Published in final edited form as:

Comput Methods Biomech Biomed Eng Imaging Vis. 2019 ; 7(3): 260–265. doi:
10.1080/21681163.2018.1427148.

The transition module: a method for preventing overfitting in convolutional neural networks

S. Akbar^a, M. Peikari^a, S. Salama^b, S. Nofech-Mozes^{a,b}, and A. L. Martel^a

^aSunnybrook Research Institute, University of Toronto, Toronto, Canada

^bSunnybrook Health Sciences Centre, Toronto, Canada

Abstract

Digital pathology has advanced substantially over the last decade with the adoption of slide scanners in pathology labs. The use of digital slides to analyse diseases at the microscopic level is both cost-effective and efficient. Identifying complex tumour patterns in digital slides is a challenging problem but holds significant importance for tumour burden assessment, grading and many other pathological assessments in cancer research. The use of convolutional neural networks (CNNs) to analyse such complex images has been well adopted in digital pathology. However, in recent years, the architecture of CNNs has altered with the introduction of inception modules which have shown great promise for classification tasks. In this paper, we propose a modified ‘transition’ module which encourages generalisation in a deep learning framework with few training samples. In the transition module, filters of varying sizes are used to encourage class-specific filters at multiple spatial resolutions followed by global average pooling. We demonstrate the performance of the transition module in AlexNet and ZFNet, for classifying breast tumours in two independent data-sets of scanned histology sections; the inclusion of the transition module in these CNNs improved performance.

Keywords

Convolutional neural networks; histology; inception; breast tumour; overfitting

1. Introduction

With the introduction of slide scanners in the pathology work-flow, microscopic examination of tissue can now be performed on specialised software (Martel et al. 2017) at extremely high resolutions (Figure 1). In cancer research, the use of digital slides to examine disease-specific biological characteristics and quantitatively assess tumour progression is common

CONTACT S. Akbar shazia.akbar@sunnybrook.ca.

Notes on contributor

Shazia Akbar is a postdoctoral fellow at Sunnybrook Research Institute in Canada is fully affiliated with the Department of Medical Biophysics at University of Toronto. She holds a PhD in Computing from the University of Dundee, U.K., and her research interests lie in computer vision and machine learning in medical image analysis.

Disclosure statement

No potential conflict of interest was reported by the authors.

practice (Gutman et al. 2013). Archival storage of slides for clinical trials has also been aided with the introduction of digitisation, enabling repositories of vast numbers of slides.

Tumour identification in cancer research is necessary for various pathological assessments including tumour burden, grading and tumour-infiltrating lymphocytes (Schwartz et al. 2014; Hendry et al. 2017). In current practice, tumours are identified and analysed by trained pathologists who can interpret the complex bio-logical changes undergone in the presence of cancer. However, such manual labour is time-consuming and cumbersome, particularly for clinical research. Furthermore, standardising measurements for reproducibility and clinical trials is challenging due to inter- and intra-observer variability. With the advancements of slide scanners, there is potential to capture slides digitally and examine them through fully- or semi-automated means. Given the difficulties associated with manual assessment, an accurate automated solution for recognising tumour in vastly heterogeneous pathology data-sets would be of great benefit, enabling high-throughput experimentation and greater standardisation.

Deep convolutional neural networks (CNNs) are now a widely adopted architecture in machine learning. Indeed, CNNs have been adopted for tumour classification in applications such as analysis of whole slide images (WSI) of breast tissue using AlexNet (Spanhol et al. 2016a) and voxel-level analysis for segmenting tumours in CT scans (Vivanti et al. 2015). Such applications of CNNs continue to grow and the traditional architecture of a CNN has also evolved since its origin in 1998 (LeCun et al. 1998). A basic CNN architecture encompasses a combination of convolution and pooling operations. As we traverse deeper in the network, the network size decreases resulting in a series of outputs, whether that be classification scores or regression outcomes. In lower layers of a typical CNN, fully connected (FC) layers are required to learn non-linear combinations of learned features. However, the transition between a series of two-dimensional convolutional layers to a one-dimensional FC layer is abrupt, making the network susceptible to overfitting (Lin et al. 2014).

Existing methods for avoiding overfitting include augmentation, regularisation layers such as Dropout (Srivastava et al. 2014) and batch normalisation (Krizhevsky et al. 2012), and L2 normalisation when computing training loss. Whilst data augmentation is commonly used in practice, it is not wholly suited for tumour identification due to high variability between tumour textures and appearances, which can only be solved by exposing the network to a wider population, i.e. collection of more data. Furthermore, augmentation increases computational costs exponentially. In this paper, we propose a method for transitioning between convolutional layers and FC layers by introducing a framework which encourages generalisation. Different from other regularisers, our method congregates high-dimensional data from feature maps produced in convolutional layers in an efficient manner before reducing them to a one-dimensional FC layer.

To ease the dimensionality reduction between convolutional and FC layers, we propose a transition module, inspired by the inception module (Szegedy et al. 2015). Here, a module is a series of layers learned in parallel, i.e. a sub-architecture, of which the output of each layer is concatenated for the proceeding FC layer. The motivation behind our proposed module

structure is to gradually reduce network size and retain rich features learned automatically. Our method encompasses convolution layers of varying filter sizes, capturing learned feature properties at multiple scales, before collapsing them to a series of global average pooling layers. We show that this configuration gives considerable performance gains for CNNs in a tumour classification problem in scanned images of breast cancer tissue. As our method is designed to reduce overfitting, we also evaluate the performance of the transition module compared to other commonly used regularisers and demonstrate that the transition module is an effective and simple addition to typical CNN architectures.

Related work in tumour classification include that of Araújo et al. (2017) who previously used a standard AlexNet-like CNN architecture to classify normal tissue and, benign, in situ and invasive breast cancers. To avoid overfitting, Araújo et al. (2017) augmented their data-set by rotating and flipping patches. A similar set-up was also used by Xu et al. (2016) in combination with a SVM classifier to distinguish between cancer subtypes in two data-sets (brain, colon). Clearly, the use of deep learning models in many digital pathology tasks has grown considerably in the last few years. In several global challenges (mit 2013,2014; gla 2015), winning entries have utilised deep learning architectures on vast amounts of histology data, which whilst enables researchers to benefit from deep architectures is difficult to recreate in clinical practice. Therefore, there is a need to build deep architectures which can adapt to limited labelled data.

2. Method

2.1. Inception module

Inception modules, originally proposed by Szegedy et al. (2015), are a method of gradually increasing feature map dimensionality thus increasing the depth of a CNN without adding extreme computational costs. In particular, the inception module enables filters to be learned at multiple scales simultaneously in a single layer, also known as a sub-network. Since its origin, there have been multiple inception networks incorporating various types of inception modules (Szegedy et al. 2016a, 2016b), including GoogleNet. The base representation of the original inception module is shown in Figure 2 (left).

Each inception module is a parallel series of convolutional layers restricted to filter sizes 1×1 , 3×3 and 5×5 . By encompassing various convolution sub-layers in a single deep layer, features can be explored at multiple scales simultaneously. During training, the combination of filter sizes which results in optimal performance is weighted accordingly. However, on its own, this configuration results in a very large network with increased complexity. Therefore, for practicality purposes, the inception module also encompasses 1×1 convolutions which act as dimensionality reduction mechanisms. The Inception *network* is defined as a stack of inception modules with occasional maxpooling operations. The original implementation of the Inception network encompasses nine inception modules (Szegedy et al. 2015).

2.2. Transition module

In this paper, we propose a modified inception module, called the ‘transition’ module, explicitly designed for the final stages of a CNN, in which learned features are mapped to FC layers. Whilst the inception module successfully captures multiscale information from input data, the bridge between learned feature maps and classification scores is still treated as a black box. To ease this transition process, we propose a method for enabling two-dimensional feature maps to be downscaled substantially before tuning FC layers. In the transition module, instead of concatenating outcomes from each filter size, as in the original inception module, independent global average pooling layers are configured after learning convolution layers which enable feature maps to be compressed via an average operation. During this operation, two-dimensional features maps are converted to a single one-dimensional layer.

Originally proposed by Lin et al. (2014), global average pooling layers were introduced as a method of enforcing correspondences between categories of the classification task (i.e. the softmax output layer) and filter maps. As the name suggests, in a global averaging pooling layer, a single output is retrieved from each filter map in the preceding convolutional layer by averaging it. For example, if given an input of $256 \times 3 \times 3$ feature maps, a global average pool layer would form an output of size 256. In the transition module, we use global average pooling to sum out spatial information at multiple scales before collapsing each averaged filter to independent one-dimensional output layers. This approach has the advantage of introducing generalizability and encourages a more gradual decrease in network size earlier in the network structure. As such, subsequent FC layers in the network are also smaller in size, making the task of delineating classification categories much easier. Furthermore, there are no additional parameters to tune. The structure of the transition module is shown in Figure 2 (right).

The transition module is placed after the last convolutional layer in the CNN. Depending on the network, additional convolution sizes may be adopted to further downsample feature maps if the resulting output from the last convolutional layer has large dimensions.

3. Experiment

We evaluated the performance of the proposed transition module on two independent data-sets. Some example patches from each data-set are shown in Figure 3. A description of each data-set is as follows.

- **In-House:** an in-house data-set acquired from the Department of Anatomic Pathology in Sunnybrook Health Sciences Centre. This data-set comprises of 1229 image patches extracted and labelled from breast WSIs scanned at x40 objective by a Scanscope XT (Aperio technologies, Leica Biosystems) scanner. Each RGB patch of size 512×512 was hand selected from 31 WSIs, each one from a single patient, by a trained pathologist. The surgical excision specimens represent sections from residual invasive and in situ breast carcinoma after presurgical systemic therapy (also known as neo-adjuvant therapy).

Each image patch was confirmed to contain either malignant or benign tissue by an expert pathologist. ‘Benign’ refers to patches which are absent of cancer cells but may contain epithelial and stromal elements with normal or spectrum of benign changes (Figure 3). 5-fold cross validation was used to evaluate the performance over 100 epochs.

- **BreaKHis:** a public data-set, BreaKHis (Spanhol et al. 2016b) which contains scanned images of benign (adenosis, fibroadenoma, phyllodes tumour, tubular adenoma) and malignant (ductal carcinoma, lobular carcinoma, mucinous carcinoma, papillary carcinoma) breast tumours at 40X magnification. Images were resampled into patches of dimensions 228×228, suitable for the CNNs we adopted, resulting in 11, 800 image patches in total. BreaKHis was validated using 2-fold cross-validation and across 30 epochs.
- **MNIST, CIFAR-10:** two publically available data-sets containing greyscale 28×28 pixel images of handwritten digits from 0 to 9 (MNIST), and RGB 32×32 pixel images of random everyday objects such as airplanes and cats divided into 10 classes (CIFAR-10). As images in these data-sets are considerably smaller than the two data-sets described above, we opted to train LeNet-5 (LeCun et al. 1998) CNN architectures which originally has two convolution layers. Each data-set has an independent train and test set and we trained the network for 30 epochs.

In Section 4.2, results are reported for three different CNN architectures (AlexNet, ZFNet, Inception-v3), of which transition modules were introduced in AlexNet and ZFNet. All CNNs were trained from scratch with no pretraining and batch normalisation was used after all convolutional layers. Transition modules in both implementations encompassed 3×3 , 5×5 and 7×7 convolutions, thus producing three average pooling layers. Each convolutional layer has a stride of 2 (except in LeNet-5), and 128, 1024 and 2048 filter units for LeNet-5, AlexNet and VFNet, respectively. Note, the number of filter units was adapted according to the size of the first FC layer proceeding the transition module.

CNNs were trained on a NVidia GeForce GTX TITAN X 12GB GPU and were implemented in Lasagne v0.2 with a Theano backend. A softmax function was used to obtain classification predictions and convolutional layers encompassed ReLU activations. 10 training instances were used in each batch in reported data-sets. We used Nesterov momentum (Sutskever 2013) to perform updates with an initial learning rate of $1e^{-5}$.

In the following section, we perform multiple comparisons to evaluate the transition module for different configurations.

- First, we compare the transition module with other popularly utilised techniques for avoiding overfitting including Dropout, L2 normalisation and cross-channel local response normalisation (Section 4.1).
- We report how the transition module performed when inserted into two difference CNN architectures (Section 4.2).

- To determine the effect of artificially increasing the size of our data-set, we also perform augmentation through a series of four 90 degree rotations (Section 4.3).

4. Results

4.1. Experiment 1: comparison with regularisers

Our first experiment evaluated the performance of the transition model when compared to other commonly used regularisers including Dropout (Srivastava et al. 2014) and cross-channel local response normalisation (Krizhevsky et al. 2012). We evaluated the performance of each regulariser in AlexNet and report results for (a) a single transition module added before the first FC layer, (b) two Dropout layers, one added after each FC layer with $p = 0.5$, (c) L2 normalisation introduced when computing training loss (weight decay = $1e^{-4}$) and lastly (d) normalisation layers added after each max-pooling operation, similar to how it was utilised in Krizhevsky et al. (2012).

The transition module achieved an overall accuracy rate of 91.5% which when compared to Dropout (86.8%), L2 normalisation (87.6%) and local response normalisation (88.5%) showed considerable improvement, suggesting the transition module makes an effective regulariser compared to existing methods. When local response normalisation was used in combination with the transition module in ZFNet (below), we achieved a slightly higher test accuracy of 91.9%.

4.2. Experiment 2: comparing CNN architectures

Next, we evaluated the performance of the transition module in two different CNN architectures: AlexNet (Krizhevsky et al. 2012) and ZFNet (Zeiler and Fergus 2014). We also report the performance of Inception-v3 which has built-in regularisers in the form of 1×1 convolutions (Szegedy et al. 2016b), for comparative purposes. ROC curves for each implementation are shown in Figure 4.

Both AlexNet and ZFNet benefited from the addition of a single transition module, improving test accuracy rates by an average of 4.3%, particularly at lower false positive rates. Smaller CNN architectures proved to be better for tumour classification in this case as overfitting was avoided, as shown by the comparison with Inception-v3. Surprisingly, the use of dimensionality reduction earlier in the architectural design does not prove to be effective for increasing classification accuracy. We also found that the incorporation of global average pooling in the transition module improved results slightly and resulted in 3.1% improvement in overall test accuracy.

4.3. Experiment 3: augmentation

When our in-house data-set was augmented, we achieved even greater performance gain at higher specificities with the transition module (Figure 5(a)). Interestingly, without the transition module, performance dropped in AlexNet when more data were incorporated suggesting increasing the data-set alone did not improve performance. Additional architectural changes which occurred when we incorporated the transition module held significant importance for gaining additional information from augmented data.

4.4. Experiment 4: BreakHis, MNIST, CIFAR-10

We used the same AlexNet architecture used above to also validate the performance of the transition module on BreakHis. ROC curves are shown in Figure 5(b). There was a noticeable improvement ($AUC+ = 0.06$) when the transition module was incorporated, suggesting that even when considerably more training data are available, a smoother network reduction can be beneficial. The transition module achieved an overall test accuracy of 82.7% which is comparable to 81.6% achieved with SVM in Spanhol et al. (2016b); however, these results are not directly comparable and should be interpreted with caution.

The authors note that in recent work by Han et al. (2017) which utilises information from multiple classes in BreakHis sur-pass outcomes reported here. However, given that our method has few adaptations to the original architecture of widely used CNNs, we expect further small changes to our methodology can improve results further.

We also validated the performance of the transition module on two other publically available data-sets: MNIST and CIFAR-10. Results on the test set for each of these data-sets are given in Table 1. Similar accuracy rates were achieved in the MNIST data-set across all 10 digits, whereas in CIFAR-10, a significant improvement was observed in the majority of classes (Figure 6). In particular, there was test accuracy rates improved in the ‘bird’ class, achieving an F1 score of 0.39, an increase of 0.22 compared to the unmodified LeNet-5 architecture.

5. Conclusions

In this paper, we propose a novel regularisation technique for CNNs called the transition module, which captures filters at multiple scales and then collapses them via global average pooling in order to ease network size reduction from convolutional layers to FC layers. We showed that in two CNNs (AlexNet, ZFNet) this design proved to be beneficial for distinguishing tumour from healthy tissue in digital slides. We also showed an improvement in a larger publically available data-set, BreakHis.

Our evaluations showed that compared to other popularly utilised regularisers, the transition module was able to adapt to a small data-set successfully achieving accuracy rates of 91.9%. The use of augmentation further improved performance at higher sensitivities whilst maintaining performance at lower sensitivities.

Acknowledgements

We gratefully acknowledge the support of NVidia Corporation with the donation of GPUs used for this research.

Funding

This work has been supported by grants from the Canadian Breast Cancer Foundation Canadian Cancer Society [grant number 703006] and the National Cancer Institute of the National Institutes of Health [grant number U24CA199374-01], and also supported by NVidia Corporation with the donation of GPUs used for this research.

References

Grand challenge on mitosis detection. 2013 MICCAI [cited 2017 Oct 18]. Available from: <http://amida13.isi.uu.nl/>

- MITOS-ATYPIA-14 challenge. 2014 ICPR [cited 2017 Oct 18]. Available from: <https://mitos-atypia-14.grand-challenge.org/>
- Gland segmentation challenge contest. 2015 MICCAI [cited 2017 Oct 18]. Available from: <https://www2.warwick.ac.uk/fac/sci/dcs/research/tia/glascontest>
- Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, Polónia A, Campilho A. 2017 Classification of breast cancer histology images using convolutional neural networks. *PloS One* 12.
- Gutman D, Cobb J, Somanna D, Park Y, Wang F, Kurc T, Saltz J, Brat D, Cooper L. 2013 Cancer digital slide archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J Amer Med Inform Assoc* 20(6):1091–1098. [PubMed: 23893318]
- Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S. 2017 Breast cancer multi-classification from histopathological images with structured deep learning model. *Sci Rep* 7.
- Hendry S, Salgado R, Gevaert T, Russell P, John T, Thapa B, et al. 2017 Assessing tumor-infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the international immunooncology biomarkers working group: part 1: assessing the host immune response, TILs in invasive breast carcinoma and ductal carcinoma in situ, metastatic tumor deposits and areas for further research. *Anat. Pathol* 24:235–251.
- Krizhevsky A, Sutskever I, Hinton GE. 2012 ImageNet classification with deep convolutional neural networks. Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. *Advances in neural information processing systems Vol 25* New York (NY): ACM; p. 1097–1105.
- LeCun Y, Bottou L, Bengio Y, Haffner P. 1998 Gradient-based learning applied to document recognition. *Proc. IEEE* 86:2278–2324.
- Lin M, Chen Q, Yan S. 2014 Network in network. *Proc ICLR Banff*, AB, Canada.
- Martel A, Hosseinzadeh D, Senaras C, Zhou Y, Yazdanpanah A, Shojaii R, Patterson E, Madabhushi A, Gurcan M. 2017 An image analysis resource for cancer research: PIIP-pathology image informatics platform for visualization, analysis, and management. *Cancer Res* 77:e83–e86. [PubMed: 29092947]
- Schwartz A, Henson D, Dechang C, Rajamarthandan S. 2014 Histologic grade remains a prognostic factor for breast cancer regardless of the number of positive lymph nodes and tumor size: a study of 161708 cases of breast cancer from the SEER program. *Arch Pathol Lab Med* 138:1048–1052. [PubMed: 25076293]
- Spanhol F, Oliveira LS, Petitjean C, Heutte L. 2016a Breast cancer histopathological image classification using convolutional neural network. In: *International Joint Conference on Neural Networks*. Vancouver, BC, Canada; p. 2560–2567.
- Spanhol FA, Oliveira LS, Petitjean C, Heutte L. 2016b A dataset for breast cancer histopathological image classification. *IEEE Trans Biomed Eng* 63:1455–1462. [PubMed: 26540668]
- Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014 Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15:1929–1958.
- Sutskever I 2013 Training recurrent neural networks [thesis] University of Toronto.
- Szegedy C, Ioffe S, Vanhoucke V, Alemi A. 2016a Inception-v4, Inception-ResNet and the impact of residual connections on learning. *arXiv preprint arXiv:160207261*.
- Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. 2015 Going deeper with convolutions. *Computer Vision and Pattern Recognition (CVPR)* Boston, MA.
- Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. 2016b Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV; p. 2818–2826.
- Vivanti R, Ephrat A, Joskowicz L, Karaaslan OA, Lev-Cohain N, Sosna J. 2015 Automatic liver tumor segmentation in follow up CT studies using convolutional neural networks. In: *Proceeding of the Patch-Based Methods in Medical Image Processing Workshop, MICCAI*. Munich, Germany.
- Xu Y, Jia Z, Wang LB, Ai Y, Zhang F, Lai M, Chang EC. 2016 Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinform* 18.
- Zeiler MD, Fergus R. 2014 Visualizing and understanding convolutional networks. In: *Proceedings of ECCV 2014: 13th European Conference*, Zurich, Switzerland; p. 818–833.

Figure 1.
Digital slide shown at multiple resolutions. Regions-of-interest outlined in red are shown at greater resolutions from left to right.

Figure 2.
Original inception module (left) and the proposed transition module (right).

Figure 3.

A subset of training data from two independent data-sets used in reported experiments. The three examples on the left are labelled as healthy and the three on the right as cancerous.

Figure 4. ROC curves for AlexNet and ZFNet, with and without the proposed transition module, and Inception-v3. ROC curves are also shown for the transition module with and without average pooling.

Figure 5. ROC curves for (a) AlexNet with and without the transition module when the in-house data-set was augmented through rotations, and (b) BreakHis data-set with and without the proposed transition module.

Figure 6. Confusion matrices for CIFAR-10 with and without the transition module. True labels are shown along the y -axis and predicted labels along the x -axis. Correctly labelled classes are shown along the diagonal (top left to bottom right).

Table 1.

Test accuracy (%) of LeNet-5 with and without the transition module on MNIST and CIFAR-10 data-sets.

Method	MNIST	CIFAR-10
LeNet-5	0.98	0.41
LeNet-5 + Transition	0.98	0.51

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript