# Feature set analysis for chess 3UNN networks Tesis de Licenciatura

#### Martín Emiliano Lombardo

Departamento de Computación Facultad de Ciencias Exactas y Naturales Universidad de Buenos Aires

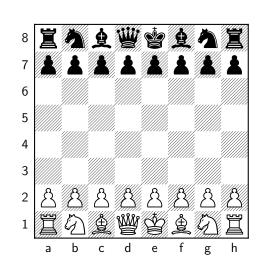
2024



## Introducción

# Ajedrez

- Dos jugadores
- Suma cero



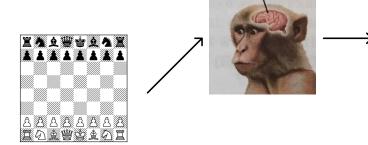
## Humano vs. Computadora



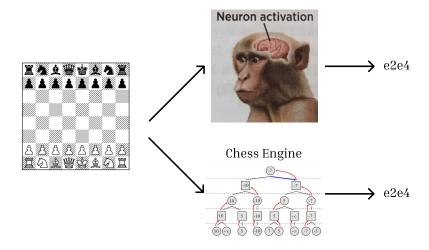
**Neuron** activation

e2e4

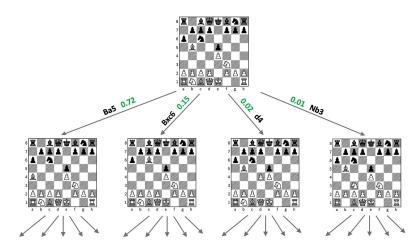
# Humano vs. Computadora



## Humano vs. Computadora

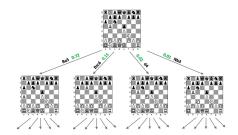


# Ajedrez como árbol



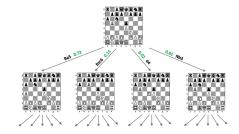
# Motores de ajedrez (Chess Engines)

 Exploran el árbol de juego (Minimax, MCTS, etc.)

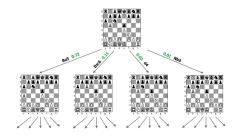


# Motores de ajedrez (Chess Engines)

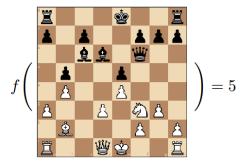
- Exploran el árbol de juego (Minimax, MCTS, etc.)
- Utilizan funciones de evaluación en las hojas



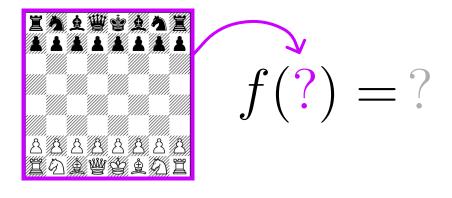
- Exploran el árbol de juego (Minimax, MCTS, etc.)
- Utilizan funciones de evaluación en las hojas
- La evaluación se propaga hacia arriba, según el algoritmo



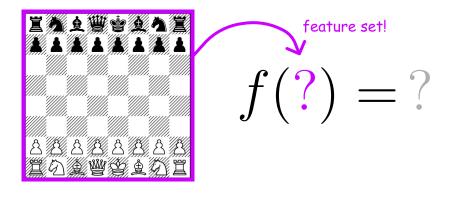
## Función de evaluación



# (adelanto) Feature set: ¿Cómo transformar la posición a un vector?



# (adelanto) Feature set: ¿Cómo transformar la posición a un vector?



# Motores de ajedrez (breve historia)

#### asdasd

■ Text visible on slide 1

#### asdasd

- Text visible on slide 1
- Text visible on slide 2

#### asdasd

- Text visible on slide 1
- Text visible on slide 2
- Text visible on slide 3

- Text visible on slide 1
- Text visible on slide 2
- Text visible on slide 4 asdasd

## Contenido

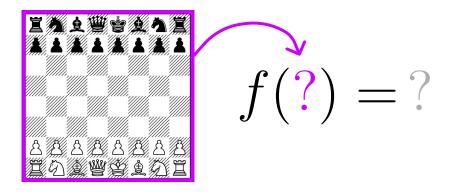
- 1 Introducción
- 2 Engine
- 3 Feature set
  - Motivación
  - Definición
  - Operadores
  - Feature sets conocidos
  - Resumen
- 4 BUNN (NNUE)
- 5 Training
- **6** Experimentos
- 7 Conclusión

# Engine

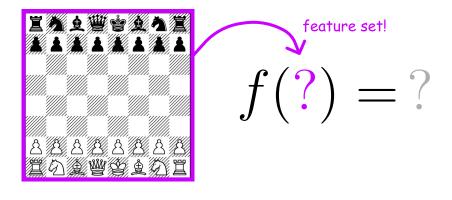
# Feature set

Motivación

# ¿Cómo transformar la posición a un vector?



# ¿Cómo transformar la posición a un vector?



## Definición

Un feature set  $S_P$  se define con un conjunto S y un predicado asociado P(e), donde:

- **S** es un conjunto de conceptos (rol, color, celda, número, etc.).
- **P**(**e**) es un predicado que determina si *e* está presente (o *activo*) en la posición (implícita).

## Definición

Un feature set  $S_P$  se define con un conjunto S y un predicado asociado P(e), donde:

- **S** es un conjunto de conceptos (rol, color, celda, número, etc.).
- P(e) es un predicado que determina si e está presente (o activo) en la posición (implícita).
- Cada elemento en  $S_P$  es un *feature*.

Definición

## Definición

Un feature set  $S_P$  se define con un conjunto S y un predicado asociado P(e), donde:

- **S** es un conjunto de conceptos (rol, color, celda, número, etc.).
- P(e) es un predicado que determina si e está presente (o activo) en la posición (implícita).
- Cada elemento en S<sub>P</sub> es un feature.
- Cada feature es un valor en el vector de entrada, valiendo 1 si está activo y 0 si no.

# Ejemplos de S

Información posicional:

Files = 
$$\{a, b, ..., h\}$$
  
Ranks =  $\{1, 2, ..., 8\}$   
Squares =  $\{a1, a2, ..., h8\}$ 

Información sobre las piezas:

Roles = { & Pawn, ② Knight, & Bishop, 罩 Rook, 豐 Queen, 堂 King  $Colors = {\bigcirc White, ullet Black}$ 

# Ejemplo completo



	Feature set	
	$(\text{Files} \times \text{Colors})_P$	$(\text{Files} \times \text{Roles})_Q$
Active features	$\langle a, \bigcirc \rangle, \langle a, \bullet \rangle, \langle c, \bullet \rangle,$	$\langle a, \& \rangle, \langle c, & \rangle, \langle c, \& \rangle,$
	$\langle c, \bigcirc \rangle, \langle d, \bigcirc \rangle, \langle h, \bullet \rangle$	$\langle d,  \mathring{\triangle} \rangle,  \langle h,  \mathring{\underline{a}} \rangle$

 $P(\langle f, c \rangle)$ : there is a piece in file f with color c.  $Q(\langle f, r \rangle)$ : there is a piece in file f with role r.

Operadores

# Operación: Suma $\oplus$ (concatenación)

Hay veces que es útil combinar información de dos feature sets

Hay veces que es útil combinar información de dos feature sets

$$S_P,\, T_Q:$$
 feature sets  $S_P\oplus T_Q=(S\cup T)_R$  where  $R(e)=egin{cases} P(e) & ext{if } e\in S \ Q(e) & ext{if } e\in T \end{cases}$ 

# Operación: Producto $\times$ (and)

$$S_P imes T_Q = (S imes T)_R$$
 where  $R(\langle e_0, e_1 
angle) = P(e_0) \ \land \ Q(e_1)$ 

Feature sets conocidos

## Feature set: ALL

La codificación más natural de una posición de ajedrez

ALL:  $(SQUARES \times ROLES \times COLORS)_P$  $P(\langle s, r, c \rangle)$ : there is a piece in square s with role r and color c Feature sets conocidos

## Feature set: ALL

La codificación más natural de una posición de ajedrez

ALL: 
$$(SQUARES \times ROLES \times COLORS)_P$$
  
 $P(\langle s, r, c \rangle)$ : there is a piece in square  $s$  with role  $r$  and color  $c$ 

■ Es pequeño:  $64 \times 6 \times 2 = 768$  *features* 

## Feature set: ALL

La codificación más natural de una posición de ajedrez

ALL: 
$$(SQUARES \times ROLES \times COLORS)_P$$
  
 $P(\langle s, r, c \rangle)$ : there is a piece in square  $s$  with role  $r$  and color  $c$ 

- Es pequeño:  $64 \times 6 \times 2 = 768$  *features*
- Es completo: contiene toda la información de la posición

## Feature set: ALL

La codificación más natural de una posición de ajedrez

ALL: 
$$(SQUARES \times ROLES \times COLORS)_P$$
  
 $P(\langle s, r, c \rangle)$ : there is a piece in square  $s$  with role  $r$  and color  $c$ 

- Es pequeño:  $64 \times 6 \times 2 = 768$  *features*
- Es completo: contiene toda la información de la posición
- Es muy rápido computar cuáles features están activas

## Feature set: KING-ALL ó "KA"

Los engines modernos usan variaciones del siguiente feature set. Permite entender la posición en relación a la posición del rey:

> $KING-ALL = SQUARE_K \times ALL$ K(s): s is the square of the king of the side to move

Los engines modernos usan variaciones del siguiente feature set. Permite entender la posición en relación a la posición del rey:

$$KING-ALL = SQUARE_K \times ALL$$
  
 $K(s)$ : s is the square of the king of the side to move

**E**s grande:  $64 \times 768 = 49152$  *features* 

Los engines modernos usan variaciones del siguiente feature set. Permite entender la posición en relación a la posición del rey:

$$ext{King-All} = ext{Square}_{K} imes ext{All}$$
  
 $K(s)$ :  $s$  is the square of the king of the side to move

- Es grande:  $64 \times 768 = 49152$  features
- Es muy rápido como All

Los engines modernos usan variaciones del siguiente feature set. Permite entender la posición en relación a la posición del rey:

$$ext{King-All} = ext{Square}_{K} imes ext{All}$$
  
 $K(s)$ :  $s$  is the square of the king of the side to move

- Es grande:  $64 \times 768 = 49152$  features
- Es muy rápido como All
- Entrenarlo require un dataset más grande y lleva más tiempo (no me meto acá)

### Feature sets: resumen

- **S**: set of concepts (roles, colors, squares, files, ranks, etc.).
- **P**(e): predicate that defines when the feature e is present in the (implicit) position.
- **S**<sub>P</sub>: a feature set. Every element in  $S_P$  is a feature. Features that satisfy P are active.
- $S_P \times T_Q = (S \times T)_R$  where  $R(\langle e_0, e_1 \rangle) = P(e_0) \wedge Q(e_1)$

# (AUNN) NNUE

**З**UИИ: **E**fficiently **U**pdatable **N**eural **N**etworks

### **З**UИИ: **N**eural **N**etworks

- El input es un vector one-hot generado por el feature set.
  - Debe tener pocos *features* activos (rala): introduce una cota superior.
- La red es una feedforward clásica con dos capas ocultas.

### Linear layer

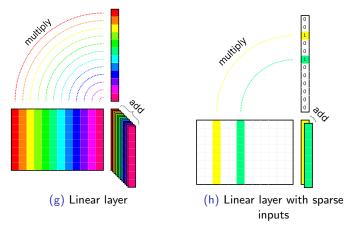


Figure: Linear layer operation comparison. Figures from [18].

## **З**UИИ: **E**fficient **U**pdates

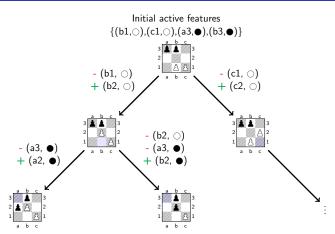


Figure: Partial tree of feature updates (removals and additions) for ( $SQUARES \times COLORS$ ) (white's point of view) in a simplified 3x3 pawn-only board.

### **З**UИИ: Tradeoff

motivacion comparacion de burns

# Training

## Experimentos

Recapitulando... ¿Qué hay que definir para entrenar una red?

■ Feature set: determina la codificación y los patrones que se pueden aprender

- Feature set: determina la codificación y los patrones que se pueden aprender
- Dataset: datos de entrenamiento, visto anteriormente

- **Feature set**: determina la codificación y los patrones que se pueden aprender
- Dataset: datos de entrenamiento, visto anteriormente
- Arquitectura de la red: el tamaño de cada capa; L<sub>1</sub> y L<sub>2</sub>

- **Feature set**: determina la codificación y los patrones que se pueden aprender
- Dataset: datos de entrenamiento, visto anteriormente
- **Arquitectura de la red**: el tamaño de cada capa;  $L_1$  y  $L_2$
- Método de entrenamiento: PQR/target scores; determina el formato de las muestras y la loss function

- Feature set: determina la codificación y los patrones que se pueden aprender
- Dataset: datos de entrenamiento, visto anteriormente
- **Arquitectura de la red**: el tamaño de cada capa;  $L_1$  y  $L_2$
- Método de entrenamiento: PQR/target scores; determina el formato de las muestras y la loss function
- **Hiperparámetros**: learning rate, batch size, epochs, etc.

- Loss (train y val.): indica la calidad de las predicciones.
  - Permite detectar overfitting y otros problemas

- Loss (train y val.): indica la calidad de las predicciones.
  - Permite detectar overfitting y otros problemas
- Puzzle accuracy: porcentaje de movimientos acertados en puzzles de Lichess.
  - Sólo hay un movimiento correcto
  - Proxy (muy malo) de la fuerza de la red

- **Loss** (train y val.): indica la calidad de las predicciones.
  - Permite detectar overfitting v otros problemas
- Puzzle accuracy: porcentaje de movimientos acertados en puzzles de Lichess.
  - Sólo hay un movimiento correcto
  - Proxy (muy malo) de la fuerza de la red
- Elo relativo: la medida más común para comparar engines.
  - Se realizan torneos de 100ms por movimiento
  - El elo es calculado a partir de Ordo

Busco fijar el setup de entrenamiento con valores razonables

Busco fijar el setup de entrenamiento con valores razonables

■ El feature set va a cambiar cada experimento

Busco fijar el setup de entrenamiento con valores razonables

- El feature set va a cambiar cada experimento
- El dataset está fijo

Busco fijar el setup de entrenamiento con valores razonables

- El feature set va a cambiar cada experimento
- El dataset está fijo
- El método de entrenamiento principal es *target scores*

Busco fijar el setup de entrenamiento con valores razonables

- El feature set va a cambiar cada experimento
- El dataset está fijo
- El método de entrenamiento principal es *target scores*

Entonces queda por determinar...

■ La arquitectura de la red  $(L_1 \ y \ L_2)$ 

Busco fijar el setup de entrenamiento con valores razonables

- El feature set va a cambiar cada experimento
- El dataset está fijo
- El método de entrenamiento principal es target scores

Entonces queda por determinar...

- La arquitectura de la red  $(L_1 \ y \ L_2)$
- Los hiperparámetros

Los hiperparámetros fueron seleccionados en base al trainer oficial de Stockfish:

Los hiperparámetros fueron seleccionados en base al trainer oficial de Stockfish:

■ Learning rate: 0.0005

Los hiperparámetros fueron seleccionados en base al trainer oficial de Stockfish:

**Learning rate**: 0.0005

**Exponential decay**: 0.99

Los hiperparámetros fueron seleccionados en base al trainer oficial de Stockfish:

**Learning rate**: 0.0005

**Exponential decay**: 0.99

■ Batch size: 16384

Los hiperparámetros fueron seleccionados en base al trainer oficial de Stockfish:

**Learning rate**: 0.0005

**Exponential decay**: 0.99

■ Batch size: 16384

**Epoch size**: 100 million

Los hiperparámetros fueron seleccionados en base al trainer oficial de Stockfish:

■ Learning rate: 0.0005

**Exponential decay**: 0.99

Batch size: 16384

■ **Epoch size**: 100 million

cada epoch realiza 6104 batches

Los hiperparámetros fueron seleccionados en base al trainer oficial de Stockfish:

■ Learning rate: 0.0005

**■ Exponential decay**: 0.99

Batch size: 16384

■ **Epoch size**: 100 million

cada epoch realiza 6104 batches

**■ Epochs**: 256

cada run observa 25.6 billion samples

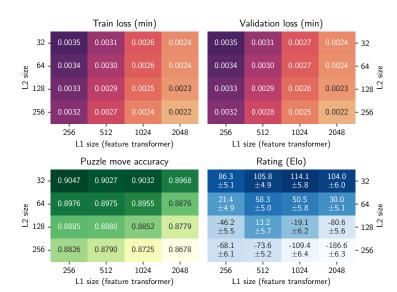
### Baseline: experimento

Sólo queda buscar parámetros  $L_1$  y  $L_2$  razonables. Realizo una búsqueda en grilla con:

- $\blacksquare \ \mathsf{L1} \in \{256, 512, 1024, 2048\}$
- L2 ∈ {32, 64, 128, 256}

El feature set a utilizar es ALL[768].

#### Baseline: resultados



#### Baseline: conclusión

- L2=32. El performance cae dramáticamente si L2 aumenta, utilizo el más bajo.
  - Sería buena idea probar valores más chicos de L2.

#### Baseline: conclusión

- L2=32. El performance cae dramáticamente si L2 aumenta, utilizo el más bajo.
  - Sería buena idea probar valores más chicos de L2.
- L1=512. Es el mejor valor para L2=64 y L2=128, y en margen de error para L2=32.
  - Además es el más rápido de entrenar.

# Axis encoding: motivación

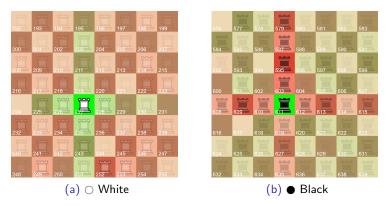


Figure: Weights of **a neuron** in the L1 layer, which are connected to features in ALL where the role is  $\Xi$  Rook. The intensity represents the weight value, and the color represents the sign (although not relevant).

## Axis encoding: motivación

La red detecta patrones parecidos a los movimientos de las piezas.

# Axis encoding: motivación

La red detecta patrones parecidos a los movimientos de las piezas. Para hacerle la vida más fácil a la red, propongo agregar features como:

"there is a ○ White 🖺 Rook in the 4th rank"

## Axis encoding: experimento



# Axis encoding: experimento



Depiction	Block name		Definition	Number of features
<del></del>	Н	(FILES	$\times$ Roles $\times$ Colors) <sub>P</sub>	96
<b>‡</b>	V	(Ranks	$\times$ Roles $\times$ Colors) <sub>P</sub>	96
	D1	(Diags1	$\times$ Roles $\times$ Colors) <sub>P</sub>	180
	D2	(Diags2	$\times$ Roles $\times$ Colors) <sub>P</sub>	180

 $P(\langle x, r, c \rangle)$ : there is a piece in x with role r and color c

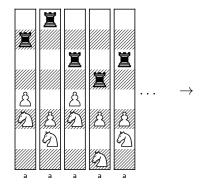
# Axis encoding: experimento

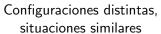
Depiction	Feature set	Number of features
$\longleftrightarrow \oplus  \updownarrow$	$\mathrm{H}\oplus\mathrm{V}$	192
✓ ⊕ 🔨	$\mathrm{D}1\oplus\mathrm{D}2$	360
↔ ⊕ ↑ ⊕ ✓ ⊕ \	$H\oplus V\oplus D1\oplus D2$	552
$All \oplus \longleftrightarrow \oplus \qquad \updownarrow$	$\mathrm{All} \oplus \mathrm{H} \oplus \mathrm{V}$	960
$\mathrm{All} \oplus \nearrow \oplus \diagdown$	$ALL \oplus D1 \oplus D2$	1128
$ALL \oplus \longleftrightarrow \oplus \ \downarrow \ \oplus \ \swarrow \ \oplus \ \diagdown$	$ALL \oplus H \oplus V \oplus D1 \oplus D2$	1320

# Axis encoding: resultados

Feature set	Number of features	Val. loss	Rating elo (rel. to All)	Puzzles move acc.
←→ ⊕ 🚶	192	0.005810	$-384.3 \pm 5.1$	0.8618
✓ ⊕ 🔨	360	0.006707	$-444.1 \pm 5.1$	0.8517
	552	0.003907	-183.5 ± 4.1	0.8748
ALL (reference)	768	0.003134	0.0	0.8865
$ALL \oplus \longleftrightarrow \oplus $	960	0.003082	$-27.1\pm4.1$	0.8851
$ALL \oplus \nearrow \oplus \searrow$	1128	0.003087	$-26.1 \pm 3.8$	0.8814
$\begin{array}{c} \text{All} \oplus \longleftrightarrow \oplus \\ \oplus \nearrow \oplus \\ \end{array}$	1320	0.003067	-58.7 ± 3.7	0.8766

#### Pairwise axes: motivación







Las mismas dos features (par rojo y par azul)

#### Pairwise axes: motivación

Comparando con el experimento anterior, es más específico en vez de más general:

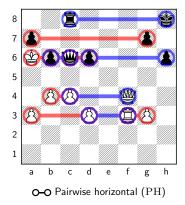
"there is a ○ White \( \begin{aligned} \begin{aligned} \text{Rook in the 4th rank} \\ \text{vs.} \end{aligned} \)

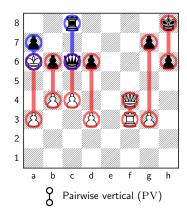
"there is a ● Black \( \mathbb{Z}\) Rook next to a \( \cap \) White \( \text{\alpha}\) Pawn in the 'a' file"

# Pairwise axes: experimento

D.	Block name	Definition	Num. of features
		$(RANKS \times (ROLES \times COLORS) \times (ROLES \times COLORS))_P$	
0-0	PH	$P(\langle r, r_1, c_1, r_2, c_2 \rangle)$ : there is a piece in rank $r$ with role $r_1$ and color $c_1$ to the left of a piece with role $r_2$ and color $c_2$	1152
_		$({\rm Files}\times({\rm Roles}\times{\rm Colors})\times({\rm Roles}\times{\rm Colors}))_Q$	
ð	PV	$Q(\langle f, r_1, c_1, r_2, c_2 \rangle)$ : there is a piece in file $f$ with role $r_1$ and color $c_1$ below a piece with role $r_2$ and color $c_2$	1152

### Pairwise axes: experimento





### Pairwise axes: experimento

#### Los feature sets a entrenar son:

- lacktriangle ALL  $\oplus$  PH (1920 features)
- lacktriangle ALL  $\oplus$  PV (1920 features)
- $\blacksquare$  ALL  $\oplus$  PH  $\oplus$  PV (3072 features)

### Pairwise axes: resultados

Feature set	Number of features	Val. loss	Rating elo (rel. to ALL)
All (reference)	768	0.003134	0.0
All ⊕ <b>0-0</b>	1920	0.003033	-38.2 ± 4.8
$ALL \oplus $	1920	0.002946	-8.4 ± 5.0
$All \oplus o - o \oplus \emptyset$	3072	0.002868	-37.6 ± 4.9

■ Reducir el número de pairs puede llevar a una mejora por sobre ALL (ej. △)

 La mobilidad en ajedrez es una medida de la cantidad de movimientos que puede hacer un jugador en una posición.

- La mobilidad en ajedrez es una medida de la cantidad de movimientos que puede hacer un jugador en una posición.
- Un paper de Eliot Slater (1950) mostró que hay una correlación entre la mobilidad de un jugador y la cantidad de partidas ganadas.

- La mobilidad en ajedrez es una medida de la cantidad de movimientos que puede hacer un jugador en una posición.
- Un paper de Eliot Slater (1950) mostró que hay una correlación entre la mobilidad de un jugador y la cantidad de partidas ganadas.
- Se usa en funciones de evaluación hechas a mano.

- La mobilidad en ajedrez es una medida de la cantidad de movimientos que puede hacer un jugador en una posición.
- Un paper de Eliot Slater (1950) mostró que hay una correlación entre la mobilidad de un jugador y la cantidad de partidas ganadas.
- Se usa en funciones de evaluación hechas a mano.
- Propongo agregar mobilidad como features en la red.

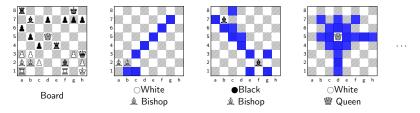
### Mobility: experimento

Hay dos maneras de codificar la mobilidad:

- Bitsets (por rol/color)
- Cantidades (por rol/color)

# Mobility: experimento (bitsets)

Los features proveen **las celdas** a las que una pieza de determinado rol/color puede moverse.



La cantidad de features es  $64 \times 6 \times 2 = 768$ , la misma que ALL.

# Mobility: experimento (counts)

Los features proveen **la cantidad de celdas** a las que una pieza de determinado rol/color puede moverse. Esto reduce la cantidad de features significativamente.

Piece role	Min	Max
å Pawn	0	8+
🛭 Knight	0	15+
🎍 Bishop	0	16+
□ Rook	0	25+
₩ Queen	0	25+
🗳 King	0	8

# Mobility: experimento (counts)

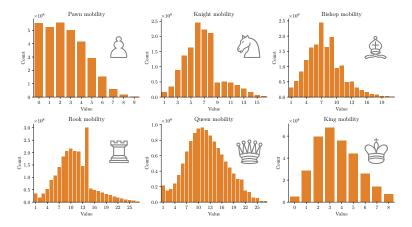


Figure: Total mobility values for each piece on the board. Computed using 2 billion boards. The value 0 for the 🖄 Knight, 🚊 Bishop, 🖺 Rook, and 👑 Queen has been excluded from the plot, as it is very common.

## Mobility: experimento

Block name	Definition	Number of features
	(Squares $\times$ Roles $\times$ Colors) <sub>P</sub>	
MB	$P(\langle s, r, c \rangle)$ : there is a piece of role $r$ and color $c$ that <b>can move to</b> square $s$	768
МС	$(\{0,1,\ldots\} \times \text{Roles} \times \text{Colors})_P$	
	$P(\langle m, r, c \rangle)$ : the value of mobility for a piece of role $r$ and color $c$ is $m$	206

Los feature sets a entrenar son:  $ALL \oplus MB$  (1536 features) y  $ALL \oplus MC$  (974 features).

### Mobility: resultados

Table: Mobility encodings results

Feature set	Number	Val. loss	Rating
	of features	mın	elo (rel. to All)
$\mathrm{All}$ (reference)	768	0.003134	0.0
$ALL \oplus MB$	1536	0.002824	$-260.9 \pm 5.4$
$ALL \oplus MC$	974	0.003032	-280.9 ± 5.6

- Las predicciones mejoran muy poco (el loss no se reduce tanto).
- Por ende, el costo de las actualizar los features es más alto al beneficio que aportan.

## Mobility: resultados

Table: Mobility encodings results

Feature set	Number of features	Val. loss	Rating elo (rel. to ALL)
ALL (reference)	768	0.003134	0.0
$ALL \oplus MB$	1536	0.002824	$-260.9 \pm 5.4$
$ALL \oplus MC$	974	0.003032	-280.9 ± 5.6

- Las predicciones mejoran muy poco (el loss no se reduce tanto).
- Por ende, el costo de las actualizar los features es más alto al beneficio que aportan.
- MB tiene más updates que MC, pero menor loss que compensa.

#### Feature set statistics

Depiction	Feature block	Number of features	Average features		
2 органоп	Non reacture block (valider of reacture)		active per position	added per move	removed per move
<b>+</b>	All	768	14.68	0.98	0.60
<b></b>	Н	96	14.68	0.60	0.43
1	V	96	14.68	0.61	0.43
	D1	180	14.68	0.77	0.52
	D2	180	14.68	0.77	0.52
0-0	РН	1152	8.23	0.92	0.57
9	PV	1152	8.30	0.83	0.53
MB	MB	768	48.93	5.68	4.35
MC	MC	206	12.00	2.34	1.48

# **PQR**

:)

### Conclusión

# Ajedrez

asdasd