CATEGORY: COMPUTER VISION & MACHINE VISION - CVMV10
POSTER P6365 | CONTACT NAME Brian Cheung: XXXX@XXX
GPU TECHNOLOGY CONFERENCE

REDWOOD CENTER for Theoretical Neuroscience

Neural Attention for Object Tracking

Brian Cheung, Eric Weiss, Shalini Gupta, Pavlo Molchanov, Stephen Tyree, Jan Kautz
{bcheung, eaweiss}@berkeley.edu, {shalinig, pmolchanov, styree, jkautz}@nvidia.com

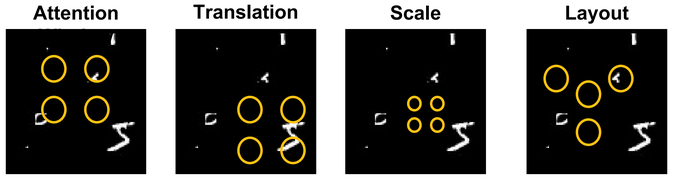
Introduction

With differentiable forms of attention being integrated into neural networks, end-to-end training with backpropagation is possible. We adopt the recently proposed attention mechanism in Spatial Transformer Networks (STNs) into a recurrent architecture to perform object tracking. We show that this attention mechanism has significant overlap with the mechanism in Deep Recurrent Attentive Writer (DRAW) networks which has have been successfully used to create generative models of images. We present an end to end trainable recurrent attention model for tracking a variety of objects in video sequences recorded by cameras mounted on an automobile. We also present several issues which arise when such recurrent attention models are scaled up to much larger and more complex images/videos. We present pretraining strategies to resolve some of these training issues.

Overview of Attention

$$V_i = \sum_n^H \sum_m^W U_{nm} k(m, n; \Phi_i)$$
$$\forall i \in [1, ..., H'W']$$

Generic Formulation of Differentiable Attention
The pixels in the input image U are mapped to a smaller output V . This can be interpreted as a form of routing where a select number of pixels from the input are connected to the output. The routing is defined by a kernel filter k(). The kernel defines which pixels in the input will contribute to a particular output.



Examples of possible routing configurations for the attention mechanism formulated in 1. Each yellow circle corresponds to a single kernel filter.

Factorized Attention
Most formulations of visual attention over the input image assume a factorization between the m and n dimensions of the input.

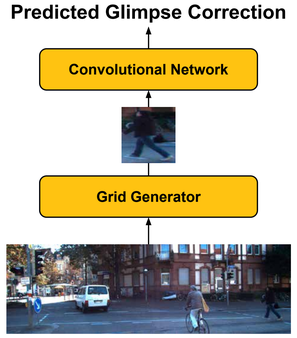
$$k(m, n; \Phi_i) = k(m, \Phi_{xi}) k(n, \Phi_{yi})$$

KITTI Tracking Dataset

- 375x1240 video
- Bounding boxes over time of cars, pedestrians, etc.

Registration Task for Pretraining Convolutional Network

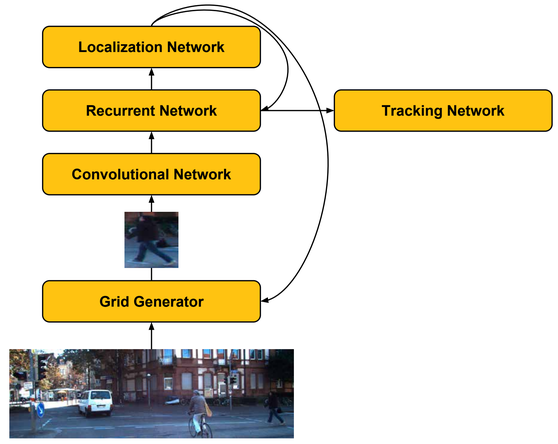
Predicted Glimpse Correction



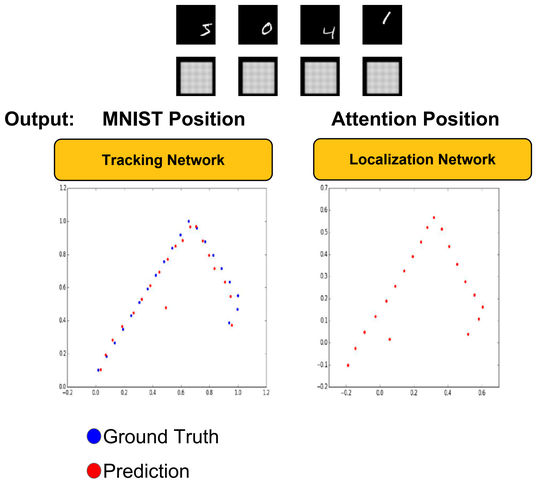
Prior to training on the tracking task, the Convolutional Network component of our model is trained on the simpler task of correcting a misaligned glimpse (leftmost image below). This misaligned window is input into the Convolutional Network which outputs the change in the glimpse parameters to correct the misalignment.

Input Glimpse **Predicted Correction** **Actual Correction**

Our Model



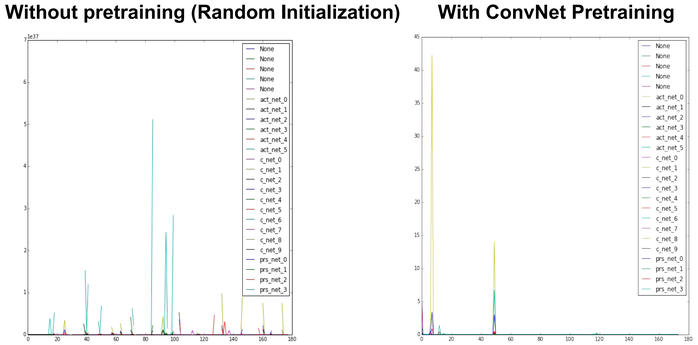
Toy Dataset: Bouncing MNIST



- **Grid Generator:** Generates a glimpse (window) of the input image.
- **Convolutional Network:** Transforms the glimpse into more abstract/semantic features.
- **Recurrent Network:** Integrate abstract/semantic features with memory of previous features/states.
- **Localization Network:** Transform recurrent network features into motor commands to create parameters for the next glimpse.
- **Tracking Network:** Transform recurrent network features into features useful for object tracking.

Results

Without pretraining (Random Initialization) **With ConvNet Pretraining**



This figure shows a comparison between the magnitudes of gradients for each component of our model during training. We show two different conditions: without pretraining (leftmost) and with pretraining (rightmost).

- Gradients magnitudes are significantly more stable with pretraining
- Gradient magnitudes are significantly smaller with pretraining

Conclusion

While we were able to train our model for the tracking task on a simple toy dataset (Bouncing MNIST), these results did not generalize to the more difficult task of tracking in natural videos. In particular, we found that backpropagation could not provide a useful training signal to the model even with the initialization provided by pretraining. In future work, we hope to explore integrating exploration mechanisms for the attention mechanism and using ImageNet pretrained networks to alleviate the issues encountered with end-to-end training with backpropagation.

References

- Brian Cheung, Eric Weiss, and Bruno Olshausen. Learning Retinal Tiling in a Model of Visual Attention. http://beta.openreview.net/pdf?id=1WvOZJ0yDTMnPB1oinGN, 2016.
- Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623, 2015.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044, 2015.
- Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In Advances in Neural Information Processing Systems, pp. 2008–2016, 2015.
- Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In Advances in Neural Information Processing Systems, pp. 2204–2212, 2014.