CSC2547 Project Proposal
Searching in Machine Translation

Zikun Chen, Jerrod Parker, Yu-Siang Wang

October 17, 2019

Abstract

Recent algorithms in machine translation have included a value network to assist the policy network when deciding which word to output at each step of the translation. This can help optimize our algorithm to perform well on evaluation metrics such as the BLEU score. After training the policy and value networks in a supervised setting, the policy and value networks can be jointly improved through common actor-critic methods. The main idea of our project is to instead leverage Monte-Carlo Tree Search (MCTS) to search for good output words with guidance from a combined policy and value network architecture in a similar fashion as AlphaGo Zero. The network serves both as a local and a global look-ahead reference that uses the result of the search to improve itself. Additionally, we attempt to improve the inference stage of any sequence prediction task which involves sequence generation such as machine translation and image captioning by replacing beam search with MCTS even without the help of a policy or a value network.

1 Introduction

Many algorithms in machine translation use an encoder-decoder model where we encode an input sentence into a feature vector and then output the translated sentence word by word. For each word output, we have a conditional distribution over the word given the input sentence and previous output words. These algorithms are often trained by conditioning on the true output so far in an attempt to maximize the probability of the correct next word. For example, if we have output $(t-1)$ words and have an input sentence $(x_1, \cdots, x_m)$ of length m, our distribution over the next output is $p(y_t|y_{t-1}, \cdots, y_1; x_1, \cdots, x_m)$. This works decently but it can often choose words that are good in the immediate future but do not lead to optimal sentences in the long run. This has led to alternative algorithms [1, 3, 4] which use both a policy and value network. The policy network is our conditional language model, such as an attention RNN that outputs a distribution of next words given what we have seen at each time step. The value network predicts the BLEU score we would obtain given our current output if we continue to follow the policy to completion of the sentence. One benefit of jointly training the policy and value networks is that it helps guide the policy network to learn to optimize for longer term rewards like the final BLEU score. Furthermore, the networks improve each other through actor-critic methods. During inference at test time, these algorithms either use only the policy outputs [1] or a combination of the value network and output of the conditional language model [3].

In our project, we aim to improve the joint training of the policy and value networks with the aid of MCTS so that our model has a better notion of global look-ahead which will help us predict optimal translations. To the best of our knowledge, the two papers that are most directly applicable to this problem are [1, 4]. The algorithm in [4] used a policy network and a value network for the image captioning task. Both of these networks are used in the testing phase. The other relevant paper was [1]. In their work, the value network is only utilized during the training stage since the network takes the ground truth sentences as input.

We attempt to improve the existing actor-critic framework by using MCTS in a way similar to AlphaGo Zero. We expect MCTS to have several advantages over the actor-critic methods used in these papers. Firstly, we will use a mixture of the value and policy at each node in our tree to guide the search which is expected to give more information than using only the policy like in [1]. We expect an improvement in performance over the algorithm used in [4] since at each time step, we use a deeper and more powerful look ahead mechanism than that of [4]. [4] uses a linear combination of the policy and value at the current state to choose the next output word whereas our MCTS performs a deeper simulation of the tree to help us decide which word to output. The trade-off is that our method can be computationally more expensive. On top of this, we have seen the benefits of using MCTS in AlphaGo Zero in conjunction with the policy network, which narrows down search expansions to high probability outputs and a value network [5], which provides longer-term reward signals in different tree nodes. If this method does help to improve the models in machine translation, it is quite possible that it can also improve other sequence prediction models such as those in image captioning.

It is common in sequence generation tasks (e.g. Machine Translation) for the inference stage to utilize beam search [2] to infer the best output. One of the main drawbacks of beam search is that it chooses the best paths based off actions that provide the highest immediate rewards even though the goal is to obtain a good longer term score. It is not guaranteed to find output words that have a maximum joint prediction probability given the input sentence. We attempt to improve this by replacing beam search with MCTS to help us choose output token at each step from a longer term reward perspective. Note that this is done without any framing from reinforcement learning that requires a policy and a value networks. Again, we expect this to improve the performance as better long term reward is our interest.

2 Work Plan

1. Extended literature search and writing related work section (4 hours)

2. Learning the basics of relevant deep learning libraries such as PyTorch and open source projects to reproduce the baselines (8 hours)

3. Implement batch MCTS to run in the inference phase for better efficiency. (8 hours)

4. Implement AlphaGo Zero MCTS to use and train a combined policy and value network (24 hours)

5. Running baseline on benchmark dataset WMT 2014 English-to-German translation and tuning hyper-parameters for the MCTS (8 hours)

6. Create plots described in section 3 and produce the project presentation and report (10 hours)

3 Description of proposed results

Model Descriptions:

- **Model 1** contains a policy and value network with architecture similar to that in [1] where policy and value networks have been trained in a supervised setting.

- **Model 2** is equivalent to **model 1** except that additionally policy and value networks are trained using MCTS similar to AlphaGo Zero.

- **Model 3** is equivalent to **model 1** except that additionally we train the policy and value networks using the actor-critic method from [4].

Proposed Experiments:

1. Evaluate the BLEU score difference when we replace beam search with MCTS in the inference phase for the pre-trained seq2seq model from [6] since it achieved strong results and used a beam search at inference time. We also include time spent on inference for each of these methods to have a fair comparison. This might help the reader to learn a new easy improvement to their own seq2seq models if MCTS does improve performance compared to beam search.

2. Additionally, we train the networks from **model 1** described above using the MCTS from AlphaGo Zero and investigate if this improves performance of test set BLEU score. The reader will learn whether MCTS can be applied in machine translation to help improve the policy and value networks after they have been trained in a supervised fashion.

3. We will compare the performance of **model 2** and **model 3**. This is an important comparison to be made because the actor-critic method in **model 3** allowed excellent results in image captioning, [4] which is a similar task to machine translation.

4. We will use the model from [6] called 'Transformer + RL' as our baseline because it achieved competitive results and it can be easily replicated exactly from their given code. This baseline will be compared to **model 2** described above by the BLEU score on the test set as well as time spent during training and inference so that we can evaluate the trade-off between computation efficiency and performance. The reader will learn whether adding a value network and using MCTS can improve current methodologies in machine translation.

Table and Figures:
For the above points 2, 3 and 4 we will create the following table that have test results. Note that the baseline trains a policy network in a supervised setting and then improves it through policy gradients afterwards. We will also produce figures for these metrics on the validation set as they improve over time during training. The reader can compare and contrast visually about the performance and efficiency of each model to have an idea of whether our attempts are successful.

Methodology	BLEU	BLEU-4	Meteor	TrainingTime
Model 1 (supervised)				
Model 2 (MCTS)				
Actor-Critic (from [4])				
Baseline (refer to 4 above)				

4 Related Work

Several previous works have leveraged a value network to compliment the policy network in machine translation [1, 3], image captioning [4] and playing Go [5].

- Paper [4] designs a model for image captioning and trains a policy and value network in a supervised manner. The model is updated through the actor-critic reinforcement learning method. The authors found that the global guidance introduced by the value network greatly improves performance over just using a policy network.
- Paper [1] trains their policy and value networks by reinforcement learning but allows the value network to also take the true output as its input. This method helps to improve the policy by allowing the policy to directly optimize for BLEU score.
- As a significant milestone in reinforcement learning, paper [5] uses self-play and MCTS to train policy and value network with a shared body, which led to efficient optimization for move predictions in the game of Go. The MCTS method was shown to be a powerful policy improvement and policy evaluation method.

One of the big differences of our method compared to previous works is the addition of a value network with reinforcement learning methods to jointly update the policy and value networks for neural machine translation. The only papers we noticed that did this were in [4, 1], an image captioning paper where the authors used actor-critic methods. In contrast, we will be using MCTS which is expected to improve performance as described in the introduction. In [5], MCTS is applied to the game of Go instead of machine translation, which we are interested in. One of our contributions would be to show that MCTS can be used successfully in this domain. Another main contribution would be using MCTS as a look-ahead module during the sequence inference stage to improve beam search which is a local greedy strategy. We have yet to find a literature that did this.

References

[1] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. C. Courville, and Y. Bengio, “An actor-critic algorithm for sequence prediction,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[2] A. Graves, “Sequence transduction with recurrent neural networks,” *CoRR*, vol. abs/1211.3711, 2012.

[3] D. He, H. Lu, Y. Xia, T. Qin, L. Wang, and T. Liu, “Decoding with value networks for neural machine translation,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 178–187. [Online]. Available: http://papers.nips.cc/paper/6622-decoding-with-value-networks-for-neural-machine-translation

[4] Z. Ren, X. Wang, N. Zhang, X. Lv, and L. Li, “Deep reinforcement learning-based image captioning with embedding reward,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 1151–1159. [Online]. Available: https://doi.org/10.1109/CVPR.2017.128

[5] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, “Mastering the game of go without human knowledge,” *Nature*, vol. 550, pp. 354–, Oct. 2017. [Online]. Available: http://dx.doi.org/10.1038/nature24270

[6] L. Wu, F. Tian, T. Qin, J. Lai, and T. Liu, “A study of reinforcement learning for neural machine translation,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 2018, pp. 3612–3621. [Online]. Available: https://www.aclweb.org/anthology/D18-1397/