# Attention Models for Image and Video Caption Generation

Submitted By: Riashat Islam; Supervisor: Professor Juergen Schmidhuber

17th January 2016

## 1 Introduction

Adaptive target detection and sequential eye movement were originally introduced by Schmidhuber and Huber ((1991)) through the study of adaptive vision with neural networks, by using an adaptive control mechanism for sequential movements for target detection. Schmidhuber and Huber ((1991)) originally introduced a system for target detection inspired by the observation that biological systems employ sequential fovea movements for target detection. Schmidhuber and Huber ((1991)) introduced the system for active perception, where the system can learn to produce sequential fovea movements to decide where to perceive next.

Adaptive neuro-controllers were introduced for learning target detection a long time back. Recently Graves ((2013)) further introduced attention based mechanisms, with Long Short Term Memory recurrent neural networks Hochreiter and Schmidhuber ((1997)) to generate complex sequences one data point at a time, on tasks such as online handwriting. Inspired from Schmidhuber and Huber ((1991)), Mnih et al. ((2014)) further introduced a policy search based reinforcement learning approach for attention models, using a novel recurrent neural network model to adaptively select sequence of regions in an image or video for processing. Attention models models with convolutional neural networks have recently further gained interest, through the introduction of Deep Attention Selective Network architecture Stollenga et al. ((2014)) for dynamically altering the convolutional filter sensitivities for classification, for sequential processing to improve classification performance. Many of the recent approaches in attention models for computer vision Stollenga et al. ((2014)), Mnih et al. ((2014)) considers allowing the network to iteratively focus internal attention, with a policy search method to train the feedback network. Such models are further inspired from vision based reinforcement learning Koutník et al. ((2013)).

## 2 Approach

Motivated by recent work based on attention models and caption generation, we will focus our work towards developing novel attention models that can automatically learn to describe contents of images and videos. Recent work has described models capable of extracting information from an image or video by adaptively selecting a sequence of regions and only processing the selected regions at high resolution. Based on this, we can further automatically generate captions of an image. Much of recent work has focused on the caption generation problem using a combination of convolutional neural networks for vectoral representations of images and recurrent neural networks to decode the representations into natural language sentences. Furthermore, since applying convolutional neural networks to large images is computationally expensive, recent work has focused on policy search reinforcement learning algorithms to focus attention selectively on parts of the visual space instead of processing the whole scene entirely at once. Such sequential decision making approaches to focus attention can substantially reduce computational task complexity as only the object of interest are placed at the center of fixation and irrelevant features are ignored.

We will consider extensions of both soft and hard attention models. We will consider soft deterministic attention mechanisms, which is based on training deep stochastic neural networks using an unbiased back-propagation algorithm. For the hard attention model, we will consider using stochastic and deterministic policy gradients instead of REINFORCE based on actor-critic and Q learning based approaches. Using variants of soft and hard attention models, we will further validate the usefulness of attention models for image and video caption generation tasks.

Xu et al. ((2015)) considers caption generation tasks from images based on attention models. The attention models in Xu et al. ((2015)) are trained in a deterministic manner using standard backpropagation and stochastically by maximizing a variational lower bound, equivalently by the REINFORCE algorithm. Furthermore, Mnih et al. ((2014)) considers the attention problem as a sequential decision process of a goal directed agent interacting with the visual scene environment. We will develop our work further from using variants in the models described in Xu et al. ((2015)) and Mnih et al. ((2014)) for tasks based on either image and video caption generation or provide an extension of the Deep Q Network algorithm (DQN) based on soft and hard attention using multiple Atari 2600 games as the testbed.

## 2.1 Deep Reinforcement Learning for Attention Models

Mnih et al. ((2014)) presents a recurrent neural network model that processes inputs sequentially attending to different locations within the images or videos one at a time.Mnih et al. ((2014)) presents an optimization procedure allowing the model to be trained directly with respect to a given task and to maximize the performance procedure measure which depends on the entire sequence of decisions made by the model. Such a procedure is based on using backpropagation to train the neural network components and policy gradient to address the non-differentiabilities.The parameters of the agent are the parameters of the glimpse network, the core network and the action network, and the goal is to maximize the total reward of the agent. The agent needs to learn a stochastic policy with parameters, such that at each time step, it maps the history of past interactions with the environment. To a distribution over actions for the current time step. The policy is defined by the RNN and the history is summarized in the state of the hidden units. The policy of the agent induces a distribution over possible interaction sequences. Mnih et al. ((2014)) uses a sample approximation to the gradient based on the REINFORCE algorithm, and it involves running the agent with its current policy to obtain samples of interaction sequences and then adjusting the parameters of the agent such that the log probability of the chosen actions that have led to high cumulative reward can be increased. The log term in the gradient based on Monte Carlo approximation is just the gradient of the RNN that defines the agent evaluated at time step t and can be further computed by standard backpropagation.

Building up from work from Mnih et al. ((2014)) for our work, we will consider computing the gradient based on both on-policy and off-policy actor-critic frameworks, based on both the stochastic and deterministic policy gradients, which is the expected gradient of the action-value function for learning continuous deterministic actions. Further to considering a gradient based approach for the policy search, we can also consider estimating the action value function using the Bellman equation as an iterative update. Such value iteration algorithms converge to the optimal Q function, with recent approaches based on using non-linear function approximators such as neural networks. We will refer to approaches based on Deep Q learning from which the optimal policy can be derived. For our work, we will therefore develop attention models based on stochastic and deterministic policy gradients to maximize the cumulative reward, or use a Deep Q Network for learning the optimal action-value function based on which an optimal attention mechanism or policy can be derived.

## 2.2 Adaptive Monte Carlo and Unbiased Backpropagation

Bahdanau et al. ((2014)) introduced a soft attention weighted annotation vector, where the whole model is smooth and differentiable under deterministic attention, learning end to end by using standard backpropagation. Bahdanau et al. ((2014)) introduced a context vector that depends on a sequence of annotations, and each annotation contains information about the whole input sequence with a strong focus on the parts surrounding the iSuch alignment model directly computes a soft alignment – whch allows the gradient of the cost function to be backpropagated through. This gradient is used to train the alignment model as well as the whole transition model jointly. The probability reflects the importance of annotation – this is a form of attention in the decoder, as the decoder decides parts of the source sentence to pay attention to.

For our work, we will consider exploring the soft attention model based on training the stochastic neural network following work from Gu et al. ((2015b)). Modelling the attention mechanism as a stochastic process – it might be make more sense to learn a model that can carry out a sequence of stochastic operations. Gu et al. ((2015b)) presents an unbiased estimator for deep stochastic neural networks based on backpropagation, handing both continuous and discrete stochastic variables. MuProp is an unbiased estimator of derivatives in stochastic computational graphs that combines the statistical efficiency of backpropagation with the correctness of the likelihood ratio method. The REINFORCE algorithm described in Mnih et al. ((2014)) is a special case of the likelihood-ration estimator. The likelihood-ratio provides a convenient method for estimating the gradient and serves as the basis for unbiased estimators such as MuProp.

Furthermore, since learning the stochastic attention requires sampling the attention location, we can also develop our work based on Gu et al. ((2015a)). we will adapt to using NASMC for automatically adapting the proposal distribution supporting both online and batch variants of training. Gu et al. ((2015a)) proposes a method that parameterizes a proposal distribution using a recurrent neural network to model long-range contextual information and further allows effective training of latent variable recurrent neural networks using SMC.

Ba et al. ((2015)) considers a method for training stochastic attention networks that can reduce the variability in the stochastic gradients by presenting a wake-sleep recurrent attention model as a step towards reducing the high variance in the stochastic gradient estimates. Our hypothesis is that, using the MuProp and NASMC based approaches, similar to Ba et al. ((2015)), we can develop novel methods for training stochastic recurrent attention models which can deal with problems of high variance gradient estimates, and compare it with a training method based on wake-sleep algorithm presented in Bornschein and Bengio ((2014)). Using our model based on unbiased backpropagation for stochastic neural networks, we will develop a novel technique for gradient estimation that can speed up training, and demonstrate results on classification and image or video captions by attending the relevant objects in the images.

# 3 Experiments

We will present our results on domains of image classification and caption generation. To measure the effectiveness of the variants introduced in our model, we will first investiage a toy classification task involving a variant of the MNIST handwritten digits dataset similar to results presented in Ba et al. ((2015)). Furthermore, similar to experimental results presented in Xu et al. ((2015)), we will report our results on the popular Flickr8k and Flickr30k dataset which has 8,000 and 30,000 images respectively. We will also use the more challenging Microsoft COCO dataset for further comparison of models. We can also experiment our proposed algorithm on several popular Atari 2600 games: Breakout, Seaquest, Space Invaders, Tutankham, and Gopher, and compare the results obtained with the corresponding results of DQN that were presented by Mnih et al. ((2013)).

# 4 Conclusion

In our work, we will therefore consider both soft and hard attention models based on the actor-critic reinforcement learning, Deep Q learning and stochastic and deterministic policy gradient approaches. We will also consider training the neural networks using MuProp instead of standard backpropagation, and use adaptive sequential Monte Carlo sampling scheme. We will compare our attention model with several other state of the art results for image and video caption generation. More ambitiously, we will explore the usefulness of attention models for playing Atari games using the Deep Q Learning framework.

# References

J. Ba, R. B. Grosse, R. Salakhutdinov, and B. J. Frey. Learning wake-sleep recurrent attention models. *CoRR*, abs/1509.06812, 2015. URL `http://arxiv.org/abs/1509.06812`.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL `http://arxiv.org/abs/1409.0473`.

J. Bornschein and Y. Bengio. Reweighted wake-sleep. *CoRR*, abs/1406.2751, 2014. URL `http://arxiv.org/abs/1406.2751`.

A. Graves. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850, 2013. URL `http://arxiv.org/abs/1308.0850`.

S. Gu, Z. Ghahramani, and R. E. Turner. Neural adaptive sequential monte carlo. *CoRR*, abs/1506.03338, 2015a. URL `http://arxiv.org/abs/1506.03338`.

S. Gu, S. Levine, I. Sutskever, and A. Mnih. Muprop: Unbiased backpropagation for stochastic neural networks. *CoRR*, abs/1511.05176, 2015b. URL `http://arxiv.org/abs/1511.05176`.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL `http://dx.doi.org/10.1162/neco.1997.9.8.1735`.

J. Koutník, G. Cuccu, J. Schmidhuber, and F. J. Gomez. Evolving large-scale neural networks for vision-based reinforcement learning. In *Genetic and Evolutionary Computation Conference, GECCO '13, Amsterdam, The Netherlands, July 6-10, 2013*, pages 1061–1068, 2013. doi: 10.1145/2463372.2463509. URL `http://doi.acm.org/10.1145/2463372.2463509`.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013. URL `http://arxiv.org/abs/1312.5602`.

V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2204–2212, 2014. URL `http://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention`.

J. Schmidhuber and R. Huber. Learning to generate artificial fovea trajectories for target detection. *Int. J. Neural Syst.*, 2(1-2):125–134, 1991. doi: 10.1142/S012906579100011X. URL `http://dx.doi.org/10.1142/S012906579100011X`.

M. Stollenga, J. Masci, F. J. Gomez, and J. Schmidhuber. Deep networks with internal selective attention through feedback connections. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3545–3553, 2014. URL `http://papers.nips.cc/paper/5276-deep-networks-with-internal-selective-attention-through-feedback-conn`

K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2048–2057, 2015. URL `http://jmlr.org/proceedings/papers/v37/xuc15.html`.