Clustering in transformer models in ML and its role in sentiment analysis

Albert Alcalde, Giovanni Fantuzzi & Enrique Zuazua

Friedrich-Alexander-Universität Erlangen-Nürnberg
Chair for Dynamics, Control, Machine Learning, and Numerics (AvH Professorship)

Benasque (August 26, 2024)

MODCONFLEX
Modelling and control of flexible structures interacting with fluids

Funded by the European Union

Motivation

New machine learning tool in everyday life: **ChatGPT**.

G for **Generative** **P** for **Pre-trained** **T** for **Transformer**

Motivation

New machine learning tool in everyday life: **ChatGPT**.

Generative: learn to predict the next word/pixel.

"A dog doing math research".[1]

[1] Source: https://pixlr.com/image-generator/

Motivation

New machine learning tool in everyday life: **ChatGPT**.

Pre-trained: trained on a massive corpus of text/images.

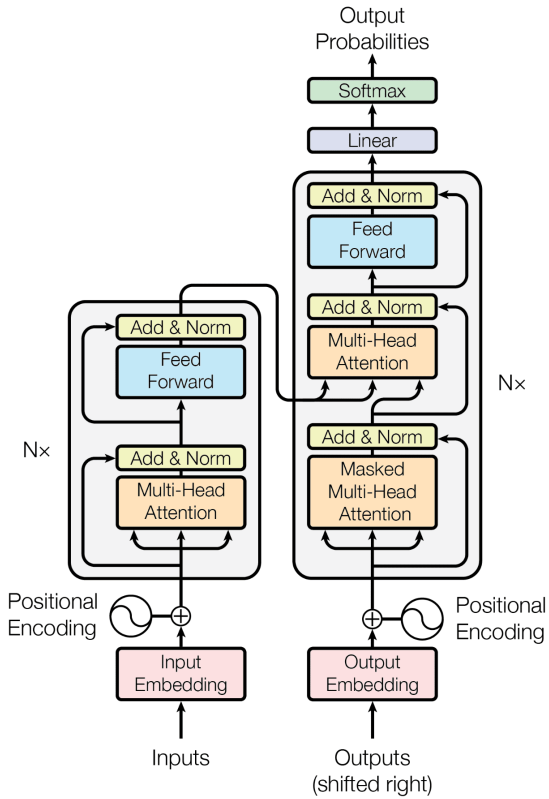
Evolution of number of words in training datasets.[2]

[2] Source: https://www.lesswrong.com/posts/asqDCb9XzXnLjSfgL/trends-in-training-dataset-sizes

Motivation

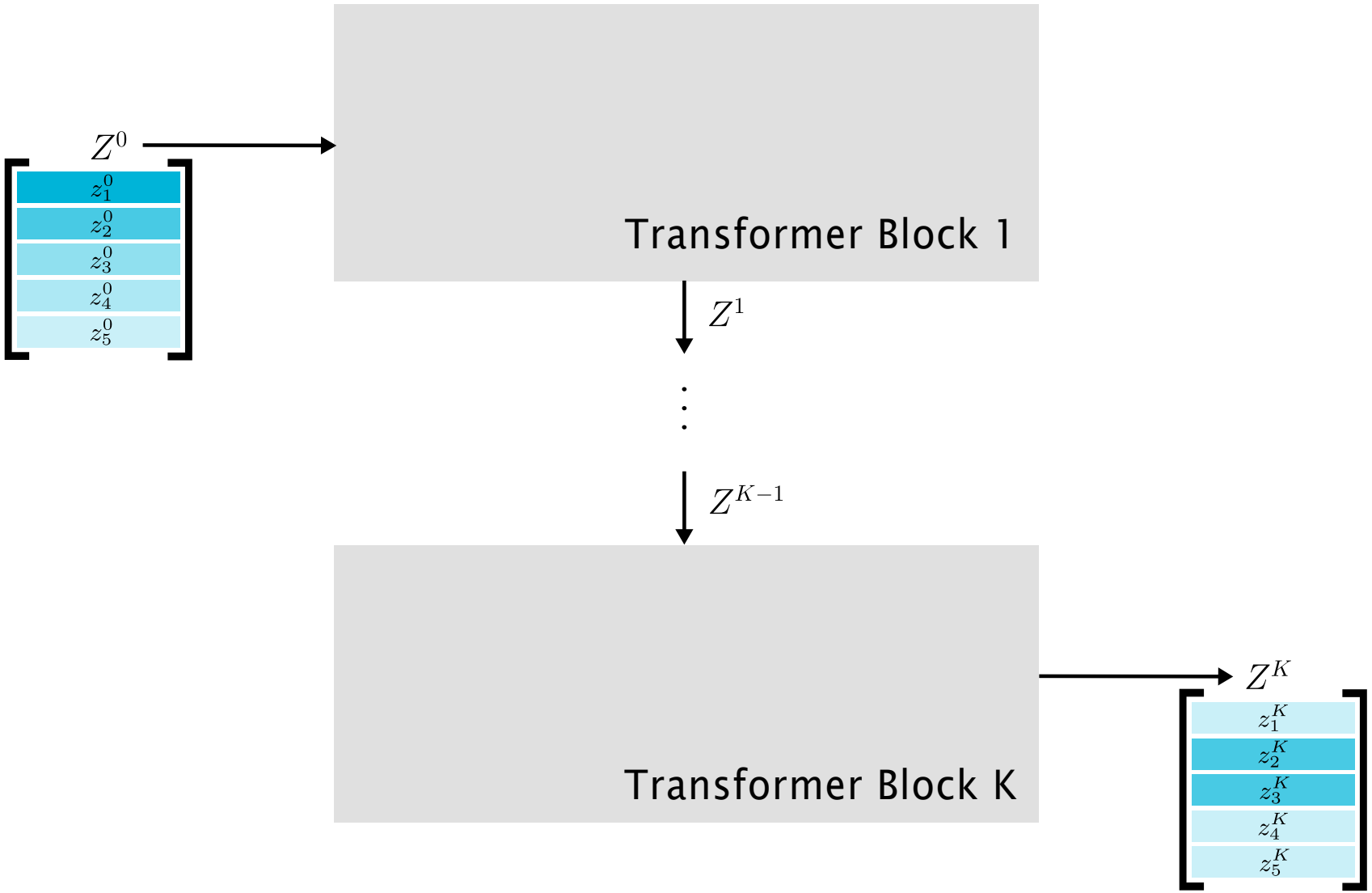
New machine learning tool in everyday life: **ChatGPT**.

Transformer: deep neural network architecture.

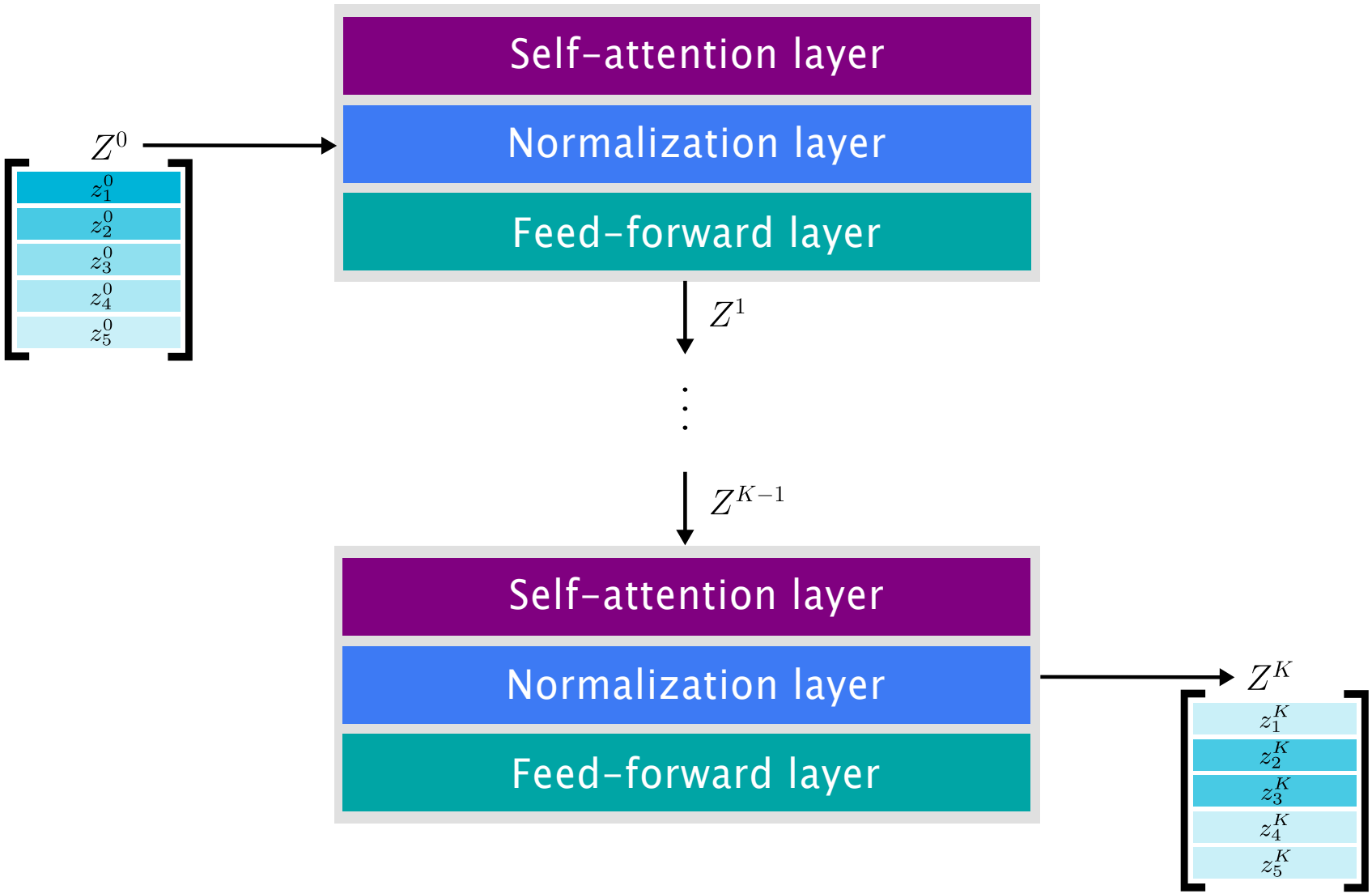


Original transformer architecture in [VSP+17].

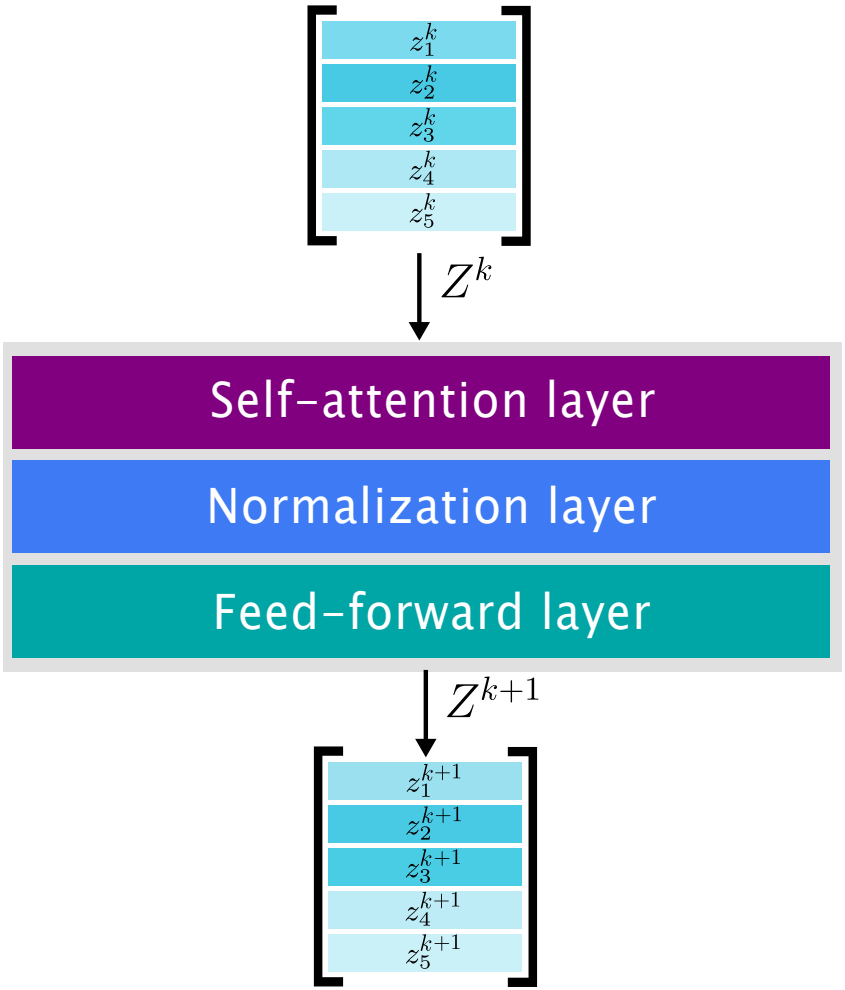
Modeling of transformers



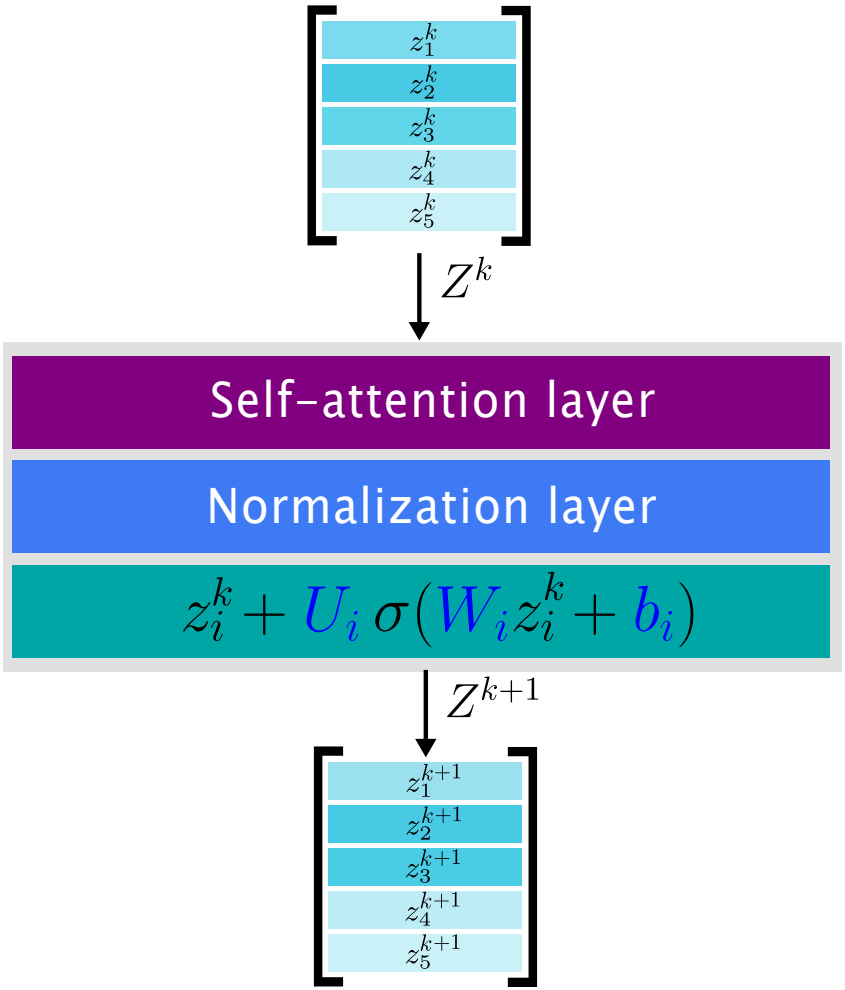
Modeling of transformers



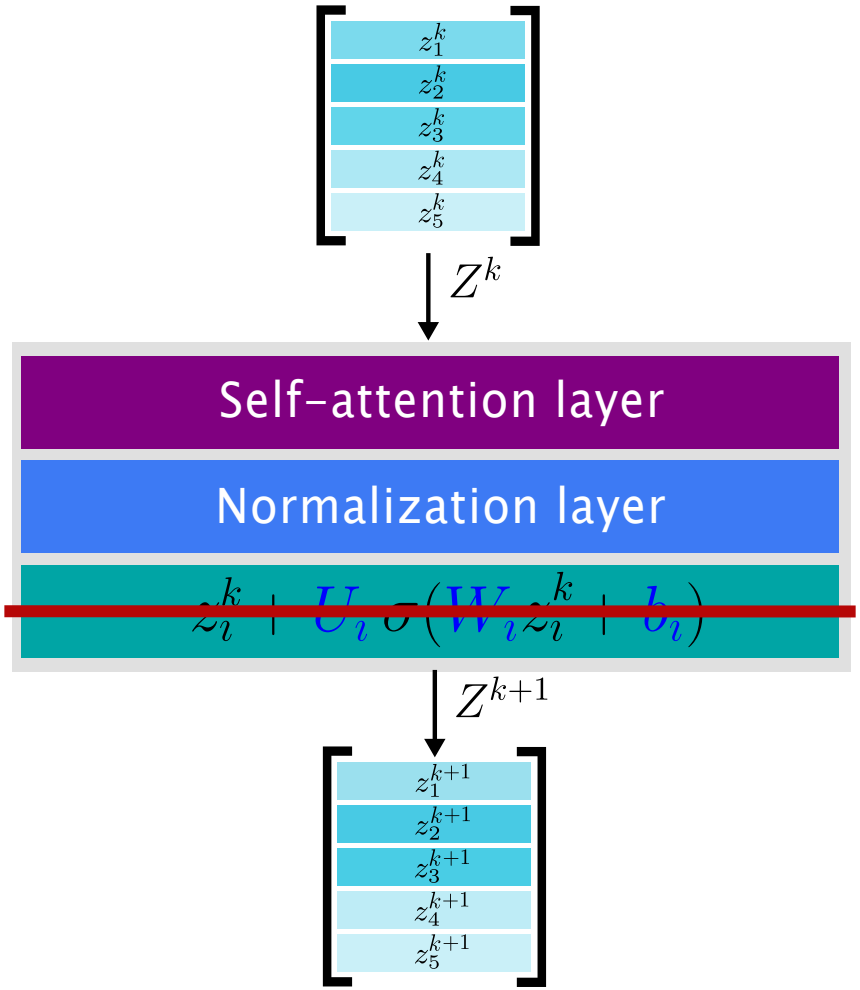
Modeling of transformers



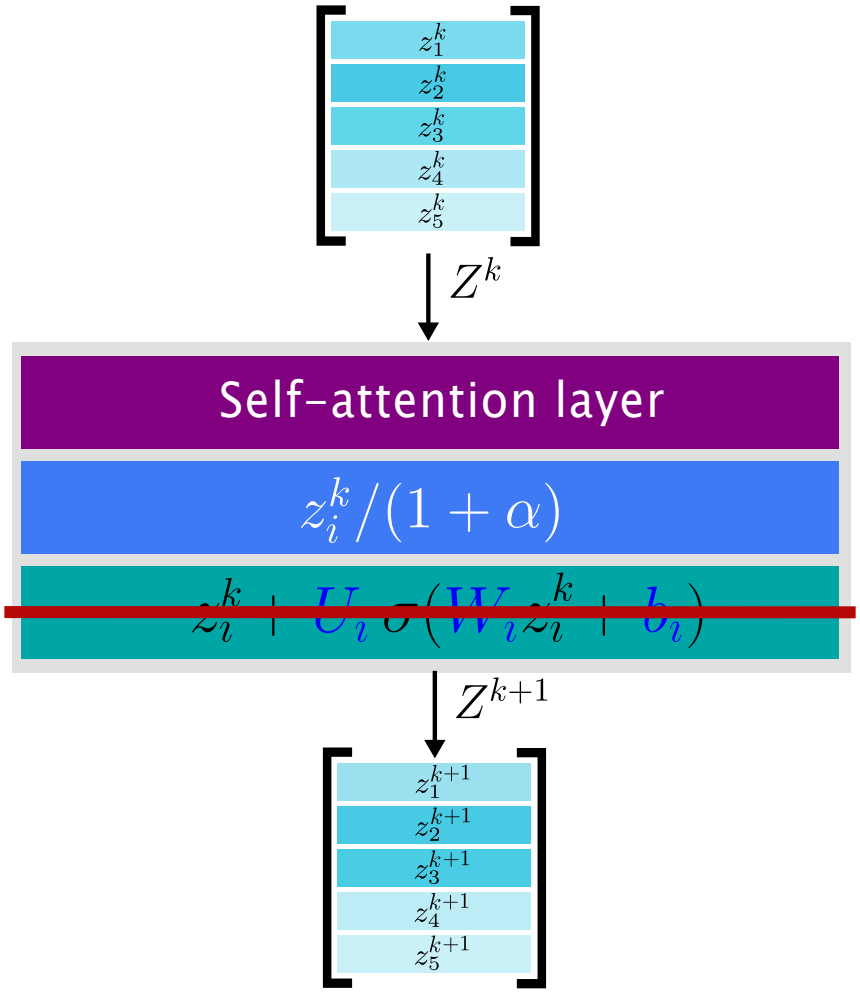
Modeling of transformers



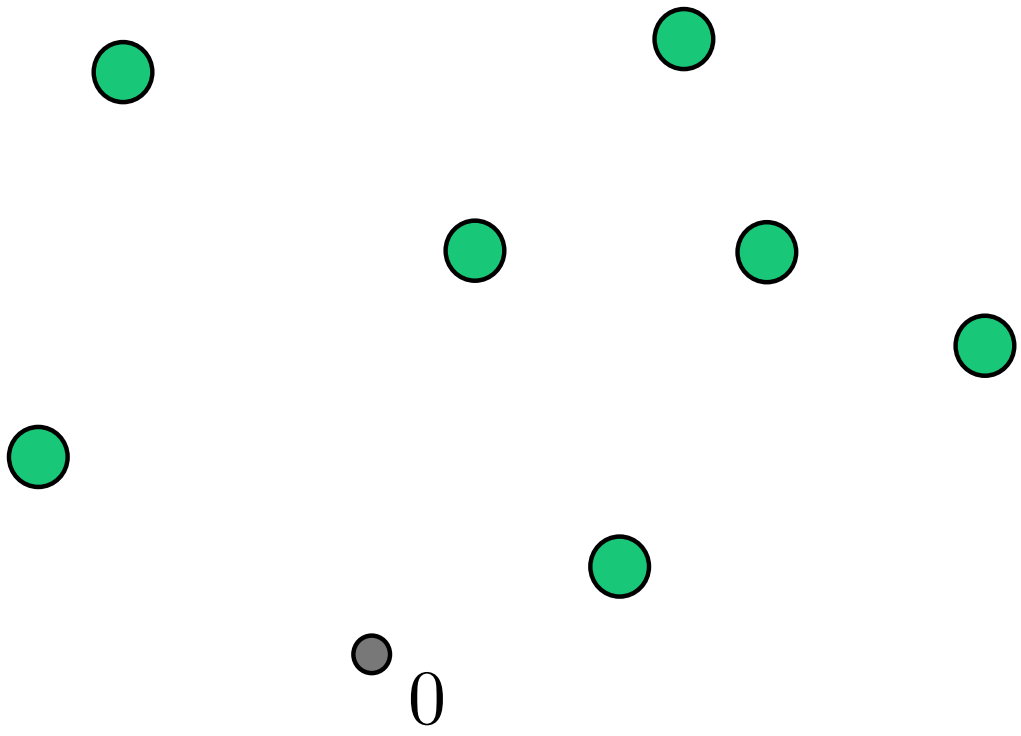
Modeling of transformers



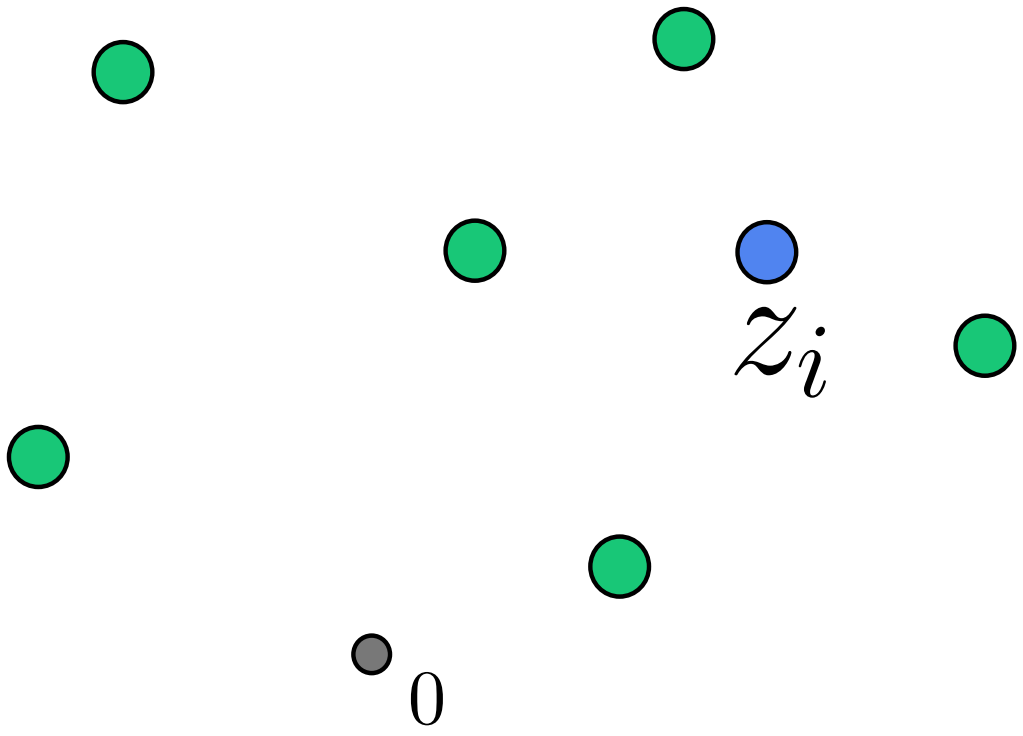
Modeling of transformers



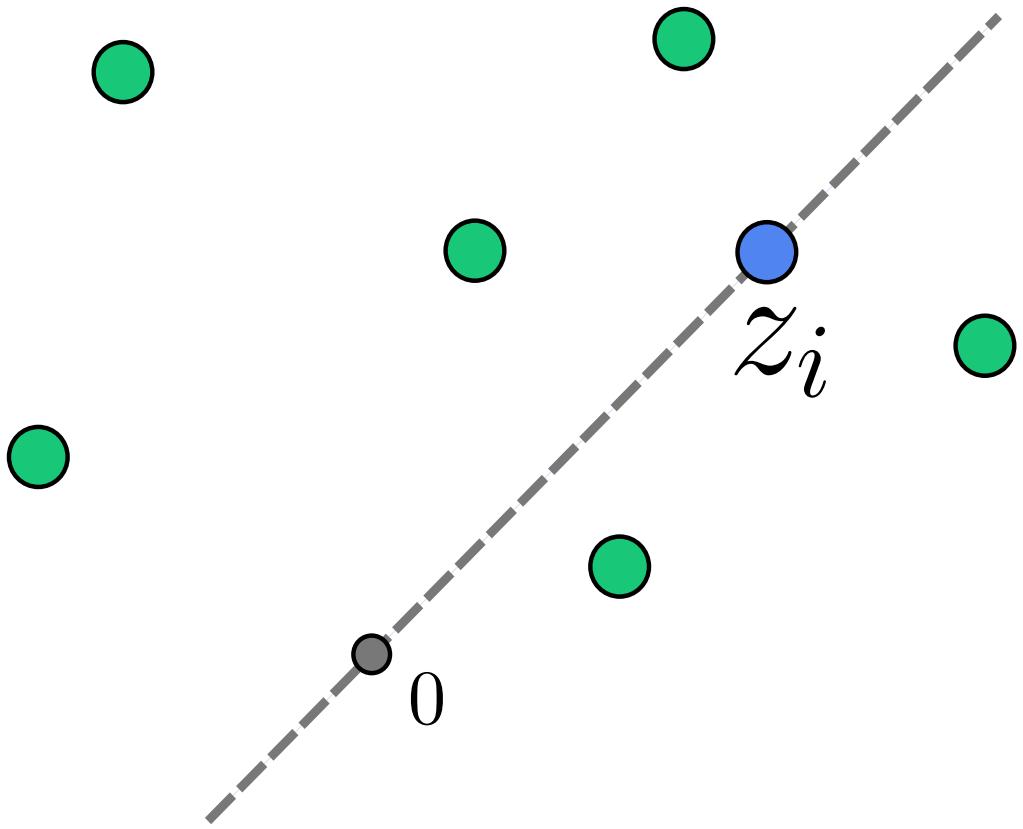
The self-attention layer



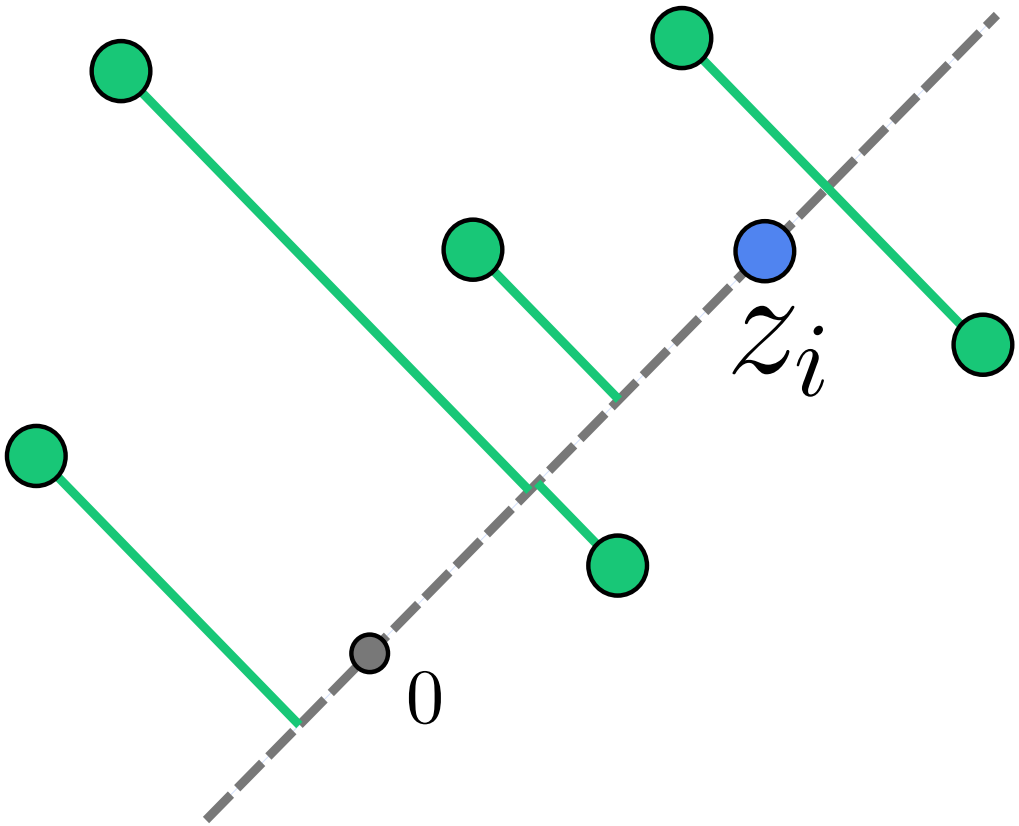
The self-attention layer



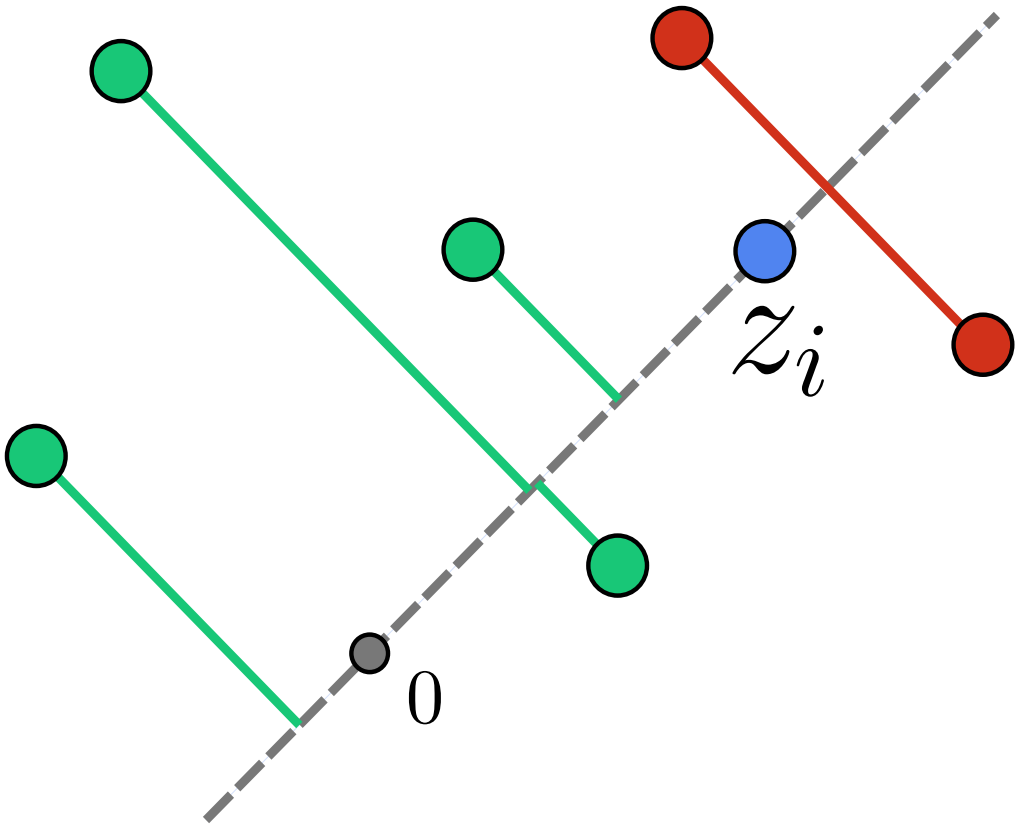
The self-attention layer



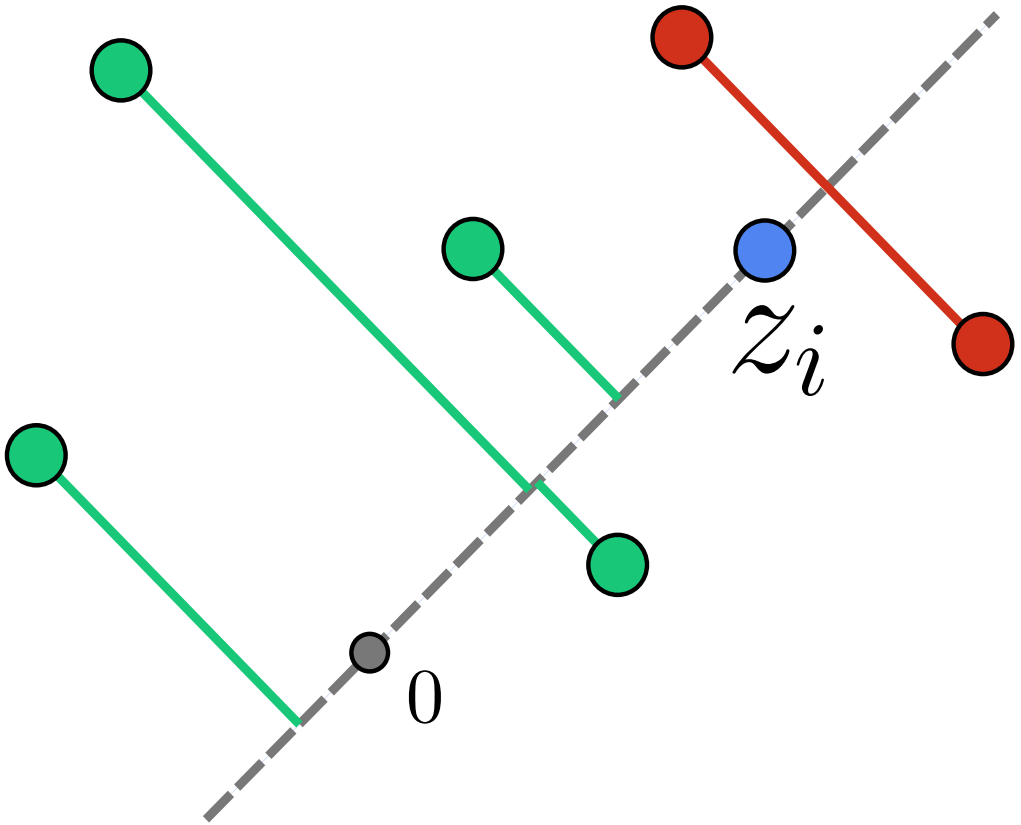
The self-attention layer



The self-attention layer



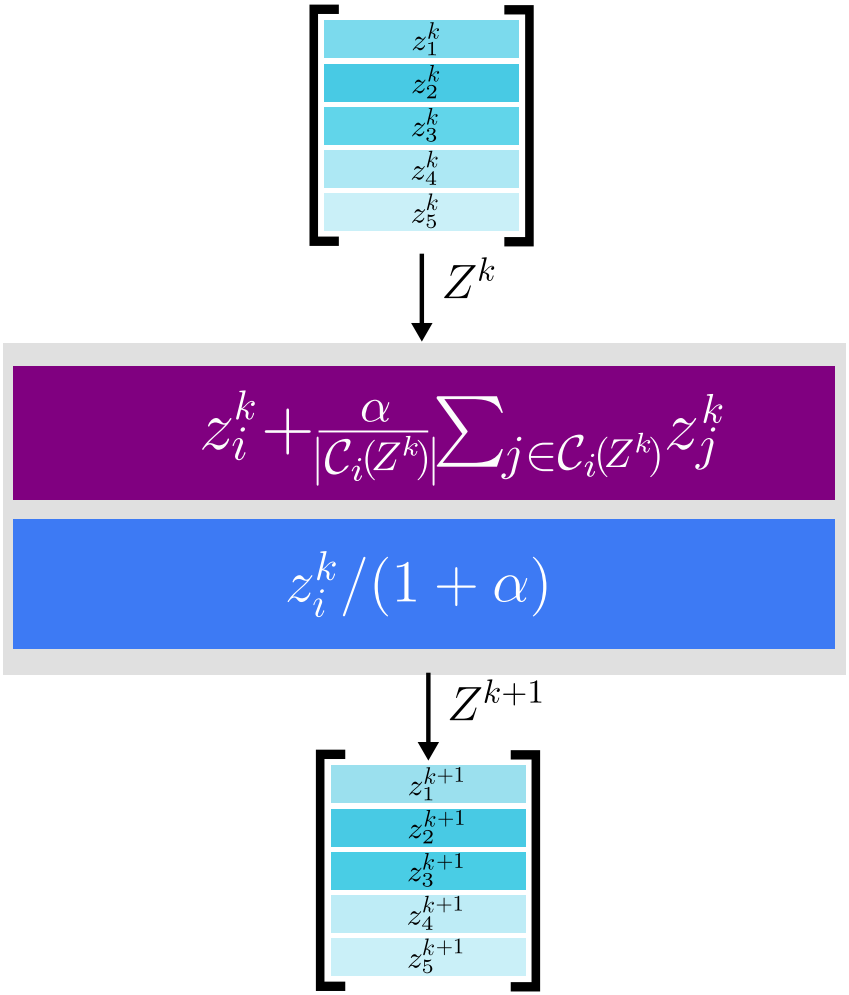
The self-attention layer



$$\mathcal{C}_i(Z) = \{j \in [n] \, : \, \langle z_i, z_j \rangle = \max_{\ell \in [n]} \langle z_i, z_\ell \rangle\}.$$

$$z_i^+ = z_i + \frac{\alpha}{|\mathcal{C}_i(Z)|} \sum_{j \in \mathcal{C}_i(Z)} z_j$$

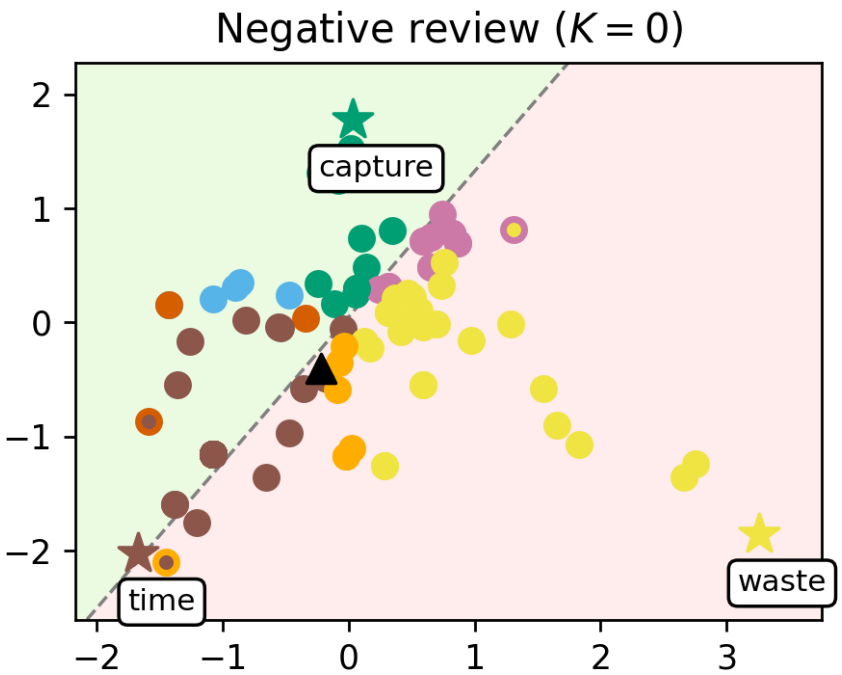
Pure-attention hardmax transformers

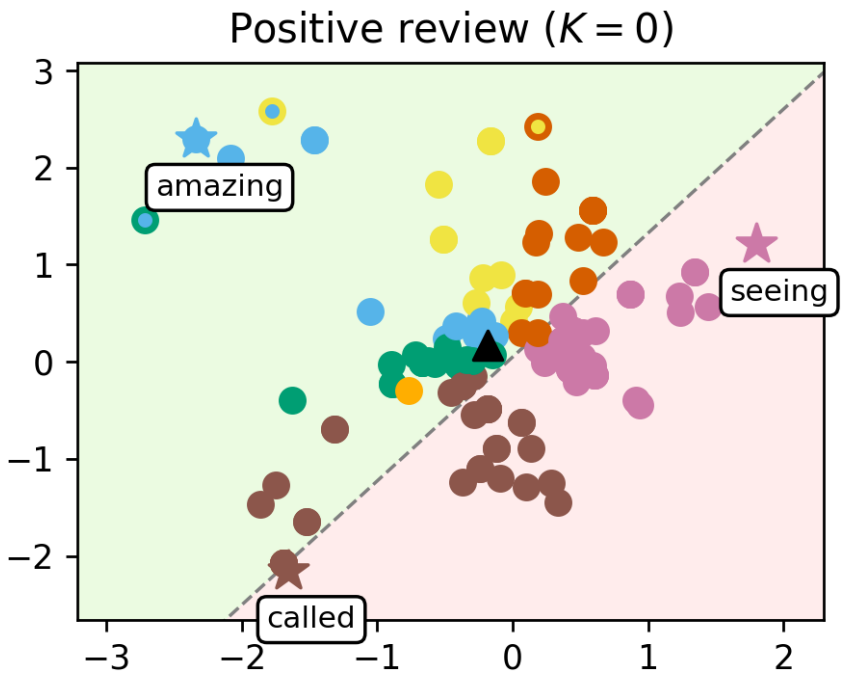


$$z_i^{k+1} = z_i^k + \frac{\alpha}{1+\alpha} \frac{1}{|\mathcal{C}_i(Z^k)|} \sum_{j \in \mathcal{C}_i(Z^k)} \left(z_j^k - z_i^k \right), \quad k \geq 0.$$

Interpretability through dynamics and control

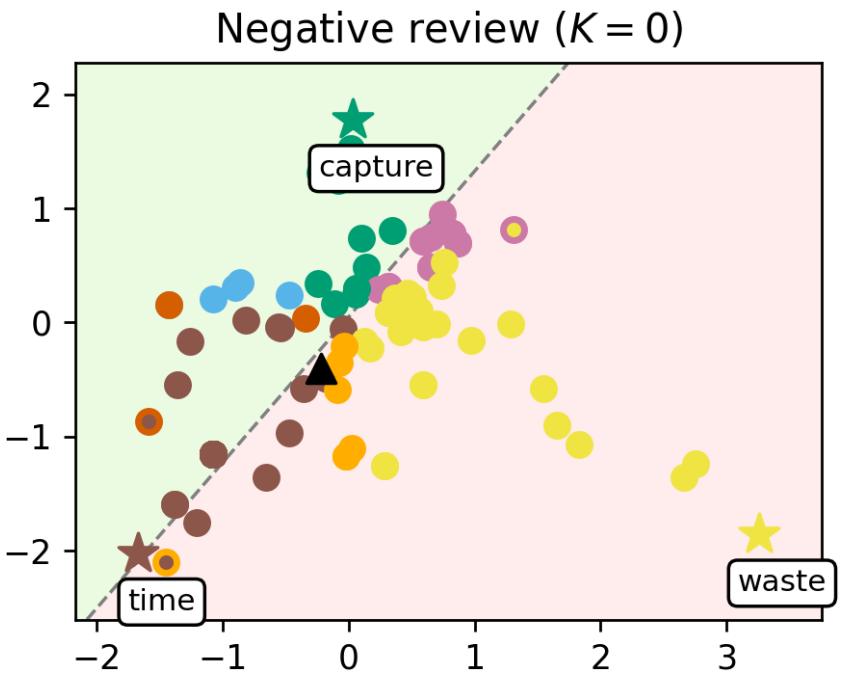
- $k \to \infty$ asymptotics: clustering first proved for similar continuous transformer models in [GLPR23a], [GLPR23b].

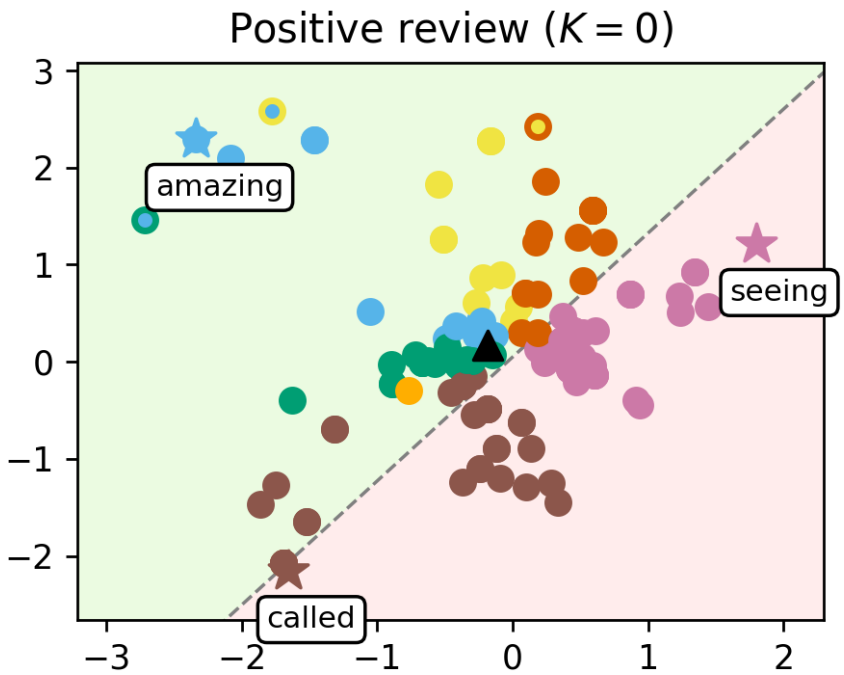




Interpretability through dynamics and control

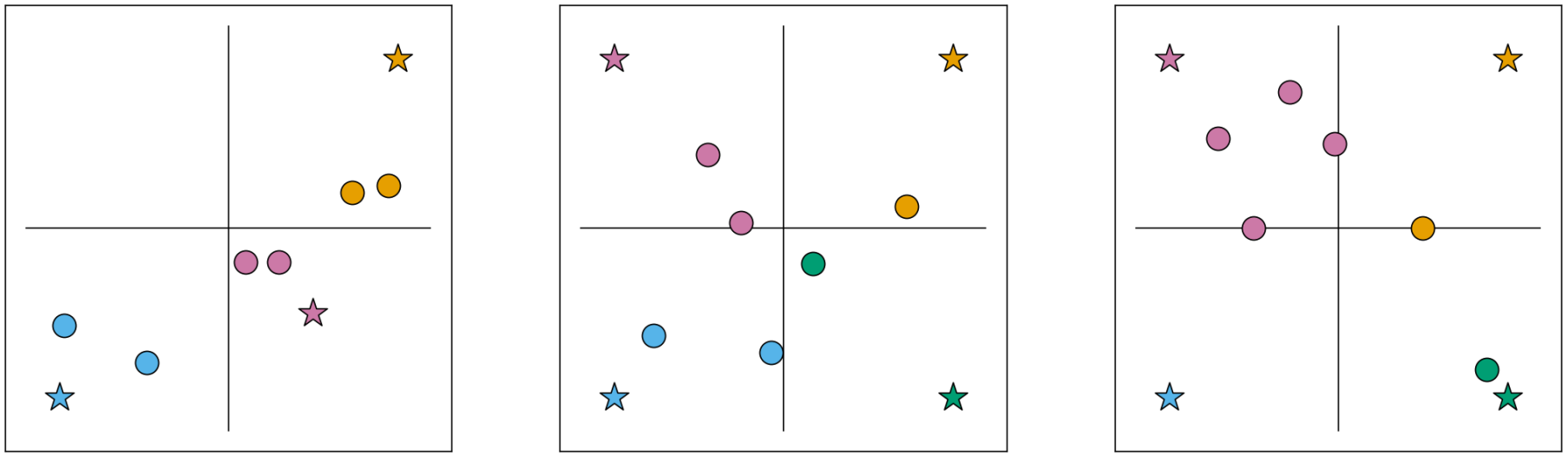
- $k \to \infty$ asymptotics: clustering first proved for similar continuous transformer models in [GLPR23a], [GLPR23b].





- **Q1:** When and how do discrete dynamics exhibit clustering?
- **Q2:** Is clustering capturing the 'context' in a text?

Asymptotic dynamics



Point z_i is a **leader** if there exists a layer $k \in \mathbb{N}$ s.t. $\mathcal{C}_i(Z^k) = \{i\}$.

Emergence and characterization of cluster points

Q1: When and how do discrete dynamics exhibit clustering?

Emergence and characterization of cluster points

Q1: When and how do discrete dynamics exhibit clustering?

Theorem (A. A, G. Fantuzzi, E. Zuazua 2024)

Assume $z_1^0, \dots, z_n^0 \in \mathbb{R}^d$ are nonzero and distinct. There exist

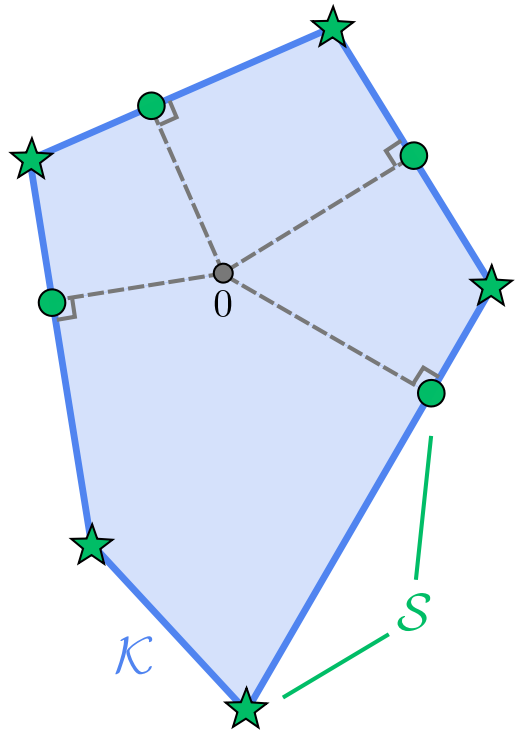
(a) a convex polytope $\mathcal{K}$, and

(b) a finite set $\mathcal{S} \subset \partial\mathcal{K}$

such that:

(i) $z_i^k \to s \in \mathcal{S}$ as $k \to \infty$ for all i.

(ii) Every $s \in \mathcal{S}$ is a limit value of a leader or a convex combination of them.

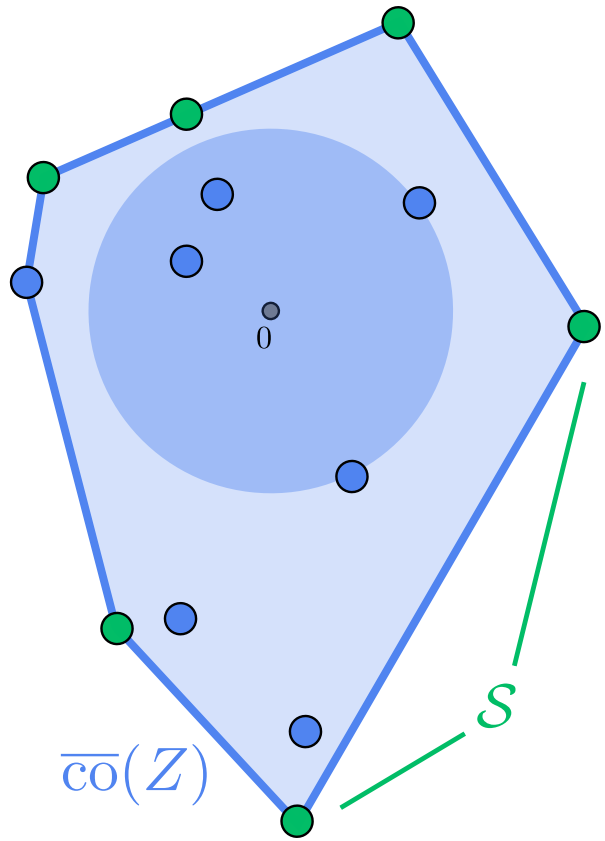


Idea of the proof

Step 1: convergence to an equilibrium at the boundary of a convex polytope $\mathcal{K}$. Two competing forces:

- Convex hull of tokens shrinks,
- Norm of tokens strictly grows when not close to finite set $\mathcal{S} \subset \partial\mathcal{K}$.

Step 2: vertices of $\mathcal{K}$ = leaders.



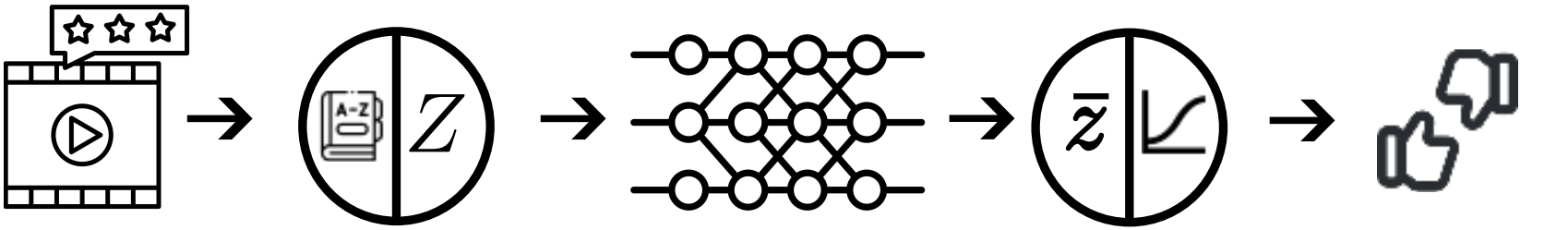
Role of clustering in sentiment analysis

Q2: What is the role of clustering in a real machine learning application? Is it capturing 'context'?

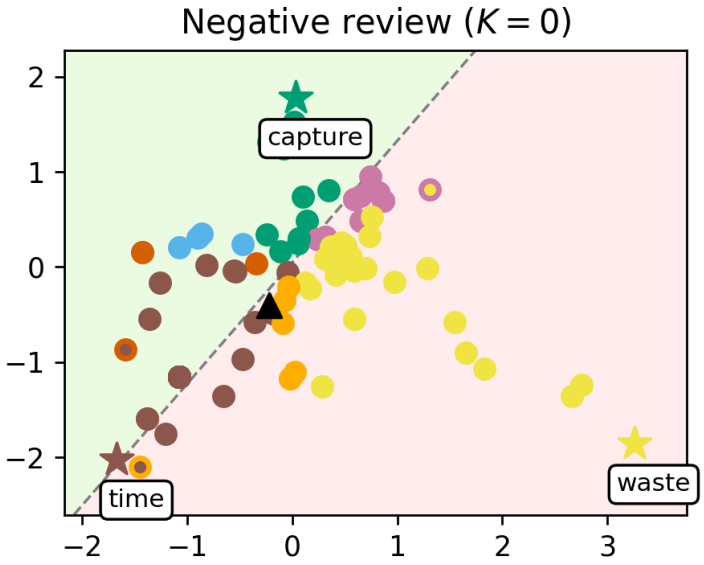
Role of clustering in sentiment analysis

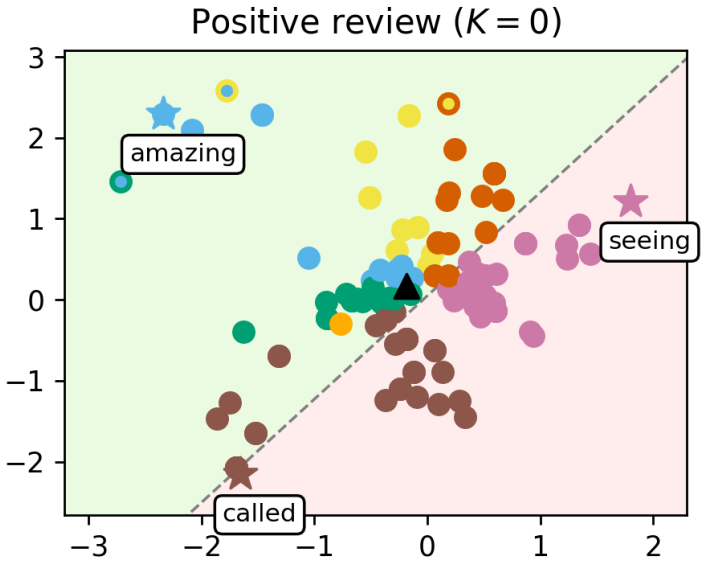
Q2: What is the role of clustering in a real machine learning application? Is it capturing 'context'?

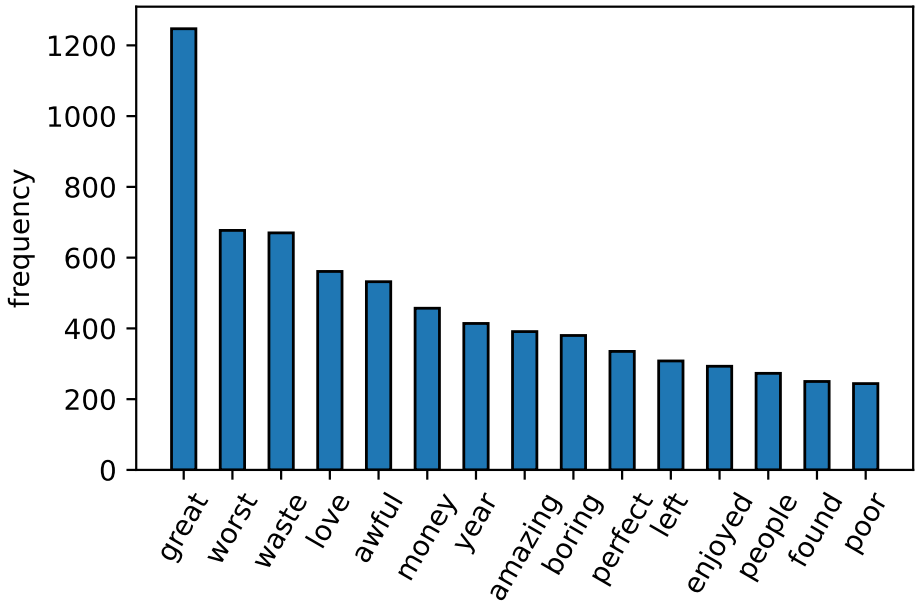
Task: Sentiment analysis of movie reviews.



Results


Negative review ($K = 0$)
capture
time
waste


Positive review ($K = 0$)
amazing
seeing
called


frequency
great
worst
waste
love
awful
money
year
amazing
boring
perfect
left
enjoyed
people
found
poor

Conclusions and perspectives

- We analyzed a transformer model, proving that it entails a clustering effect.
- Related clustering to emergence of context in the ML application of sentiment analysis.
- **Open questions:**
  - *Controlling* the leaders by appropriately choosing the matrix $A \in \mathbb{R}^{d\times d}$ in

$$\mathcal{C}_i(Z) = \{j \in [n] \,:\, \langle Az_i, z_j\rangle = \max_{\ell\in[n]} \langle Az_i, z_\ell\rangle\}.$$

  - What does clustering imply for physical systems, what are the *leaders* in, say, a flow past a cylinder?

Conclusions and perspectives

- We analyzed a transformer model, proving that it entails a clustering effect.
- Related clustering to emergence of context in the ML application of sentiment analysis.
- **Open questions:**
  - *Controlling* the leaders by appropriately choosing the matrix $A \in \mathbb{R}^{d\times d}$ in

$$\mathcal{C}_i(Z) = \{j \in [n] \,:\, \langle Az_i, z_j\rangle = \max_{\ell\in[n]} \langle Az_i, z_\ell\rangle\}.$$

  - What does clustering imply for physical systems, what are the *leaders* in, say, a flow past a cylinder?

Thank you for your

$$\text{“}z_i + \frac{\alpha}{|\mathcal{C}_i(Z)|}\sum_{j\in\mathcal{C}_i(Z)} z_j\text{”}$$

arXiv:2407.01602