Improving Transformers with Probabilistic Attention Keys

Tam Nguyen (co-first author), Tan M. Nguyen (co-first author),
Dung Le, Khuong Nguyen, Anh Tran,
Richard G. Baraniuk, Nhat Ho, Stanley J. Osher

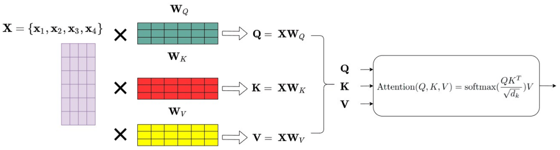
UCLA
Fpt Software
VINUNIVERSITY
RICE Unconventional Wisdom
TEXAS The University of Texas at Austin

Tam Nguyen (co-first author), Tan M. Nguyen (co-first author), Dung Le, Khuong Nguyen, Anh Tran, Richard G. Baraniuk, Nhat Ho, Stanley J. Osher. "Improving Transformers with Probabilistic Attention Keys". International Conference on Machine Learning (ICML), 2022.

Transformer and Self-Attention

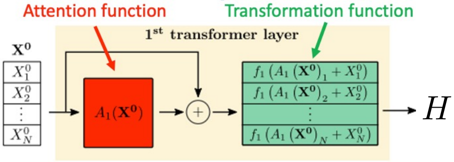
Self-attention transforms sequences $\boldsymbol{X} := [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D_x}$ using $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{D \times D_x}$ and $\mathbf{W}_V \in \mathbb{R}^{D_v \times D_x}$ as follows:

$$Q = XW_Q^\top$$
$$K = XW_K^\top$$
$$V = XW_V^\top$$
$$H = \underbrace{\text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)}_{\text{attention matrix}} V := AV.$$



A Transformer Layer

$$H = f_\ell\left(X + AV\right)$$
$$= f_\ell\left(X + \text{softmax}\left(\frac{XW_Q^\top W_K X^\top}{\sqrt{D}}\right) XW_V^\top\right)$$



Self-Attention: The Current Problems

A good understanding of the self-attention mechanism is missing.

Transformers for practical tasks learn redundant heads, limiting their representation capacity while wasting parameters, memory and computation

Self-attention from a Probabilistic Perspective

Gaussian Mixture Model for Self-Attention

Consider a query $\boldsymbol{q}_i \in \mathbf{Q}$ and a key $\boldsymbol{k}_j \in \mathbf{K}$. Let $\boldsymbol{t}$ be a K-dimensional binary random variable having 1-of-K representation. Our GMM is defined as follows

$$p(\boldsymbol{q}) = \sum_{j=1}^{N} \pi_j N(\boldsymbol{q} \mid \boldsymbol{k}_j, \sigma_j^2 \mathbf{I}) \tag{1}$$

where π_j is the prior $p(\boldsymbol{t}_j = 1)$.

In our mixture model, each key $\boldsymbol{k}_j$ is the cluster mean. The query data $\boldsymbol{q}_i$ is assigned to those clusters.

Attention Score as a Posterior Distribution

$$p(\boldsymbol{t}_j = 1|\boldsymbol{q}_i) = \frac{\pi_j N(\boldsymbol{q}_i \mid \boldsymbol{k}_j, \sigma_j^2)}{\sum_{j'} \pi_{j'} N(\boldsymbol{q}_i \mid \boldsymbol{k}_{j'}, \sigma_{j'}^2)}$$

$$= \frac{\pi_j \exp\left[-\left(\|\boldsymbol{q}_i\|^2 + \|\boldsymbol{k}_j\|^2\right)/2\sigma_j^2\right] \exp\left(\boldsymbol{q}_i \boldsymbol{k}_j^\top / \sigma_j^2\right)}{\sum_{j'} \pi_{j'} \exp\left[-\left(\|\boldsymbol{q}_i\|^2 + \|\boldsymbol{k}_{j'}\|^2\right)/2\sigma_{j'}^2\right] \exp\left(\boldsymbol{q}_i \boldsymbol{k}_{j'}^\top / \sigma_{j'}^2\right)}.$$

Assuming that $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$ are normalized, uniform priors, and $\sigma_j^2 = \sigma^2$, the posterior becomes

$$p(\boldsymbol{t}_j = 1|\boldsymbol{q}_i) = \frac{\exp\left(\boldsymbol{q}_i \boldsymbol{k}_j^\top / \sigma^2\right)}{\sum_{j'} \exp\left(\boldsymbol{q}_i \boldsymbol{k}_{j'}^\top / \sigma^2\right)} = a_{ij}. \tag{2}$$

a_{ij} is the attention score deciding how much the token at location i attends to the token at location j.

Attention score in self-attention is secretly a posterior.

Attention with a Mixture of Gaussian Keys

We model each key $\boldsymbol{k}_j$ as a mixture of M Gaussians $N(\boldsymbol{k}_{jr}, \sigma_{jr}^2 \mathbf{I})$, $r = 1, \ldots, M$. The Mixture of Gaussian Keys (MGK) is defined as

$$p(\boldsymbol{q}_i|\boldsymbol{t}_j = 1) = \sum_r \pi_{jr} \mathcal{N}(\boldsymbol{q}_i \mid \boldsymbol{k}_{jr}, \sigma_{jr}^2 \mathbf{I}). \tag{3}$$

Then the posterior is given by

$$p(\boldsymbol{t}_j = 1|\boldsymbol{q}_i) = \frac{\sum_r \pi_{jr} \exp\left(-\|\boldsymbol{q}_i - \boldsymbol{k}_{jr}\|^2 / 2\sigma_{jr}^2\right)}{\sum_{j'} \sum_r \pi_{j'r} \exp\left(-\|\boldsymbol{q}_i - \boldsymbol{k}_{j'r}\|^2 / 2\sigma_{j'r}^2\right)}. \tag{4}$$

MGK uses multiple M keys at each position j
and allows the number of head to be reduced by M times.

Mixture of Keys: Approximation Guarantee

Theorem

Assume that P is probability distribution on $[-a, a]^d$ for some $a > 0$ and admits density function p such that p is differentiable and bounded. Then, for any given variance $\sigma > 0$ and for any $\epsilon > 0$, there exists a mixture of K components $\sum_{i=1}^{K} \pi_i \mathcal{N}(\theta_i, \sigma^2 \mathbf{I})$ where $K \leq (C \log(1/\epsilon))^d$ for some universal constant C such that

$$\sup_{x \in \mathbb{R}^d} |p(x) - \sum_{i=1}^{K} \pi_i \phi(x|\theta_i, \sigma^2 \mathbf{I})| \leq \epsilon,$$

where $\phi(x|\theta, \sigma^2 \mathbf{I})$ is the density function of multivariate Gaussian distribution with mean θ and covariance matrix $\sigma^2 \mathbf{I}$.

MGK can approximate any distribution of the queries.

Inference and Learning

Soft E-step

$$\gamma_{ir} = \frac{\pi_{jr} \exp\left(-\|\boldsymbol{q}_i - \boldsymbol{k}_{jr}\|^2/2\sigma_{jr}^2\right)}{\sum_{r'} \pi_{jr'} \exp\left(-\|\boldsymbol{q}_i - \boldsymbol{k}_{jr'}\|^2/2\sigma_{jr'}^2\right)}, \quad N_{jr} = \sum_{i=1}^{N} \gamma_{ir}, \quad \pi_{jr} = \frac{N_{jr}}{N}.$$

Hard E-step

$$p(\boldsymbol{t}_j = 1|\boldsymbol{q}_i) = \frac{\max_r \ \exp\left(-\|\boldsymbol{q}_i - \boldsymbol{k}_{jr}\|^2/2\sigma_{jr}^2\right)}{\sum_{j'} \max_r \ \exp\left(-\|\boldsymbol{q}_i - \boldsymbol{k}_{j'r}\|^2/2\sigma_{j'r}^2\right)}.$$

M-step

$$\boldsymbol{k}_{jr}^{\text{new}} = \frac{1}{N_{jr}} \sum_{i=1}^{N} \gamma_{ir} \boldsymbol{q}_i, \quad \sigma_{jr}^{2\,\text{new}} = \frac{1}{N_{jr}} \sum_{i=1}^{N} \gamma_{ir} (\boldsymbol{q}_i - \boldsymbol{k}_{jr}^{\text{new}})^\top (\boldsymbol{q}_i - \boldsymbol{k}_{jr}^{\text{new}}).$$

This M-step can be replaced by a generalized M-step that takes the advantage of SGD and backpropagation

Attention with a Mixture of Linear Keys

Output of attention with a Mixture of Gaussian Keys (MGK)

$$\boldsymbol{h}_i = \sum_j \left(\frac{\sum_r \pi_{jr} \exp\left(-\|\boldsymbol{q}_i - \boldsymbol{k}_{jr}\|^2/2\sigma_{jr}^2\right)}{\sum_{j'} \sum_r \pi_{j'r} \exp\left(-\|\boldsymbol{q}_i - \boldsymbol{k}_{j'r}\|^2/2\sigma_{j'r}^2\right)} \right) \boldsymbol{v}_j.$$

Output of attention with a Mixture of Linear Keys (MLK)

$$\begin{aligned} \boldsymbol{h}_i &= \frac{\sum_j \sum_r \pi_{jr} \phi(\boldsymbol{q}_i)^\top \phi(\boldsymbol{k}_{jr}) \boldsymbol{v}_j}{\sum_j \sum_r \pi_{jr} \phi(\boldsymbol{q}_i)^\top \phi(\boldsymbol{k}_{jr})} \\ &= \frac{\phi(\boldsymbol{q}_i)^\top \sum_j \sum_r \pi_{jr} \phi(\boldsymbol{k}_{jr}) \boldsymbol{v}_j^\top}{\phi(\boldsymbol{q}_i)^\top \sum_j \sum_r \pi_{jr} \phi(\boldsymbol{k}_{jr})}. \end{aligned}$$

MLK increases the capacity of linear attention while maintaining the linear complexity of $\mathcal{O}(N)$.

Generalization on Large Scale Tasks

Language Modeling

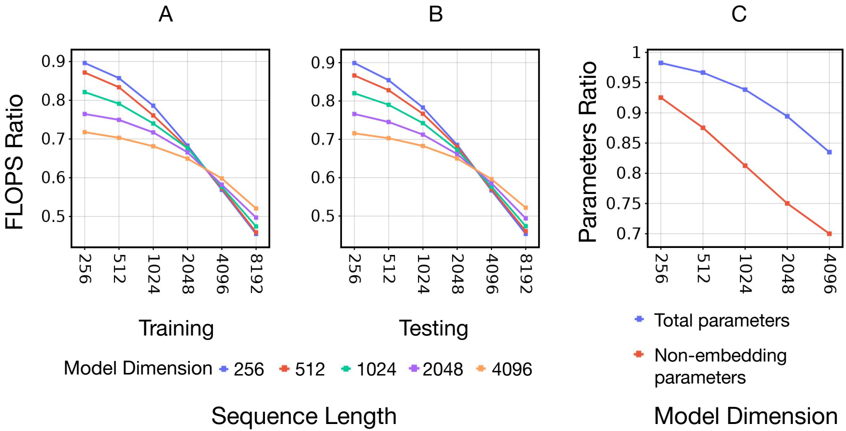
Method	Valid PPL	Test PPL
Softmax 8 heads (small)	33.15	34.29
MGK 4 heads (small)	33.28	34.21
sMGK 8 heads (small)	32.92	33.99
MGK 8 heads (small)	32.74	**33.93**
Softmax 4 heads (small)	34.80	35.85
Linear 8 heads (small)	38.07	39.08
MLK 4 heads (small)	38.49	39.46
MLK 8 heads (small)	37.78	**38.99**
Linear 4 heads (small)	39.32	40.17
Softmax 8 heads (medium)	27.90	29.60
MGK 4 heads (medium)	27.58	**28.86**

Machine Translation

Method	BLEU score
Softmax 4 heads	34.42
Transformer sMGK 2 head	**34.69**
Transformer MGK 2 head	34.34

MGK/MLK still has advantage over softmax/linear attention in large-scale tasks.

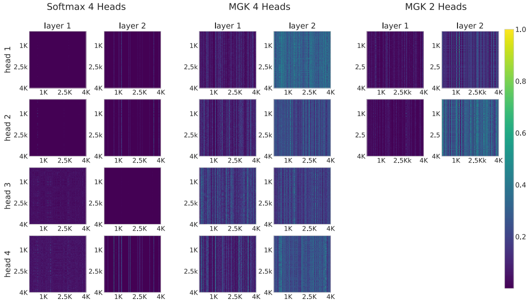
Efficiency Analysis



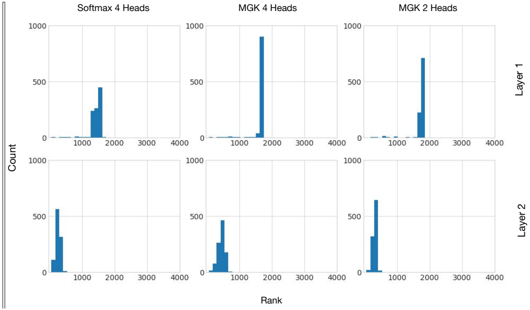
The advantage in efficiency of MGK/MLK grows with the sequence length and the model size.

Redundancy Reduction

Attention Matrices at Each Head



Rank Distribution of Attention Matrices



MGK attention has more representation capacity and is able to capture more diverse attention patterns than softmax attention.

Conclusions

We construct **a Gaussian mixture model underlying the self-attention mechanism**.

We show that the attention score in the attention matrix corresponds to a posterior distribution in our mixture model.

Using our model, we propose a new attention mechanism that uses a mixture of Gaussian and linear keys to increase the efficiency and reduces the redundancy in multi-head self-attention.

Current/Future Work: Understanding transformers via nonparametric regression and the applications of the Fourier integral attention.

Tan M Nguyen, Minh Pham, Tam Nguyen, Khai Nguyen, Stanley J. Osher, Nhat Ho. "Transformer with Fourier Integral Attentions". arXiv:2206.00206, 2022.