{1}{2m}\sum_{i=1}^{m}\left(\hat{y}(x^{(i)}) - y^{(i)}\right)^2 \tag{4}$$

A variety of other 'regularisation techniques' are employed to improve this process, but further details will be omitted for the preliminary report. If further details are sought on the Deep Learning techniques discussed and beyond, refer to Goodfellow et al. [3].

1.1.1 Interpretability

Despite their great performance success, there is little knowledge of the inner-workings of DNNs that lead to the output classification. This introduces the field of DNN Interpretability, which seeks to achieve model assurance by extracting the features that inform a DNN's decisions for reasons ranging from model improvement to algorithmic fairness - for example, if the DNN makes decisions based on factors a human would deem arbitrary or unfair.

1.1.2 Hierarchy

A second problem in the field of Deep Learning is introducing classification hierarchies. In many problems, an output belongs to a class of outputs; as an example: an image of a Jack Russell Terrier should output dog, but humans are able to specify *with higher confidence* that the image belongs to a set containing all dogs, which itself belongs to the set of all animals. Current DNNs perform badly at hierarchical modelling, which can make extracting physical meaning from classification problems difficult.

1.2 Information Theory

Information Theory is a long established field within probability theory dealing with data compression and transmission. The mathematics in this project closely reflect concepts found in statistical physics, given the parallels between the flow of state variables as information in a fluid. Some key results are extracted from Cover, Thomas (1991) [4] that will be particularly relevant to this project.

Mutual information

The Mutual Information between random variables X and Y with a joint distribution $p(x, y)$ is:

$$I\left(X;Y\right) = \sum_{x\in X, y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) = H(X) - H(X|Y) \tag{5}$$

Given the definition of the Shannon Entropy for random variable Z:

$$H(Z) = \sum_{z\in Z} p(Z)\log\left(p(Z)\right) \tag{6}$$

Data Processing Inequality (DPI)

Another key result is that the loss of information about an input variable X cannot be recovered in a later system-state. That is, for the Markov Chain $\boldsymbol{h}_0 \rightarrow \boldsymbol{h}_1 \rightarrow ...$ and any $i \geq j$:

$$I(X; \boldsymbol{h}_j) \geq I(X; \boldsymbol{h}_i) \tag{7}$$

1.2.1 Information Bottleneck Theory

The field of Information Theory in the context of DNNs was heavliy influenced by Tishby, Pereira, Bialek (1999) [5]. Taking this work further, Tishby, Zaslavsky (2015) [6] show that *any* DNN can be quantified by the *mutual information* between layers, and an optimal *information theoretic limit* is introduced by application of the DPI.

Mutual Information

By recalling the Mutual Information between random variables X and Y, $I(X;Y)$ (equation 5), the *relevant information* of the input X held in the output Y is defined in the context of a DNN.

Optimal Representation

An *optimal representation* of X would capture all the relevant features of the input, and dismiss the irrelevant parts that do not inform the output Y [5]. The problem of keeping a fixed amount of meaningful information about the output Y, while minimising the information from (or compressing) the input X leads to a Lagrangian minimisation problem:

$$\mathcal{L}\left[p(\hat{x}|z)\right] = I\left(X; \hat{X}\right) - \beta I\left(\hat{X}; Y\right) \tag{8}$$

Where $\hat{X}$ is the simplest mapping of X that captures the mutual information $I(X;Y)$. Variation of β trades-off between the need to preserve meaningful information about the output (the second term) and achieving optimal compression (the second term), i.e. such that only the most relevant features are kept in the network.

Learning and Generalisation Conditions

Combining this analysis with the DPI (equation 7), it is found that the most efficient learning and generalisation is achieved when each layer in a DNN attempts to maximise the mutual information between the representation held in layer i and the output Y, $I(Y; \boldsymbol{h}_i)$ while minimising $I(\boldsymbol{h}_{i-1}; \boldsymbol{h}_i)$ [6].

Information Theoretic Bound

Due to the fact that DNN training only has access to a finite sample of the input space $X \in \mathcal{X}$ (i.e. the number of labelled training examples not-in the test sample), it has been shown that there exists a guaranteed bound as follows [7]:

$$I\left(\hat{X}; Y\right) \leq \hat{I}\left(\hat{X}; Y\right) + O\left(\frac{|\hat{\mathcal{X}}||\mathcal{Y}|}{\sqrt{n}}\right) \tag{9}$$

and

$$I\left(X; \hat{X}\right) \leq \hat{I}\left(X; \hat{X}\right) + O\left(\frac{|\hat{\mathcal{X}}|}{\sqrt{n}}\right) \tag{10}$$

That is, the worst case upper bound on the mutual information depends on the cardinality of $\hat{X} \in \hat{\mathcal{X}}$ and $Y \in \mathcal{Y}$, and the number of training examples n.

2 Method

2.1 System Architectures of Interest

2.1.1 Interpretable Model

This investigation will focus on the information flow through the groundbreaking model laid out in De Fauw et al. (2018) [1]. The model acts to inject physicality into the problem by designing its architecture to specify the way in which the problem should be approached; first segmenting the tissues present and secondly making a diagnosis. The result is generalisable, and follows a logical decision making process that is interpretable to a human expert. This process (requiring two points of supervision) is exemplified in figure 1 and is summarised below:

1. *Segment* the tissues present in the image - supervised by expertly-segmented scans. Produces a tissue map as a viewable, intermediate output.
2. *Classify* the output tissue-map: predict diagnosis probabilities and optimal referral decisions based on only information retained in the tissue map. Supervised by human-expert diagnoses associated with the original image.

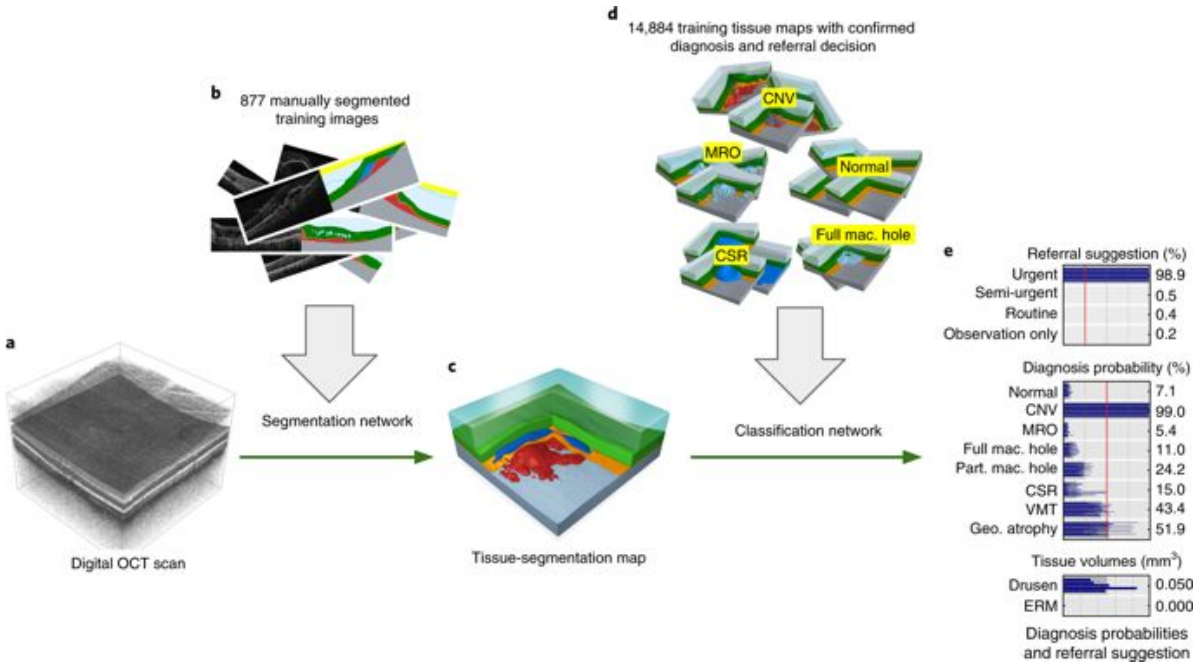


Figure 1: The flow of diagnosis from raw scan (a), segmented (b) to produce a tissue map (c), followed by classification of the tissue map (d) to output optimal referral decision, with diagnosis probability (e). Image taken from De Fauw et al. (2018) [1].

This project will introduce variants of this model to compare how the evolution of mutual information between layers in the training regime depends on the system architecture. In section 3, I outline some of the questions this project hopes to probe by taking this approach.

2.1.2 Model Variants for this Project

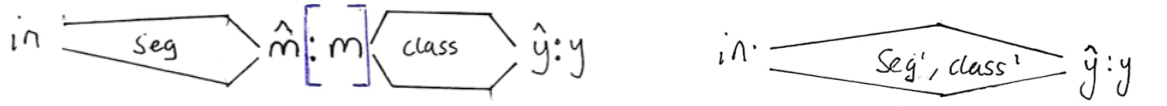
This project will investigate three specific architectures (figure 2) to make a comparison of the mutual information evolution for different forms of the *intermediate representation* (where 'intermediate' is defined to be in position c in figure 1). The first architecture will mirror that found in De Fauw et al. (2018) [1], or figure 1, as a reference point. The other two models will be variants of this model as follows.

Variant 1

The first variant will make no architectural changes to the model, but will remove the supervision on the intermediate representation. The intermediate representation will still be reduced in size, but instead be passed through the network as a hidden layer that is not penalised for deviating from the human-segmented tissue-maps. An architecturally-imposed information bottleneck (section 1.2.1) still exists in the network by specifying the form of the intermediate representation.

Variant 2

The second variant will change the architecture within the model; in addition to removing supervision on the intermediate representation, the imposed information bottleneck will be removed by maintaining a high number of features throughout the network. Instead, the network will be free to train its own information bottleneck as proven necessary in section 1.2.1.



(a) Original and Variant 1 (omit blue) (b) Variant 2

Figure 2: A graphical representation of the architectures to be used. $\hat{y} : y$ symbolises supervision of output $\hat{y}$ with the training data y. This is omitted in Variant 1 (e.g. omit blue bracket).

2.1.3 Hierarchical Model (backup / extension)

Should the above methods be quick to implement and extract insight from, I propose an extension to the project in investigating the same effects in a model designed to extract problem *hierarchy* (see section 1.1.2).

One architecture I wish to focus on specifically is represented in figure 3 [8].

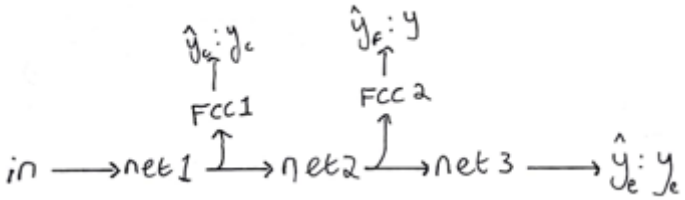


Figure 3: A DNN architecture that attempts to account for class hierarchy.

Again, removing supervision at the various FCCs in figure 3 changes the form of the intermediate representations see how performance and Mutual Information evolution is varied. Four variations would be generated by: i) leaving the network as-is, ii) removing FCC 1, iii) removing FCC 2, and iv) removing both FCCs.

2.2 Datasets

The suitability of the retinal image data to the architecture described in figure 1 will require access to that same dataset, given the time constraints. This data has been requested from Moorfields Eye Hospital, however should this request not be approved in time this experiment will resort to the hierarchical model (section 1.1.2 for which the data is already available in Cambridge.

Should the back-up be resorted to, an alternative extension plan will be devised based on other networks using data held in Cambridge [9]. This will be known before the start of the bulk of the work in mid-January (see appendix A.1 for time line).

3 Proposed Analysis

3.1 Predicted Comparative Performance of Models

The first step in comparing the information flow between network architectures (section 2.1.2) will be to establish their comparative performance. Here, references to *speed* correspond to the number of 'epochs' elapsed (or passes of the full training dataset).

It is predicted that variant 1 (e.g. no segmentation supervision) will train **faster** than the unaffected architecture due to the network's reduced emphasis on correct segmentation allowing an increased focus purely on diagnostic error reduction. However, by removing the supervision step, this variant is more susceptible to over-fitting on certain features of the input which would **reduce** the diagnostic accuracy reached.

Variant 2, which removes the architecturally-induced Information Bottleneck altogether, has high capacity to focus fully on reducing diagnostic error. Therefore, this network is likely to reach its optimal accuracy **fastest** of the three but, as we have the prior that diagnosis is best approached by firstly segmenting the input, the accuracy is likely to be **lowest**.

These predictions are based on the prior that:

$$I(Y;\hat{X}) \approx I(Y;\hat{X}')$$

Where $\hat{X}$ is the theoretical optimal representation of X (the representation immediately before reduction to the intermediate representation), and X' is the intermediate representation. In other words, it's predicted that little pertinent information will be lost in enforcing segmentation, and we hope that noise is de-emphasised in the representation reduction also (i.e. some information compression is achieved manually).

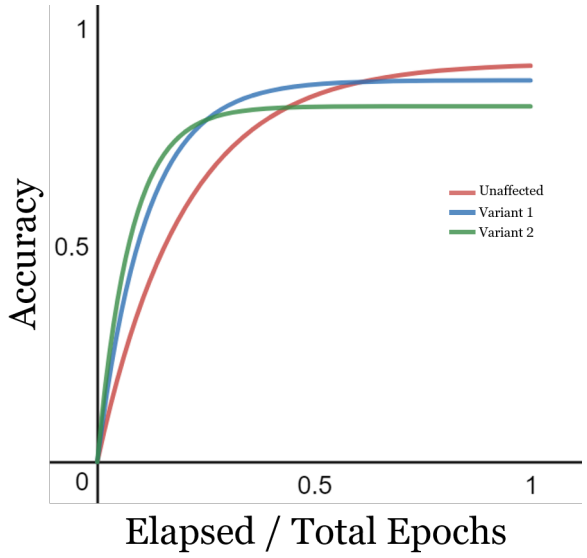


Figure 4: Predicted evolution of accuracy with fraction of total number of epochs elapsed. Predictions are made that speed of training trades-off on final accuracy for these architectures.

While these predicted performances will be interesting if proven wrong, these are not the most interesting results - the remainder of the investigation will focus on information flow, which these results will help to inform our discussion of in the final report.

3.2 Mutual Information Evolution Comparison

3.2.1 General Analysis

This project will primarily investigate the shape of the evolution of Mutual Information through training epochs for different network architectures. The classification network (see figure 1 will be focused on, thus as the intermediate representation varies, it will be the form of the input that is changed in this study (outlined in section 2.1.2).

Initially, with regard to figure 5, it will be informative to obtain the bound on the Mutual Information $I(Y;X)$ which corresponds to the information theoretic bound. Should this vary with the form of the intermediate representation, we will be informed of the appropriateness of the compression in $X \rightarrow X'$ in forming the intermediate representation. A better compression (e.g. losing less useful information in the forced Information Bottleneck) will lead to a higher bound on the Mutual Information.

3.2.2 Mapping the Phase Transition

Secondly, this study will map the point of the phase transition observed along the green line in figure 5. We will define the phase transition to be the turning point in the evolution of $I(T;Y)$ with $I(X;T)$, and is the transition between the *drift* phase - where the network is learning which features should be passed through layers T to most inform output Y, thus $I(T,Y)$ is improving most rapidly with epochs - and the *information compression* phase - where the network is compressing the representations passed through T to disregard unimportant features in X, thus $I(X;T)$ decreases.

We are primarily interested in whether the network progresses through each phase more quickly or slowly, dependent on the forced Information Bottleneck on the intermediate representation, which should inform how good the choice of each representation is for solving the problem at-hand.

Of course, the network in question is not necessarily designed to maximise the accuracy on the choice of intermediate representation, rather the interpretability of the system to human experts.

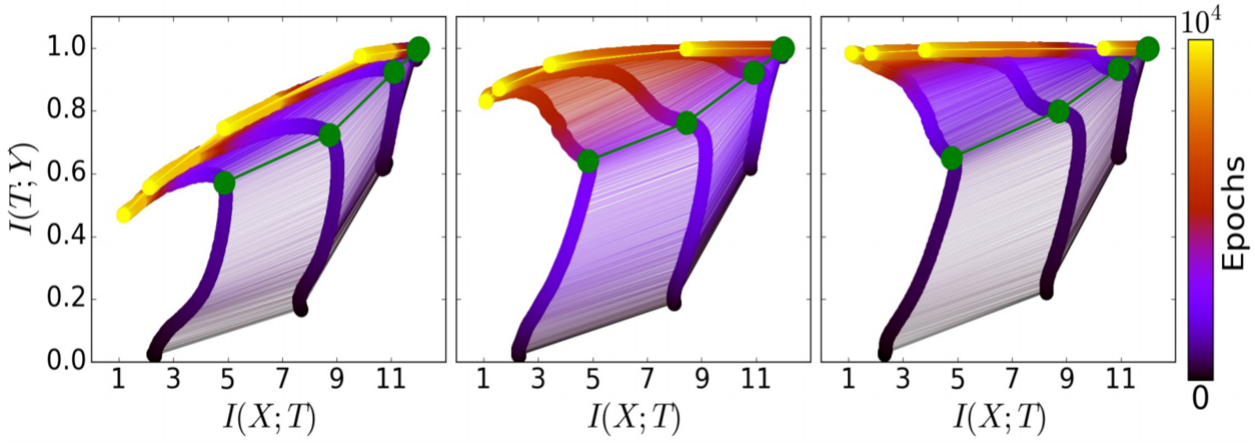


Figure 5: The evolution of Mutual Information held between input X and output Y with hidden layers T, for different samples of the training data, from 5%, 45% and 85%, from left to right. The colors indicate the number of training epochs (complete passes through the training data) with Stochastic Gradient Descent, from 0 to 10000. Example taken from Schwartz-Ziv, Tishby (2017) [10]

3.2.3 Verifying Previous Results

Finally, there has been some controversy over the generality of the Schwartz-Ziv, Tishby (2017) results. Saxe et al. [11] argue that the information flow conclusions made do not apply to all network architectures, so a final interesting conclusion will be to compare the results of this investigation with the points Saxe et al. raise.

References

[1] Jeffrey De Fauw and et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24:1342, 1350, 2018. https://doi.org/10.1038/s41591-018-0107-6.

[2] Ervin T. Rodin and S. Massoud Amin. Maneuver prediction in air combat via artificial neural networks. *Computers and Mathematics with Applications*, 24(3):95 – 112, 1992. https://doi.org/10.1016/0898-1221(92)90217-6.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[4] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006.

[5] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, 49, 07 1999.

[6] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *CoRR*, abs/1503.02406, 2015. http://arxiv.org/abs/1503.02406.

[7] Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29):2696 – 2711, 2010. Algorithmic Learning Theory (ALT 2008).

[8] Xinqi Zhu and Michael Bain. B-CNN: branch convolutional neural network for hierarchical classification. *CoRR*, abs/1709.09890, 2017. http://arxiv.org/abs/1709.09890.

[9] Duo Wang, Rui Zhang, Zhongzhao Teng, Yuan Huang, Filippo Spiga, Michael Hong-Fei Du, Jonathan Gillard, Qingsheng Lu, Pietro Lio, and Jin Zhu. Neural network fusion: a novel ct-mr aortic aneurysm image segmentation method. *Proceedings of SPIE–the International Society for Optical Engineering*, 10574(75), 2018.

[10] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017.

[11] Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*, 2018.

A Appendices

A.1 Time-line