t-vMF Similarity For Regularizing Intra-Class Feature Distribution

Takumi Kobayashi

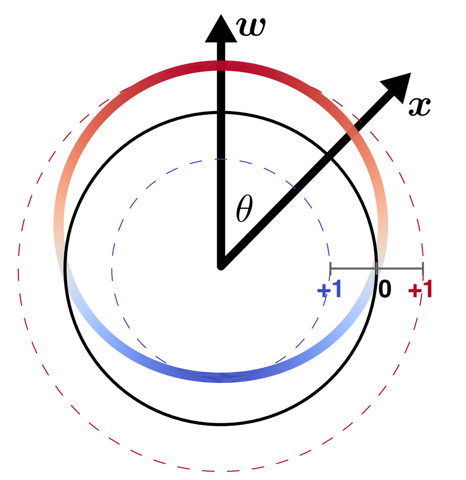
National Institute of AIST, Japan

code
github.com/tk1980/tvMF

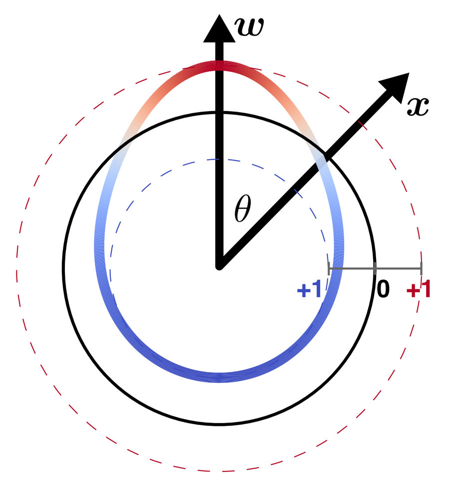
CVPR VIRTUAL JUNE 19-25

Contributions

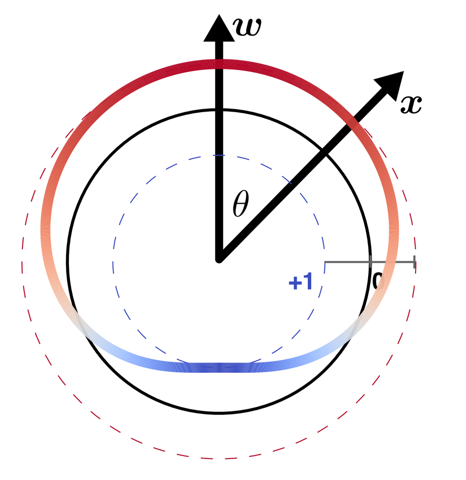
- ✓ We propose novel metric, **t-vMF similarity,** beyond cosine.
- ✓ It *naturally* regularizes **feature distribution within class** for high generalization *in a classification (softmax) loss*.
- ✓ It is simply implemented by **only one-line code**, and improves performance of such as imbalanced and small-scale learning.



Cosine Similarity



t-vMF Similarity
(narrow $\kappa = 4$)
compact intra-class



t-vMF Similarity
(broad $\kappa = -0.3$)
separate inter-class

vMF Similarity Beyond Cosine

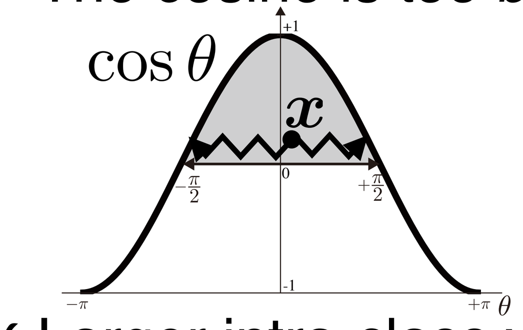
Cosine Similarity

Linear (FC) classifier is characterized by cosine similarity.

$$z = \boldsymbol{w}^T\boldsymbol{x} = \|\boldsymbol{w}\|\|\boldsymbol{x}\|\cos\theta$$

Drawback

✓ The cosine is too *broad*, permitting features of *large variance*.



Broad support region $\theta \in (-\frac{\pi}{2}, +\frac{\pi}{2})$
in which features are distributed *diversely*.

✓ Larger intra-class variance degrades generalization performance.

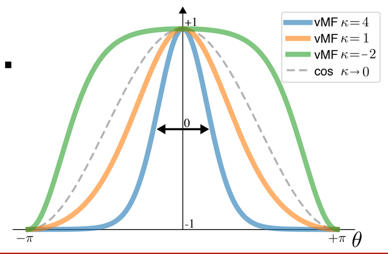
vMF Similarity

We leverage **von Mises-Fisher distribution** to model the similarity.

$$\mathrm{p}(\tilde{\boldsymbol{x}}; \tilde{\boldsymbol{w}}, \kappa) = C_\kappa \exp(\kappa \tilde{\boldsymbol{w}}^\top \tilde{\boldsymbol{x}}) = C_\kappa \exp(\kappa \cos\theta)$$ {Gaussian on sphere}

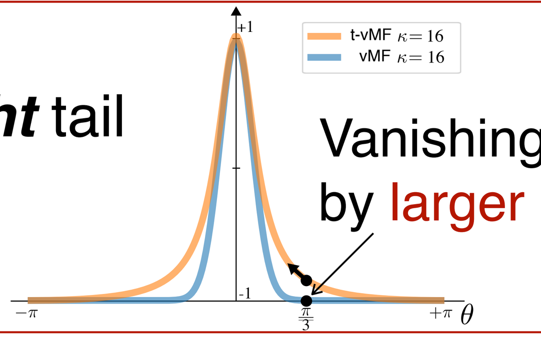
$$\textbf{vMF Similarity:}\quad 2\frac{\exp(\kappa\cos\theta) - \exp(-\kappa)}{\exp(\kappa) - \exp(-\kappa)} - 1 \in [-1, 1]$$

The parameter κ controls the support region.



Drawback

✓ Exponential function provides ***light*** tail which hinders learning.



Vanishing gradient by larger κ

t-vMF Similarity

Following the success of t-SNE, we can employ **Student's t distribution** of ***compact* support** and ***heavy* tail** function to improve vMF similarity.

Student-t function: $\mathrm{f}_t(d; \kappa) = \frac{1}{1 + \frac{1}{2}\kappa d^2}$ where $d = \|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{w}}\|$

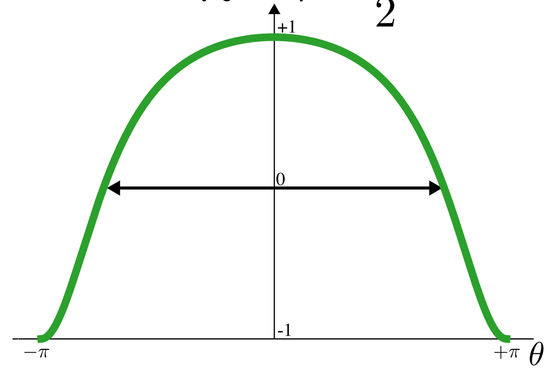
$$\textbf{t-vMF Similarity:}\quad \frac{1 + \cos\theta}{1 + \kappa(1 - \cos\theta)} - 1 \in [-1, 1]$$

Advantage

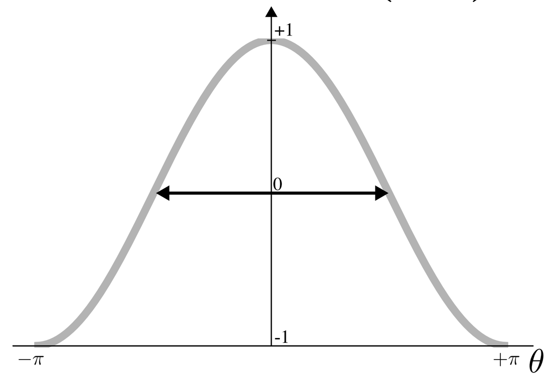
- ✓ Compact support naturally regularizes the feature distribution into **compact**. ↔ Cosine similarity
- ✓ Heavy tail **mitigates vanishing gradients.** ↔ vMF similarity
- ✓ Simple formulation (**one-line** code)

The parameter κ controls the support region.

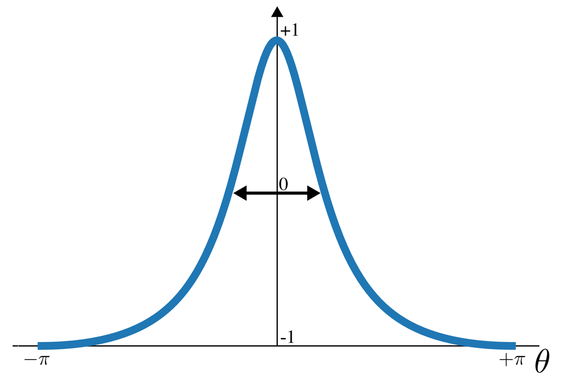
$\kappa \to -\frac{1}{2}$



$\kappa = 0$ (cos)



$\kappa \gg 1$



We can embed the t-vMF similarity into a classification loss, e.g., by

$$\mathrm{l}(\boldsymbol{x}, y) = -\log\frac{\exp\{s\phi(\frac{\boldsymbol{w}_y^\top\boldsymbol{x}}{\|\boldsymbol{w}_y\|\|\boldsymbol{x}\|}; \kappa)\}}{\sum_{c=1}^C \exp\{s\phi(\frac{\boldsymbol{w}_c^\top\boldsymbol{x}}{\|\boldsymbol{w}_c\|\|\boldsymbol{x}\|}; \kappa)\}}$$

✓ Without introducing additional regularization term (such as center loss), **feature distribution is implicitly regularized in the classification loss.**

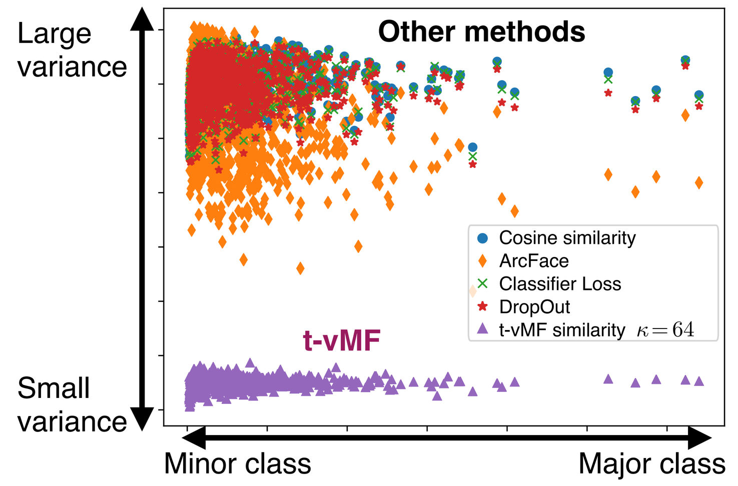
Experimental Results

✓ Performance on **Imbalanced**, **small-scale** and **noisy** datasets where *feature regularization would work* to avoid the issues such as over-fitting.

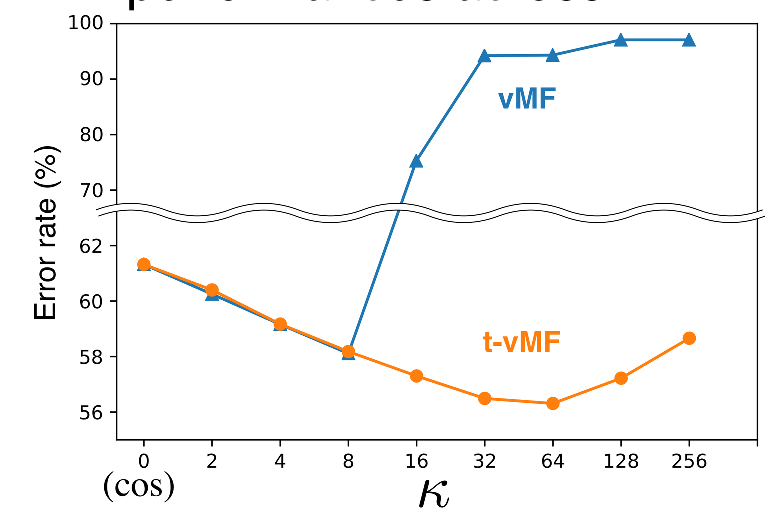
Error rates (%)	(a) Imbalanced			(b) Small-scale		(c) Noisy
Dataset	ImageNet-LT [32]	iNat2018 [22]	iNat2019 [23]	ImageNet-S	ImageNet-SS	ImageNet-N
CNN	ResNet-10	ResNet-50	ResNet-50	ResNet-10	ResNet-10	ResNet-10
Softmax	61.32 38.44	35.95 17.28	27.23 7.95	55.53 31.58	70.52 48.47	82.34 67.61
L-Softmax [31]	60.27 37.13	35.32 16.77	26.70 7.89	53.41 29.60	65.83 41.74	77.42 58.87
ArcFace [11]	59.46 35.29	33.56 14.73	26.83 8.28	53.95 29.68	65.18 40.69	73.17 48.40
Center Loss [46]	60.82 37.79	35.17 16.94	27.53 7.82	55.11 31.24	70.03 47.72	81.80 66.17
Classifier Loss [20]	60.96 37.81	35.49 16.85	26.93 7.89	55.36 31.55	70.21 48.05	82.19 66.59
Virtual Softmax [7]	61.72 35.23	43.83 20.17	30.36 8.78	60.85 33.30	70.90 43.93	72.40 47.72
DropOut [40]	59.17 35.68	32.20 14.53	26.34 7.46	52.69 28.21	66.41 42.78	75.72 55.56
t-vMF (7) ($\kappa = 4$)	59.17 35.98	31.57 13.56	25.22 6.70	53.58 29.36	67.32 43.82	77.28 58.53
t-vMF (7) ($\kappa = 16$)	57.30 32.92	**28.92** 11.75	25.64 6.53	**52.06** 27.54	**64.77** 40.67	71.46 49.19
t-vMF (7) ($\kappa = 64$)	**56.31** 31.78	29.69 11.90	**25.08** 7.10	52.51 28.09	65.73 40.86	**69.19** 45.66
Some SOTAs	58.2 [Kang+20]	32.00 [Cao+19]				

✓ Performance analysis on ImageNet-LT (imbalance)

In-class feature variance

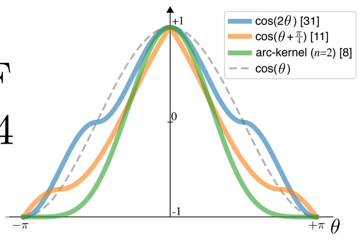


performances across κ



Similarity function

$\cos(k\theta)$			$\cos(\theta + m)$			Arc-kernel		t-vMF
$k = 2$	4	8	$m = \pi/8$	$\pi/4$	$\pi/2$	$n = 1$	2	$\kappa = 64$
59.67	61.02	61.54	60.93	60.81	57.83	61.63	60.80	56.31



✓ Performance on *Healthy* dataset (Large-scale ImageNet)

Negative κ works well due to the **large-margin** effect.