

ELF OpenGo: An Analysis and Open Reimplementation of AlphaZero

Yuandong Tian¹ Jerry Ma^{*1} Qucheng Gong^{*1} Shubho Sengupta^{*1} Zhuoyuan Chen¹ James Pinkerton¹
C. Lawrence Zitnick¹

Abstract

The AlphaGo, AlphaGo Zero, and AlphaZero series of algorithms are remarkable demonstrations of deep reinforcement learning’s capabilities, achieving superhuman performance in the complex game of Go with progressively increasing autonomy. However, many obstacles remain in the understanding of and usability of these promising approaches by the research community. Toward elucidating unresolved mysteries and facilitating future research, we propose ELF OpenGo, an open-source reimplementation of the AlphaZero algorithm. ELF OpenGo is the first open-source Go AI to convincingly demonstrate superhuman performance with a perfect (20:0) record against global top professionals. We apply ELF OpenGo to conduct extensive ablation studies, and to identify and analyze numerous interesting phenomena in both the model training and in the gameplay inference procedures. Our code, models, selfplay datasets, and auxiliary data are publicly available.¹

1. Introduction

The game of Go has a storied history spanning over 4,000 years and is viewed as one of the most complex turn-based board games with complete information. The emergence of AlphaGo (Silver et al., 2016) and its descendants AlphaGo Zero (Silver et al., 2017) and AlphaZero (Silver et al., 2018) demonstrated the remarkable result that deep reinforcement learning (deep RL) can achieve superhuman performance even without supervision on human gameplay datasets.

^{*}Equal contribution ¹Facebook AI Research, Menlo Park, California, USA. Correspondence to: Yuandong Tian <yuandong@fb.com>, Jerry Ma <jmaj@fb.com>, Larry Zitnick <zitnick@fb.com>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

¹Resources available at <https://facebook.ai/developers/tools/elf-opengo>. Additionally, the supplementary appendix for this paper is available at <https://arxiv.org/pdf/1902.04522.pdf>.

However, these advances in playing ability come at significant computational expense. A single training run requires millions of selfplay games and days of training on thousands of TPUs, which is an unattainable level of compute for the majority of the research community. When combined with the unavailability of code and models, the result is that the approach is very difficult, if not impossible, to reproduce, study, improve upon, and extend.

In this paper, we propose ELF OpenGo, an open-source reimplementation of the AlphaZero (Silver et al., 2018) algorithm for the game of Go. We then apply ELF OpenGo toward the following three additional contributions.

First, we train a superhuman model for ELF OpenGo. After running our AlphaZero-style training software on 2,000 GPUs for 9 days, our 20-block model has achieved superhuman performance that is arguably comparable to the 20-block models described in Silver et al. (2017) and Silver et al. (2018). To aid research in this area we provide pre-trained superhuman models, code used to train the models, a comprehensive training trajectory dataset featuring 20 million selfplay games over 1.5 million training minibatches, and auxiliary data.² We describe the system and software design in depth and we relate many practical lessons learned from developing and training our model, in the hope that the community can better understand many of the considerations for large-scale deep RL.

Second, we provide analyses of the model’s behavior during training. **(1)** As training progresses, we observe high variance in the model’s strength when compared to other models. This property holds even if the learning rates are reduced. **(2)** Moves that require significant lookahead to determine whether they should be played, such as “ladder” moves, are learned slowly by the model and are never fully mastered. **(3)** We explore how quickly the model learns high-quality moves at different stages of the game. In contrast to tabular RL’s typical behavior, the rate of progression for learning both mid-game and end-game moves is nearly identical in training ELF OpenGo.

Third, we perform extensive ablation experiments to study

²Auxiliary data comprises a test suite for difficult “ladder” game scenarios, comparative selfplay datasets, and performance validation match logs (both vs. humans and vs. other Go AIs).

the properties of AlphaZero-style algorithms. We identify several important parameters that were left ambiguous in Silver et al. (2018) and provide insight into their roles in successful training. We briefly compare the AlphaGo Zero and AlphaZero training processes. Finally, we find that even for the final model, doubling the rollouts in gameplay still boosts its strength by ≈ 200 ELO³, indicating that the strength of the AI is constrained by the model capacity.

Our ultimate goal is to provide the resources and the exploratory insights necessary to allow both the AI research and Go communities to study, improve upon, and test against these promising state-of-the-art approaches.

2. Related work

In this section, we briefly review early work in AI for Go. We then describe the AlphaGo Zero (Silver et al., 2017) and AlphaZero (Silver et al., 2018) algorithms, and the various contemporary attempts to reproduce them.

The game of Go Go is a two-player board game traditionally played on a 19-by-19 square grid. The players alternate turns, with the black player taking the first turn and the white player taking the second. Each turn, the player places a stone on the board. A player can capture groups of enemy stones by occupying adjacent locations, or “liberties”. Players can choose to “pass” their turn, and the game ends upon consecutive passes or resignation. In our setting, the game is scored at the end using Chinese rules, which award players one point for each position occupied or surrounded by the player. Traditionally, a bonus (“komi”) is given to white as compensation for going second. The higher score wins, and komi is typically a half-integer (most commonly 7.5) in order to avoid ties.

2.1. Early work

Classical search Before the advent of practical deep learning, classical search methods enjoyed initial success in AI for games. Many older AI players for board games use minimax search over a game tree, typically augmented with alpha-beta pruning (Knuth & Moore, 1975) and game-specific heuristics. A notable early result was DeepBlue (Campbell et al., 2002), a 1997 computer chess program based on alpha-beta pruning that defeated then-world champion Garry Kasparov in a six-game match. Even today, the predominant computer chess engine Stockfish uses alpha-beta pruning as its workhorse, decisively achieving superhuman performance on commodity hardware.

However, the game of Go is typically considered to be quite impervious to these classical search techniques, due to the

³Elo (1978)’s method is a commonly-used performance rating system in competitive game communities.

game tree’s high branching factor (up to $19 \times 19 + 1 = 362$) and high depth (typically hundreds of moves per game).

Monte Carlo Tree Search (MCTS) While some early Go AIs, such as GNUGo (GNU Go Team, 2009), rely on classical search techniques, most pre-deep learning AIs adopt a technique called Monte Carlo Tree Search (MCTS; Browne et al., 2012). MCTS treats game tree traversal as an exploitation/exploration tradeoff. At each state, it prioritizes visiting child nodes that provide a high value (estimated utility), or that have not been thoroughly explored. A common exploitation/exploration heuristic is “upper confidence bounds applied to trees” (“UCT”; Kocsis & Szepesvári, 2006); briefly, UCT provides an exploration bonus proportional to the inverse square root of a game state’s visit frequency. Go AIs employing MCTS include Leela (Pascutto, 2016), Pachi (Baudis & Gailly, 2008), and Fuego (Enzenberger et al., 2010).

Early deep learning for Go Early attempts at applying deep learning to Go introduced neural networks toward understanding individual game states, usually by predicting win probabilities and best actions from a given state. Go’s square grid game board and the spatial locality of moves naturally suggest the use of convolutional architectures, trained on historical human games (Clark & Storkey, 2015; Maddison et al., 2015; Tian & Zhu, 2015). AlphaGo (Silver et al., 2016) employs policy networks trained with human games and RL, value networks trained via selfplay, and distributed MCTS, achieving a remarkable 4:1 match victory against professional player Lee Sedol in 2016.

2.2. AlphaGo Zero and AlphaZero

The AlphaGo Zero (“AGZ”; Silver et al., 2017) and AlphaZero (“AZ”; Silver et al., 2018) algorithms train a Go AI using no external information except the rules of the game. We provide a high-level overview of AGZ, then briefly describe the similar AZ algorithm.

Move generation algorithm The workhorse of AGZ is a residual network model (He et al., 2016). The model accepts as input a spatially encoded representation of the game state. It then produces a scalar value prediction, representing the probability of the current player winning, and a policy prediction, representing the model’s priors on available moves given the current board situation.

AGZ combines this network with MCTS, which is used as a *policy improvement operator*. Initially informed by the network’s current-move policy, the MCTS operator explores the game tree, visiting a new game state at each iteration and evaluating the network policy. It uses a variant of PUCT (Rosin, 2011) to balance exploration (i.e. visiting game states suggested by the prior policy) and exploitation

(i.e. visiting game states that have a high value), trading off between the two with a c_{puct} constant.

MCTS terminates after a certain number of iterations and produces a new policy based on the visit frequencies of its children in the MCTS game tree. It selects the next move based on this policy, either proportionally (for early-stage training moves), or greedily (for late-stage training moves and all gameplay inference). MCTS can be multithreaded using the virtual loss technique (Chaslot et al., 2008), and MCTS’s performance intuitively improves as the number of iterations (“rollouts”) increases.

Training AGZ trains the model using randomly sampled data from a replay buffer (filled with selfplay games). The optimization objective is defined as follows, where V and \mathbf{p} are outputs of a neural network with parameters θ , and z and π are respectively the game outcome and the saved MCTS-augmented policy from the game record:

$$J(\theta) = |V(s; \theta) - z|^2 - \pi^T \log \mathbf{p}(\cdot | \theta) + c \|\theta\|^2 \quad (1)$$

There are four major components of AGZ’s training:

- The **replay buffer** is a fixed-capacity FIFO queue of game records. Each game record consists of the game outcome, the move history, and the MCTS-augmented policies at each move.
- The **selfplay workers** continually take the latest model, play out an AI vs. AI (selfplay) game using the model, and send the game record to the replay buffer.
- The **training worker** continually samples minibatches of moves from the replay buffer and performs stochastic gradient descent (SGD) to fit the model. Every 1,000 minibatches, it sends the model to the evaluator.
- The **evaluator** receives proposed new models. It plays out 400 AI vs. AI games between the new model and the current model and accepts any new model with a 55% win rate or higher. Accepted models are published to the selfplay workers as the latest model.

Performance Using Tensor Processing Units (TPUs) for selfplays and GPUs for training, AGZ is able to train a 256-filter, 20-block residual network model to superhuman performance in 3 days, and a 256-filter, 40-block model to an estimated 5185 ELO in 40 days. The total computational cost is unknown, however, as the paper does not elaborate on the number of selfplay workers, which we conjecture to be the costliest component of AGZ’s training.

AlphaZero While most details of the subsequently proposed AlphaZero (AZ) algorithm (Silver et al., 2018) are

similar to those of AGZ, AZ eliminates AGZ’s evaluation requirement, thus greatly simplifying the system and allowing new models to be immediately deployed to selfplay workers. AZ provides a remarkable speedup over AGZ, reaching in just eight hours the performance of the aforementioned three-day AGZ model.⁴ AZ uses 5,000 first-generation TPU selfplay workers; however, unlike AGZ, AZ uses TPUs for training as well. Beyond Go, the AZ algorithm can be used to train strong chess and shogi AIs.

2.3. Contemporary implementations

There are a number of contemporary open-source works that aim to reproduce the performance of AGZ and AZ, also with no external data. LeelaZero (Pascutto, 2017) leverages crowd-sourcing to achieve superhuman skill. PhoenixGo (Zeng et al., 2018), AQ (Yamaguchi, 2018), and MiniGo (MiniGo Team, 2018) achieve similar performance to that of LeelaZero. There are also numerous proprietary implementations, including “FineArt”, “Galaxy”, “DeepZen”, “Dolbaram”, “Baduki”, and of course the original implementations of DeepMind (Silver et al., 2017; 2018).

To our knowledge, ELF OpenGo is the strongest open-source Go AI at the time of writing (under equal hardware constraints), and it has been publicly verified as superhuman via professional evaluation.

2.4. Understanding AlphaGo Zero and AlphaZero

The release of AGZ and AZ has motivated a line of preliminary work (Addanki et al., 2019; Dong et al., 2017; Wang & Rompf, 2018; Wang et al., 2018) which aims to analyze and understand the algorithms, as well as to apply similar algorithms to other domains. We offer ELF OpenGo as an accelerator for such research.

3. ELF OpenGo

Our proposed ELF OpenGo aims to faithfully reimplement AGZ and AZ, modulo certain ambiguities in the original papers and various innovations that enable our system to operate entirely on commodity hardware. For brevity, we thoroughly discuss the system and software design of ELF OpenGo in Appendix A; highlights include (1) the colocation of multiple selfplay workers on GPUs to improve throughput, and (2) an asynchronous selfplay workflow to handle the increased per-game latency. Both our training and inference use NVIDIA Tesla V100 GPUs with 16 GB of memory.⁵

⁴A caveat here that hardware details are provided for AZ but not AGZ, making it difficult to compare total resource usage.

⁵One can expect comparable performance on most NVIDIA GPUs with Tensor Cores (e.g. the RTX 2060 commodity GPU).

Comparison of training details In Appendix A’s Table S1, we provide a comprehensive list of hyperparameters, resource provisioning, and other training details of AGZ, AZ, and ELF OpenGo.

To summarize, we largely adhere to AZ’s training details. Instead of 5,000 selfplay TPUs and 64 training TPUs, we use 2,000 selfplay GPUs and 8 training GPUs. Since AZ’s replay buffer size is unspecified in Silver et al. (2018), we use the AGZ setting of 500,000 games. We use the AGZ selfplay rollout setting of 1,600 per move. Finally, we use a c_{puct} constant of 1.5 and a virtual loss constant of 1.0; these settings are unspecified in Silver et al. (2017) and Silver et al. (2018), and we discuss these choices in greater detail in Section 5.

Silver et al. (2017) establish that during selfplay, the MCTS temperature is set to zero after 30 moves; that is, the policy becomes a Dirac (one-hot) distribution that selects the most visited move from MCTS. However, it is left ambiguous whether a similar temperature decrease is used during training. ELF OpenGo’s training worker uses an MCTS temperature of 1 for all moves (i.e. policy proportional to MCTS visit frequency).

Silver et al. (2018) suggest an alternative, dynamic variant of the PUCT heuristic which we do not explore. We use the same PUCT rule as that of AGZ and the initial December 2017 version of AZ.

Model training specification Our main training run constructs a 256-filter, 20-block model (starting from random initialization). First, we run our ELF OpenGo training system for 500,000 minibatches at learning rate 10^{-2} . Subsequently, we stop and restart the training system twice (at learning rates 10^{-3} , and 10^{-4}), each time for an additional 500,000 training minibatches. Thus, the total training process involves 1.5 million training minibatches, or roughly 3 billion game states. We observe a selfplay generation to training minibatch ratio of roughly 13:1 (see Appendix A). Thus, the selfplay workers collectively generate around 20 million selfplay games in total during training. This training run takes around 16 days of wall clock time, achieving superhuman performance in 9 days.

Practical lessons We learned a number of practical lessons in the course of training ELF OpenGo (e.g. staleness concerns with typical implementations of batch normalization). For brevity, we relate these lessons in Appendix C.

4. Validating model strength

We validate ELF OpenGo’s performance via (1) direct evaluation against humans and other AIs, and (2) indirect evaluation using various human and computer-generated game

datasets. Unless stated otherwise, ELF OpenGo uses a single NVIDIA Tesla V100 GPU in these validation studies, performing 1,600 rollouts per move.

Based on pairwise model comparisons, we extract our “final” model from the main training run after 1.29 million training minibatches. The final model is approximately 150 ELO stronger than the prototype model, which we immediately introduce. Both the prototype and final model are publicly available as pretrained models.

4.1. Prototype model

Human benchmarking is essential for determining the strength of a model. For this purpose, we use an early model trained in April 2018 and subsequently evaluated against human professional players. We trained this 224-filter, 20-block model using an experimental hybrid of the AGZ and AZ training processes. We refrain from providing additional details on the model, since the code was modified during training resulting in the model not being reproducible. However, it provides a valuable benchmark. We refer to this model as the “prototype model”.

We evaluate our prototype model against 4 top 30 professional players.⁶ ELF OpenGo plays under 50 seconds per move ($\approx 80,000$ rollouts), with no pondering during the opponent’s turn, while the humans play under no time limit. These evaluation games typically last for 3-4 hours, with the longest game lasting over 6 hours. Using the prototype model, ELF OpenGo won every game for a final record of 20:0.

Fig. 1 depicts the model’s predicted value during eight selected human games; this value indicates the perceived advantage of ELF OpenGo versus the professionals over the course of the game. Sudden drops of the predicted value indicate that ELF OpenGo has various weaknesses (e.g. ladder exploitation); however, the human players ultimately proved unable to maintain the consequent advantages for the remainder of the game.

Evaluation versus other AIs We further evaluate our prototype model against LeelaZero (Pascutto, 2017), which at the time of evaluation was the strongest open-source Go AI.⁷ Both LeelaZero and ELF OpenGo play under a time limit of 50 seconds per move. ELF OpenGo achieves an overall record of 980:18, corresponding to a win rate of

⁶The players (and global ranks as of game date) include Kim Ji-seok (#3), Shin Jin-seo (#5), Park Yeonghun (#23), and Choi Cheolhan (#30). All four players were fairly compensated for their expertise, with additional and significant incentives for winning versus ELF OpenGo. Each player played 5 games, for a total of 20 human evaluation games.

⁷2018 April 25 model with 192 filters and 15 residual blocks; public hash 158603eb.

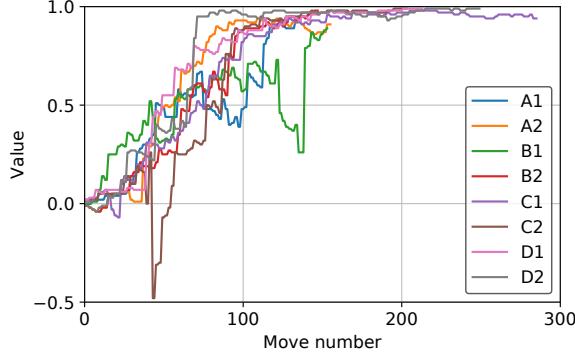


Figure 1. Predicted values after each move for 8 selected games versus human professional players (2 games per player). A positive value indicates that the model believes it is more likely to win than the human. Per players’ request, we anonymize the games by arbitrarily labeling the players with letters ‘A’ to ‘D’.

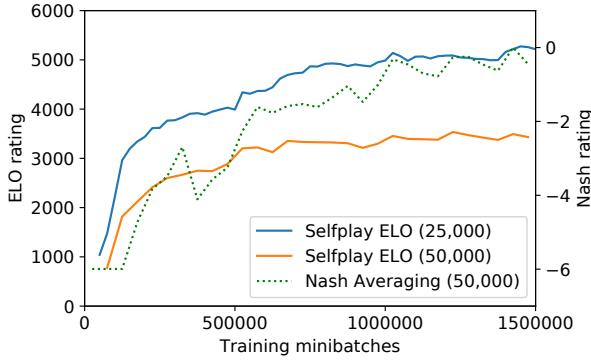


Figure 2. Progression of model skill during training. “Selfplay ELO 25,000” and “Selfplay ELO 50,000” refer to the unnormalized selfplay ELO rating, calculated based on consecutive model pairs at intervals of 25,000 and 50,000 training minibatches, respectively. “Nash Averaging 50,000” refers to the Nash averaging rating (Balduzzi et al., 2018), calculated based on a pairwise tournament among the same models as in “Selfplay ELO 50,000”.

98.2% and a skill differential of approximately 700 ELO.

4.2. Training analysis

Throughout our main training run, we measure the progression of our model against the human-validated prototype model. We also measure the agreement of the model’s move predictions with those made by humans. Finally, we explore the rate at which the model learns different stages of the game and complex moves, such as ladders.

Training progression: selfplay metrics In examining the progression of model training, we consider two selfplay rating metrics: selfplay ELO and Nash averaging. Selfplay ELO uses Elo (1978)’s rating formula in order to determine the rating difference between consecutive pairs of models,

where each pair’s winrate is known. Nash averaging (Balduzzi et al., 2018) calculates a logit-based payoff of each model against a mixed Nash equilibrium, represented as a discrete probability distribution over each model.

Intuitively, the selfplay ELO can be viewed as an “inflated” metric for two main reasons. First, the model’s rating will increase as long as it can beat the immediately preceding model, without regard to its performance against earlier models. Second, the rating is sensitive to the number of consecutive pairs compared; increasing this will tend to boost the rating of each model due to the nonlinear logit form of the ELO calculation.

Fig. 2 shows the selfplay ELO rating using every 25,000-th model and every 50,000-th model, and the Nash averaging rating using every 50,000-th model. The ratings are consistent with the above intuitions; the ELO ratings follow a mostly consistent upward trend, and the denser comparisons lead to a more inflated rating. Note that the Nash averaging rating captures model skill degradations (particularly between minibatches 300,000 and 400,000) that selfplay ELO fails to identify.

Training progression: training objective **Fig. 3a** shows the progression of the policy and value losses. Note the initial dip in the value loss. This is due to the model overestimating the white win rate, causing the black player to resign prematurely, which reduces the diversity of the games in the replay buffer. This could result in a negative feedback loop and overfitting. ELF OpenGo automatically corrects for this by evenly sampling black-win and white-win games. With this diverse (qualitatively, “healthy”) set of replays, the value loss recovers and stays constant throughout the training, showing there is always new information to learn from the replay buffer.

Performance versus prototype model **Fig. 3b** shows the progression of the model’s win rates against two weaker variants taken earlier in training of the prototype model (“prototype- α ” and “prototype- β ”) as well as the main human-validated prototype model (simply “prototype”).⁸ We observe that the model trends stronger as training progresses, achieving parity with the prototype after roughly 1 million minibatches, and achieving a 60% win rate against the prototype at the end of training. Similar trends emerge with prototype- α and prototype- β , demonstrating the robustness of the trend to choice of opponent. Note that while the strength of the model trends stronger, there is significant variance in the model’s strength as training progresses, even with weaker models. Surprisingly, this variance does not decrease even with a decrease in learning rate.

⁸Prototype- α is roughly at the advanced amateur level, and prototype- β is roughly at a typical professional level.

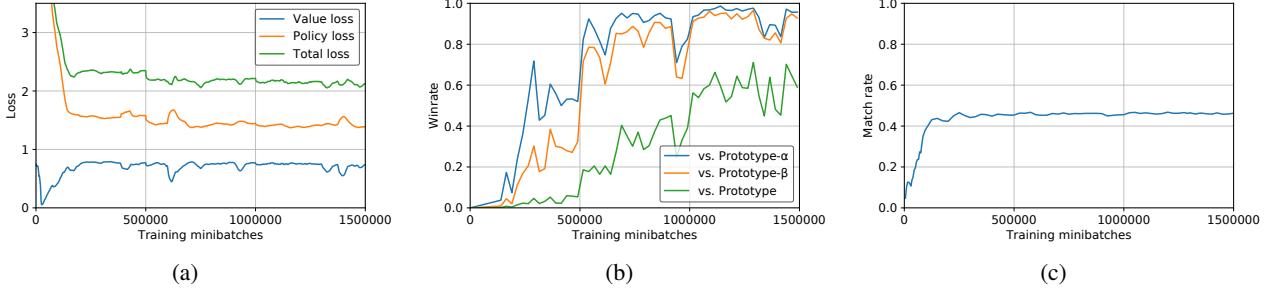


Figure 3. (a) Model’s training loss (value, policy, and sum) during training. (b) Win rates vs. the prototype model during training. (c) Match rate between trained model and human players on moves from professional games (as described in Appendix B.1). The learning rate was decreased every 500,000 minibatches.

Examining the win rates versus other models in Fig. 3b, the number of minibatches could potentially be reduced to 250,000 at each learning rate, and still similar performance can be achieved.

Comparison with human games Since our model is significantly stronger than the prototype model that showed superhuman strength, we hypothesize that our model is also of superhuman strength. In Fig. 3c, we show the agreement of predicted moves with those made by professional human players. The moves are extracted from 1,000 professional games played from 2011 to 2015. The model quickly converges to a human move match rate of $\approx 46\%$ around minibatch 125,000. This indicates that the strengthening of the model beyond this point may not be due to better human professional predictions, and as demonstrated by Silver et al. (2016), there may be limitations to supervised training from human games.

Learning game stages An interesting question is whether the model learns different stages of the game at different rates. During training, is the model initially stronger at opening or endgame moves?

We hypothesize that the endgame should be learned earlier than the opening. With a random model, MCTS behaves randomly except for the last few endgame moves, in which the actual game score is used to determine the winner of the game. Then, after some initial training, the learned endgame signals inform MCTS in the immediately preceding moves as well. As training progresses, these signals flow to earlier and earlier move numbers via MCTS, until finally the opening is learned.

In Fig. 4a and Fig. 4b, we show the agreement of predicted moves from the model during training, and the prototype model and humans respectively. Fig. 4a shows the percentage of moves at three stages of gameplay (moves 1-60, 61-120, and 121-180) from training selfplay games that agree with those predicted by the prototype model. Fig. 4b

is similar, but it uses moves from professional human games, and measures the match rate between the trained model and professional players. Consistent with our hypothesis, the progression of early games moves (1-60) lags behind that of the mid-game moves (61-120) and the end-game moves (121-180). Upon further exploration, we observed that this lag resulted from some initial overfitting before eventually recovering. Counter to conventional wisdom, the progression of match rates of mid-game and end-game moves are nearly identical. Note that more ambiguity exists for earlier moves than later moves, so after sufficient training the match rates converge to different match rates.

Ladder moves “Ladder” scenarios are among the earliest concepts of Go learned by beginning human players. Ladders create a predictable sequence of moves for each player that can span the entire board as shown in Fig. 5. In contrast to humans, Silver et al. (2017) observes that deep RL models learn these tactics very late in training. To gain further insight into this phenomenon, we curate a dataset of 100 ladder scenarios (as described in Appendix B.2) and evaluate the model’s ladder handling abilities throughout training.

As shown in Fig. 4c, we observe that the ability of the network to correctly handle ladder moves fluctuates significantly over the course of training. In general, ladder play improves with a fixed learning rate and degrades after the learning rate is reduced before once again recovering. While the network improves on ladder moves, it still makes mistakes even late in training. In general, increasing the number of MCTS rollouts from 1,600 to 6,400 improves ladder play, but mistakes are still made.

5. Ablation study

We employ ELF OpenGo to perform various ablation experiments, toward demystifying the inner workings of MCTS and AZ-style training.

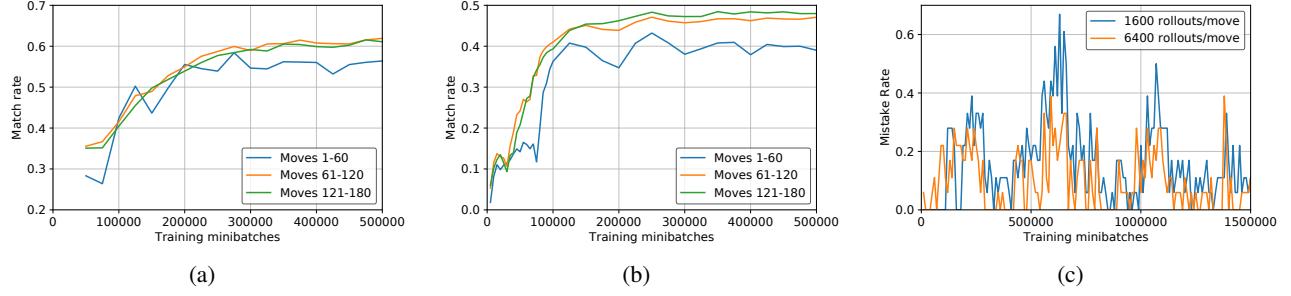


Figure 4. (a) Move agreement rates with the prototype model at different stages of the game during training. Note that the model learns better end-game moves before learning opening moves. The first part of the figure is clipped due to short-game dominance cased by initial overfitting issues. (b) Move agreement rates with human games at different stages of the game during the first stage of training. (c) Model’s rate of “mistake moves” on the ladder scenario dataset during training, associated with vulnerability to ladder scenarios.

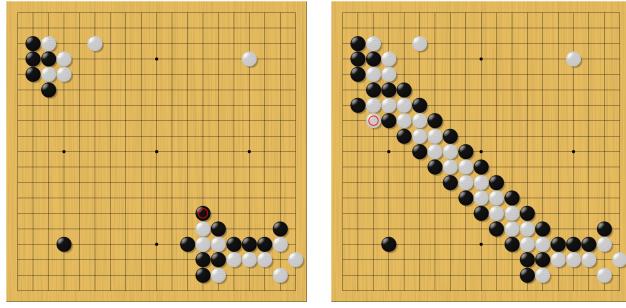


Figure 5. The beginning (left) and end (right) of a ladder scenario in which OpenGo (black) mistakenly thought it could gain an advantage by playing the ladder. Whether a ladder is advantageous or not is commonly dependent on stones distant from its start.

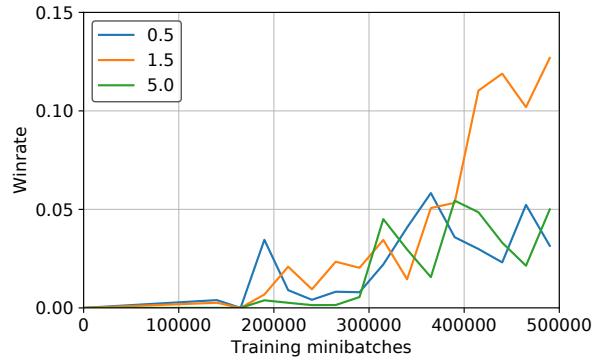


Figure 6. Win rates vs. the prototype model during an abbreviated training run for three different c_{puct} values. The learning rate is decayed 10-fold after 290,000 minibatches.

5.1. PUCT constant

Both AGZ (Silver et al., 2017) and AZ (Silver et al., 2018) leave c_{puct} unspecified. Recall that c_{puct} controls the balance of exploration vs. exploitation in the MCTS algorithm. Thus, it is a crucial hyperparameter for the overall behavior and effectiveness of the algorithm. Setting c_{puct} too low results in insufficient exploration, while setting c_{puct} too high reduces the effective depth (i.e. long-term planning capacity) of the MCTS algorithm. In preliminary experimentation, we found $c_{\text{puct}} = 1.5$ to be a suitable choice. Fig. 6 depicts the ELF OpenGo model’s training trajectory under various values of c_{puct} ; among the tested values, $c_{\text{puct}} = 1.5$ is plainly the most performant.

5.2. MCTS virtual loss

0.1	0.2	0.5	1.0	2.0	3.0	5.0
22%	37%	49%	50%	32%	36%	30%

Table 1. Win rates of various virtual loss constants versus the default setting of 1.

Virtual loss (Chaslot et al., 2008) is a technique used to ac-

celerate multithreaded MCTS. It adds a temporary and fixed amount of loss to a node to be visited by a thread, to prevent other threads from concurrently visiting the same node and causing congestion. The virtual loss is parameterized by a constant. For AGZ, AZ, and ELF OpenGo’s value function, a virtual loss constant of 1 is intuitively interpretable as each thread temporarily assuming a loss for the moves along the current rollout. This motivates our choice of 1 as ELF OpenGo’s virtual loss constant.

We perform a sweep over the virtual loss constant using the prototype model, comparing each setting against the default setting of 1. The results, presented in Table 1, suggest that a virtual loss constant of 1 is indeed reasonable.

5.3. AlphaGo Zero vs. AlphaZero

We hypothesize that AZ training vastly outperforms AGZ training (given equivalent hardware) due to the former’s asynchrony and consequent improved throughput. We train 64-filter, 5-block models from scratch with both AGZ and AZ for 12 hours and compare their final performance. The model trained with AZ wins 100:0 versus the model trained

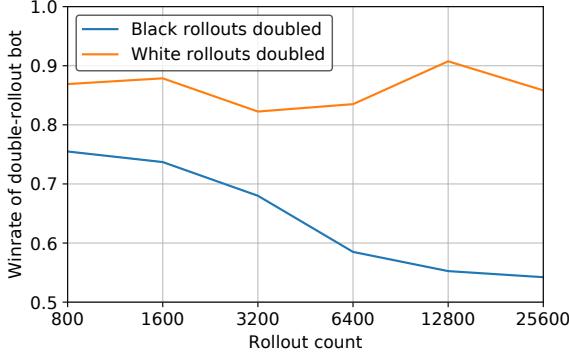


Figure 7. Win rate when playing with 2x rollouts against the same model with 1x rollouts.

with AGZ, indicating that AZ is indeed much more performant.

5.4. MCTS rollout count

Intuitively, increasing the number of MCTS iterations (rollout count) improves the AI’s strength by exploring more of the game tree. Toward better understanding the rollout count’s impact on strength, we perform a selfplay analysis with the final model, in which one player uses twice as many MCTS rollouts as the other. We perform this analysis across a wide range of rollout counts (800-25,600).

From the results shown in Fig. 7, we find that ELF OpenGo consistently enjoys an 80-90% win rate ($\approx 250\text{-}400$ additional ELO) from doubling the number of rollouts as the white player. On the other hand, as the black player, ELF OpenGo only enjoys a 55-75% win rate ($\approx 35\text{-}200$ additional ELO). Moreover, the incremental benefit for black of doubling rollouts shrinks to nearly 50% as the rollout count is increased, suggesting that our model has a skill ceiling with respect to rollout count as the black player. That this skill ceiling is not present as the white player suggests that a 7.5 komi (white score bonus) can be quite significant for black.

Since using the same model on both sides could introduce bias (the player with more rollouts sees all the branches explored by the opponent), we also experiment with the prototype/final model and still observe a similar trend that doubling the rollouts gives ≈ 200 ELO boost.

6. Discussion

ELF OpenGo learns to play the game of Go differently from humans. It requires orders of magnitude more games than professional human players to achieve the same level of performance. Notably, ladder moves, which are easy for human beginners to understand after a few examples, are difficult for the model to learn and never fully mastered. We suspect

that this is because convolutional networks lack the right inductive bias to handle ladders and resort to memorizing specific situations. Finally, we observe significant variance during training, which indicates the method is still not fully understood and may require much tuning. We hope this paper serves as a starting point for improving upon AGZ/AZ to achieve efficient and stable learning.

Surprisingly, further reduction of the learning rate does not improve training. We trained to 2 million minibatches with a learning rate of 10^{-5} , but noticed minimal improvement in model strength. Furthermore, the training becomes unstable with high variance. This may be due to a lack of diversity in the selfplay games, since either the model has saturated or the lower learning rate of 10^{-5} resulted in the selfplay games coming from nearly identical models. It may be necessary to increase the selfplay game buffer when using lower learning rates to maintain stability.

Finally, RL methods typically learn from the states that are close to the terminal state (end game) where there is a sparse reward. Knowledge from the reward is then propagated towards the beginning of the game. However, from Fig. 4a and Fig. 4b, such a trend is weak (moves 61-180 are learned only slightly faster than moves 1-60). This brings about the question of why AGZ/AZ methods behave differently, and what kind of role the inductive bias of the model plays during the training. It is likely that due to the inductive bias, the model quickly predicts the correct moves and values of easy situations, and then focuses on the difficult cases.

7. Conclusion

We provide a reimplementation of AlphaZero (Silver et al., 2018) and a resulting Go engine capable of superhuman gameplay. Our code, models, selfplay datasets and auxiliary data are publicly available. We offer insights into the model’s behavior during training. Notably, we examine the variance of the model’s strength, its ability to learn ladder moves, and the rate of improvement at different stages of the game. Finally, through a series of ablation studies, we shed light on parameters that were previously ambiguous. Interestingly, we observe significant and sustained improvement with the number of rollouts performed during MCTS when playing white, but diminishing returns when playing black. This indicates that the model’s strength could still be significantly improved.

Our goal is to provide the insights, code, and datasets necessary for the research community to explore large-scale deep reinforcement learning. As demonstrated through our experiments, exciting opportunities lie ahead in exploring sample efficiency, reducing training volatility, and numerous additional directions.

References

- Addanki, R., Alizadeh, M., Venkatakrishnan, S. B., Shah, D., Xie, Q., and Xu, Z. Understanding & generalizing alphago zero, 2019. URL <https://openreview.net/forum?id=rkxtl3C5YX>.
- Balduzzi, D., Tuyls, K., Pérolat, J., and Graepel, T. Re-evaluating evaluation. *CoRR*, abs/1806.02643, 2018. URL <http://arxiv.org/abs/1806.02643>.
- Baudis, P. and Gailly, J. Pachi. <https://github.com/pasky/pachi>, 2008.
- Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.
- Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfsen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.
- Campbell, M., Hoane Jr, A. J., and Hsu, F.-h. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.
- Chaslot, G. M.-B., Winands, M. H., and van Den Herik, H. J. Parallel monte-carlo tree search. In *International Conference on Computers and Games*, pp. 60–71. Springer, 2008.
- Clark, C. and Storkey, A. Training deep convolutional neural networks to play go. In *International Conference on Machine Learning*, pp. 1766–1774, 2015.
- Coleman, C., Narayanan, D., Kang, D., Zhao, T., Zhang, J., Nardi, L., Bailis, P., Olukotun, K., Ré, C., and Zaharia, M. Dawnbench: An end-to-end deep learning benchmark and competition. 2017.
- Dong, X., Wu, J., and Zhou, L. Demystifying alphago zero as alphago gan. *arXiv preprint arXiv:1711.09091*, 2017.
- Elo, A. E. The rating of chess players, past and present. *Ishi Press International*, 1978.
- Enzenberger, M., Muller, M., Arneson, B., and Segal, R. Fuego: an open-source framework for board games and go engine based on monte carlo tree search. *IEEE Transactions on Computational Intelligence and AI in Games*, 2(4):259–270, 2010.
- GNU Go Team. Gnu go. <https://www.gnu.org/software/gnugo/>, 2009.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 448–456, 2015. URL <http://jmlr.org/proceedings/papers/v37/ioffe15.html>.
- Knuth, D. E. and Moore, R. W. An analysis of alpha-beta pruning. *Artificial intelligence*, 6(4):293–326, 1975.
- Kocsis, L. and Szepesvári, C. Bandit based monte-carlo planning. In *European conference on machine learning*, pp. 282–293. Springer, 2006.
- Maddison, C. J., Huang, A., Sutskever, I., and Silver, D. Move evaluation in go using deep convolutional neural networks. *International Conference on Learning Representations*, 2015.
- Minigo Team. Minigo. <https://github.com/tensorflow/minigo>, 2018.
- Pascutto, G.-C. Leela. <https://www.sjeng.org/leela.html>, 2016.
- Pascutto, G.-C. Leelazero. <https://github.com/gcp/leela-zero>, 2017.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. In *NIPS-W*, 2017.
- Rosin, C. D. Multi-armed bandits with episode context. *Annals of Mathematics and Artificial Intelligence*, 61(3):203–230, 2011.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. ISSN 0036-8075. doi: 10.1126/science.aar6404. URL <http://science.sciencemag.org/content/362/6419/1140>.

Tian, Y. and Zhu, Y. Better computer go player with neural network and long-term prediction. *CoRR*, abs/1511.06410, 2015. URL <http://arxiv.org/abs/1511.06410>.

Tian, Y., Gong, Q., Shang, W., Wu, Y., and Zitnick, C. L. ELF: an extensive, lightweight and flexible research platform for real-time strategy games. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 2656–2666, 2017.

Tromp, J. <http://tromp.github.io/go.html>, 2014.

Wang, F. and Rompf, T. From gameplay to symbolic reasoning: Learning sat solver heuristics in the style of alpha (go) zero. *arXiv preprint arXiv:1802.05340*, 2018.

Wang, L., Zhao, Y., and Jinnai, Y. Alphax: exploring neural architectures with deep neural networks and monte carlo tree search. *CoRR*, abs/1805.07440, 2018. URL <http://arxiv.org/abs/1805.07440>.

Yamaguchi, Y. Aq. <https://github.com/ymgao/AQ>, 2018.

Zeng, Q., Zhang, J., Zeng, Z., Li, Y., Chen, M., and Liu, S. Phoenixgo. <https://github.com/Tencent/PhoenixGo>, 2018.

A. Detailed system and software design of ELF OpenGo

We now describe the system and software design of ELF OpenGo, which builds upon the DarkForest Go engine (Tian & Zhu, 2015) and the ELF reinforcement learning platform (Tian et al., 2017).

A.1. Distributed platform

ELF OpenGo is backed by a modified version of ELF (Tian et al., 2017), an Extensive, Lightweight and Flexible platform for reinforcement learning research; we refer to the new system as ELF⁺⁺. Notable improvements include:

- **Distributed support.** ELF⁺⁺ adds support for distributed asynchronous workflows, supporting up to 2,000 clients.
- **Increased Python flexibility.** We provide a distributed adapter that can connect any environment with a Python API, such as OpenAI Gym (Brockman et al., 2016).
- **Batching support.** We provide a simple autobatching registry to facilitate shared neural network inference across multiple game simulators.

A.2. Go engine

We have integrated the DarkForest Go engine (Tian & Zhu, 2015) into ELF OpenGo, which provides efficient handling of game dynamics (approximately $1.2 \mu\text{s}$ per move). We use the engine to execute game logic and to score the terminal game state using Tromp-Taylor rules (Tromp, 2014).

A.3. Training system

ELF OpenGo largely replicates the training system architecture of AlphaZero. However, there are a number of differences motivated by our compute capabilities.

Hardware details We use NVIDIA V100 GPUs for our selfplay workers. Every group of eight GPUs shares two Intel Xeon E5-2686v4 processors.

We use a single training worker machine powered by eight V100 GPUs. Increasing the training throughput via distributed training does not yield considerable benefit for ELF OpenGo, as our selfplay throughput is much less than that of AlphaZero.

To elaborate on this point, our ratio of selfplay games to training minibatches with a single training worker is roughly 13:1. For comparison, AlphaZero’s ratio is 30:1 and AlphaGo Zero’s ratio is 7:1. We found that decreasing this ratio significantly below 10:1 hinders training (likely due

to severe overfitting). Thus, since adding more training workers proportionally decreases this ratio, we refrain from multi-worker training.

GPU colocation of selfplay workers We use GPUs instead of TPUs to evaluate the residual network model. The primary difference between the two is that GPUs are much slower for residual networks.⁹ Neural networks on GPUs also benefit from batching the inputs; in our case, we observed near-linear throughput improvements from increasing the batch size up to 16, and sublinear but still significant improvements between 16 and 128.

Thus, to close the gap between GPU and TPU, we co-locate 32 selfplay workers on each GPU, allowing the GPU to process inputs from multiple workers in a single batch. Since each worker has 8 game threads, this implies a theoretical maximum of 256 evaluations per batch. In practice, we limit the batch size to 128.

This design, along with the use of half-precision floating point computation, increases the throughput of each ELF OpenGo GPU selfplay worker to roughly half the throughput of a AlphaGo Zero TPU selfplay worker. While AlphaGo Zero reports throughput of 2.5 moves per second for a 256-filter, 20-block model with 1,600 MCTS rollouts, ELF OpenGo’s throughput is roughly 1 move per second.

Asynchronous, heterogenous-model selfplays This colocation, along with the inherent slowness of GPUs relative to TPUs, results in much higher latency for game generation (on the order of an hour). Since our training worker typically produces a new model (i.e. processes 1,000 minibatches) every 10 to 15 minutes, new models are published faster than a single selfplay can be produced.

There are two approaches to handling this:

Synchronous (AlphaGo Zero) mode In this mode, there are two different kinds of clients: Selfplay clients and Eval clients. Once the server has a new model, all the Selfplay clients discard the current game being played, reload the new model and restart selfplays, until a given number of selfplays have been generated. Then the server starts to update the current model according to Equation 1. Every 1,000 minibatches, the server updates the model and notifies Eval clients to compare the new model with the old one. If the new model is better than the current one by 55%, then the server notifies all the clients to discard current games, and restart the loop. On the server side, the selfplay games from the previous model can either be removed from

⁹According to December 2018 DawnBench (Coleman et al., 2017) results available at <https://dawn.cs.stanford.edu/benchmark/ImageNet/train.html>, one TPU has near-equivalent 50-block throughput to that of eight V100 GPUs

Parameter/detail	AGZ	AZ	ELF OpenGo
c_{puct} (PUCT constant)	?	?	1.5
MCTS virtual loss constant	?	?	1.0
MCTS rollouts (selfplay)	1,600	800	1,600
Training algorithm	SGD with momentum = 0.9		
Training objective	value squared error + policy cross entropy + $10^{-4} \cdot L_2$		
Learning rate	$10^{\{-2,-3,-4\}}$	$2 \cdot 10^{\{-2,-3,-4\}}$	$10^{\{-2,-3,-4\}}$
Replay buffer size	500,000	?	500,000
Training minibatch size	2048	4096	2048
Selfplay hardware	?	5,000 TPUs	2,000 GPUs
Training hardware	64 GPUs	64 TPUs	8 GPUs
Evaluation criteria	55% win rate	none	none

Table S1. Hyperparameters and training details of AGZ, AZ, and ELF OpenGo. “?” denotes a detail that was ambiguous or unspecified in Silver et al. (2017) or Silver et al. (2018)

the replay buffer or be retained. Otherwise, the `Selfplay` clients send more selfplays of the current model until an additional number of selfplays are collected by the server, and then the server starts another 1,000 batches of training. This procedure repeats until a new model passes the win rate threshold.

Asynchronous (AlphaZero) mode Note that AlphaGo Zero mode involves a lot of synchronization and is not efficient in terms of boosting the strength of the trained model. In AlphaZero mode, we release all the synchronization locks and remove `Eval` clients. Moves are always generated using the latest models and `Selfplay` clients do not terminate their current games upon receiving new models. It is possible that for a given selfplay game record, the first part of the game is played by model A while the second part of the game is played by model B.

We initially started with the synchronous approach before switching to the asynchronous approach. Switching offered two benefits: (1) Both selfplay generation and training realized a drastic speedup. The asynchronous approach achieves over 5x the selfplay throughput of the synchronous approach on our hardware setup. (2) The ratio of selfplay games to training minibatches increased by roughly 1.5x, thus helping to prevent overfitting.

The downside of the asynchronous approach is losing homogeneity of selfplays – each selfplay is now the product of many consecutive models, reducing the internal coherency of the moves. However, we note that the replay buffer that provides training data is already extremely heterogeneous, typically containing games from over 25 different models. Consequently, we suspect and have empirically verified that the effect of within-selfplay heterogeneity is mild.

A.4. Miscellany

Replay buffer On the server side, we use a large replay buffer (500,000 games) to collect game records by clients. Consistent with Silver et al. (2017), who also use a replay buffer of 500,000 games, we found that a large replay buffer yields good performance. To increase concurrency (reading from multiple threads of feature extraction), the replay buffer is split into 50 queues, each with a maximal size of 10,000 games and a minimal size of 200 games. Note that the minimal size is important, otherwise the model often starts training on a very small set of games and quickly overfits before more games arrive.

Fairness of model evaluation In synchronous mode, the server deals with various issues (e.g., clients die or taking too long to evaluate) and makes sure evaluations are done in an unbiased manner. Note that a typically biased estimation is to send 1,000 requests to `Eval` clients and conclusively calculate the win rate using the first 400 finished games. This biases the metric toward shorter games, to training’s detriment.

Game resignation Resigning from a hopeless game is very important in the training process. This not only saves much computation but also shifts the selfplay distribution so that the model focuses more on the midgame and the opening after learning the basics of Go. As such, the model uses the bulk of its capacity for the most critical parts of the game, thus becoming stronger. As in Silver et al. (2017), we dynamically calibrate our resignation threshold to have a 5% false positive rate; we employ a simple sliding window quantile tracker.

B. Auxiliary dataset details

B.1. Human games dataset

To construct the human game dataset, we randomly sample 1,000 professional games from the Gogod database from 2011 to 2015.¹⁰

B.2. Ladder dataset

We collect 100 games containing ladder scenarios from the online CGOS (Computer Go Server) service, where we deployed our prototype model.¹¹ For each game, we extract the decisive game state related to the ladder. We then augment the dataset 8-fold via rotations and reflections.

C. Practical lessons

ELF OpenGo was developed through much iteration and bug-fixing on both the systems and algorithm/modeling side. Here, we relate some interesting findings and lessons learned from developing and training the AI.

Batch normalization moment staleness Our residual network model, like that of AGZ and AZ, uses batch normalization (Ioffe & Szegedy, 2015). Most practical implementations of batch normalization use an exponentially weighted buffer, parameterized by a “momentum constant”, to track the per-channel moments. We found that even with relatively low values of the momentum constant, the buffers would often be stale (biased), resulting in subpar performance.

Thus, we adopt a variant of the postprocessing moment calculation scheme originally suggested by Ioffe & Szegedy (2015). Specifically, after every 1,000 training minibatches, we evaluate the model on 50 minibatches and store the simple average of the activation moments in the batch normalization layers. This eliminates the bias in the moment estimators, resulting in noticeably improved and consistent performance during training. We have added support for this technique to the PyTorch framework (Paszke et al., 2017).¹²

Dominating value gradients We performed an unintentional ablation study in which we set the cross entropy coefficient to $\frac{1}{362}$ during backpropagation. This change results in optimizing the value network much faster than the policy network. We observe that with this modification, ELF OpenGo can still achieve a strength of around amateur dan level. Further progress is extremely slow, likely due to the minimal gradient provided by the policy network. This sug-

gests that any MCTS augmented with only a value heuristic has a relatively low skill ceiling in Go.

¹⁰<https://gogodonline.co.uk/>

¹¹<http://www.yss-aya.com/cgos/19x19/standings.html>

¹²As of December 2018, this is configurable by setting momentum=None in the BatchNorm layer constructor.