Contrastive Sparse Autoencoders for Interpreting Planning of Chess-Playing Agents

Yoann Poupart [1,2]

[1]ENS Lyon [2]Sorbonne Université, CNRS LIP6

ENS DE LYON

TL;DR

We propose contrastive sparse autoencoders (CSAE), a novel feature extraction framework based on pairs of activations. Our preliminary study shows qualitative and quantitative results attesting that CSAE can extract meaningful planning concepts.

Introduction

In this work, we focus on the open-source version of Alpha Zero, Leela Chess Zero [1], interpreting the neural network heuristic in combination with the tree search algorithm.

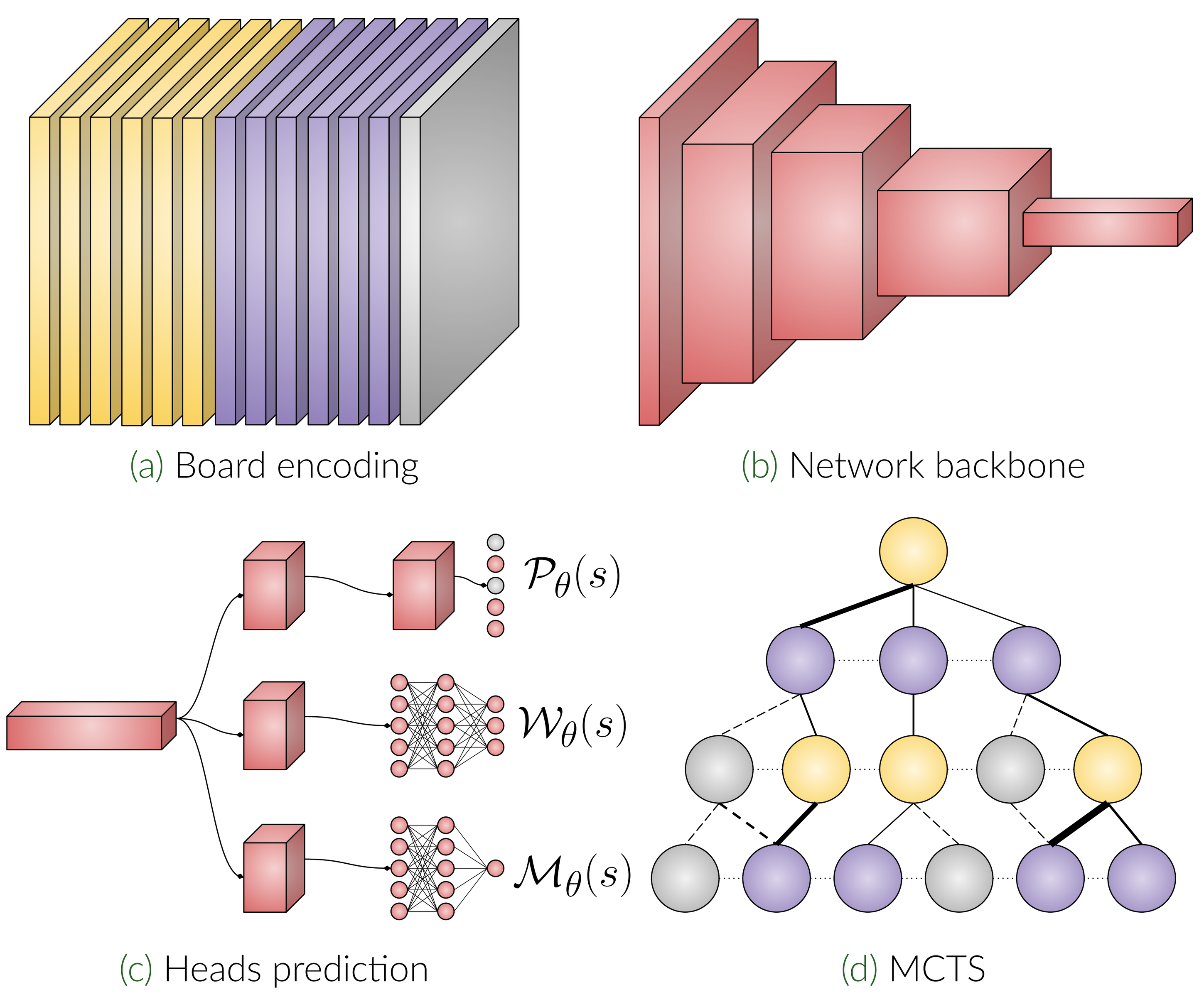

Figure 1. Modelling components; first, the boards are encoded into planes (a) and fed to the network backbone (b). The different heads use the extracted features to make heuristic predictions (c) guiding the MCTS when encountering new nodes (d).

Sparse Autoencoders

For discovering concepts we use sparse autoencoders (SAE) [2, 3], in their simplest form as described by equations 1 and 2. The base training loss uses an MSE reconstruction loss with l_1 penalisation to incentivise sparsity, equation 3.

$$f = \text{ReLU}(W_e h + b_e), \quad (1)$$

$$\hat{h} = W_d f + b_d. \quad (2)$$

$$\mathcal{L}_{\text{SAE}} = \mathbb{E}_h \left[||h - \hat{h}||_2^2 + \lambda ||f||_1 \right] \quad (3)$$

Contrastive Sparse Autoencoders

We propose contrastive sparse autoencoders (CSAE), an extension of the dynamic concepts introduced in [4], base on SAE. We illustrate their architecture in figure 2, which is trained using the base SAE loss, equation 3, augmented by a contrastive loss, equation 4.

$$\mathcal{L}_{\text{contrast}} = \mathbb{E}_h \left[||c^+ - c^-||_1 + ||d^+ \odot d^-||_1 \right] \quad (4)$$

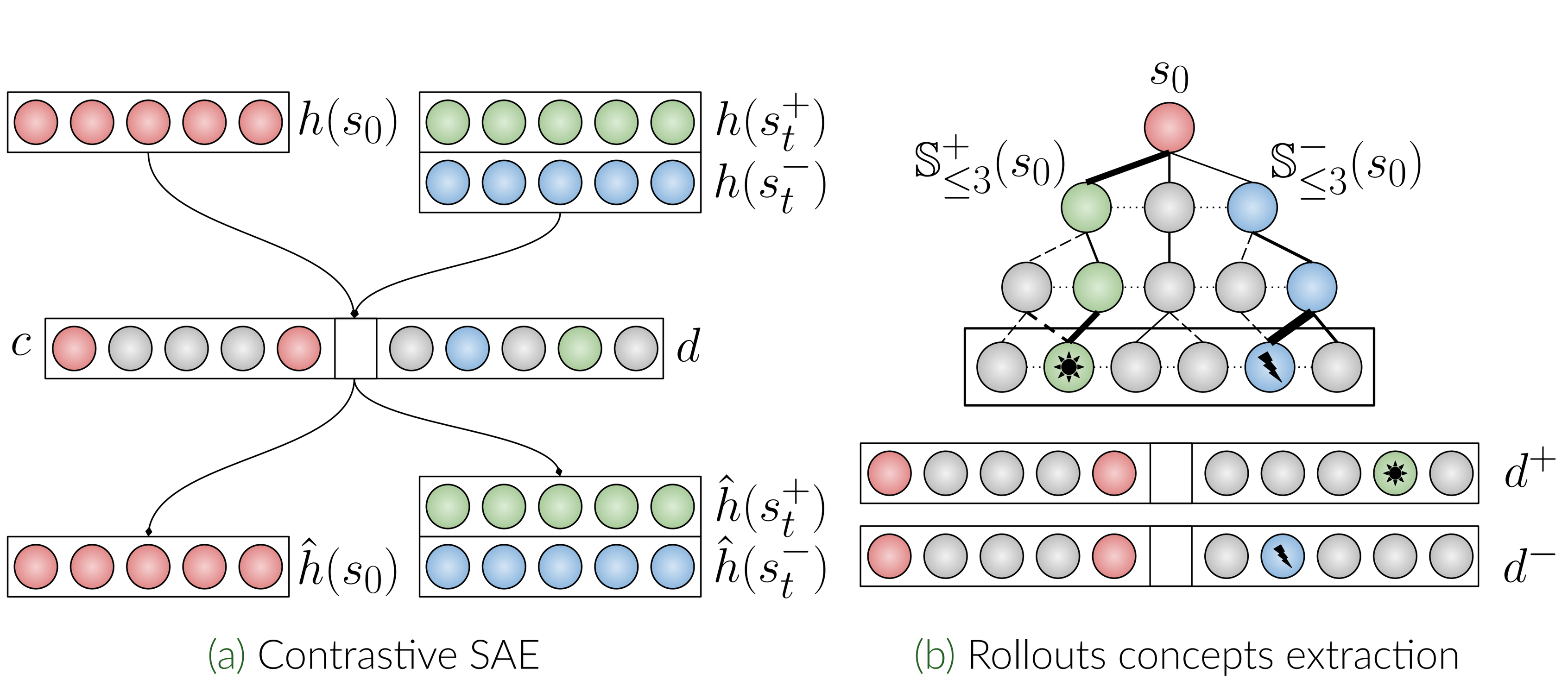

Figure 2. (a) CSAEs are trained using a contrast of an optimal trajectory (green) and suboptimal trajectories (blue). (b) Schematic view of concepts extraction from different rollouts ($\mathbb{S}^+_{\leq 3}(s_0)$ and $\mathbb{S}^-_{\leq 3}(s_0)$).

Activation Maximisation

Figure 3 illustrates a feature linked with the concept of rook threat. The shown board were picked among the samples that most activated the feature.

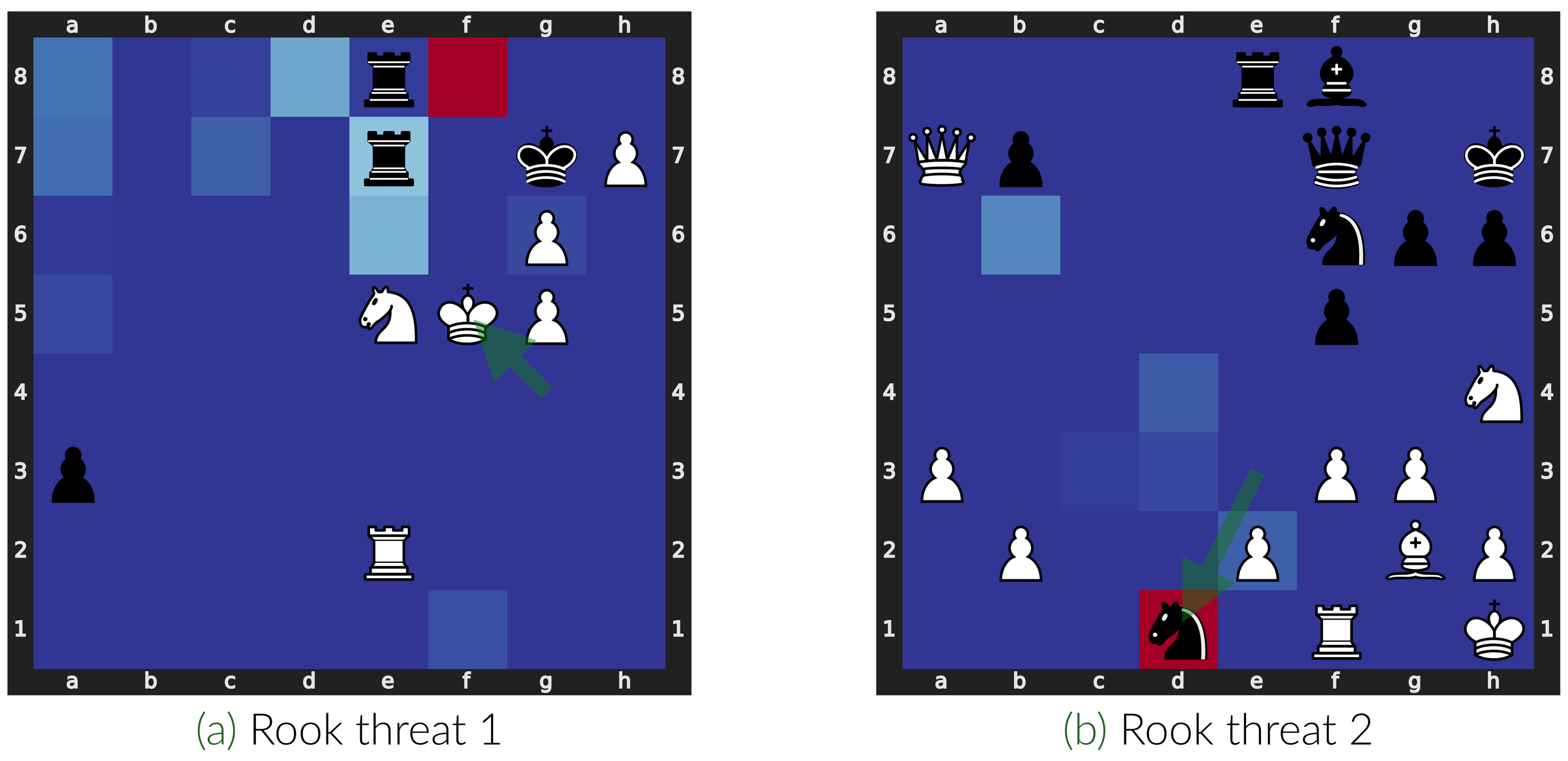

Figure 3. The feature activates for both black and white. In (a), the black rook should move to the red square to check the king, while in (b), the white rook should take the knight.

Ablation Study

In order to quantitatively assess the extraction process we ran an ablation study on a puzzle dataset from Lichess. On all puzzles we first record the model perplexity, P_{model}, then we patch the activations h by their estimates $\hat{h}$, while recording the coefficient R^2. We then intervene and ablate each feature independently to obtain P_f and R^2_f.

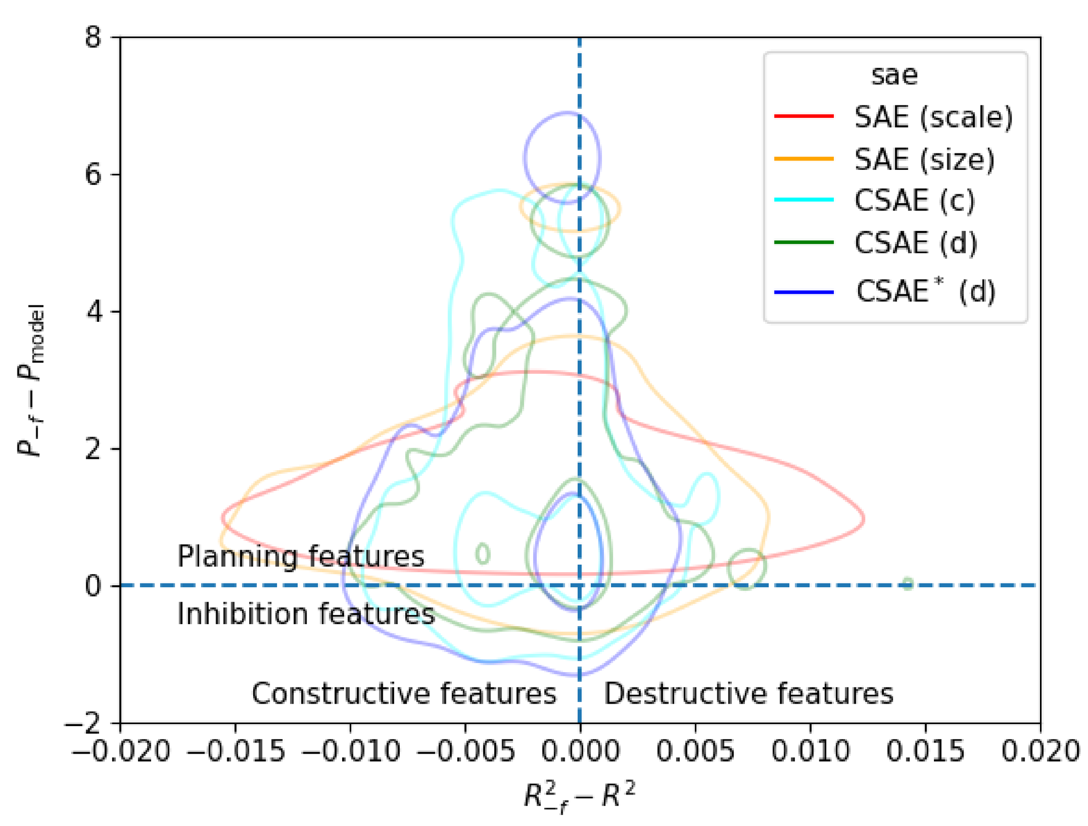

Figure 4. Density estimation for the feature ablation study. In abscissa the reconstruction impact, in ordinate the performance impact. CSAE features have higher planning impact and SAE features higher reconstruction impact

Limitations

- SAE generalisation issues
- Shallow feature interpretation
- Shallow feature relevance analysis

What's Next?

- Train better SAE
- Explore variations of models, layers and CSAE
- Feature extraction benchmark

HF Space

Project

Library

References

[1] Pascutto, Gian-Carlo and Linscott, Gary, "Leela chess zero," 2019.

[2] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey, "Sparse autoencoders find highly interpretable features in language models," *ArXiv*, vol. abs/2309.08600, 2023.

[3] T. Bricken *et al.*, "Towards monosemanticity: Decomposing language models with dictionary learning," *Transformer Circuits Thread*, 2023.

[4] L. Schut, N. Tomasev, T. McGrath, D. Hassabis, U. Paquet, and B. Kim, "Bridging the human-ai knowledge gap: Concept discovery and transfer in alphazero," 2023.