CS230

Exploring the Use of Multiprong Attention in the Classification of Human Actions from Still Images

Bohong (Danny) Du
Department of Computer Science
Stanford University
dannydu@stanford.edu

Abstract

Modern day state-of-the-art deep neural networks used in the field of computer vision are largely feed-forward convolutional neural networks (CNNs). While deep learning as a field in general was heavily inspired by the biological structure of the human brain, most modern models have since then significantly deviated from their biological counterparts; while the human ventral visual pathway is 4-8 levels, modern start-of-the-art computer vision networks often exceed 100 layers in depth, albeit with the use of skip connections and residual blocks [14]. One important aspect of the human visual system that has not be since popularly translated, however, are the existence of recurrent circuits between the ventral stream cortical areas [2,14]. In this project, I explore the augmentation of ResNet18 with one timestep of recurrent attention, in which the final global vector representation is used to generate three queries fro three prior activation maps in the network. While the model's performance is subpar compared to that of ResNet18 and MobileNet V2, the attention heatmaps show promise for reccurent visual attention models of the future.

1 Introduction

Currently, there exists a few differing interpretations for the role of these crucial recurrent circuits. One possible interpretation follows in the footsteps of ResNet and DenseNet [27][28], but leveraging skip connections recurrently within a convolutional block, as opposed to sending information forward [11]. Another interpretation focuses on the use of a variety of forms of visual attention to query and attend to the information processed by convolutional blocks before it [3,4,5,6,22]. I choose to focus on the latter interpretation of recurrence in the human ventral stream as it is my intuition that current attention gives a model the ability to query different uncertainties within its changing internal representation of the image at different timesteps. Constrained by computational resources, I will attempt to use a multiprong attention mechanism to simulate one timestep of current attention.

The task I will be approaching with this novel model is action classification framed as a general image classification problem, where the input is the entire image, and the output is one of 40 classes (see Dataset). I chose this task, along with the accompanying dataset, primarily to compare the performance of my proposed model against deep modern neural networks (e.g. ResNet, DenseNet) [27][28]. My goal with this project is to see whether or not modern network architectures can be augmented by the biologically mechanism of recurrent attention. Action classification is a task of considerable difficulty, and one that involves some level of understanding of the interaction between two objects in an image. It is my hope that such a setting will allow the proposed mutliprong attention mechanism to combine fine-grained information from different parts of the image (i.e. the human

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

pose and the object being interacted with) in a mutually reinforcing manner. With the ability to querying backwards, I hope that the model will choose to attend more closely to areas of uncertainty, and thus recover the nuanced details of human actions in the form of hand gestures and poses.

2 Related work

2.1 Biological Recurrence

When we use our human visual system, we are only capable of attending to a select portion of our visual field at any given time. Evidence suggests that there is both a bottom-up (i.e. finding saliencies in the image) and a top-down (i.e. based on our task/motivation at the time) control of where we attend to [29]. Furthermore, research has shown that there are feedback loops present at nearly every stage of visual processing in the human brain [2]. We know that the outputs of higher-level neurons actively affect the signal processing of the lower-level neurons, but we are unaware of exactly how such recurrent relationships affect the signal processing from end to end [2]. There is hope that the use of neural networks to help uncover the uncertainty behind biological recurrence. One such attempt, CORnet, is regarded as a lightweight alternative to modern, state-of-the-art neural networks for the task of image classification, while adhering strictly to the cortical areas of the human visual cortex and leveraging recurrent neural networks to simulate the action of biological recurrence [11].

2.2 Visual Attention

Attention in neural networks was invented and popularized in the realm of natural language processing (NLP) [30]. Specifically, attention has helped augment modern neural language models by providing a mechanism through which the model can query the input, given some piece of context. Visual attention is most popular in visual question-answering tasks, some 2-dimensional input (the raw image or, more commonly, some CNN-produced feature map) is attended to in order to answer a language-based question [7,9].

In its most basic form, visual attention can apply methods from 1-dimensional attention by treating each pixel location as a vector with length equal to the number of channels [9,22]. Under this assumption, we can apply dot product [22], addative [31], parametric [22], hard [9], soft[9], (etc.) attention as we would like; we use the query vector and the channel vector taken from the input to create an attention map, which is subsequently multiplied element-wise with the input, broadcasted across the channels dimension. In order to better retain the spatial aspect of 2-dimensional data, grid attention was introduced that allowed the query to also be 2-dimensional, as opposed to 1-dimensional [31]. CoordConv attempts to further augment the spatial ambiguity of convlutional filters by concatenating additional channels of spatial information [24]. Squeeze and excitation networks offers a complex visual attention mechanisms taking into account not only spatial information, but also the explicit modelling of interdependencies between channels [23].

Multihead attention has remained a rather sparse field of research up to this point. Unsurprisingly, it has been applied to NLP-computer vision hybrid tasks such as image captioning [32] and image search [33].

2.3 Recurrent Visual Attention

There predominant approach for recurrent visual attention has been the top-down, hierarchical approach. A specific class of glimpse networks aims for especially lightweight understanding of visual data. Using a RNN controller, the such models queries a 'glimpse' - a small crop of the total input image - to process at every time step [1,3,6]. Such glimpses resemble a form of hard attention. RAM introduced the idea of glimpse networks, using reinforcement learning to determine where to glimpse at every time step [6]. DRAM extended RAM in the task of multiple object recognition [3]. EDRAM produced a fully-differentiate form of glimpse network to be train end-to-end using a dataset with bounding box labels of the desirable glimpse location [1]. Other, soft attention-based approaches include MAM-RNN [8] and S3TA [13].

2.4 Novelty

Multiprong attention uses separate queries to attend to different inputs, whereas multihead attention uses one query vector to attend to different positions. This approach this paper takes is somewhat similar to [22], in which the result of the global average pool of VGG is used as a query to attend to prior feature maps in the network. The novelty we offer comes in the form of 1) separate queries per feature map attended to generated via a 2-layer perceptron and 2) propagating the attended to feature map through the remaining convolutional filters of the network. We hope that by simultaneously attending to the network with unique queries, we are able to simulate the behavior of human recurrent attention without using recurrent networks.

3 Dataset and Features

I will be using the Stanford 40 Actions dataset [21], which was introduced in 2011 as a larger human action dataset than the then most commonly used PASCAL VOC Action Images dataset [20]. The Stanford 40 Actions dataset is comprised of 40 human actions, such as "applauding" and "throwing a frisbee", and there are 180-300 images per class [21]. The biggest advantage is the size of the Stanford 40 dataset, with 9532 total images to PASCAL VOC's 1221 images. Like PASCAL, however, the Stanford 40 dataset is also of substantial difficulty due to the varying poses, degree of visibility, and amount of clutter in the images [21]. The original authors of the paper elected to use a training set of 100 images from each action class (4000 training images) with the remaining 5532 images in the test dataset [21]. I have decided to randomly split the test dataset into equally sized test and validation datasets, while maintaining training set as the suggested 4000 images [21].

Figure 1: Three images from the Stanford 40 Actions dataset corresponding to the actions "applauding", "climbing", and "cooking.

4 Methods

4.1 Model Iteration

The original proposal for this project was to explore the mechanism of recurrent attention. The first network I trained (ThreePassNet from Milestone 2) was intended to serve as a preliminary survey of the recurrent attention mechanism. What I saw by examining the attention heatmaps from that network was that the attention mechanism essentially removed crucial information, which the model then sought to recover via its attention in the next timestep.

Despite this, I constructed and trained ReflectNet, which incorporated an LSTM controller of attention, closely in-line with an heirarchical model of recurrent attention. What I found, however, was that the network simply repeatedly attended to the same regions of interest over any number of timesteps. I suspect that this is due to the difficulty of the training task at hand, with only 4000 available training images to train the LSTM-guided attention mechanism.

As such, I created the final version of my network architecture, MultiProngNet, which is designed to examine recurrent attention in the sense of higher level layers querying lower levels, without using any RNN or LSTM blocks in the model architecture.

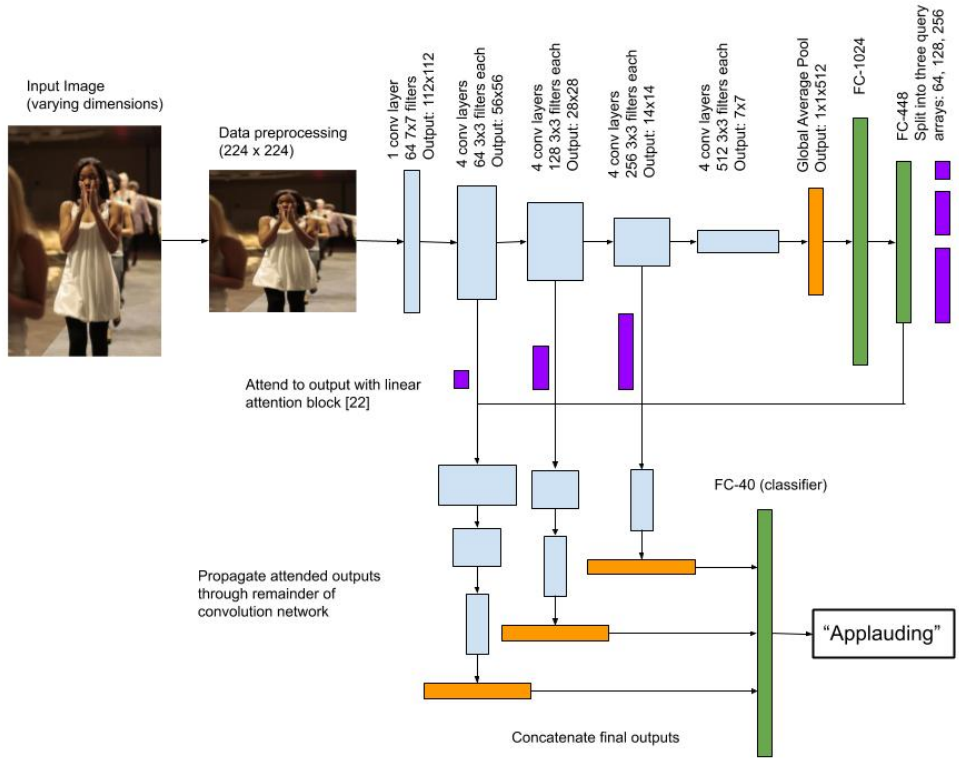

Figure 2: The proposed MultiProngNet is based on four convolutional blocks taken with pretrained weights from ResNet18. The input is first processed through the ResNet architecture, with the final fully connected layer removed. The output vector of the global average pool is then passed through a query network (2-layer perceptron), the output of which is split into three distinct query vectors. These three query vectors then query the outputs of the first, second, and third convolutional blocks to create a new set of feature maps. Each new feature map is then passed through the remainder of the ResNet18 architecture before being the resulting outputs vectors of each feature map is concatenated together and passed through a final classifier with 40 output classes.

5 Experiments/Results/Discussion

5.1 Experiments

ThreePassNet, ReflectNet, and MultiProngNet was each trained for 50 epochs using the Adam optimizer (lr = $1e^{-3}$, momentum = 0.9, weight decay = $1e^{-5}$). A learning rate scheduler exponentially reduced the learning rate by a factor of 0.1 every 10 epochs.

All convolutional layers were taken with pretrained weights from Torchvision's repository of ResNet18. All layers were then finetuned using the 4000 image training set from the Stanford 40 dataset.

5.2 Results & Discussion

Model	Stanford 40 Top-1 Accuracy
ResNet18	75.2
MobileNet V2	**75.8**
ThreePassNet	62.8
ReflectNet	56.0
MultiProngNet	**66.7**

Figure 4: Attention heatmaps for three validation set examples

In terms of performance, none of the three network architectures I proposed and implemented were able to surpass the baselines set forth by ResNet and MobileNet. While 66.7% top-1 accuracy leaves much to be desired, it shows that there is some merit to this approach. What is less encouraging, however, is that the use of attention on the ResNet18 architecture actually lowered the performance of ResNet quite noticeably, despite using the same pretrained layers.

Looking at the attention heatmaps, we see that the final layer of attention (on the most abstract feature map) is far more concentrated compared to the earlier layers of attention. Interestingly, we see that the model definitely is able to pick out salient regions of the image in all three cases. This is, however, the model's greatest strength and weakness. By focusing on the salient region, the model suppresses background information to focus on certain areas of the image. In doing so, it is able to classify such information on a more fine grained level. At the same time, the model might miss out on crucial details it does not deem as salient at first.

The first example shows us a case in which the model correctly focses on both the person and the object of interaction, and results in a correct classification. In the second example, we see the model, particularly in the first level of attention, closely examine the pose of the actor. However, the object of interaction (frisbee) is not attended to. Finally, we see an interesting example where the model appears to correctly identify both the pose and the object, but wrongly classifies the image. In this case, it is most likely due to the similarities between playing the violine and shooting an arrow. However, we would hope that the fine-grained classification enabled by supressing background information augment our ability to distinguish such cases.

6 Conclusion/Future Work

There is much more to be done in the field of visual attention. Ultimately, it is hoped that the use of visual attention can lower computational costs by only examining certain salient regions of the image, as opposed to its current state, which is very computationally heavy. For my line of model architectures specifically, I would love to experiment more closely with the LSTM-guided attention network. From what I examined in the heatmaps for ReflectNet, it appeared that the LSTM-

guided attention would not change the areas of the image it would attend to between timesteps, and thus negating the advantage of recurrent attention to start with. Given more time and computational resources, I believe that recurrent attention can indeed be used to augment existing deep, convolutional networks in classification involving fine-grained details, such as human actions.

References

[1] Artsiom Ablavatski, Shijian Lu, and Jianfei Cai. Enriched deep recurrent visual attention model for multiple object recognition. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference* on, pages 971–978. IEEE, 2017.

[2] Bruno A Olshausen. 20 years of learning about vision: Questions answered, questions unanswered, and questions not yet asked. In *20 Years of Computational Neuroscience*, pages 243–270. Springer, 2013.

[3] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014.

[4] Minki Chung and Sungzoon Cho. Cram: Clued recurrent attention model. *arXiv preprint arXiv:1804.10844*, 2018

[5] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, volume 2, page 3, 2017

[6] Volodymyr Mnih, Nicolas Heess, Alex Graves, Koray Kavukcuoglu. Recurrent Models of Visual Attention. In *arXiv preprint arXiv:1406.6247*, 2014.

[7] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *arXiv preprint arXiv:1707.07998*, 2018.

[8] XuelongLi,BinZhao,andXiaoqiangLu.MAM-RNN:multi- level attention model based RNN for video captioning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *arXiv preprint arXiv:1502.03044*, 2016.

[10] Mo Shan, Nikolay Atanasov. A spatiotemporal model with visual attention for video classification. In *arXiv preprint arXiv:1707.02069*, 2017.

[11] Liu, Wei Zhao, Bin Lu, Xiaoqiang. (2017). MAM-RNN: Multi-level Attention Model Based RNN for Video Captioning. 2208-2214. 10.24963/ijcai.2017/307. [12] Kubilius, J., et al. CORnet: modeling the neural mechanisms of core object recognition. Preprint at *biorXiv* https://www.biorxiv.org/ content/10.1101/408385v1 (2018).

[13] Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, Pushmeet Kohl. Towards Robust Image Classification Using Sequential Attention Models. In *arXiv preprint arXiv:1912.02184*, 2019.

[14] Kar, K., Kubilius, J., Schmidt, K. et al. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat Neurosci* 22, 974–983 (2019). https://doi.org/10.1038/s41593-019-0392-5

[15] Yazan Abu Farha and Jurgen Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, Zsolt Kira. Action Segmentation with Joint Self-Supervised Temporal Domain Adaptation. In *arXiv preprint arXiv:2003.02824*, 2020.

[17] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3281–3288, 2011.

[18] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 780– 787, 2014.

[19] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 729–738, 2013.

[20] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Chal- lenge 2010 (VOC2010) Results

[21] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. In *International Conference on Computer Vision*, pages 1331-1338, 2011.

[22] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, Philip H.S. Torr. Learn To Pay Attention. In *arXiv preprint arXiv:1804.02391*, 2018.

[23] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, Enhua Wu. Squeeze-and-Excitation Networks. In *arXiv preprint arXiv:1709.01507*, 2018.

[24] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, Jason Yosinski. An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution. In *arXiv preprint arXiv:1807.03247*, 2018.

[25] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, Jonathon Shlens. Stand-Alone Self-Attention in Vision Models. In *arXiv preprint arXiv:1906.05909*, 2019.

[26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *arXiv preprint arXiv:1801.04381*, 2018.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. In *arXiv preprint arXiv:1512.03385*, 2015

[28] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *arXiv preprint arXiv:1608.06993*, 2017

[29] Laurent Itti and Christof Koch. Computational Modelling of Visual Attention. In *Nature Reviews - Neuroscience, Vol 2*, 2001

[30] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *arXiv preprint arXiv:1409.0473*, 2014

[31] Jo Schlemper, Ozan Oktay, Liang Chen, Jacqueline Matthew, Caroline Knight, Bernhard Kainz, Ben Glocker, Daniel Rueckert. Attention-Gated Networks for Improving Ultrasound Scan Plane Detection. In *arXiv preprint arXiv:1804.05338*, 2018

[32] Geondo Park, Chihye Han, Wonjun Yoon, Daeshik Kim. MHSAN: Multi-Head Self-Attention Network for Visual Semantic Embedding. In *arXiv preprint arXiv:2001.03712*, 2020

[33] Po-Yao Huang, Xiaojun Chang, Alexander Hauptmann. Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations. In *arXiv preprint arXiv:1910.00058*, 2019