Probing as a Technique to Understand Abstract Spaces

Ashlen Plasek
University of Minnesota Morris

Talk Outline

- Linear Algebra
  - Vectors
  - Vector Spaces
  - Linearity
- Machine Learning
  - Training
  - Single Layers
  - Neural Networks
- NLP & Word Embeddings
  - Character Encodings
  - Higher Dimensionality
  - Encoders and Decoders

Talk Outline (Cont'd)

- Evaluating Word Embeddings using Probing
  - A Different Result
- Criticisms of Probing
  - No Control
  - Model Variety
  - Correlation and Causation
- Experiments on Large Language Models

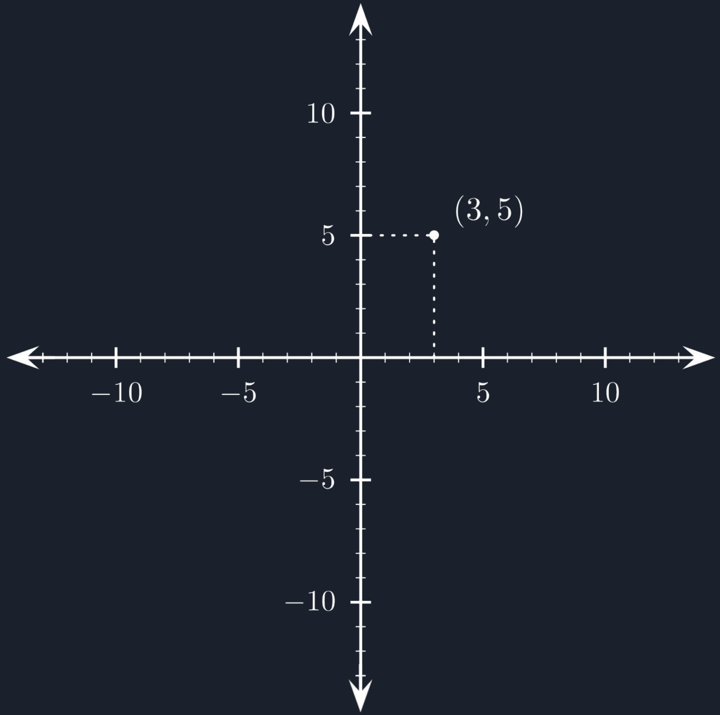
Linear Algebra

Background - Vectors and Vector Spaces

- We can think of vectors as lists of numbers

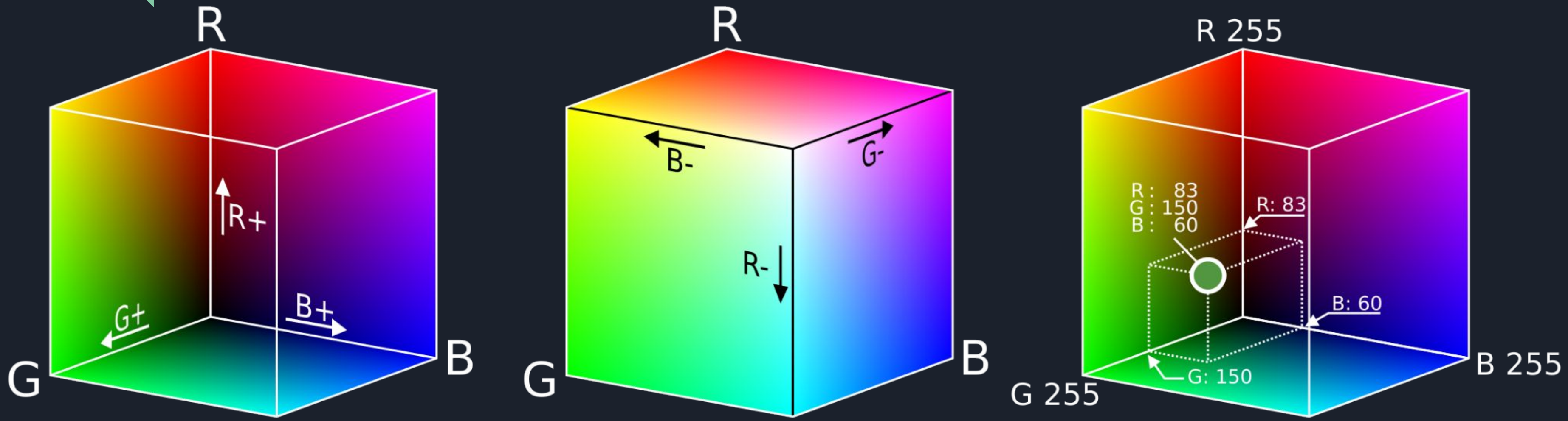
$$\begin{bmatrix} 0.24 \\ 8.02 \\ -3.4 \\ 3.14 \end{bmatrix}$$

Background - Vectors and Vector Spaces



Adapted From: [WikiMedia]

Background - Vectors and Vector Spaces

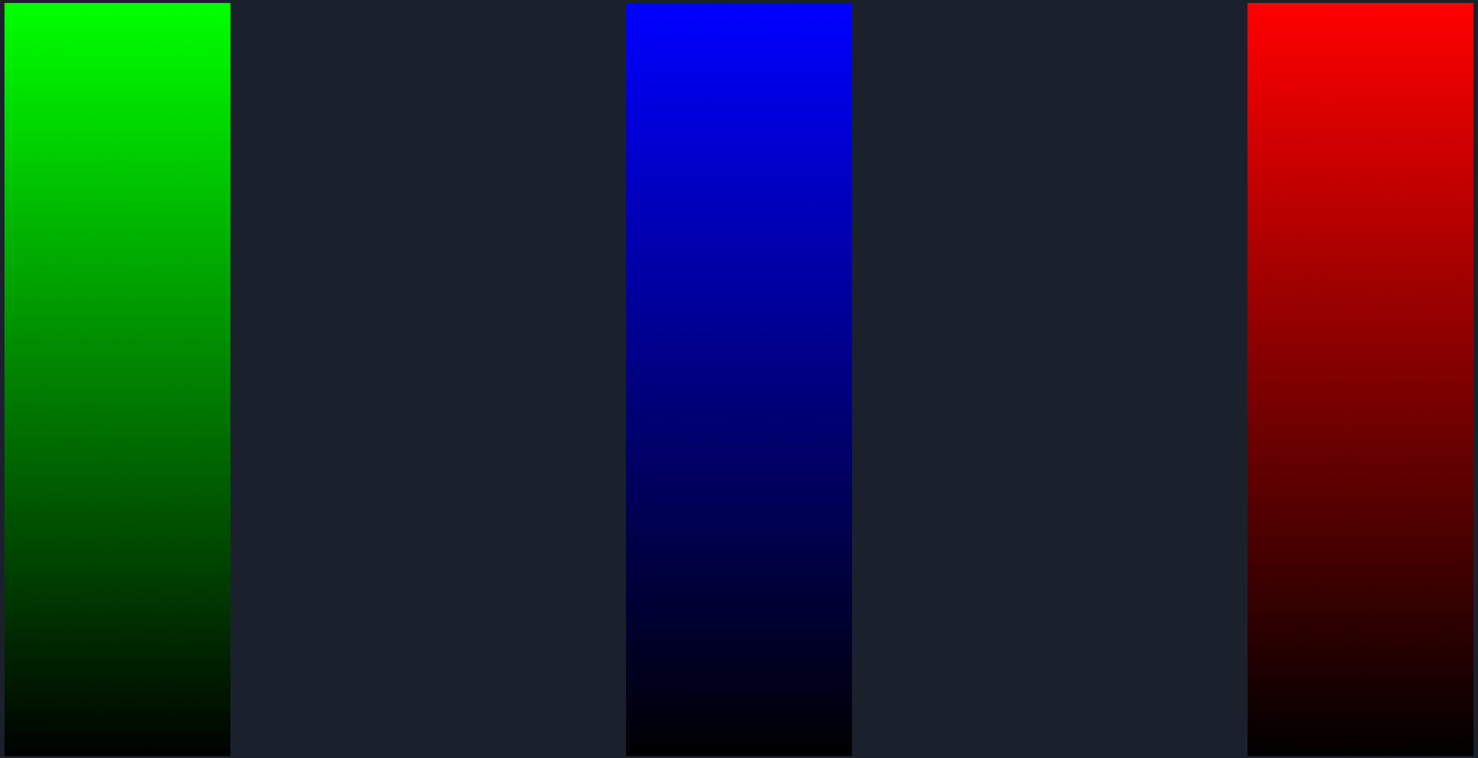


Adapted From: WikiMedia

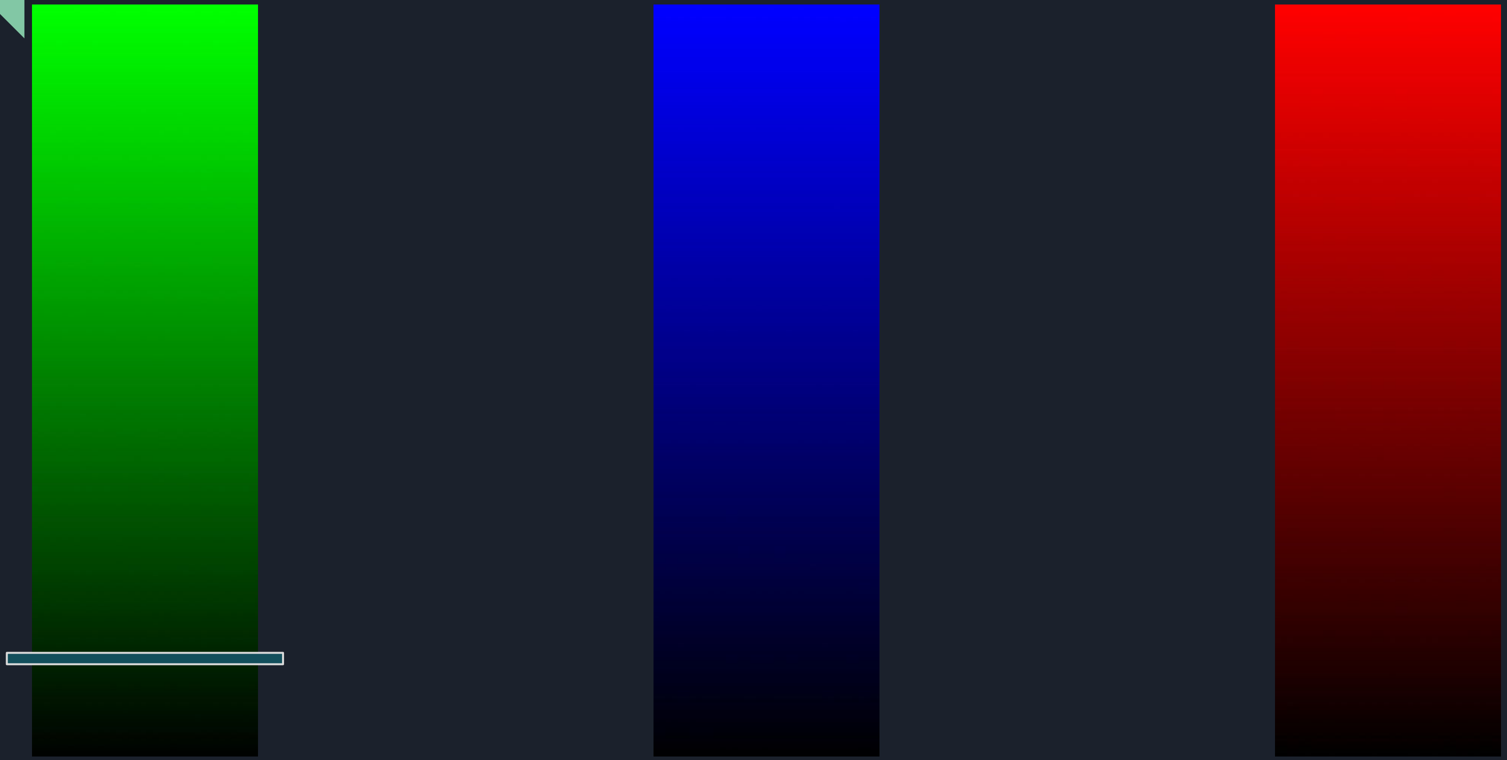
Background - Linear Transformations

- Treating dimensions individually
- Combining dimensions individually by summing

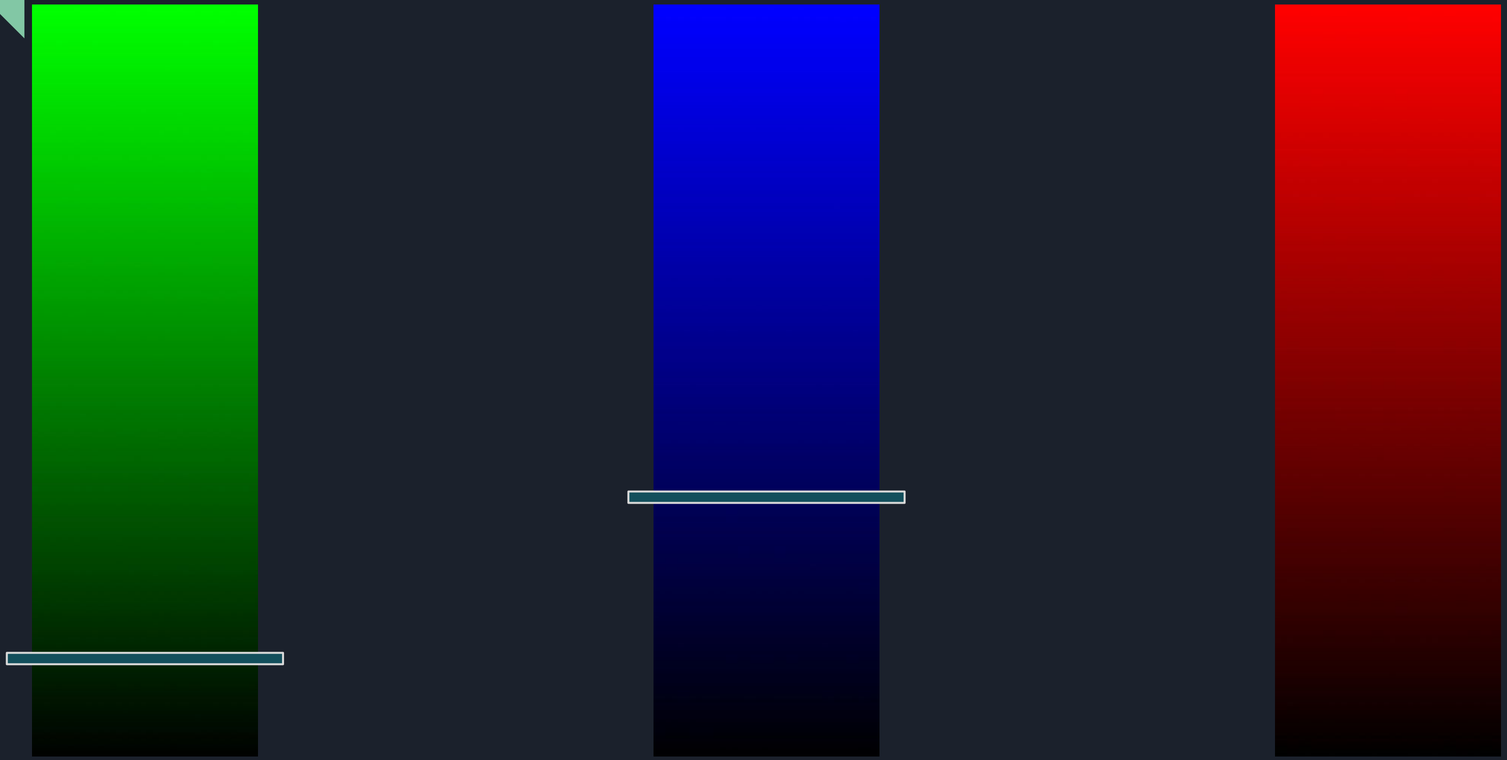
Background - Linear Transformations



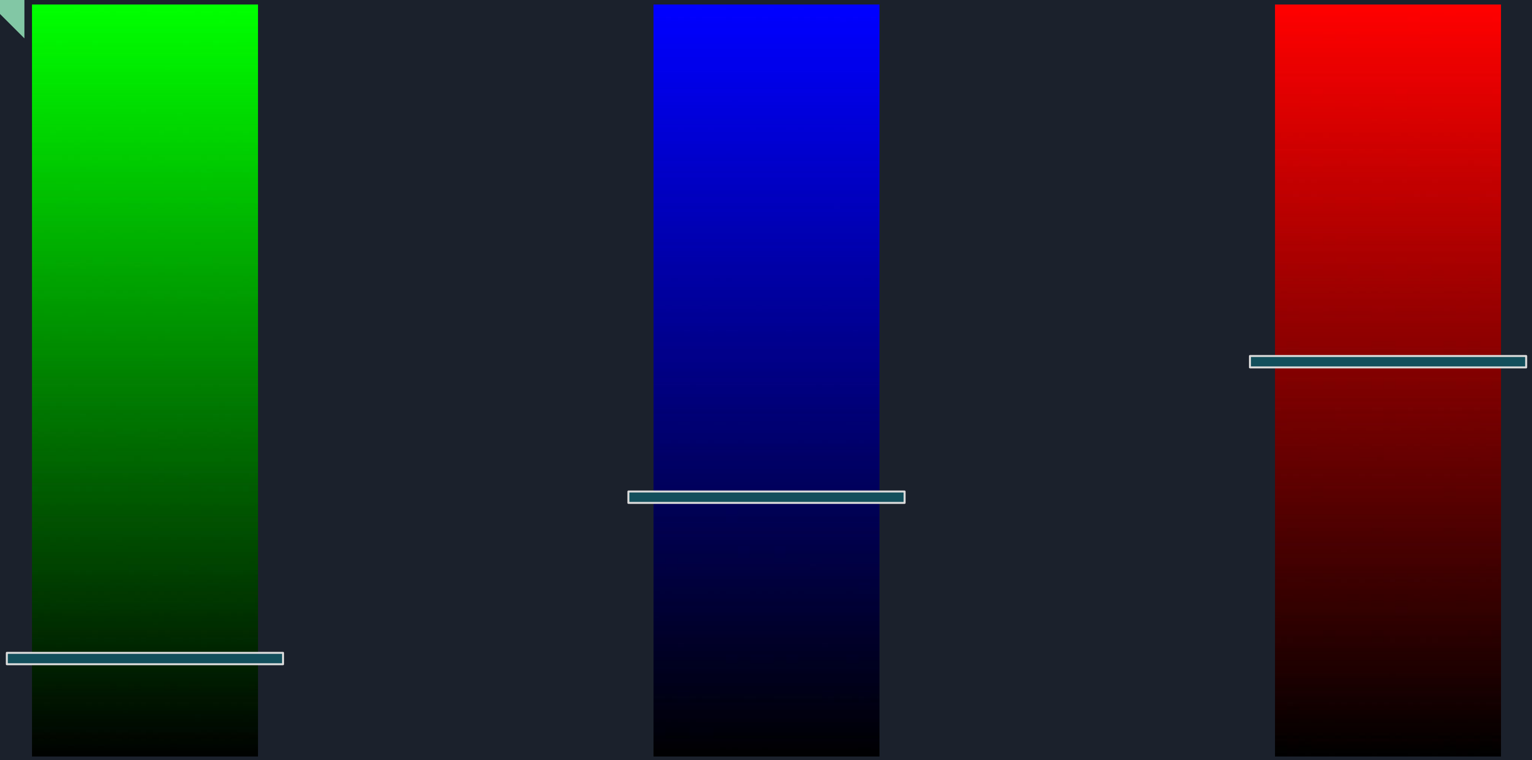
Background - Linear Transformations



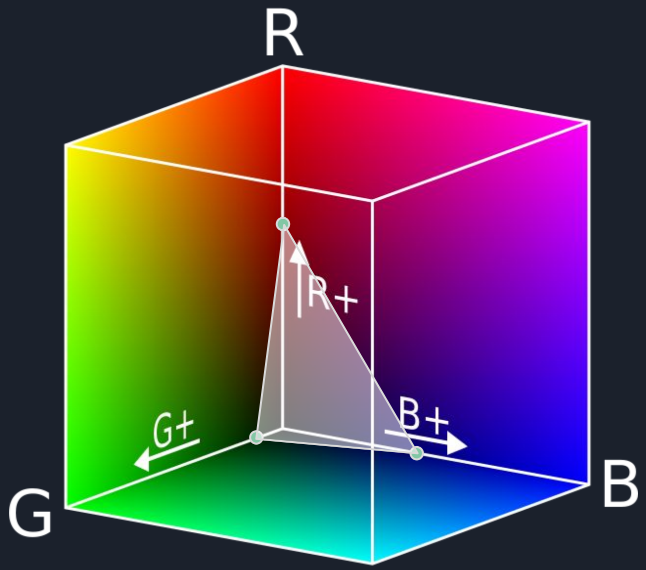
Background - Linear Transformations



Background - Linear Transformations



Background - Linear Transformations



Background - Linear Transformations

```
Luma(Blue) = 1

Luma(Red) = 3

Luma(Green) = 4
```

Background - Linear Transformations

```
Luma(Blue) = 1
Luma(Red) = 3
Luma(Green) = 4
```

```
Luma(Blue) = 1/8
Luma(Red) = 3/8
Luma(Green) = 4/8
```

Background - Linear Transformations

$$\mathrm{Luma}(r \cdot \mathrm{Red} + g \cdot \mathrm{Green} + b \cdot \mathrm{Blue})$$

$$= \tfrac{3}{8} \cdot r + \tfrac{1}{2} \cdot g + \tfrac{1}{8} \cdot b$$

Background - Matrix Multiplication

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix}$$

Background - Matrix Multiplication

$$\begin{bmatrix} A \\ B \\ C \\ D \\ E \end{bmatrix} \longrightarrow \begin{bmatrix} a \\ b \end{bmatrix}$$

Machine Learning

Background - Machine Learning

$$\begin{array}{c} \text{Parameters} \\ \downarrow \\ \text{Input} \longrightarrow \ [\text{Model}] \ \longrightarrow \text{Output} \end{array}$$

Background - Machine Learning

- Trained by providing pairs of input and output

Background - Machine Learning

- Trained by providing pairs of input and output
  - Apply model to the input

Background - Machine Learning

- Trained by providing pairs of input and output
    - Apply model to the input
    - Compare the output with the expected output

Background - Machine Learning

- Trained by providing pairs of input and output
  - Apply model to the input
  - Compare the output with the expected output
  - Use that information to update parameters

Background - Machine Learning

$$
\begin{tikzcd}
 & \text{Parameters} \arrow[d] & \text{Feedback} \arrow[l] \\
\text{Input} \arrow[r] & {[\text{Model}]} \arrow[r] & \text{Output} \arrow[u]
\end{tikzcd}
$$

Background - Single Layer

- Matrix Multiplication
- Linear Classifier

Background - Single Layer

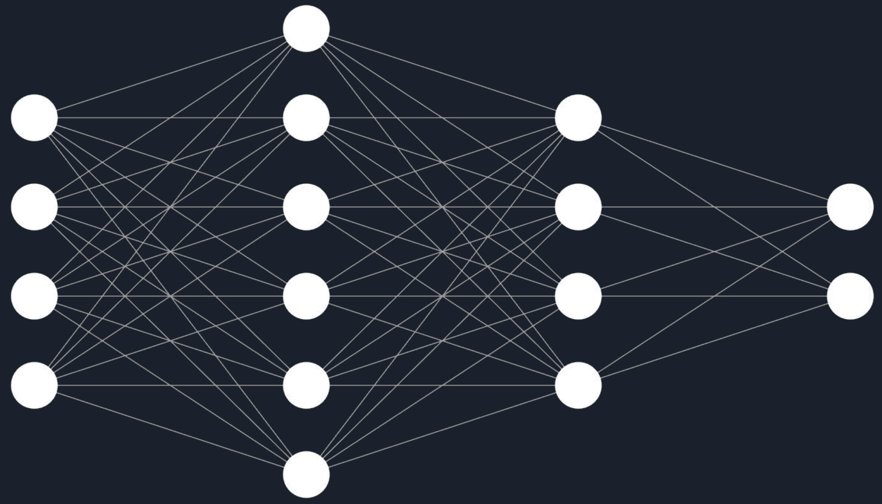
Linear Separation

Background - Neural Networks

Activation Functions

Background - Neural Networks

$$\begin{bmatrix}a\\b\\c\\d\end{bmatrix} \xrightarrow{[\,]\,\sigma} \begin{bmatrix}a\\b\\c\\d\\e\\f\end{bmatrix} \xrightarrow{[\,]\,\sigma} \begin{bmatrix}a\\b\\c\\d\end{bmatrix} \xrightarrow{[\,]\,\sigma} \begin{bmatrix}a\\b\end{bmatrix}$$



Natural Language Processing

NLP - Word Embeddings

How do we represent a word as a vector?

- Character Encodings?
- Many, Many Dimensions
- Encoders and Decoders

NLP - Character Encoding?

$$\begin{bmatrix} 86 & V \\ 101 & e \\ 99 & c \\ 116 & t \\ 111 & o \\ 114 & r \\ 0 & \\ \vdots & \\ 0 & \end{bmatrix}$$

Figure 1. ASCII-based embedding of the word "vector"

NLP - Character Encoding?

Limitations:

- Limited in length
- Doesn’t play well as a vector

[6]

NLP - A Dimension for Every Word

$$\begin{bmatrix} 0 & a \\ 0 & Aalenian \\ \vdots & \\ 0 & vectitory \\ 1 & vector \\ 0 & vectorcardiogram \\ \vdots & \\ 0 & zythum \\ 0 & Zyzzyva \end{bmatrix}$$

Figure 2. Simple embedding for the word "vector"

NLP - A Dimension for Every Word

Advantages:

- Plenty of space for machine learning

Disadvantages:

- 1.3 Million dimensions

[4]

NLP - A Useless Transformation

$$\begin{array}{ccccc} & & \text{Parameters} & & \\ & & \downarrow & & \\ \text{Input} & \longrightarrow & [\text{Model}] & \longrightarrow & \text{Output} \end{array}$$

NLP - A Useless Transformation

$$\text{Parameters}$$
$$\downarrow$$

“Vector”

$$\text{Input} \longrightarrow [\text{Model}] \longrightarrow \text{Output}$$

[4]

NLP - A Useless Transformation

$$\text{Parameters}$$

$$\downarrow$$

“Vector” “Vector”

$$\text{Input} \longrightarrow [\text{Model}] \longrightarrow \text{Output}$$

NLP - Encoders and Decoders

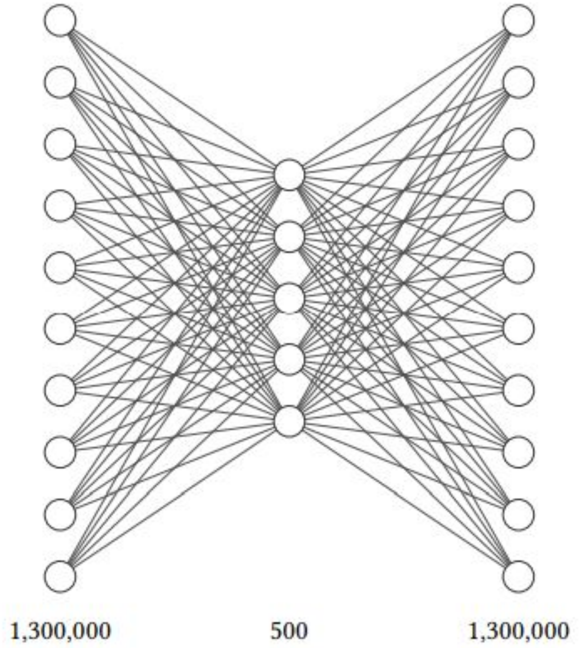


Figure 3. Encoder Decoder Pair https://alexlenail.me/NN-SVG/index.html

NLP - Encoders and Decoders

$$\begin{array}{ccccc} & & \text{Parameters} & & \\ & & \downarrow & & \\ \text{Input} & \longrightarrow & \left[\text{Model}\right] & \longrightarrow & \text{Output} \end{array}$$

Probing

Probing - Assessing Word Embeddings

- Part of Speech?
- Plurality?
- Ends-in-’s’-ness?

[1]

Probing - A Different Application

“Traditional” Machine Learning:

The parameters are the result, the accuracy is just a measurement

Probing:

The accuracy in training is the result, the parameters are coincidental

Criticisms

[2]

Criticisms

No Control Reference

Models Can Vary

Correlation and Causation

Criticisms - No Control

What is the baseline?

Criticisms - Model Variety

When does one use a Linear Classifier?

What activation function should be used if any?

Criticisms - Correlation and Causation

How does one remove a property?

[2]

Criticisms - Correlation and Causation

How do we know the property we want is present?

- Retrain the encoder
- Modifying the embedded vector

Using Our Toolkit

[5]

Experiments on Large Language Models

$$\begin{array}{ccc} & \text{Parameters} & \\ & \downarrow & \\ \text{Input} \longrightarrow & [\text{Model}] & \longrightarrow \text{Output} \end{array}$$

[5]

Experiments on Large Language Models

Parameters

$\downarrow$

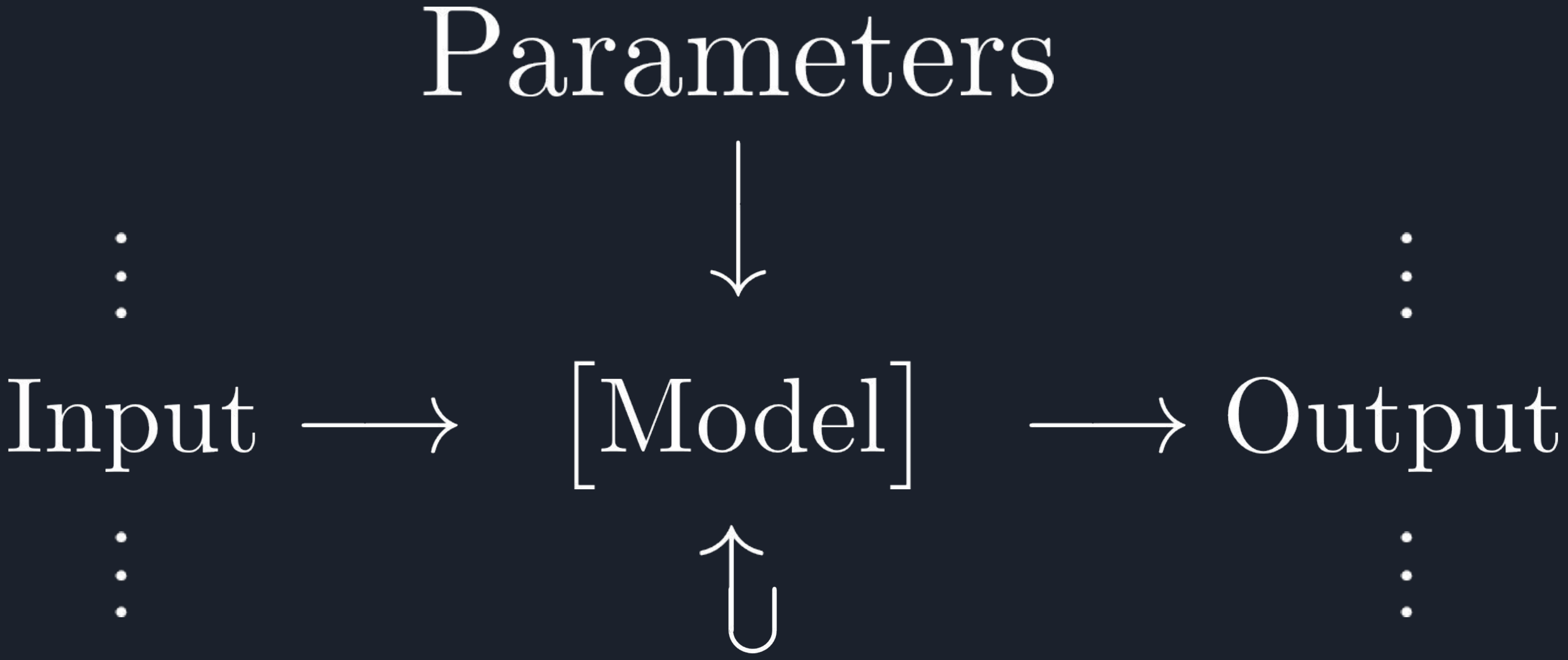
$\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots$

Input $\longrightarrow$ [Model] $\longrightarrow$ Output

$\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots$

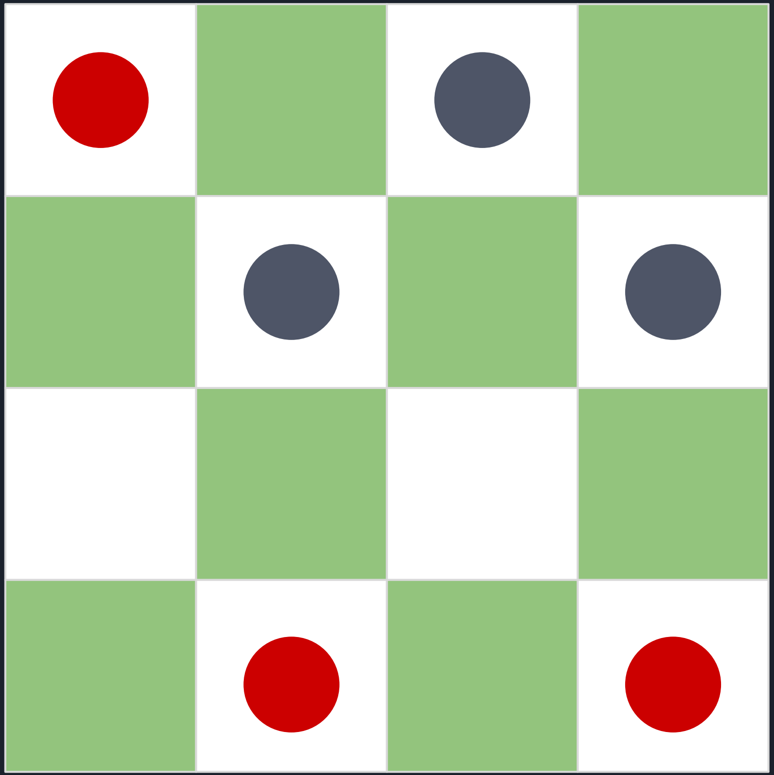
[5]

Experiments on Large Language Models



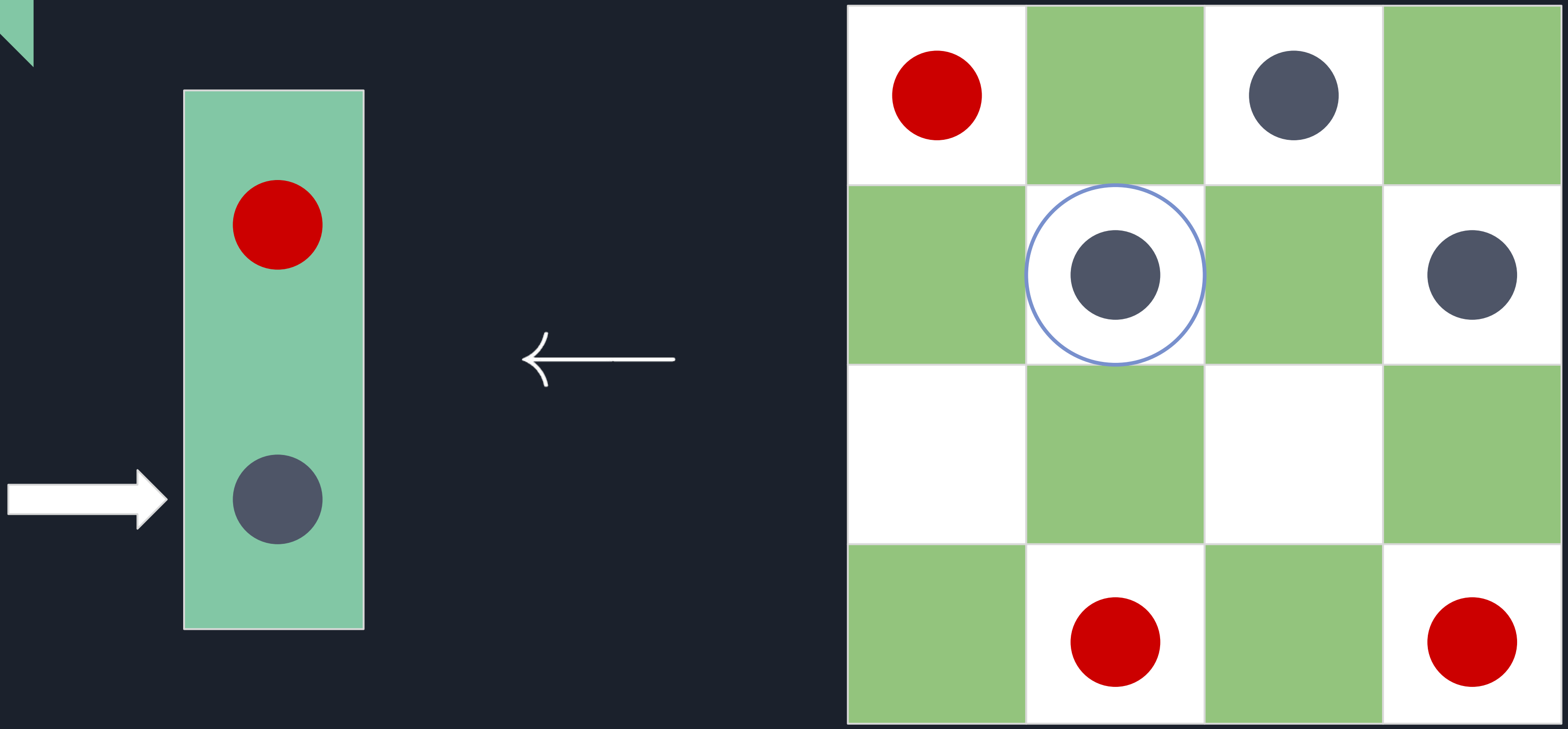
[5]

Experiments on Large Language Models

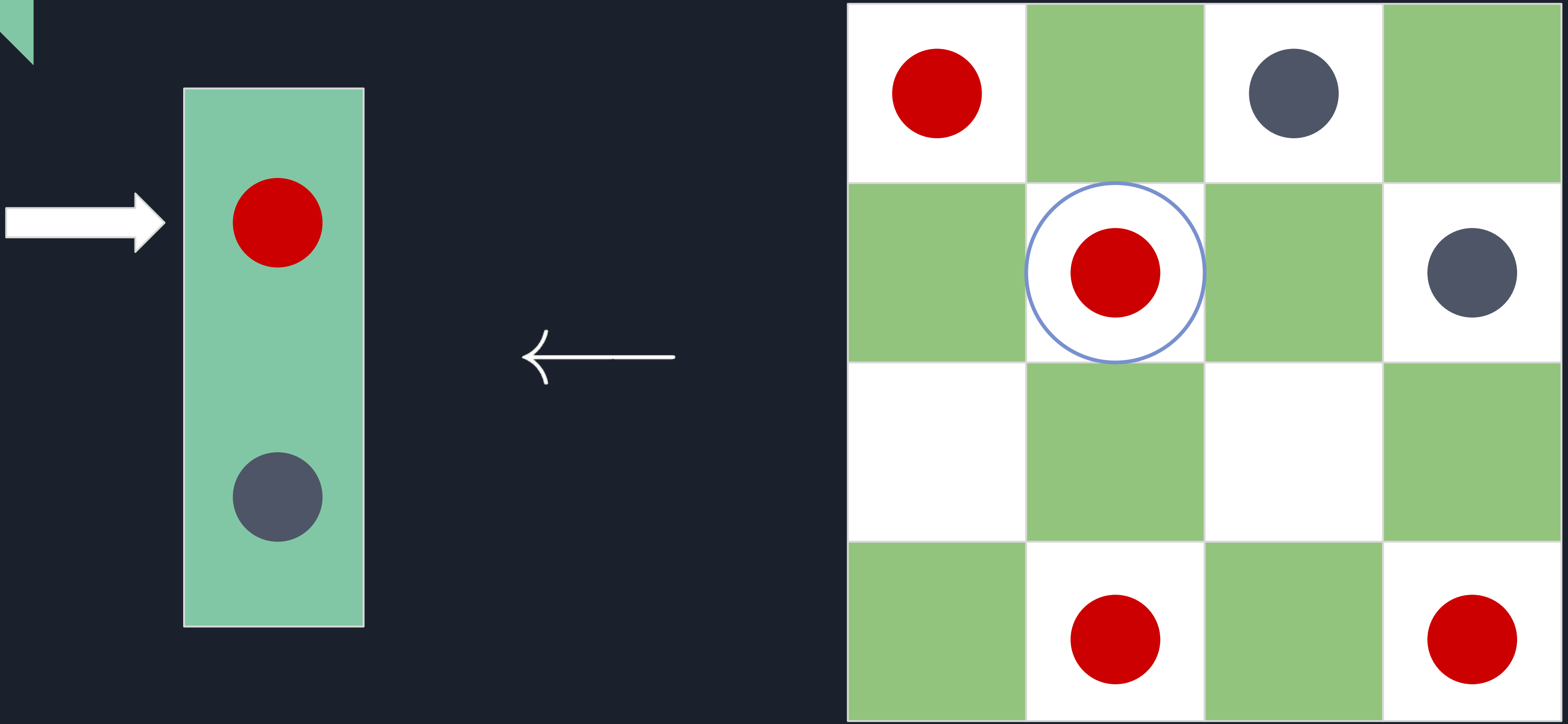


[5]

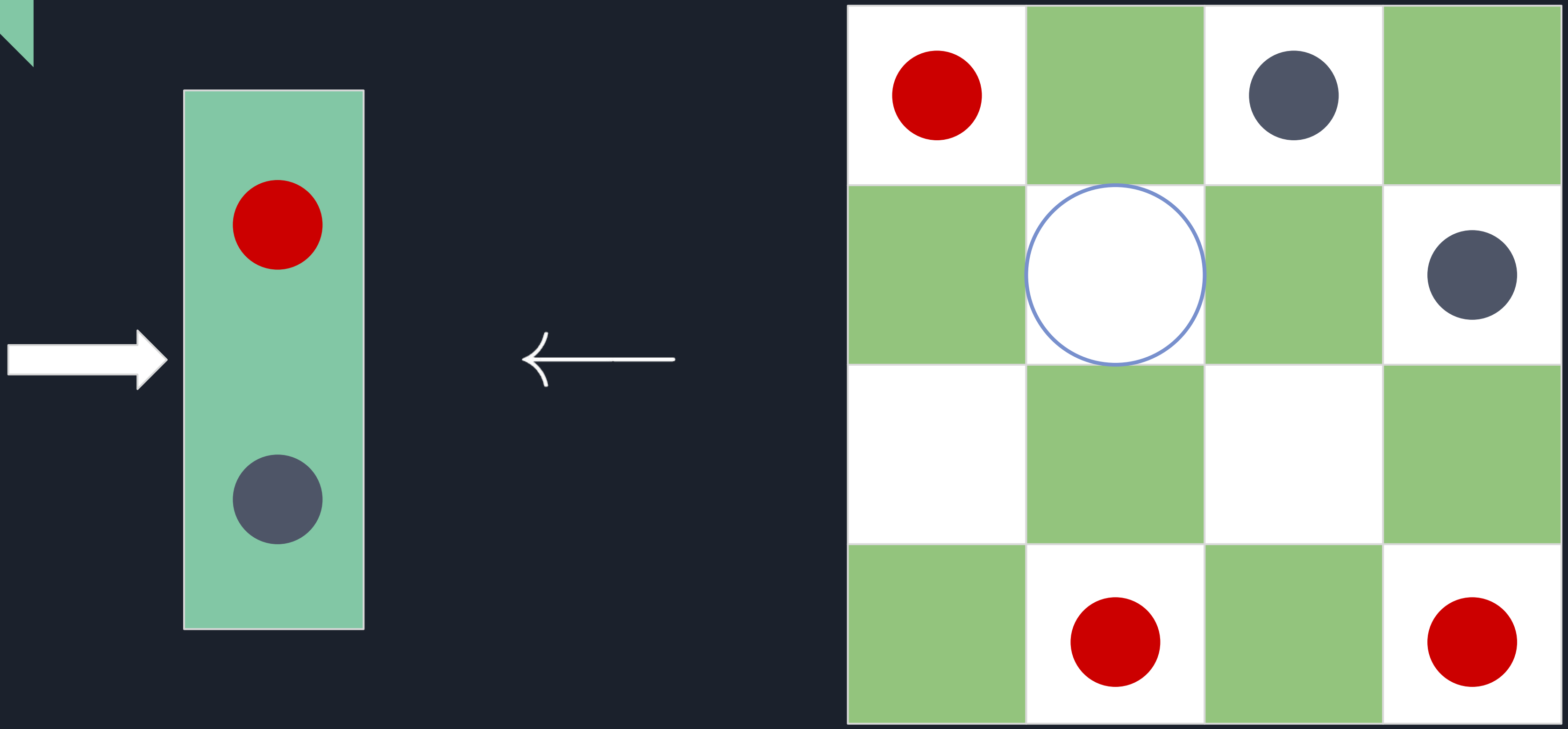
Experiments on Large Language Models



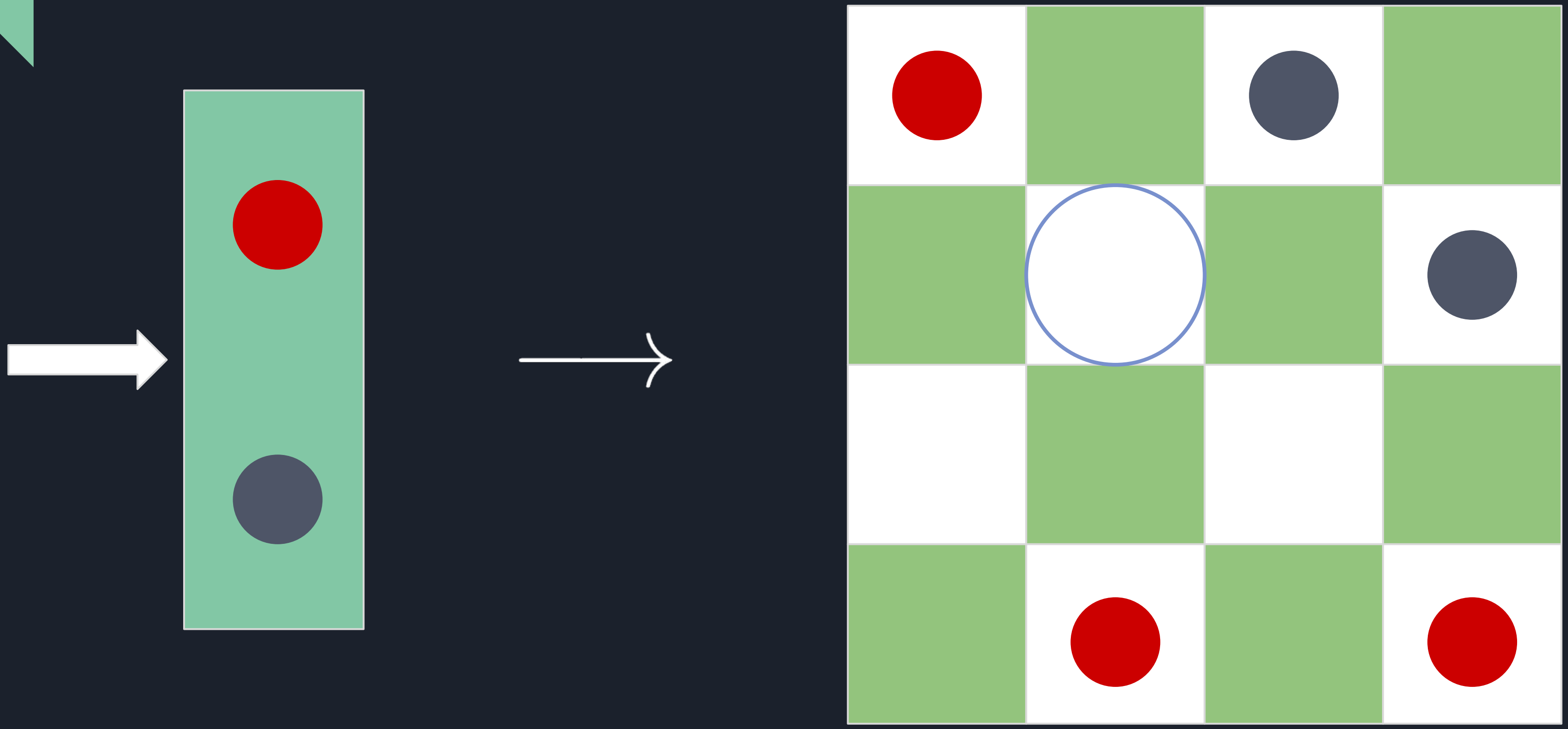
Experiments on Large Language Models



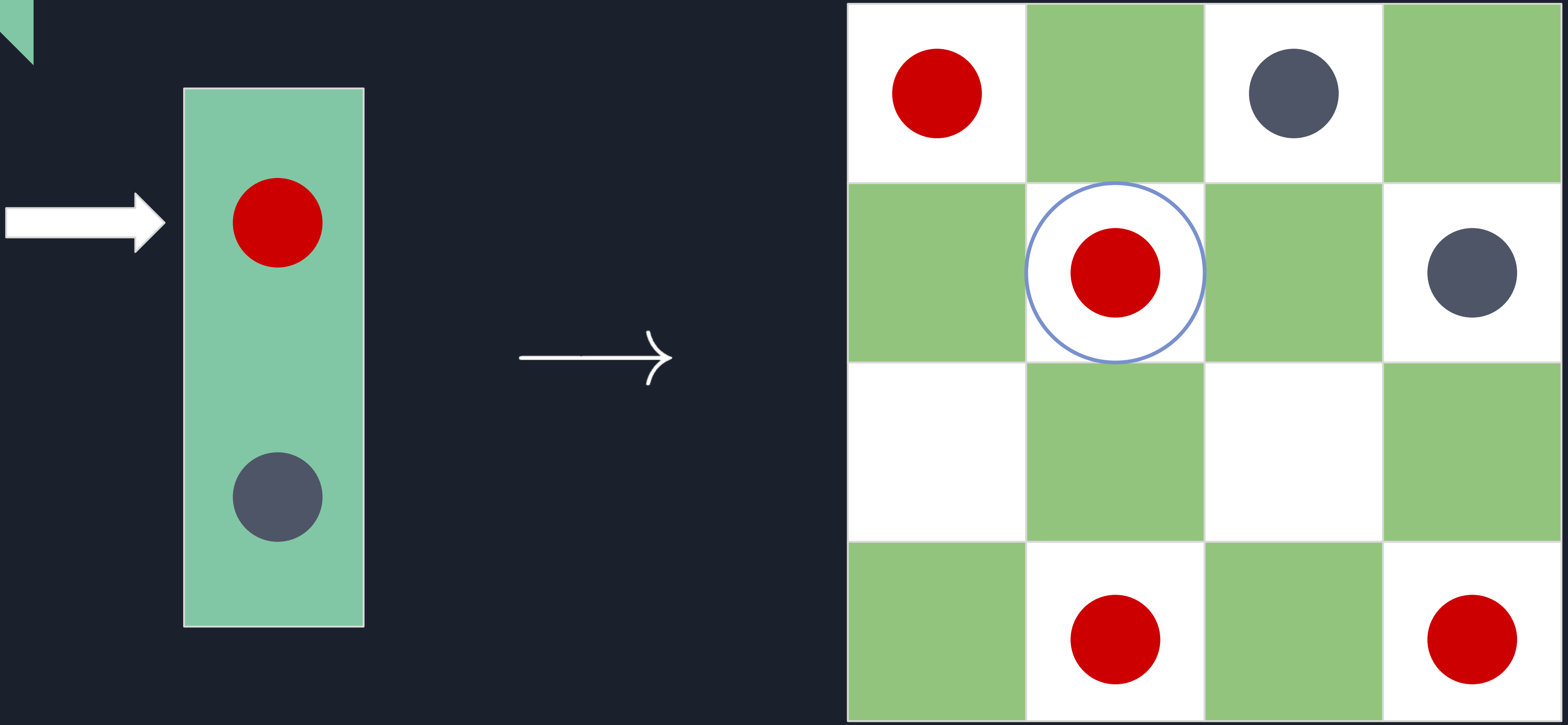
Experiments on Large Language Models



Experiments on Large Language Models

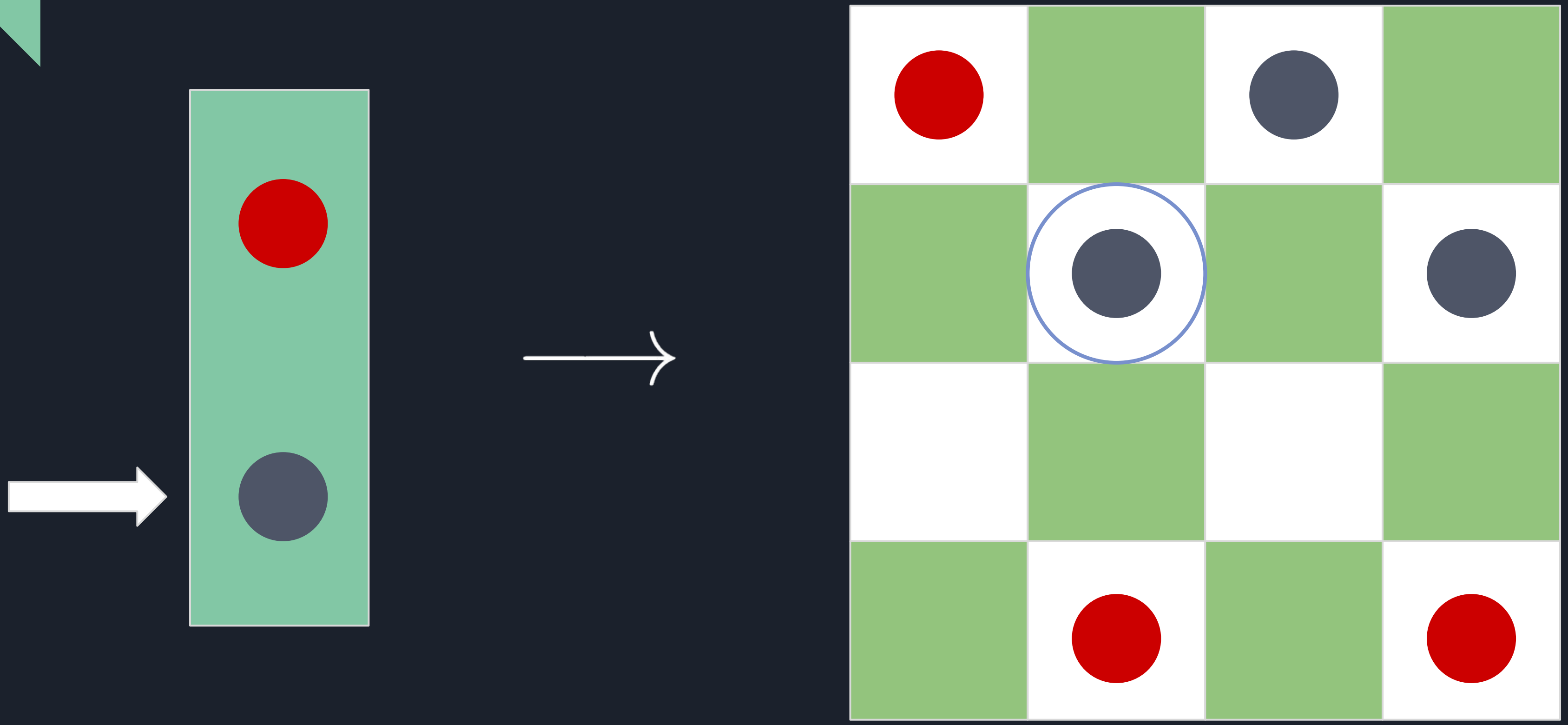


Experiments on Large Language Models



[5]

Experiments on Large Language Models



Conclusion

Conclusion

- Linear Algebra
- Machine Learning
- NLP & Word Embeddings
- Evaluating Word Embeddings using Probing
- Criticisms of Probing
- Experiments on Large Language Models

References

[1] Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. (2016). https://doi.org/10.48550/ ARXIV.1610.01644

[2] Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. Computational Linguistics 48, 1 (04 2022), 207–219. https://doi.org/10.1162/coli_a_00422 arXiv: https://direct.mit.edu/coli/article-pdf/48/1/207/2006605/coli_a_00422.pdf

[3] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, 2126–2136. https://doi.org/10.18653/v1/P18-1198

References

[4] Arne Köhn. 2015. What's in an Embedding? Analyzing Word Embeddings through Multilingual Evaluation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, 2067–2073. https://doi.org/10.18653/v1/D15-1246

[5] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. (2022). https://doi.org/10.48550/ARXIV.2210.13382

[6] Wiktionary. 2023. Statistics — Wiktionary. https://en.wiktionary.org/wiki/Special:Statistics. [Online; accessed 01-March-2023].

Questions?