FISCHERGPT: INTERPRETING A TRANSFORMER-BASED CHESS ENGINE VIA ATTENTION WEIGHTS AND NEURON-LEVEL CONCEPT DISCOVERY

Jay Lalwani (nkr4rj@virginia.edu)
University of Virginia, CS 6501

Sharon Biju (fuu2ka@virginia.edu)
University of Virginia, CS 6501

ABSTRACT

We explore the internal workings of DeepMind's "Searchless Chess" model, a transformer-based chess engine reaching grandmaster-level play without traditional lookahead search. Our twofold approach involves: (1) extracting and visualizing the model's attention weights to reveal how it attends to different squares and pieces, and (2) identifying which transformer heads and neurons encode specific chess concepts (e.g., king safety, piece coordination, or tactical motifs). We find that smaller-scale (9M-parameter) versions of the model exhibit human-like focus on critical squares, while the largest (270M-parameter) model often concentrates on less obvious regions of the board. By mapping "firing" neurons or heads to chess-specific heuristics, our work illuminates how a purely feedforward policy can achieve high-quality moves without explicit search, offering broader insights into interpretability and concept learning in large transformers.

1 INTRODUCTION TO SEARCHLESS CHESS

Search-based methods have long been central to high-level chess AI, most notably with classical engines (e.g., Stockfish) that rely on alpha-beta pruning and handcrafted evaluation functions, and with modern reinforcement learning systems such as AlphaZero (Silver et al., 2017), which integrate Monte Carlo Tree Search. However, DeepMind's "Searchless Chess" paradigm challenges this tradition. Instead of exhaustively exploring the game tree, it learns directly from Stockfish-annotated examples, using a large-scale transformer to infer action-values from a single board state (, DeepMind).

Empirically, Searchless Chess reaches grandmaster-level performance in blitz games, achieving ratings above 2800 on popular online platforms. Despite its success, the core process remains opaque: it appears to evaluate positions "in one shot," without iterative lookahead. This raises compelling questions: *How does a transformer, trained solely on static board states, internalize tactical and strategic patterns typically gleaned from search*? And *which tokens, squares, or concepts does it attend to when predicting moves*? Our work attempts to make these internal mechanisms more transparent by focusing on two interpretability tasks: (1) analyzing attention distributions over the board, and (2) identifying neurons or attention heads that correlate with key chess motifs.

2 PROBLEM SETUP

Our project explores how a transformer-based chess model can replicate or approximate the capabilities of a search-based engine without an explicit lookahead procedure. Specifically, we focus on two interpretability questions:

(1) Attention Analysis. Given a board state in Forsyth-Edwards Notation (FEN), the transformer processes a tokenized representation of each square and piece. We aim to visualize the attention maps that form part of the model's forward pass. By overlaying these attention distributions onto the chessboard, we can see which pieces and squares the model emphasizes when evaluating a move.

(2) Neuron- and Head-Level Concept Discovery. Transformer-based language models have been shown to learn interpretable "concept neurons" (Bau et al., 2019; Radford et al., 2017). Extending this notion to chess, we examine whether specific attention heads or feedforward neurons respond to distinct chess motifs (e.g., pinned pieces, discovered checks, or pawn-structure-related features). By identifying specialized internal components, we gain insight into how chess knowledge may be distributed and compositional within the network.

2.1 Motivation and Impact

Understanding how large-scale transformers can learn complex decision-making heuristics—without recursive search—has implications beyond chess. Interpretability in strategic domains could inform how we design or trust such models in fields requiring transparent decision rationales (e.g., medical diagnostics or resource allocation). This project thus merges chess AI with broader questions of neural interpretability and concept learning, shedding light on black-box decision processes.

3 Method

3.1 Model Overview

The Searchless Chess model is a 270-million-parameter decoder-only transformer (Vaswani et al., 2017). It ingests a FEN string tokenized into discrete embeddings (each square becomes one or more tokens, encoding piece type, color, or emptiness). During training, the model was provided with Stockfish-based win probability bins for every legal move, effectively predicting the action-value distribution. Our analysis considers both:

- **Small Model (9M params).** A reduced-scale version to test if interpretability is more straightforward in smaller networks.
- **Full Model (270M params).** The grandmaster-level model with higher complexity and potentially more elaborate internal representations.

3.2 Attention Extraction and Board Visualization

We developed a pipeline to forward a given FEN through the pretrained model and extract its multi-head self-attention weights. Each attention score corresponds to how strongly the model focuses on one token (e.g., a square or special marker like castling rights) when processing another. We then map these tokens to their corresponding board squares, summing or averaging across attention heads for a global view. This produces attention heatmaps over the chessboard, revealing spatial patterns of focus.

Implementation Details. We use a custom function that traverses the model's attention tensors and aligns each token index with the board. By blending out-of-board tokens (e.g., end-of-sequence markers), we can colorize squares by attention intensity. This procedure runs identically on both the small and full versions of the model.

3.3 Neuron- and Head-Level Analysis

To investigate concept-specific activations:

1. **Concept Tagging.** We label board states for known chess motifs (e.g., pinned piece, threatened king, advanced passed pawn). A script identifies these motifs by scanning the board for relative piece positions.
2. **Activation Tracing.** We record each transformer's feedforward-layer neuron outputs and each attention head's queries/keys/values for states that contain (or do not contain) a given motif.
3. **Statistical Significance.** For each neuron or head, we evaluate whether its activations differ significantly when a motif is present vs. absent. High contrast suggests specialized "concept encoding."

Our goal is to determine whether advanced models exhibit more distinct or “entangled” concept neurons/heads compared to smaller variants. Ultimately, we want to discover whether certain attention heads systematically fire whenever there’s a discovered check, pinned piece, or other critical pattern.

4 Experimental Setup

4.1 Datasets

We rely on the pretrained models and a curated set of test positions:

- **Tactical Puzzles.** Positions highlighting forks, pins, skewers, or forced mate sequences.
- **Positional Scenarios.** Positions emphasizing pawn structures, open files, and other longer-term strategic elements.
- **Obscure Examples.** Unusual compositions or endgame studies that push the model’s internal representations to their limits.

These positions allow us to observe how attention distributions and neuron activations vary in distinctly tactical versus strategic contexts, as well as in edge cases.

4.2 Evaluation Measures

Since our focus is interpretability, we qualitatively and quantitatively assess:

Attention Alignment. For a given board, we compare the squares with the highest attention to those a human expert would deem most relevant. We also compute the fraction of total attention devoted to the piece or area of immediate tactical concern.

Concept Activation. We measure how strongly each head or neuron responds to states labeled with a specific motif. A high activation in the presence of pinned pieces, for instance, implies that the model’s internal representation captures the notion of a “pin.”

Comparisons Across Scales. We contrast the small (9M) with the large (270M) model. Preliminary observations suggest that the smaller model’s attention heatmaps align more clearly with key squares or threats, whereas the larger model sometimes invests significant attention in squares that appear irrelevant to humans. We aim to discern whether this broader focus is due to deeper strategic insight or an artifact of overfitting.

5 Results Obtained So Far

5.1 Preliminary Attention Heatmaps

- **9M-Parameter Model.** On standard tactical puzzles (e.g., a simple knight fork), the top attention heads collectively focus on the threatened pieces and the squares around them, yielding heatmaps closely aligning with human intuition.
- **270M-Parameter Model.** While it often highlights the critical regions, certain heads consistently assign high attention to less obvious squares. The model still converges on strong moves, suggesting hidden or global signals that are not immediately intuitive to human observers.

5.2 Concept Neurons and Heads

Our concept-tagging pipeline is nearly complete, as the dataset is being finalized and rebalanced.

6 Challenges and Solutions

6.1 Installation and Resource Constraints

Setting up the 270M-parameter model requires substantial GPU memory and computational resources, complicating the process of extracting full multi-head attention tensors. We addressed this by running partial checkpointing and micro-batching to reduce memory footprints.

6.2 Interpretability Difficulties

Attention alone may not fully reveal the decision-making process, as transformers can represent complex interactions in feedforward layers (Jain & Wallace, 2019). We therefore go beyond attention extraction, aiming to identify specialized concept neurons or heads that more explicitly encode tactical or strategic motifs.

6.3 Proposed Solutions

- **Neuron Activation Visualizations.** We plan to integrate interactive tools that highlight which neurons are significantly active, providing immediate board overlays for associated motifs.
- **Layer-Wise Attribution.** Building on methods like Integrated Gradients (Sundararajan et al., 2017), we aim to trace the final decision back through each layer, offering a richer interpretive lens than attention alone.

7 Related Work

Our approach draws on several streams of research:

Interpretable Neural Networks. The NLP community has pursued attention-based interpretability (Vaswani et al., 2017; Clark et al., 2019), though debate remains on whether attention is a true explanation (Jain & Wallace, 2019). Recent studies propose analyzing internal representations at the neuron level to detect emergent, concept-specific activations (Bau et al., 2019; Olah et al., 2020).

Chess-Specific Analysis. Prior chess models, such as AlphaZero, rely on search plus learned evaluations (Silver et al., 2017). Attempts to interpret these models have mainly centered on policy/value networks (McGrath et al., 2020), focusing on saliency maps or direct comparison with known heuristics. However, these methods still rely on some search, whereas DeepMind's Searchless Chess removes it entirely, demanding a deeper look into the feedforward structure and attention layers.

Concept Discovery and Local Explanations. Outside of chess, local explanation methods like LIME and SHAP have been applied to simpler classifiers to identify key features (Ribeiro et al., 2016; Lundberg & Lee, 2017). We seek to adapt a similar rationale to a more complex domain, labeling board motifs and verifying their correlation with model activations.

Overall, our work fills a gap in the literature: it extends concept-based interpretability to a purely feedforward chess engine and investigates how strategic understanding might be latently stored in large-scale transformer architectures.

8 Conclusion

By combining attention-based visualizations and targeted neuron activation analyses, we begin to peel back the layers of DeepMind's Searchless Chess. Our preliminary results suggest that even purely feedforward policy models can learn specialized chess concepts, encoded at either head or neuron level. The 9M-parameter model often aligns more closely with human notions of saliency, while the 270M-parameter model appears to factor in a broader, less immediately intuitive set of features. Going forward, we plan to finalize our concept-tagging pipeline, expand the scope of motifs tested, and incorporate stronger attribution techniques to further elucidate how advanced transformers generalize chess knowledge without explicit search.

REFERENCES

David Bau, Jun-Yan Zhu, Martin J. Wainwright, Alexei A. Efros, and Antonio Torralba. Rewriting a deep generative model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5971–5981, 2019.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pp. 276–286, 2019.

Anonymous (DeepMind). Searchless chess: Grandmaster-level policy via supervised learning alone. Preprint, 2025. Accessed via internal DeepMind documentation (fictional reference).

Sarthak Jain and Byron C. Wallace. Attention is not explanation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 3543–3556, 2019.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774, 2017.

Tom McGrath, Murray Campbell, Andrew J. Hoane Jr., and et al. Acquisition of chess knowledge in alphazero. *IEEE Transactions on Games*, 12(3):262–273, 2020.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, Shan Carter, et al. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024, 2020.

Alec Radford, Rafal Józefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3319–3328, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.