Friedrich-Alexander-Universität
Naturwissenschaftliche Fakultät

Friedrich-Alexander-Universität
DYNAMICS, CONTROL, MACHINE LEARNING AND NUMERICS

CLUSTERING IN PURE-ATTENTION HARDMAX TRANSFORMERS AND ITS ROLE IN SENTIMENT ANALYSIS

Albert Alcalde† **Giovanni Fantuzzi**† **Enrique Zuazua**†

†Chair for Dynamics, Control, Machine Learning and Numerics – AvH Professorship, FAU Erlangen-Nürnberg, Germany.

Introduction

The transformer is a fundamental model in contemporary machine learning whose power is attributed to **self-attention layers**. Heuristically, they capture 'context' by identifying relations between components of the transformer's input data, which is observed to facilitate prediction tasks. Our work [1] rigorously justifies these heuristics by proving that:

- self-attention layers entail a clustering effect in the infinite-depth limit
- transformers leverage the clustering to capture 'context' in sentiment analysis.

Pure-attention hardmax transformers

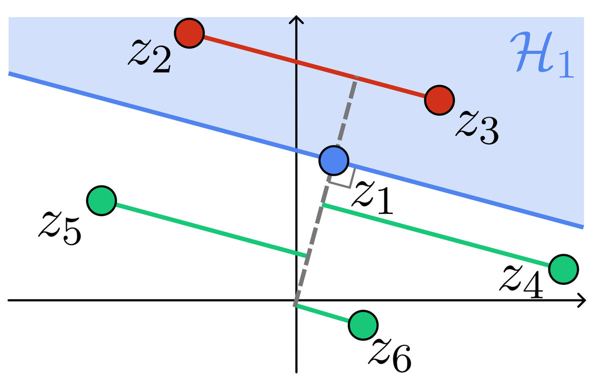
We study **pure-attention hardmax transformers**, parameterized by a symmetric positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and a scalar $\boldsymbol{\alpha} > 0$. They act on a collection of tokens $z_1, \ldots, z_n \in \mathbb{R}^d$. Given initial token values $z_1^0, \ldots, z_n^0$, the value z_i^{k+1} of token z_i returned by the k^{th} layer of our transformer model is

$$z_i^{k+1} = z_i^k + \frac{\boldsymbol{\alpha}}{1+\boldsymbol{\alpha}} \frac{1}{|\mathcal{C}_i^k|} \sum_{j\in\mathcal{C}_i^k} \left(z_j^k - z_i^k\right), \tag{1a}$$

$$\mathcal{C}_i^k = \left\{ j \in [n] \ : \ \langle \boldsymbol{A} z_i^k, z_j^k \rangle = \max_{\ell\in[n]} \langle \boldsymbol{A} z_i^k, z_\ell^k \rangle \right\}, \tag{1b}$$

where $|\mathcal{C}_i^k|$ is the cardinality of the index set $\mathcal{C}_i^k$, and $\langle\cdot,\cdot\rangle$ denote the standard inner product in $\mathbb{R}^d$.

As in [2,3], we view (1) as a discrete-time dynamical system describing the evolution of tokens. Our system has a simple geometric interpretation: token z_i is attracted to the tokens with largest orthogonal projection in the direction of $\boldsymbol{A} z_i$, $\boldsymbol{\alpha}$ acting as an intensity regulator.



Example for $i = 1$ with $\boldsymbol{A} = \boldsymbol{I}$.

Theoretical contribution

Motivated by the increasing depth of transformers, we study the asymptotic behavior of tokens evolving according to (1). We prove that, as $k \to \infty$, tokens converge to a clustered equilibrium where the cluster points are either special tokens that we call **leaders**, or particular convex combinations thereof.

Definition. Token z_i is a *leader* if $\exists\, k \in \mathbb{N}$ such that $\mathcal{C}_i^k = \{i\}$.

Theorem. Assume the initial token values $z_1^0, \ldots, z_n^0 \in \mathbb{R}^d$ are nonzero and distinct, and $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is symmetric positive definite. Then, the set of leaders $\mathcal{L}$ is not empty, and there exist a convex polytope $\mathcal{K}$ with $|\mathcal{L}|$ vertices and a finite set $\mathcal{S} \subset \partial\mathcal{K}$ such that:

(i) Every token converges to a point in $\mathcal{S}$.
(ii) Every leader converges to a distinct vertex of $\mathcal{K}$ in finite layers.
(iii) If $s \in \mathcal{S}$ is not a vertex of $\mathcal{K}$, then it is a projection of the origin onto a face of $\mathcal{K}$ w.r.t. the norm associated with $\boldsymbol{A}$.

Selected publications

[1] Alcalde, A., Fantuzzi G., Zuazua E. (2024). **Clustering in pure-attention hardmax transformers and its role in sentiment analysis.** arXiv:2407.01602.

[2] Geshkovski, B., Letrouit, C., Polyanskiy, Y., Rigollet, P. (2024). **The emergence of clusters in self-attention dynamics.** Advances in NeurIPS, 36.

[3] Geshkovski, B., Letrouit, C., Polyanskiy, Y., Rigollet, P. (2023). **A mathematical perspective on transformers.** arXiv:2312.10794.

[4] Maas A. L. *et al.* (2011). **Learning word vectors for sentiment analysis.** In Proceedings of HLT '11, pages 142–150.

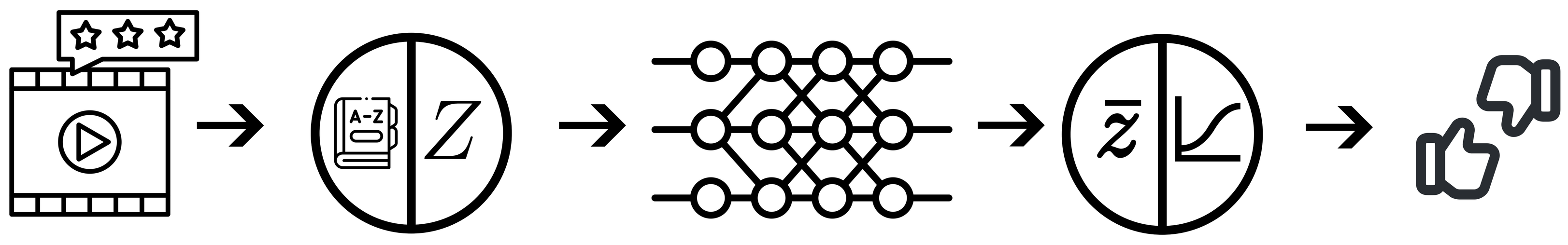
Computational contribution

We leverage our clustering results to design an interpretable transformer-based model to solve the supervised learning task of sentiment analysis:

"predict the sentiment of $50\,000$ movie reviews labeled as positive (1) or negative (0) from the benchmark IMDb dataset [4]."

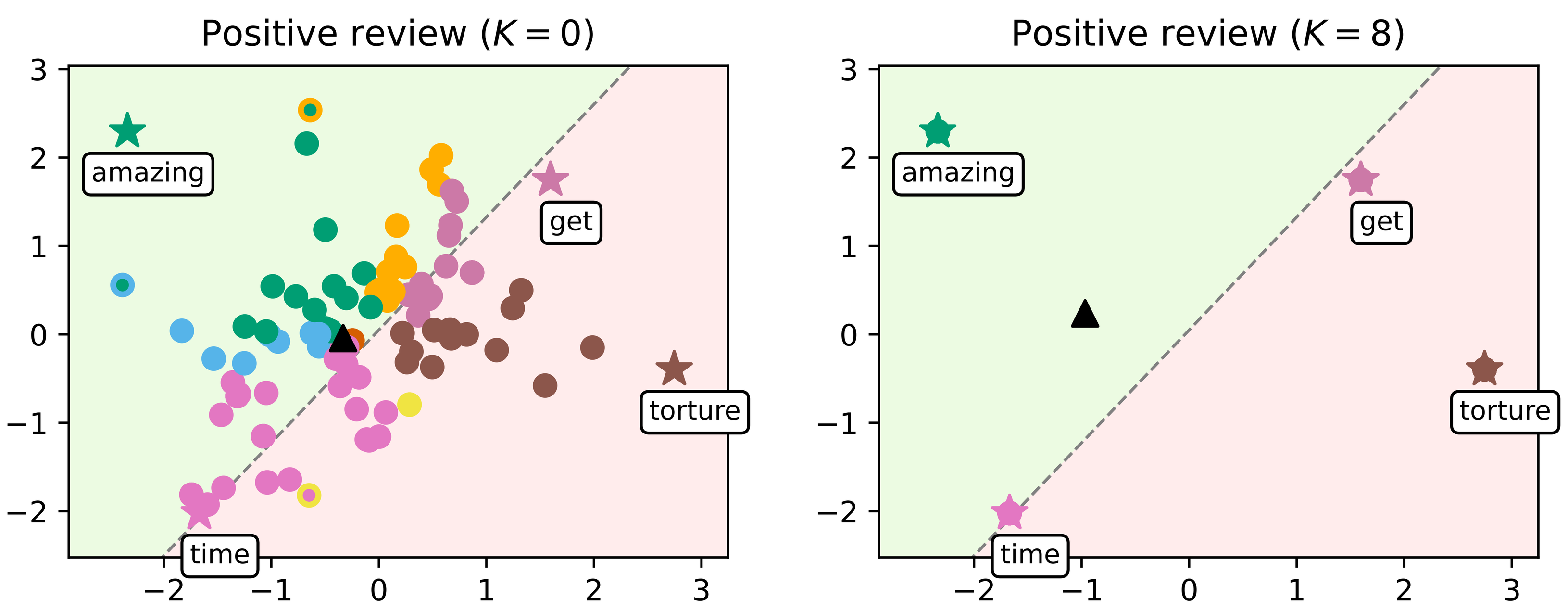
Each review has $n = 128$ words, uniquely identified with a basis vector in $\mathbb{R}^W$. The proposed model inputs a movie review $X \in \mathbb{R}^{n\times W}$ and outputs a prediction $\hat{y} \in \mathbb{R}$, using three components with explainable roles:

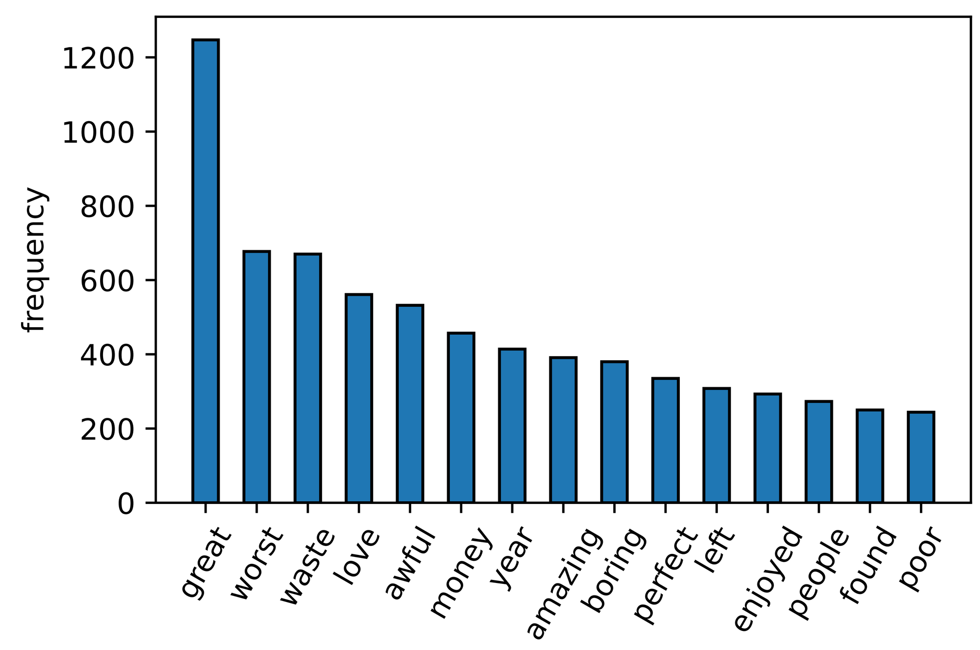
- **Embedding:** select meaningful words as leaders.
- **Transformer:** capture 'context' and reduce dimensionality by clustering tokens around leaders.
- **Decoder:** project clustered token values to a positive or negative sentiment prediction.



Training and results. We fix depth $K = 8$ and encoder dimension $d = 2$, and train embedding $\boldsymbol{E} \in \mathbb{R}^{W\times d}$, decoder vector $\boldsymbol{w} \in \mathbb{R}^d$, decoder bias $\boldsymbol{v} \in \mathbb{R}$, and step-size $\boldsymbol{\alpha} > 0$. Our results confirm:

- $K = 8$ layers are enough to approximate the asymptotic clustered state
- the leaders furthest away from decision hyperplane (dashed black line) indeed convey sentiment: **amazing**, **torture**
- the 15 most frequent leaders in correctly classified test reviews furthest from decision hyperplane mostly related with sentiment.





Conclusions and further work

- Clustering has been rigorously proven for a class of transformers, and directly related to 'context' emergence in a language modelling application.
- It remains to extend the analysis for general $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, specially relevant in applications where this matrix is typically low-rank.
- As further research, one could constructively choose $\boldsymbol{A}$ and $\boldsymbol{\alpha}$ to control a full transformer. The dimensionality reduction of self-attention should help obtain a simpler control.

https://dcn.nat.fau.eu/
09/2024

Funded by the European Union

Department MATHEMATIK

Unterstützt von / Supported by
Alexander von Humboldt Stiftung/Foundation