MultiMax: Sparse and Multi-Modal Attention Learning

Yuxuan Zhou[1,2], Mario Fritz [2], Margret Keuper [1,3]

[1] University of Mannheim, [2] CISPA Helmholz Center for Information Security, [3] Max Planck Institute for Informatics

Sparsity versus Multi-modality

We first define sparsity and multi-modality metrics for SoftMax and its variants.

Definition 1. Let x_{max} be the largest entry and $x_{max} > x_n > \epsilon$, where ϵ could be any reasonable threshold for an entry to be considered relevant and N is the counts of such entries. Multi-Modality is defined as:

$$\mathcal{M}(\boldsymbol{x}) = 1 - \frac{1}{N} \sum_{\epsilon<x_n<x_{max}}^{N} (\phi(\boldsymbol{x})_{max} - \phi(\boldsymbol{x})_n),$$

Definition 2. Let s∈[0,1] be any reference value for a non-linear scaling of the sparisty score, e.g., the probability of the smallest entry x_{min} after SoftMax ($\text{SoftMax}_{t=1}\,(x)_{min}$). With the exponential term, S(x) results in a smooth approximation of a step function, where larger values indicate stronger degrees of sparsity.

$$\mathcal{S}(\boldsymbol{x}) = \frac{1}{L} \sum_{x_l<\epsilon}^{L} \exp\left(\frac{s-\phi(\boldsymbol{x})_l}{s} - 1\right),$$

Then we revealed and proved the following trade-off:

Proposition 1. For a given input x, the following statements hold w.r.t. temperature t.
1. Multi-modality of SoftMax is monotonically increasing.
2. Sparsity of SoftMax is monotonically decreasing.

This trade-off leads to the inefficacy of temperature tuning in the attention mechanism:

Table 1. Classification accuracy on ImageNet1K using Deit-small with Global Average Pooling (GAP) and classification token (CLS).

Model	Head	Temperature $\frac{1}{t}$					
		0.1	0.5	1	2	10	trainable
Deit-small	CLS	5.1	79.9	79.9	**80.0**	79.5	79.7
	GAP	4.7	80.3	**80.4**	80.0	79.9	80.2

MultiMax: Improved Sparsity and Multi-modality

Definition 3. Let b and d be two control parameters, we apply two corresponding temperatures t_b and t_d only to the entries smaller than b and larger than d, respectively. We construct a piece-wise linear function σ to modulate the SoftMax input $\boldsymbol{x}$, which brings forth our MultiMax:

$$\phi_{MultiMax}(\boldsymbol{x})_i = \frac{\exp(\sigma(x_i))}{\sum_{k=1}^{K} \exp(\sigma(x_k))}, \quad \text{where}$$

$$\sigma(x) = x + \underbrace{(1-t_b)Max(b-x,0)}_{\text{term(1)}} + \underbrace{(t_d-1)Max(x-d,0)}_{\text{term(2)}},$$

We further proved the following for the first-order MultiMax:

Proposition 2. The following properties hold for t_d<1 and t_b>1:
1.MultiMax achieves better sparsity than SoftMax with temperature 1.
2.MultiMax achieves better multi-modality than SoftMax with temperature 1.

Thus MutiMax has higher Pareto Efficiency than SoftMax: Although SoftMax is as sparse as MultiMax with a sufficiently large temperature, but its sparsity will be further decreased, and vice versa.

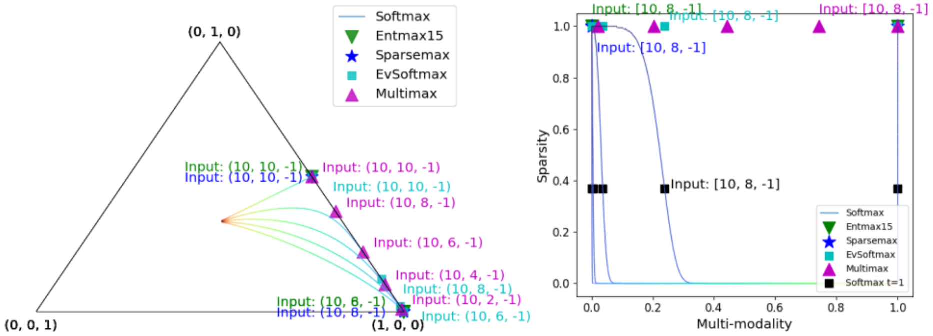

Figure 1. We evaluate SoftMax, SparseMax, EntMax and MultiMax (ours) functions on example input points $v \in \mathbb{R}^3$ and project the resulting distribution on a simplexΔ^2. Informally, the interior of the simplex stands for trimodal distributions, the edges constitute the set of bimodal distributions, and the vertices are the unimodal distributions.

Image Classification

Table 2. Comparing SoftMax to its variants in the attention and/or output layers.

Model	Method	Parameters	Epochs	Modulation		Acc. (%)
				Output	Attention	
Deit-tiny	SoftMax	5M	300	N/A	N/A	72.8
	MultiMax		300	✓	✓	**73.4**
Deit-small	Softmax	22M	300	N/A	N/A	80.4
	Top-k (Wang et al., 2022a)		300	✓	N/A	80.6
	Ev-SoftMax (Chen et al., 2021)		300	-	✓	80.0
	MultiMax		300	✓	-	80.7
			300	-	✓	80.7
			300	✓	✓	**81.0**
Deit-base	SoftMax	86M	300	N/A	N/A	82.1
	MultiMax		300	✓	✓	**82.6**

Language Modeling

Table 3. Evaluation of the performance on WikiText-103 language modeling task by test perplexity.

Method	Attention	Output	Perplexity ↓
SoftMax	-	-	29.4
Top-k (Gupta et al., 2021)	✓	N/A	29.1
MultiMax	✓	-	29.0
	✓	✓	**28.7**

Visualization

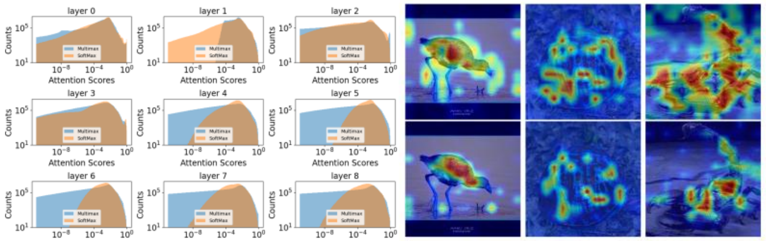

Figure 2. Histograms of the attention scores at each layer (Left) and Grad-CAM using SoftMax (top right row) and MultiMax (bottom right row). From the left, small scores are pushed closer to zero and more scores lie above 0.1 with MultiMax. From the right, MultiMax attention maps are better localized on the objects and are close to zero in most background areas, indicating stronger sparsity.

[1] Wang, et al. Kvt: k-nn attention for boosting vision transformers.
[2] Chen, et al. Evidential softmax for sparse multimodal distributions in deep generative models.
[3] Gupta, et al. Memory-efficient transformers via top-k attention.