Emergence of Clustering in Self-Attention

Anton Sugolov and Murdock Aubry

MAT1510, Deep Learning: Theory and (or) Data Science

November 27, 2023

Table of Contents

What is Attention?

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

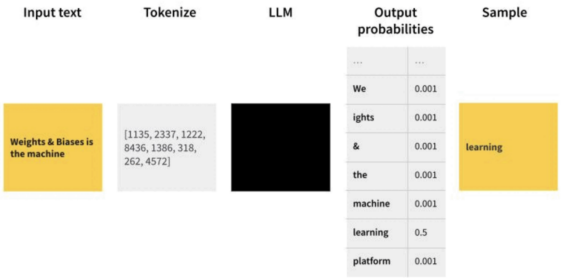
Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

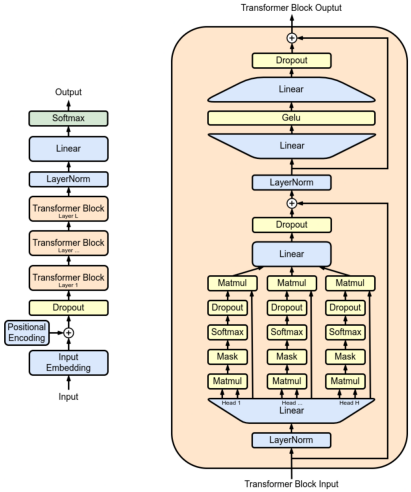
- Proposed by Vaswani et al. in 2017
- Central role in transformers
- Key in LLMs like ChatGPT

Large Language Models



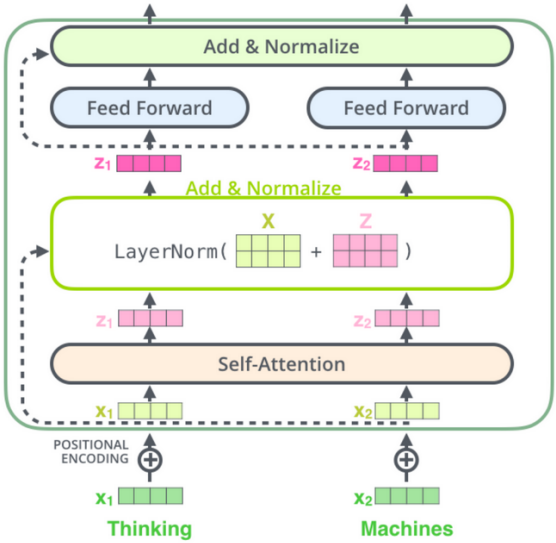
1. words $\rightarrow$ tokens
2. tokens $\rightarrow$ vectors
3. vectors $\rightarrow$ model
4. model $\rightarrow$ next token

GPT Transformer



- Repeated application of tranformer blocks

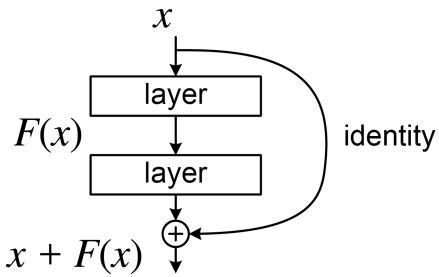
GPT Block



2 key features

- residual connections
- self-attention

Key Feature 1: Residual Connections



- very important in neural networks
- add output of layers onto input
- Feedforward: $x \mapsto f_\theta(x)$
- ResNet: $x \mapsto x + f_\theta(x)$

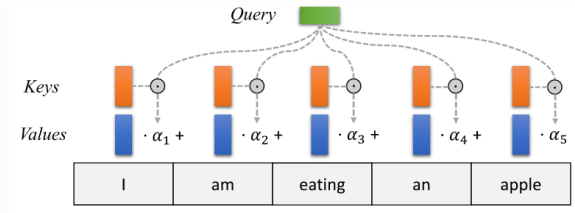
Modeling a change in the input?

Key Feature 2: Self-Attention

Attention is built to identify what part of the input is 'important'

$$\text{Attention}(Q, K, V)$$

is a function of **queries**, **keys**, and **values**



Think of a dictionary:

- **Queries** : words you're looking up
- **Keys** : words in the dictionary
- **Values**: meaning of words in the dictionary

Compare queries and keys (dot product), give a weighted sum of the values.

Key Feature 2: Self-Attention

Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- QK^T: Compare keys and queries with dot products
- softmax: get row-wise probabilities (**stochastic matrix!**)
- V: apply to values

What is softmax? For a vector $x \in \mathbb{R}^n$, define a normalized probability vector $(p_1, \ldots, p_n) \in \mathbb{R}^n$ with

$$p_i = \frac{\exp(x_i)}{\sum_{\ell=1}^{n} \exp(x_\ell)}$$

Turn a vector into a probability vector, larger entries given large probabilities.

Key Feature 2: Self-Attention

Self-attention applied to some input X:

$$\text{Attention}(W_Q X, W_K X, W_V X)$$

- Attention on weighted transformations of X
- Trainable combination of input used for keys, queries, values

Multihead Self-attention is a combination of many self-attentions

$$\begin{bmatrix} H_1 & H_2 & \cdots & H_k \end{bmatrix} \cdot W_0$$

where $H_i = \text{Attention}(W_{Q_i} X, W_{K_i} X, W_{V_i} X)$ and W_0 are some overall combination weights.

Question

- Token embeddings X
- ResNet and Attention give rise to 'iterated dynamics'

$$X \mapsto X + \text{Attention}(W_Q X, W_K X, W_V X)$$

- What is the **effect of attention dynamics on information representation**?

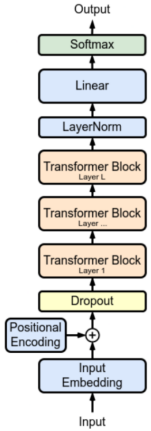


Table of Contents

THE EMERGENCE OF CLUSTERS IN SELF-ATTENTION DYNAMICS

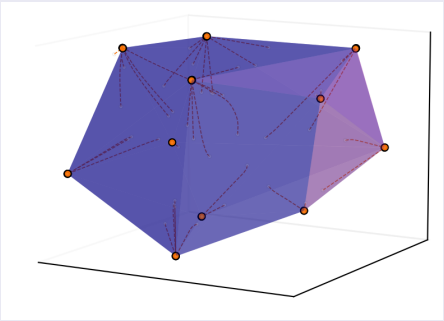
BORJAN GESHKOVSKI, CYRIL LETROUIT, YURY POLYANSKIY,
AND PHILIPPE RIGOLLET

Abstract. Viewing Transformers as interacting particle systems, we describe the geometry of learned representations when the weights are not time dependent. We show that particles, representing tokens, tend to cluster toward particular limiting objects as time tends to infinity. Cluster locations are determined by the initial tokens, confirming context-awareness of representations learned by Transformers. Using techniques from dynamical systems and partial differential equations, we show that the type of limiting object that emerges depends on the spectrum of the value matrix. Additionally, in the one-dimensional case we prove that the self-attention matrix converges to a low-rank Boolean matrix. The combination of these results mathematically confirms the empirical observation made by Vaswani et al. [29] that *leaders* appear in a sequence of tokens when processed by Transformers.

Geshkovski et al.

The emergence of clusters in self-attention dynamics. B. Geshkovski, C. Letrouit, Y. Polyanskiy, and Philippe Rigollet. (2023).

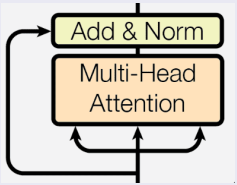
Token flows



- View token embeddings as particles
- Self-attention defines 'dynamics' on particles
- Study clustering of geometric representations after repetition of dynamics

Setting

Dynamical framework



$$x(t+1) = x(t) + f_\theta(x(t))$$
$$= x(t) + \dot{x}(t)$$

$$\dot{x}_i(t) = \sum_{j=1}^{n} \underbrace{P_{ij}(t)}_{\text{attention mat.}} \quad V x_j(t)$$

- Consider reptitions of self-attention
- Study change in embeddings (particles) as time variable
- Residual connection modifies input to self-attention matrix
- Defines 'dynamics' on particles

Setting

Self-attention matrix

$$P_{ij}(t) = \frac{e^{\langle Qx_i(t), Kx_j(t)\rangle}}{\sum_{\ell=1}^{n} e^{\langle Qx_i(t), Kx_\ell(t)\rangle}} \quad (i,j) \in [n]^2$$

- $P(t) = \text{softmax}\left(Qx(t)(Kx(t))^T\right)$
- $x(t) = (x_1(t), \ldots, x_n(t)) \in \mathbb{R}^{n \times d}$ tokens
- Q, K are usually denoted W_Q, W_K

- Iterated dynamics of self-attention matrix
- Under what conditions on Q, K, V can we describe dynamics as $t \to \infty$?

Summary

Theorem assumptions on Q, K, V and limiting geometric shapes

$\mathbf{V}$	$\mathbf{Q}, \mathbf{K}$	**Limiting Representations**
$V = -I$	$Q = K = I$	cluster at origin
$V = I$	$Q^T K > 0$	vertices of convex polytope
$\lambda_1(V) > 0$	$\langle Q\varphi_1, K\varphi_1 \rangle > 0$	3 parallel hyperplanes
$\left\| V^2 x \right\| \geq \|Vx\|^2$	$Q^T K > 0$	product of polytope and subspaces

Theorem

For $x_i(t) \in \mathbb{R}$ and as $t \to \infty$, $P(t) \to P^*$ where P^* is a low-rank boolean matrix.

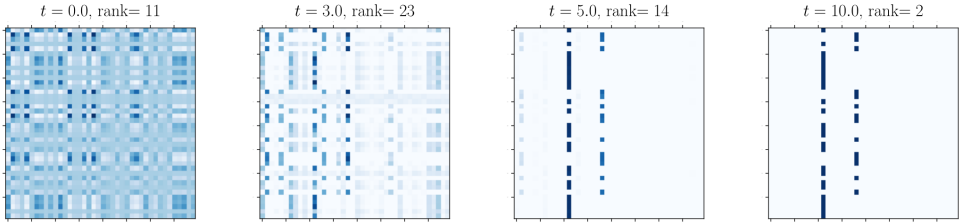


Figure: Example of Theorem 1 result when $Q = K = V = I$ with $n = 40$ tokens.

Geskovski et al. Convex Polytopes

Theorem

When $V = I$ and $Q^T K > 0$ (positive matrix) then points flow to corners of convex polytope.

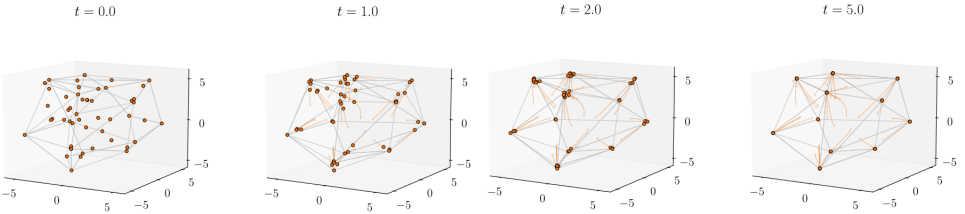


Figure: Example of Theorem 2 result when $Q = K = V = I \in \mathbb{R}^{3\times 3}$ with $n = 40$ tokens.

Geskovski et al. Hyperplanes

Theorem

Suppose the eigenvalue λ of V with largest modulus is real, positive, and has algebraic multiplicity 1 (simple). Suppose $\langle Q\phi_1, K\phi_1 \rangle > 0$ for any ϕ_1 in the eigenspace of λ. There exist at most three parallel hyperplanes that the tokens converge to.

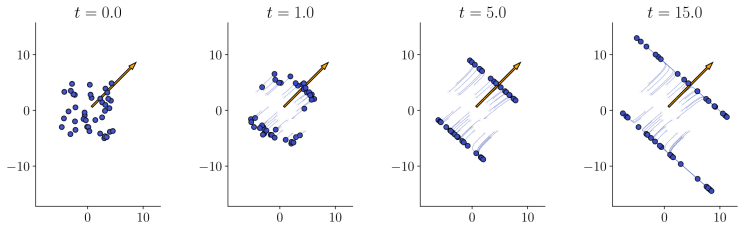


Figure: Example of result when $Q = K = I \in \mathbb{R}^{2\times 2}$, V is random with eigenvalues $\lambda \in \{1.35, -0.07\}$.

Geskovski et al. Well-Posedness

Theorem (Well-posedness)

Solutions to ODEs defined by self-attention dynamics exist, are unique, and are stable.

A major contribution of this work is proving this.

Questions

- Similar dynamics occur for trained weights from real transformers?
- Effect of multihead self-attention?
- How does number of heads affect dynamics?
- How does token initialization affect dynamics?
- Does the number of tokens affect dynamics?

Table of Contents

ALBERT Transformer Weights by Lan et. al

Implementation of multihead dynamics for the weights of a real trained transformer.

Observation

Multiple cluster points may emerge with trained weights.

ALBERT Transformer Weights by Lan et. al

Repeated weight sharing between multi-head layers. Trained value matrix eigenvalue for head 5 is positive and real.

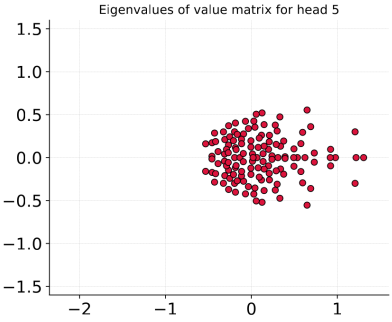


Figure: Top eigenvalue of V is real.

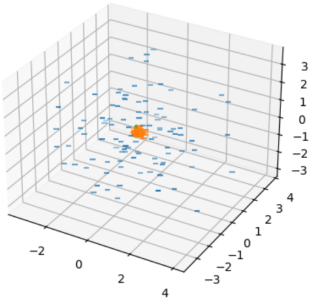
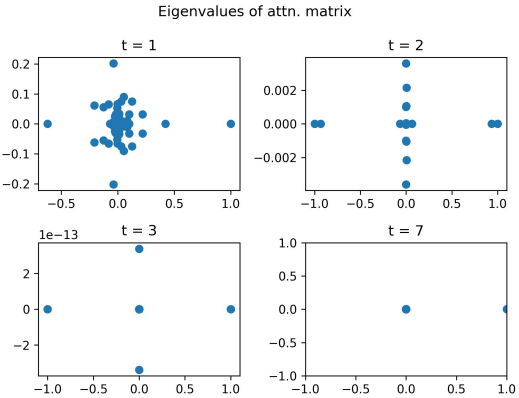
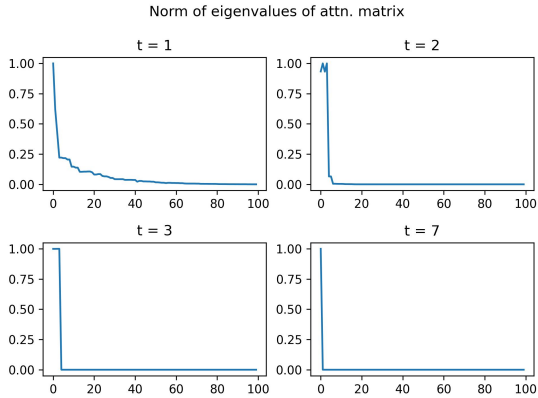


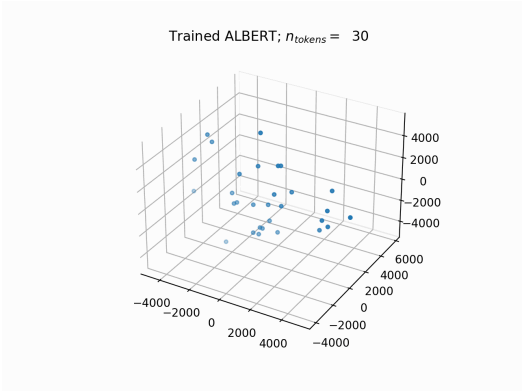
Figure: PCA of flows with head 5 weights shows clustering.

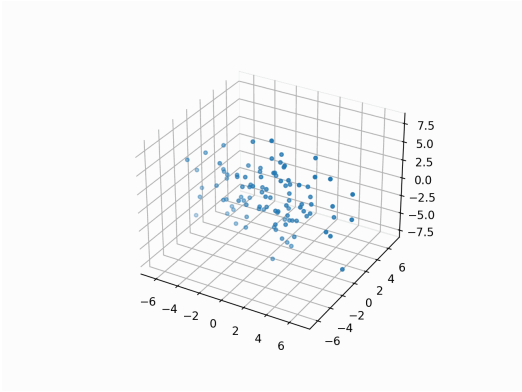
ALBERT Transformer Weights by Lan et. al.





ALBERT Transformer Weights by Lan et. al. Multihead Dynamics





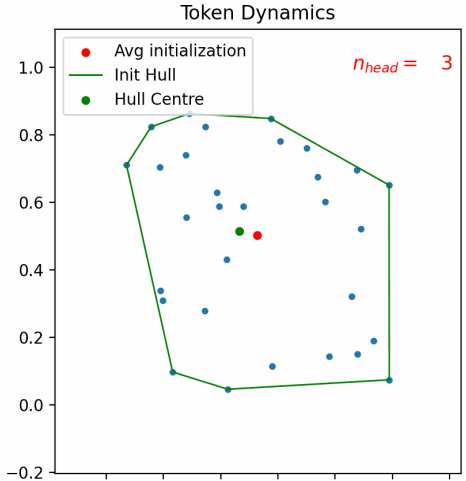
Affect of Token Initialization

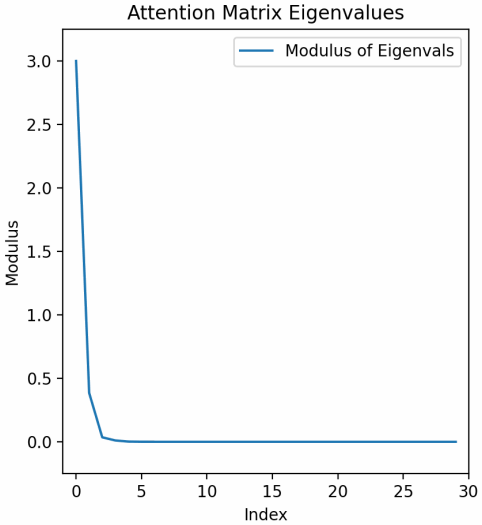
How do the distribution of the initized tokens affect the long-term dynamics.

Observation

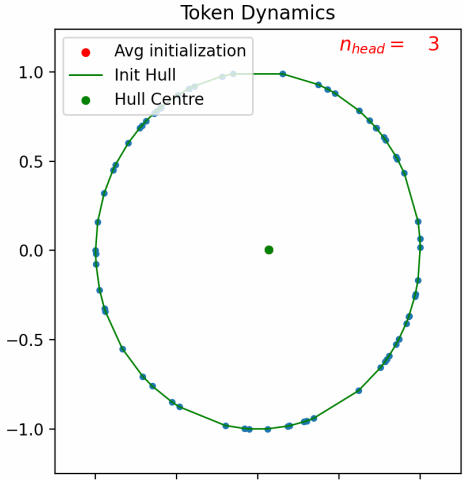
The initialization greatly affects the long-term clustering.

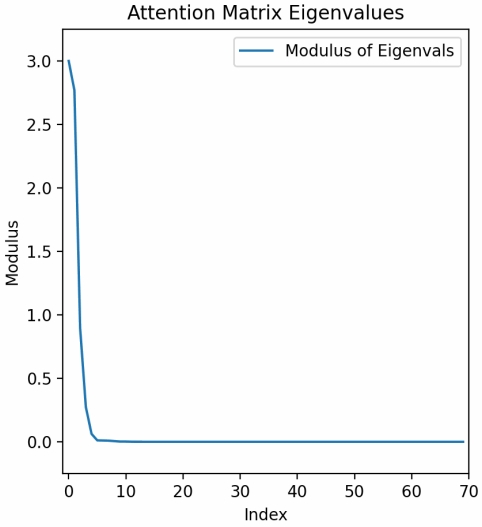
Multihead Dynamics - Eigenvalue Analysis





Multihead Dynamics - Unit Circle Initialization





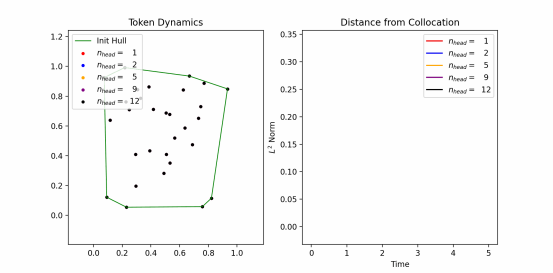
Multihead Dynamics - Number of Heads

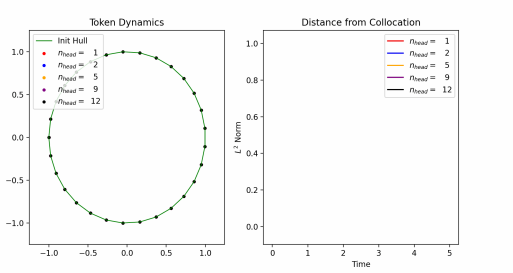
How does the number of heads affect the dynamics of each token?

Observation

The number of heads is directly related to the rate of convergence of the tokens.

Multihead Dynamics - Rate of Convergence





Rate of Convergence - Justification

Multihead self-attention dynamics:

$$\dot{X} = \text{Multihead}(X, X, X)X$$

where

$$\text{Multihead}(X, X, X) = \text{Cat}[h_1 \quad h_2 \quad \dots \quad h_n]W_0$$

and

$$h_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V).$$

If $\text{Multihead}(X, X, X) \sim$ constant, then solutions approximately take the form

$$X(t) = X_0 \exp\left(\text{Multihead(X, X, X)t}\right) \approx X_0 \exp\left(n\text{Avg}_i\text{Attention}(XW_i^Q, XW_i^K, XW_i^V)t\right)$$

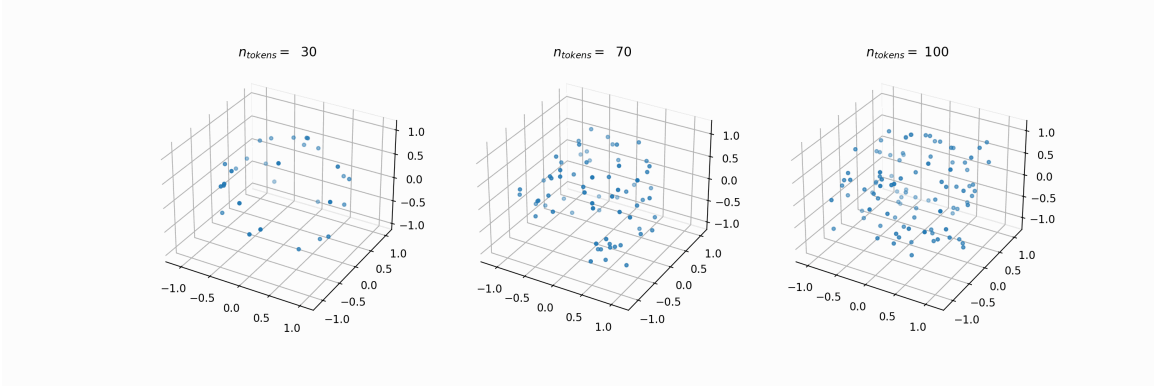
Multihead Dynamics - Number of Tokens

How does the number of tokens affect the clustering patterns?

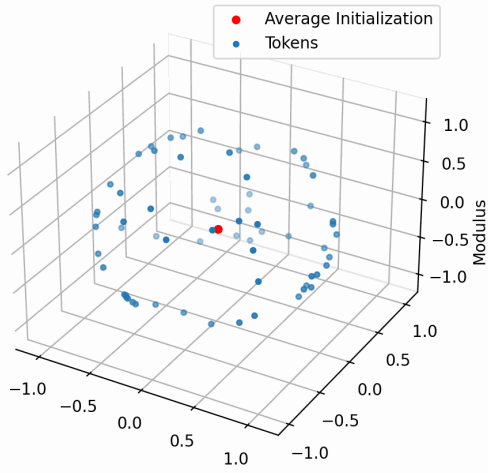
Observation

Increasing the number of tokens can result in multiple clustering points. In particular, larger number of tokens leads to multiple stable non-zero eigenvalues of the attention matrix.

Multihead Dynamics - 3 Dimensions



Multihead Dynamics - Eigenvalue Analysis



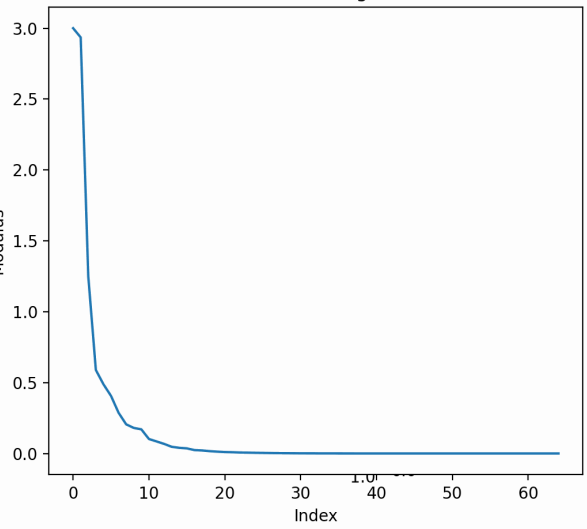


Table of Contents

Next Experiments

- Test dynamics for on the weights of more trained transformer models. Iterpret dependence of dynamics on the model architecture.
- Observe dynamics when a real tokenized sentence is passed.
- Explore relationship between Neural collapse.
- Quantify the relationship between limiting structure and number of tokens and token initialization.
- Comparison between dynamics predicted by the Master equation.

The Master Equation

- The dynamics of the tokens are governed by the discrete-time versions of

$$\dot{X}(t) = f_\theta(X(t)) = P(X(t))X(t)$$

  where $P(t)$ is the attention matrix.
- Analogy with the time-dependent Master equation:

$$\frac{d\vec{P}}{dt} = A(t)\vec{P}(t)$$

- If A is approximately constant, then the solutions are given by

$$\vec{P}(t) = \sum_{i=1}^{n} c_i e^{\lambda_i t}\vec{v}_i(t)$$