A STUDY OF RECURRENT AND CONVOLUTIONAL NEURAL NETWORKS FOR VISUAL ATTENTION

FARHAD NAWAZ [FARHADN@SEAS.UPENN.EDU], KAUSIK SIVAKUMAR [KAUSIK@SEAS.UPENN.EDU],
HARSH GOEL [HARSHG99@SEAS.UPENN.EDU],

ABSTRACT. Several CNN based network architectures have been used for image classification for a very long time. However, using convolutions is computationally very expensive for very large images. Thus, we aim to study a new model based on visual attention (Recurrent Attention Model(RAM)) that is able to extract information from an image by adaptively selecting a sequence of patches/images that are smaller than the actual image. We would like to study the model and compare it with the current SOTA such as RESNET for classifying digits in large MNIST images. Since the recurrent model is non-differentiable, it uses a reinforcement learning mechanism to learn the next region of interest. We would like to investigate and explore possible performance improvements with a secondary reward structure for the network to select good informative regions from the image.

1. BACKGROUND

Human vision often focuses on specific regions of an image and processes only those regions to efficiently understand the objects in the scene. This idea has been adapted to neural networks that learn to process only specific locations in the image. A novel recurrent neural network model has been developed [Mnih et al., 2014] to learn an internal representation of the image through a sequence of specific patches/glimpses extracted from certain locations of a given image. The paper successfully applies the model to MNIST images for classification. Roughly a recurrent attention model(RAM) does the following:

(1) The glimpse network would generate sub-samples of images with a certain resolution. The location of where the sampling is done is stochastic. The glimpse network builds an MLP/CNN from the image samples and locations to generate a learned feature vector
(2) This feature vector is fed as an input to a recurrent layer, which would generate class predictions at each time-step as well as the next location to sample based upon both the previous hidden state and the learnt glimpse features
(3) A reinforced learning mechanism is used to make this stochastic search for the next location to sample more directed based upon the expected reward of classifying the image correctly (this is one of the outputs of the recurrent layer at each time-step)

We aim to compare and study it's performance against state of the art convolution architectures that have been used to recognize digits in MNIST. images. [He et al., 2016, Goodfellow et al., 2014] .

2. Project Aims

Our primary aim is to study and compare the computational aspects, and accuracy by implementing a low parameter RESNET model and a RAM for the MNIST data set. We expect that the RAM would perform much worse than the RESNET model as the full image is not observable in the case of RAM. Each recurrent step in RAM output's a policy that is essentially mean and variance of the future location to see. The next location is sampled from this policy which is optimised using REINFORCE algorithm. The reward structure,however, is very sparse and in the paper they use the final classification score as the reward. Thus, our secondary aim is to study the effect of a secondary reward structure on the classification accuracy. Roughly, we would like to improve upon the policy of the location of the next patch by assigning a reward on how informative the patch is.

We have 2 reward structures in mind. A log-likelihood function that outputs the classification given the current aggregated hidden state t based on t glimpses. Essentially, the agent is rewarded if it's classification is much clearer compared to what it was before in the prior time step. The second reward structure aims to location data and actual bounding boxes of the digits to hard penalise the policy layer from looking too far away from the MNIST image.

Thus, in summary we would like to address the following questions:

(1) What makes RAM models computationally lesser expensive compared to a RESNET Model? Is the model robust to translations or data augmentations otherwise unseen?
(2) How can the accuracy of such a model be improved by tweaking the reward structure for selecting locations using Reinforcement Learning?

3. Timeline and splitting of work

- Week 1: Pull up the code for RAM and begin training the model (Harsh) Pull up a RESNET Model and begin training the model(Kausik and Farhad) [Mnih et al., 2014]
- Week 2: Model evaluation for accuracy under translational data augmentations,etc and computational costs.
- Week 3: Implement reward structure and retrain RAM to observe any improved performances
- Week 4: Consolidate and write project report

References

[Goodfellow et al., 2014] Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., and Shet, V. (2014). Multi-digit number recognition from street view imagery using deep convolutional neural networks.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Mnih et al., 2014] Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). Recurrent models of visual attention.