TallyFormer: Replacing the KV Cache with a Token Histogram

Andras Szecsenyi, Thomas L. Lee, Vaisakh Saj, Amos J. Storkey

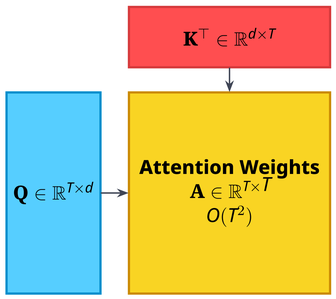
TL;DR

Transformers represent the context with T=context length **continuous** vectors. This history is initially **discrete**: each token is one of V vocabulary elements. Two drop-in upgrades follow:

1. **Coded KV** - only cache the integer vocab index, not the continuous vector representing it.

2. **Token Tally** - sum the attenuated contributions over a vocabulary histogram. instead of T for recurrent attention.

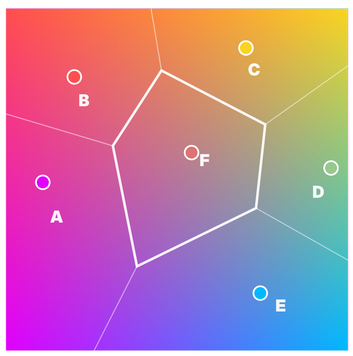
Preliminaries



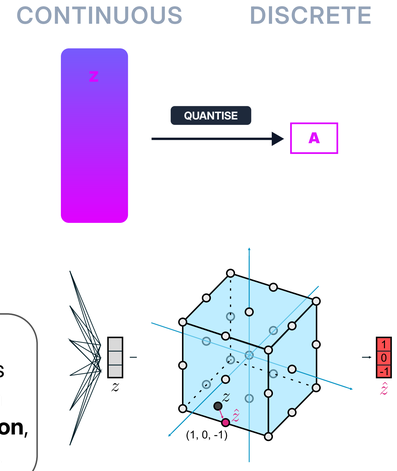
LLM Inference is **memory-bound**, and storing large context represetations is prohibitive, as they are unbounded.

Furthermore, LLM training with long contexts is **compute-bound** due to the $O(T^2)$ complexity of softmax attention.

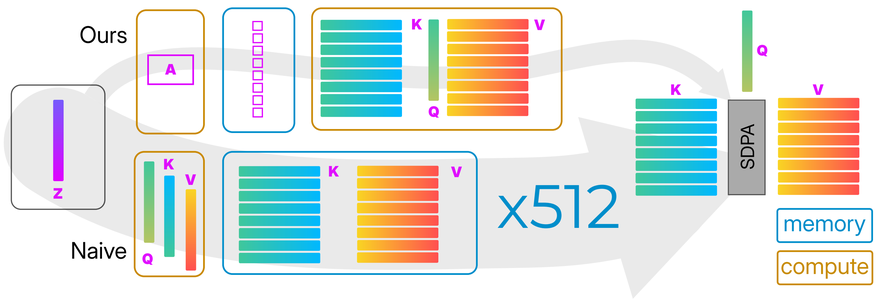
We are looking for ways to **compress the context** representation and to find **recurrent** solutions to the attenuation task.



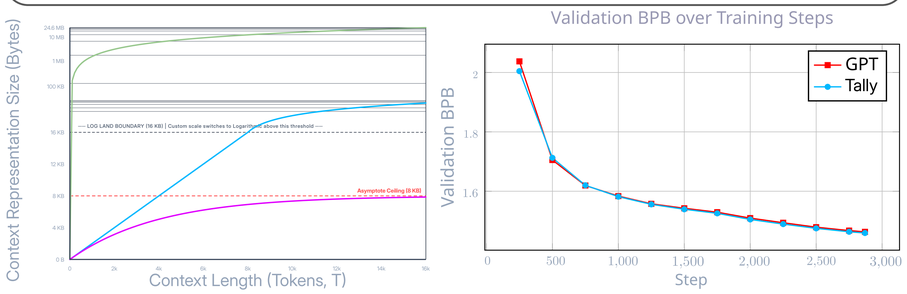
Vector quantisation allows us to **categorise** continuous vector spaces into discrete groups. We can assign a code to each cell in a **Voronoi partition**, and identify all points in that cell by it.



Coded KV

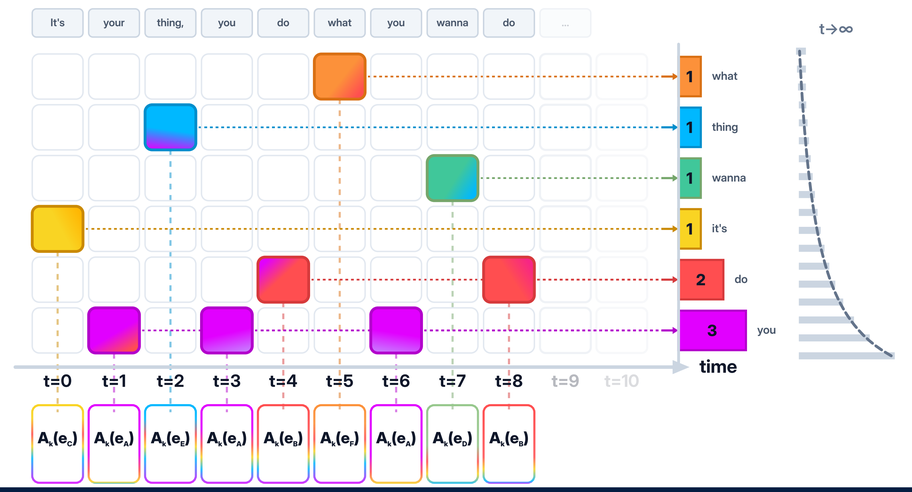


We can apply discretise the input latents and save only the code to cache, replacing the d_model vectors with integers. This reduces required memory bandwidth, KV cache size. For a model with d256 latent, this turns a 10GB KV cache into 2MB, an 512x reduction.



Token Tally

With discrete tokens, we can have a bounded representation of the context without loss of information by collecting a histogram of seen tokens with an **exponential decay**.



Mathematical equivalence to SDPA

Starting from self-attention for head k at layer 0, let $\mathbf{x}_i \in \{0,1\}^{\mathcal{V}}$ be the one-hot token vector at step i. The attention output is:

$$\mathbf{z}_k^t = \sum_{s=0}^{t-1} \frac{1}{z_k^t} \exp\ \mathbf{x}_{t-s}^\top \mathbf{K}_k \mathbf{Q}_k^\top \mathbf{x}_t / \sqrt{d_h}\ \ \mathbf{V}_k \mathbf{x}_{t-s}$$

Where $z_k^t = \sum_{s=0}^{t-1} \exp\ \mathbf{x}_{t-s}^\top \mathbf{K}_k \mathbf{Q}_k^\top \mathbf{x}_t / \sqrt{d_h}$.

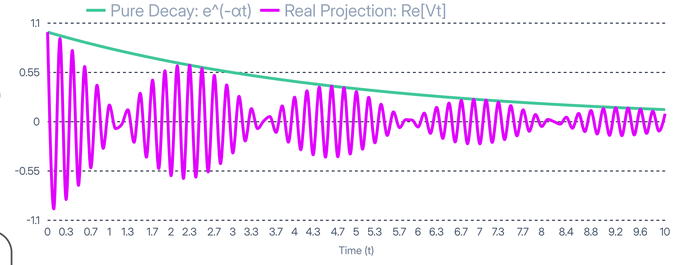
Rewrite using the standard basis vector $\mathbf{e}_c \in \{0,1\}^{\mathcal{V}}$ and the indicator function $\mathbb{I}(\cdot)$ over vocabulary $\mathcal{V}$:

$$\mathbf{z}_k^t = \sum_{s=0}^{t-1} \sum_{c=1}^{\mathcal{V}} \frac{1}{z_k^t} \mathbb{I}(\mathbf{x}_{t-s} = \mathbf{e}_c) \exp\ \mathbf{e}_c^\top \mathbf{K}_k \mathbf{Q}_k^\top \mathbf{x}_t / \sqrt{d_h}\ \ \mathbf{V}_k \mathbf{e}_c$$

Simplifying by swapping summations:

$$\mathbf{z}_k^t = \frac{1}{z_k^t} \sum_{c=1}^{\mathcal{V}} \underbrace{\left[\sum_{s=0}^{t-1} \mathbb{I}(\mathbf{x}_{t-s} = \mathbf{e}_c)\right]}_{\text{Count } N_t(\mathbf{e}_c)} \underbrace{\exp\ \mathbf{e}_c^\top \mathbf{K}_k \mathbf{Q}_k^\top \mathbf{x}_t / \sqrt{d_h}}_{\text{Score } A_k(\mathbf{e}_c, \mathbf{x}_t)} \mathbf{V}_k \mathbf{e}_c$$

Recurrence: $N_{t,k}(\mathbf{e}_c)$ is efficiently updated via: $\mathbf{N}_{t,k}(\mathbf{e}_c) = \gamma_k \mathbf{N}_{t-1,k}(\mathbf{e}_c) + D_k \mathbb{I}(\mathbf{x}_t = \mathbf{e}_c)$



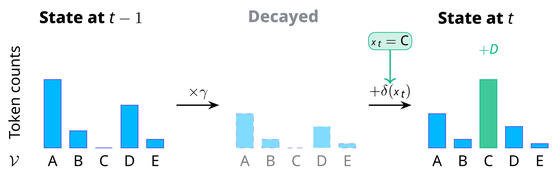
A complex decay can be used for more expressivity:

$$e^{-\alpha t + (2\pi + \epsilon) k t i}$$

Efficient Implementation

During inference, the histogram can be efficiently updated without looking at the previous tokens, in $O(V_{Act.})$ time.
At training time, a work-efficient **associative scan** can parse the sequence in O(log(T)) time.

Due to the Zipfian distribution of tokens, most histogram buckets are expected to be zero. This allows us to work with a **sparse tensor state**, and further reduce the state size to the number of seen unique tokens.



Property	Standard Attention	Linear SSMs	RVA (Ours)
History State	Full KV matrix ($T \times d$)	Continuous ($d_h \times d_h$)	Discrete Histogram ($\mathcal{V} \times 1$)
Formulation	Exact Softmax	Approximation	Exact rewriting at Layer 0
Update Cost	$O(T \cdot d)$ per token	$O(d_h^2)$ per token	$O(\mathcal{V})$ per token
State Bound	Grows linearly	Fixed (d_h)	Fixed (Vocabulary $\mathcal{V}$)

THE UNIVERSITY of EDINBURGH informatics
SYSTEMS EPSRC Centre For Doctoral Training in Machine Learning Systems
UKRI Engineering and Physical Sciences Research Council
andras.szecsenyi@ed.ac.uk