Acquisition of Chess Knowledge in AlphaZero

Authored By:

Thomas McGrath, Andrei Kapishnikov, Nenad Tomasev, Adam Pearce, Demis Hassabis,
Been Kim, Ulrich Paquet, and Vladimir Kramnik

Presented By:

Eric Hansen
Robin Na
Prayaag Venkat

Learning from machines (AlphaZero)

- Most methods we have looked so far try to interpret algorithms trained on human-generated data and labels
  - These interpretations may resemble human-understandable representations only because they learned from such data
- Can we interpret what the algorithms has been learning through its self-play training process?

Learning from machines - three pronged acquisition

- Probe for concepts
  - How closely is AlphaZero's internal representation **related to** chess concepts humans have already created?
  - Detection of human concepts from network activations
- Study behavioral changes
  - How do changing representations give rise to changing behaviors?
  - Evolution of AlphaZero vs. human's strategy in openings
- Investigate activations directly
  - Unsupervised methods using non-negative matrix factorisation (NMF) and direct measure of covariance to discover concepts not represented by humans in the past

Interpretability

- Concept-based (post-hoc) interpretability
  - Network probing / important features / mechanistic understanding
  - Challenges: correlation not causal

- Explainability in reinforcement learning
  - Structural causal models / reward difference explanations
  - Identify interesting points in behavioral trajectories

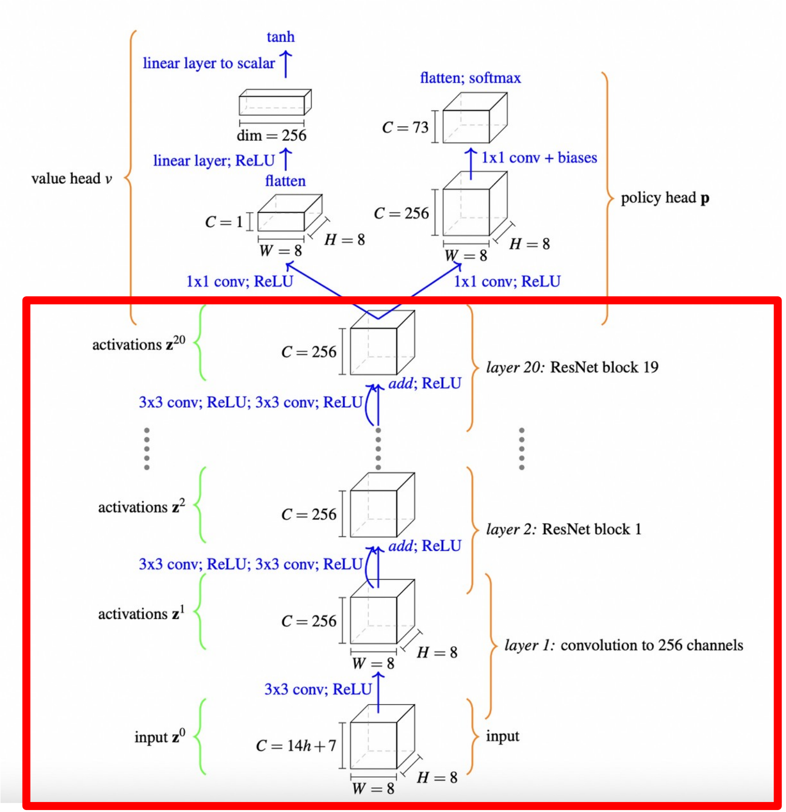
Chess as testing ground for AI interpretability

Can human learn the machine’s strategy?

- Does not rely on human-labeled data
- Tree-based organization / saliency maps
- Natural language processing to generate move-by-move commentary
- **This paper:** captures “intuitive” aspect of chess play by understanding networks that produce value assessment (*v*) and candidate move (**p**)

$$\mathbf{p}, v = f_{\boldsymbol{\theta}}(\mathbf{z}^0)$$

AlphaZero: Network structure and training



$$\mathbf{p}, v = f_{\boldsymbol{\theta}}(\mathbf{z}^0)$$

$$\mathbf{z}^l = f^l(\mathbf{z}^{l-1}) = \mathrm{ReLU}(\mathbf{z}^{l-1} + g^l(\mathbf{z}^{l-1}))$$

$$\mathbf{z}^l = f^{1:l}(\mathbf{z}^0) = f^l \circ \cdots \circ f^2 \circ f^1(\mathbf{z}^0)$$

Probing for concepts

Encoding of human conceptual knowledge

Question: Can human concepts be easily predicted from network's internal representation?

Concept: User defined function mapping network input to real line.

$$c(\mathbf{z}^0) = \begin{cases} 1 & \text{if } \mathbf{z}^0 \text{ contains a ♗-pair for the playing side} \\ 0 & \text{otherwise} \end{cases}$$

Approach: Train a sparse linear regression model from activations $\mathbf{z}^l$ at layer l and training step t to human concept j

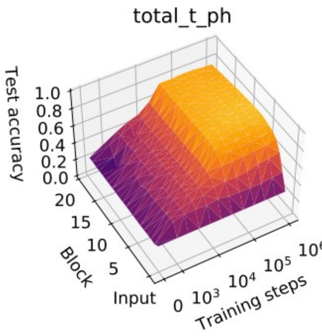
Probing concept learning: details

- **Data:** Randomly sample training, validation, test data from ChessBase archive, compute concept values and AlphaZero activations
- **Procedure:** For concept j, layer l, training step t, solve

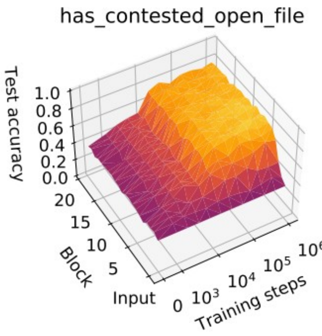
$$\mathbf{w}_{jlt}, b_{jlt} = \min_{\mathbf{w}, b} \frac{1}{N} \left\| \mathbf{w}^T \mathbf{Z}_t^l + b\mathbf{1} - \mathbf{c}_j \right\|_2^2 + \lambda \|\mathbf{w}\|_1 + \lambda |b|$$

- **Controls:** Regression from $\mathbf{z}^0$ and random concept regression.
- **Evaluation:** R^2 value, fraction of variance in concept explained by network activation

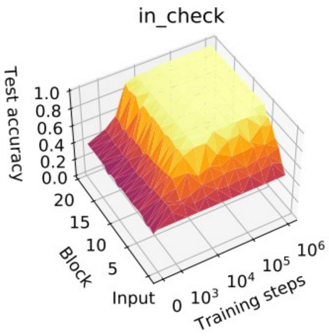
Evolution of human concepts in AlphaZero



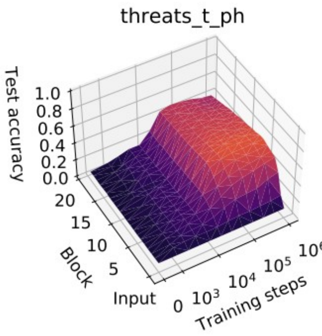
(a) Stockfish 8's total score



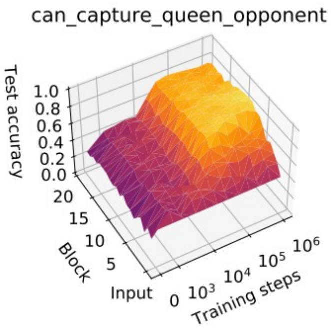
(b) A contested open file is occupied by rooks and/or queens of opposite colours



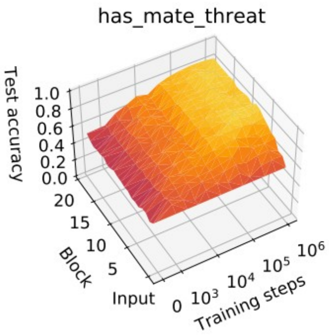
(c) Is the playing side in check?



(d) Stockfish 8's evaluation of threats.



(e) Can the playing side capture their opponent's queen?



(f) Could the opposing side checkmate the playing side in one move?

Key findings

1) **Grokking:** Many concepts begin to increase in accuracy around 32,000 steps

2) Drop in linearly-available information in later layers for some concepts

3) Some concepts **cannot** be regressed: sparsity partially obstructs ability to relate **highly-distributed representations** to concepts

4) **Learning from prediction errors**: regression errors may point to a “difference of opinion” with Stockfish

Challenges for concept probing

1) What's the right **probing architecture**?
2) How should we interpret **complex or subjective concepts**?
3) When can we definitively say a **concept is represented**?
4) When we train a probe, we cannot tell if we are getting a **confounder or the concept** itself

Progression through AlphaZero and human history

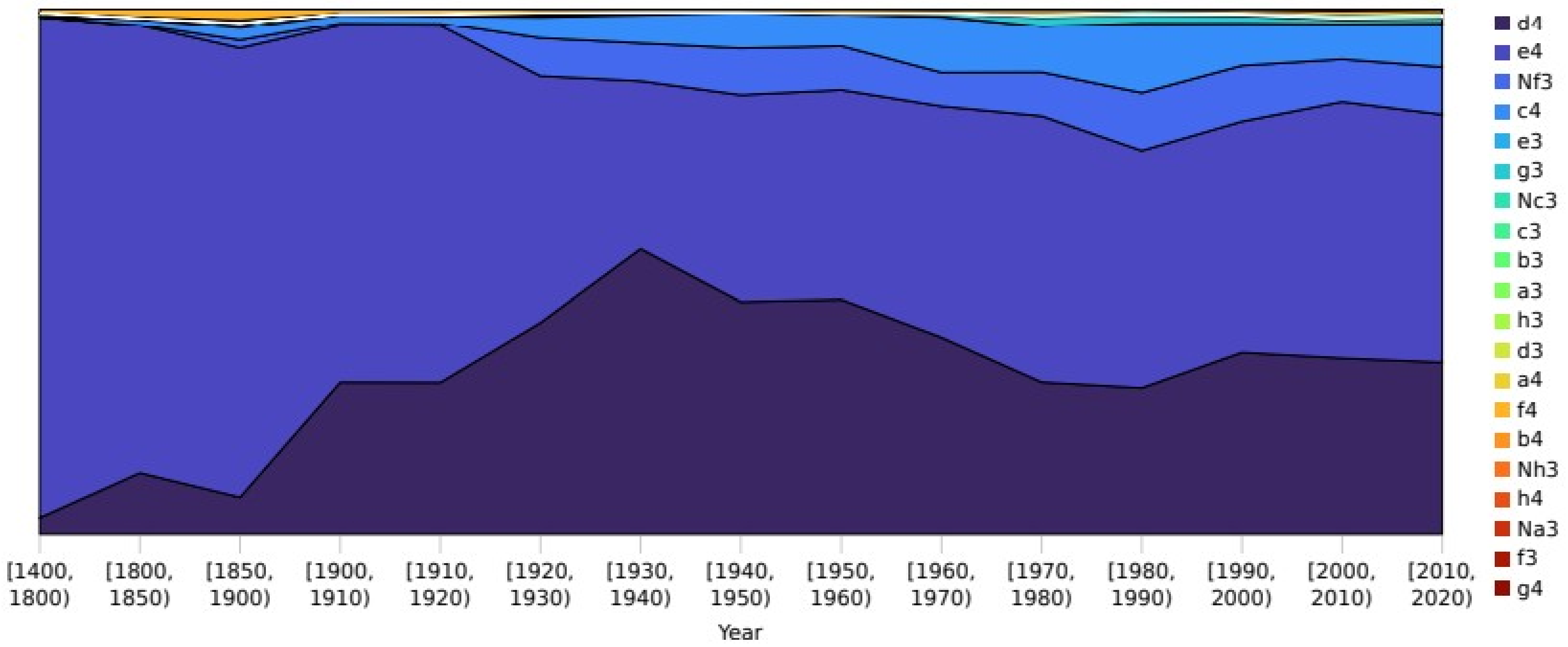
Progression of knowledge

- Recall:

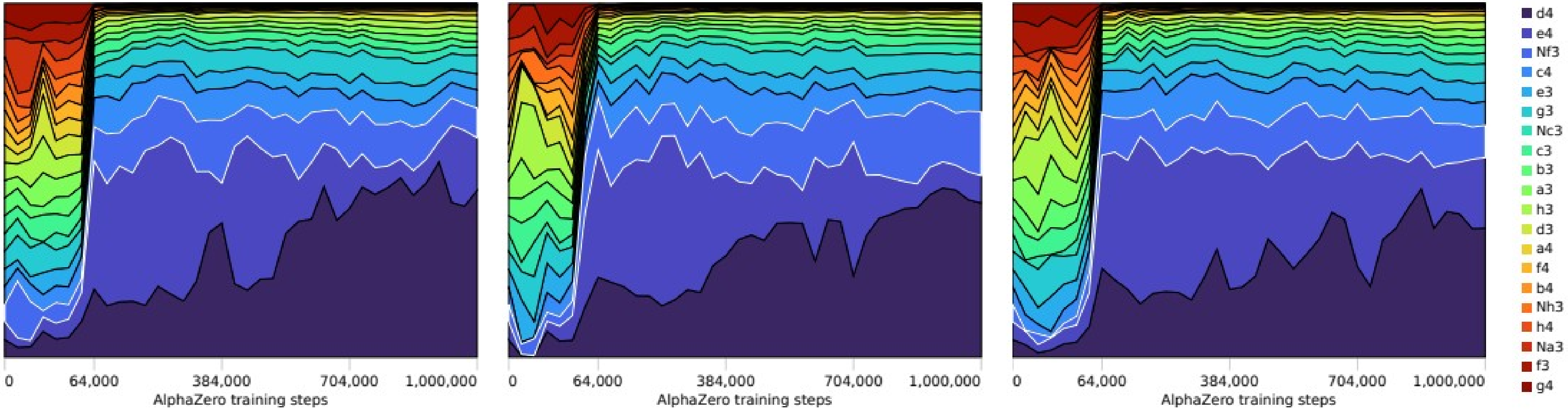
$$\mathbf{p}, v = f_{\boldsymbol{\theta}}(\mathbf{z}^0)$$

- **Question:** How does the progression of AlphaZero's knowledge compare to that of humans?
- **Key Findings:**
  - AlphaZero: Start from a uniform prior, then narrow down.
  - Humans: Start from a concentrated prior, then expand.

Progression through human history



Progression through AlphaZero history



Progression of AlphaZero’s Chess Knowledge

Methodology

- At different training steps (up to 128k), examine:
  - AlphaZero’s move tendencies
  - AlphaZero’s concept encodings

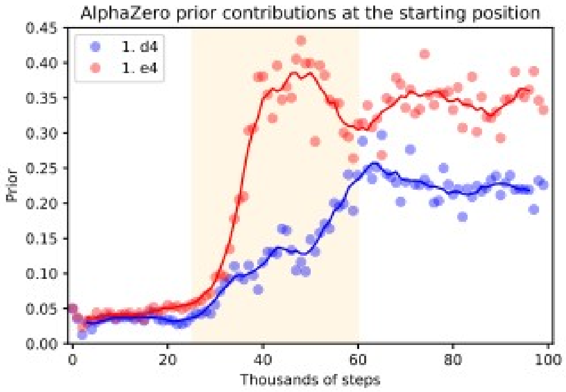
Primary Takeaways

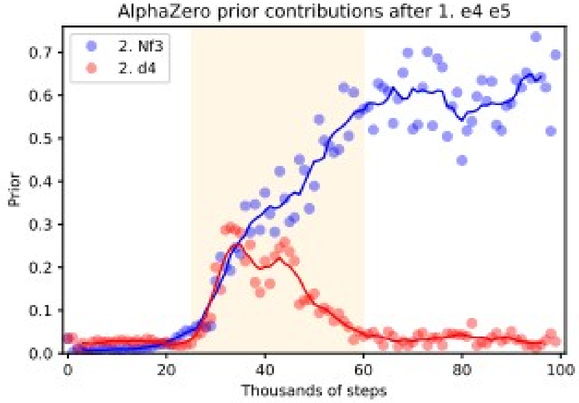
1) AlphaZero’s learns standard opening theory early on
2) AlphaZero learns material values before more complex positional concepts

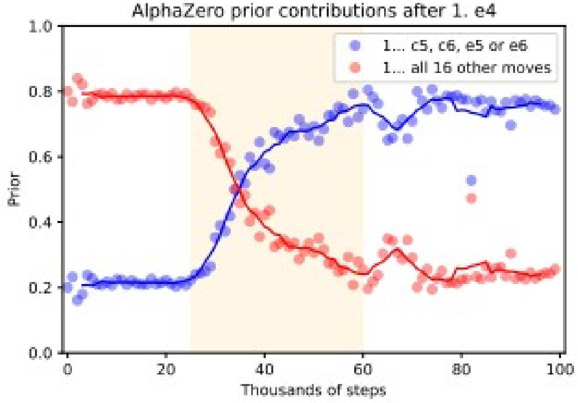
Both reinforce idea that AlphaZero learns basic human chess concepts first

Opening Theory Knowledge

1) ~30-60k: AlphaZero plays *1. e4/d4* the majority of the time
2) ~45k: AlphaZero considers *2. d4* before opting for *2. Nf3*
3) ~45k: AlphaZero plays standard responses to *1. e4*

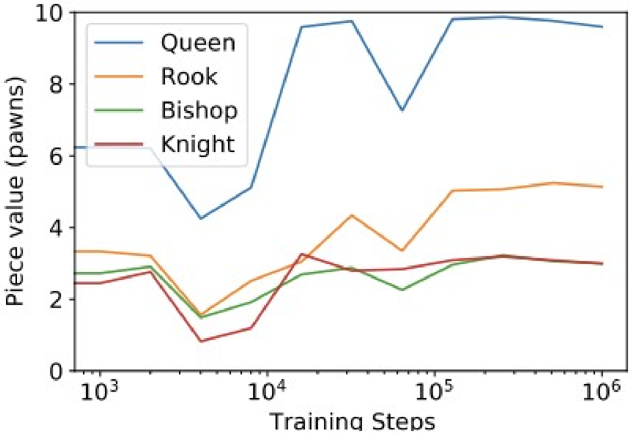


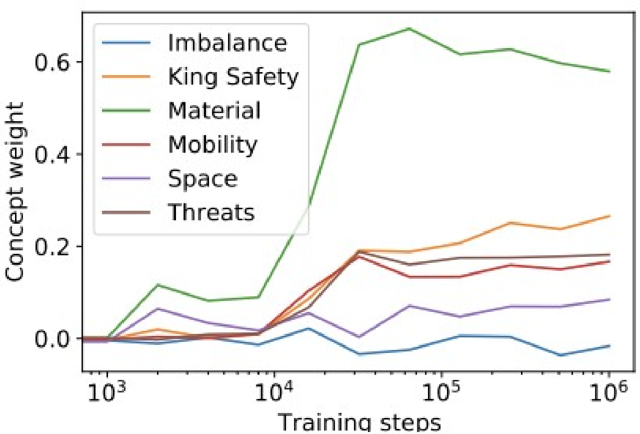




Material vs. Positional Knowledge

1) ~30k training steps: Piece Values develop, converge ~100k
2) King Safety, Mobility concepts emerge after Material
    a) More complex concepts require more training time





$$\hat{v}_{\mathbf{w},b}(\mathbf{z}^0) = \tanh\left(\mathbf{w}^T \mathbf{c}(\mathbf{z}^0) + b\right)$$

$$\mathbf{w}_t, b_t = \min_{\mathbf{w},b} \frac{1}{N} \sum_n \left| \hat{v}_{\mathbf{w},b}(\mathbf{z}_n^0) - v_{\boldsymbol{\theta}_t}(\mathbf{z}_n^0) \right|$$

Training Progression Assessment: GM Vladimir Kraminik

1) **16k to 32k:** Material Value in Complex Positions

2) **32k to 64k:** King Safety in Imbalanced Positions

3) **64k to 128k**: King Safety & Material Sacrifices in Complex Positions

- ❖ Tactical skills appear to precede positional skills as AlphaZero learns

Exploring Additional Feature Detectors in AlphaZero Network

Exploring Activations with Unsupervised Methods

1) **Goal:** Find Feature Detectors embedded within Network

2) **Methods:**
    a) Non-Negative Matrix Factorization of each layer’s channels
    b) Correlation of Input Board with each channel’s activations

Primary Takeaway

- Individual network layers & channels encode *feature detectors* related to human-recognizable chess concepts

Approach #1: NN Matrix Factorization Analysis

For each layer l with C channels,

1) Compute a matrix factorization **Ω x F** using K < C columns

For each factor k (1…K) and input n (1…N)

2) Visualize activations on Chess Board to find *feature detectors*

$$\hat{\mathbf{Z}}^l \in \mathbb{R}^{NHW \times C}$$

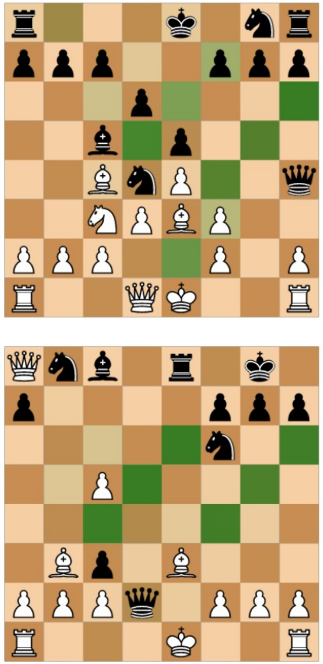
$$\mathbf{\Omega}_{\text{all}} \in \mathbb{R}^{NHW \times K}$$

$$\mathbf{F} \in \mathbb{R}^{K \times C}$$

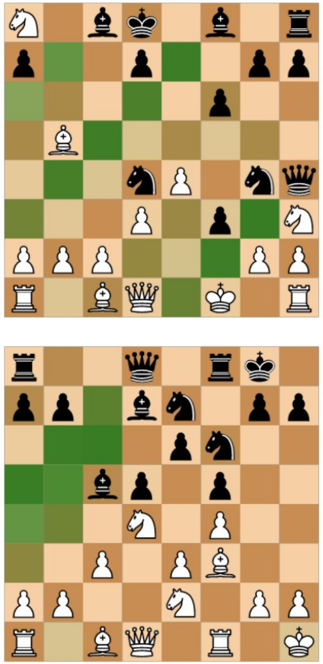
$$\mathbf{F}^*, \mathbf{\Omega}^*_{\text{all}} = \min_{\mathbf{F}, \mathbf{\Omega}_{\text{all}}} \left|\left|\hat{\mathbf{Z}}^l - \mathbf{\Omega}_{\text{all}}\mathbf{F}\right|\right|_2^2$$

$$\mathbf{F}, \mathbf{\Omega}_{\text{all}} \geq \mathbf{0}\ .$$

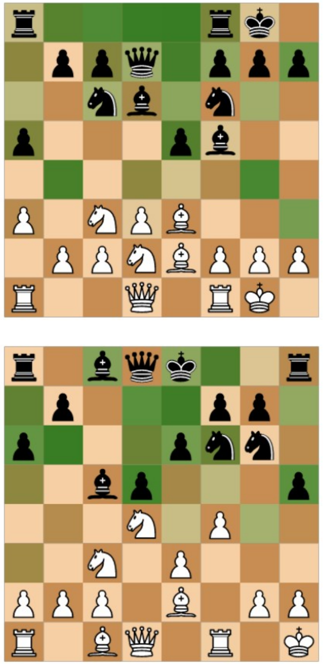
Results: NN Matrix Factorization Analysis



Diagonal Moves
(1st Layer)



Diagonal Moves
(2nd Layer)



# of Pieces that could
move to each square

Approach #2: Input-Activation Covariance Analysis

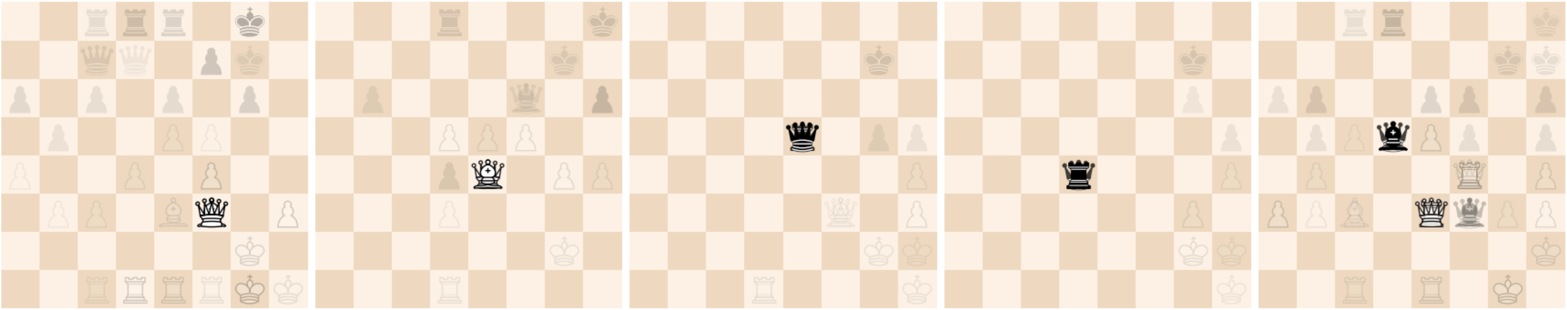
For each layer l and channel i,

1) Compute the covariance between input z0 and position activations
2) Visualize covariances on Chess Board to find *feature detectors*

$$\mathbf{cov}\left(z_i^l, \mathbf{z}^0\right) = \mathbb{E}\left[z_i^l\, \mathbf{z}^0\right] - \mathbb{E}\left[z_i^l\right] \mathbb{E}\left[\mathbf{z}^0\right]$$

Results: Input-Activation Covariance Analysis

Detecting Move-Types from a Square
1-2) Diagonal-Attacking Pieces (Queen, Bishop)
3-4) Horizontally-Attacking Pieces (Queen, Rook)
5) Both



5 covariances with different channels for the square (5, 4)

Conclusion - Key Findings

1) Human-Defined Concepts can be regressed from the AlphaZero network
    a) Despite never being trained on a human game of chess

2) As training progresses, AlphaZero understands basic concepts (openings, material) before more complex ones (king safety, mobility)

3) Feature Detectors of human-recognizable chess concepts are encoded by individual layers + channels in AZ Network
    a) Can be found using supervised & unsupervised techniques

Limitations

1) Challenges for Concept Probing (previous slide)

2) Knowledge acquisition is not complete - only a very small part of the model

3) Interpreting “Feature Detectors” found via Unsupervised techniques
    a) Inherently subjective

4) No Causal Insight into the Learned Concepts (only correlation)

Future Research Areas

1) Addressing Concept Probing Limitations (more than sparse LR)

2) Can we go beyond finding human knowledge embedded in AlphaZero and understand what new concepts are learned?
   a) Further analysis of feature detectors found with unsupervised techniques

3) How do we generalize these findings to other machine learning settings? Could we find human-recognizable concepts in different trained models?

Discussion

- How is this paper different from methods we discussed so far?
    - In terms of goals / techniques / specific models to explain
- Does this approach (or anything that can be built up on this) have potential applications for a more general setting other than board games?
    - What are the pros/cons of using games/artificial environments as baseline for evaluating RL algorithms?
- Can this be applied to our practice of doing science?
    - Can we recreate or extract theories using similar frameworks (AI for science)?
- How should we define or understand “knowledge” in these settings?
    - How is it different from similar to concepts/models/explanations/information?
- How convinced are you convinced about the “interpretability” aspect?
    - Who would be potential audience that could benefit from such analysis?