Adaptive Parametric Activation

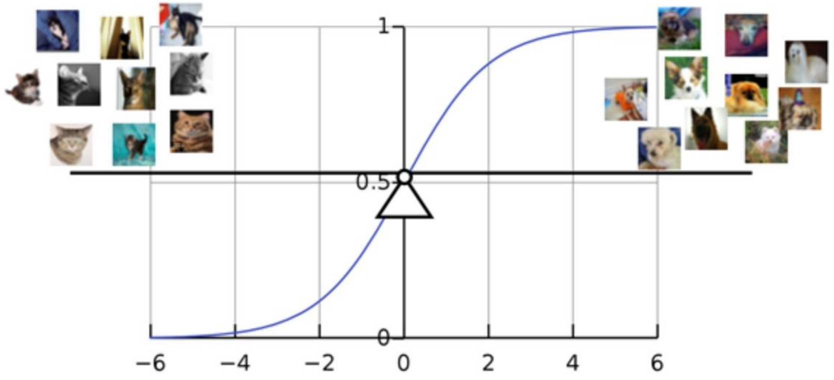
Konstantinos Alexandridis, Jiankang Deng, Anh Nguyen and Shan Luo

Poster #24 - Tue 1 Oct 16:30 - 18:30
Paper: https://arxiv.org/abs/2407.08567
Code: https://github.com/kostas1515/AGLU

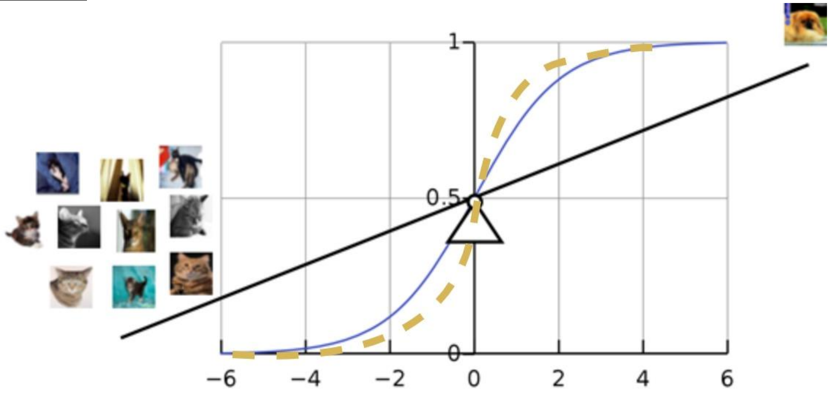
Summary

- We study the importance of the activation function in **balanced** and **imbalanced** classification problems;
- In balanced classification the Sigmoid or Softmax activations are used primarily, however, in imbalanced classification these activations impede rare class learning and other activations are used like Balanced Softmax [1] or Gumbel activation [2] .
- We propose the novel Adaptive Parametric Activation (**APA**) function that **unifies** most common activation functions under a single formula;
- We have validated the efficacy of APA in both **long-tailed**, and **balanced** benchmarks surpassing the state of the art;

Idea



(a) Balanced

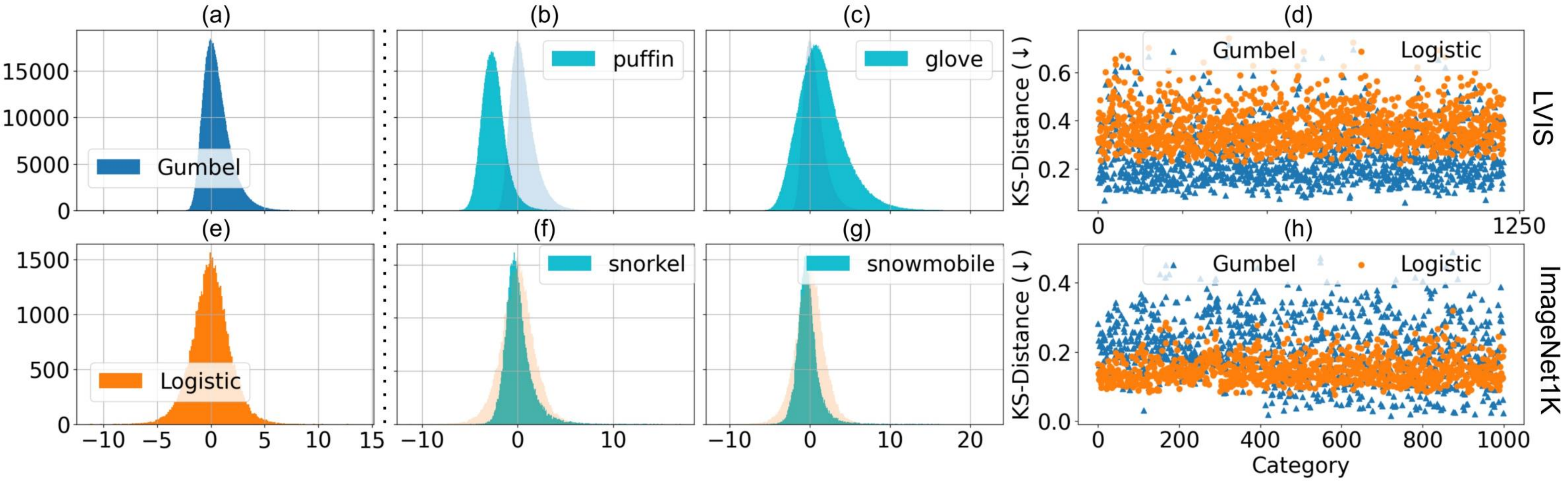


(b) Imbalanced

[1] Ren et. al. 'Balanced Meta-Softmax for Long-Tailed Visual Recognition' in NeurIPS2020
[2] Alexandridis et. al. 'Gumbel Optimised Loss for long-tailed instance Segmentation' in ECCV2022

Background (1)

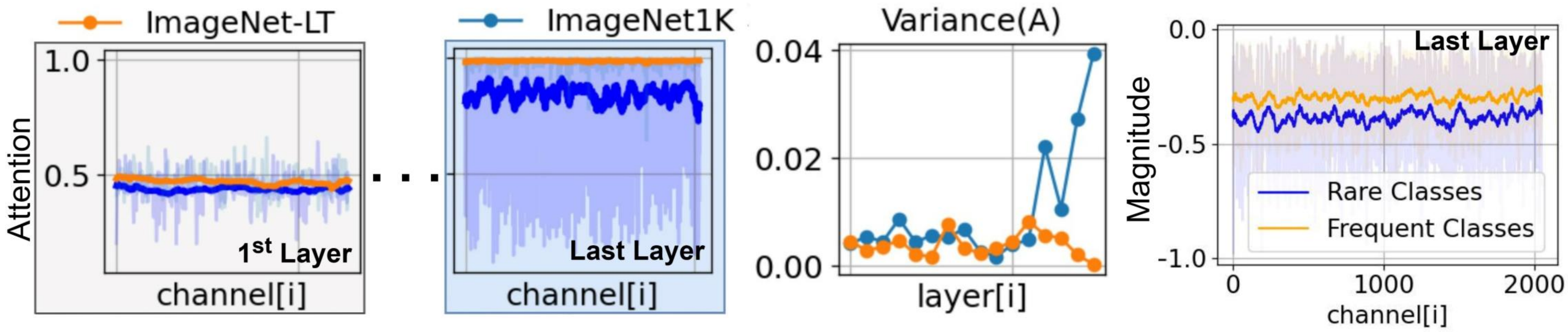
The class imbalance affects the activations of the neural network:



1) learned logit distributions under imbalanced (top) and balanced training (bottom)

Background (2)

The class imbalance affects the channel attention signals of the neural network:



2) Intermediate activations and channel attention signals

Adaptive Parametric Activation (APA)

- Adaptive Parametric Activation (APA) replaces the Sigmoid layer and the RELU layer with a learnable activation function.

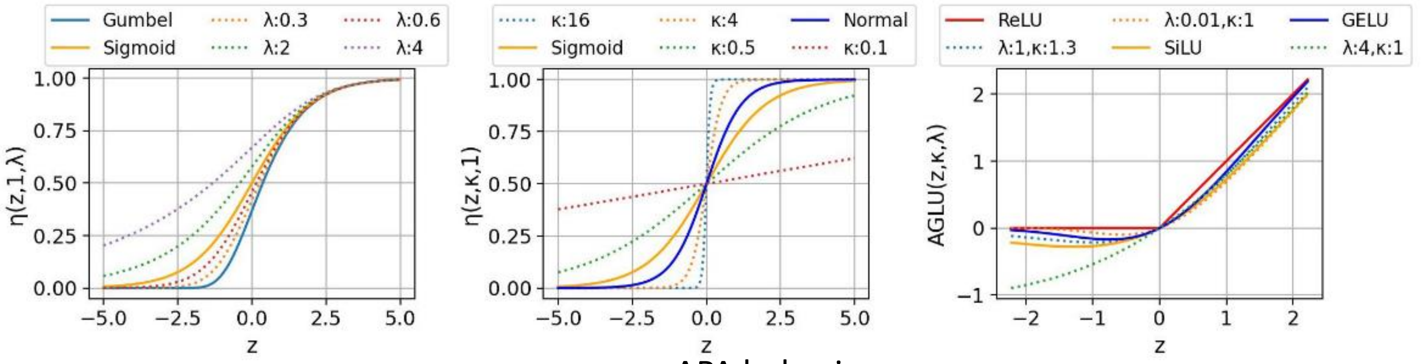
Formula:

Eq.1: $$\eta_{ad}(z, \kappa, \lambda) = (\lambda \exp(-\kappa z) + 1)^{\frac{1}{-\lambda}}.$$

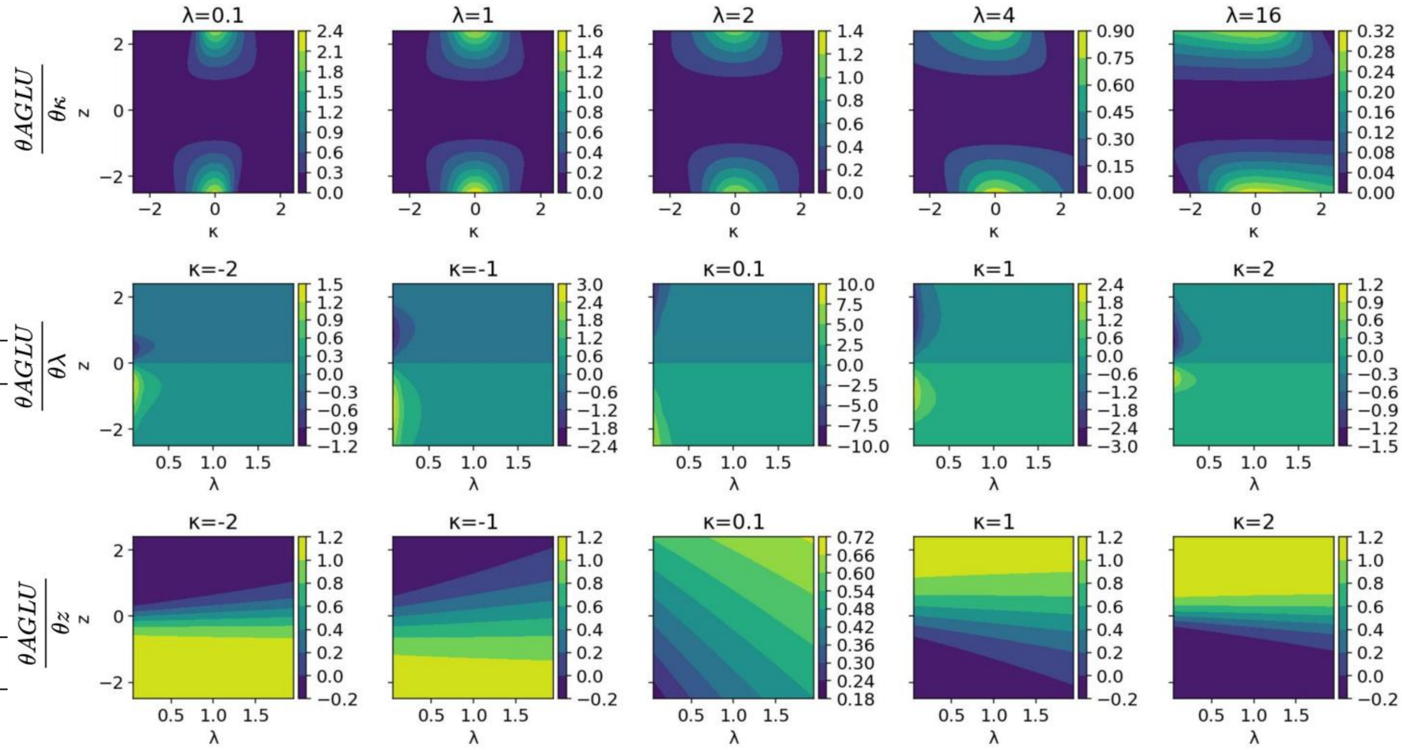
Eq.2: $$AGLU(z, \kappa, \lambda) = z \cdot \eta_{ad}(z, \kappa, \lambda)$$

Table 1: Comparison of different activation functions.

Name	Formula	Range
RELU [21]	$\eta(z) = \max(0, z)$	$(0, \infty)$
Gaussian Unit [31]	$\eta(z) = z\sigma(1.702z)$	$(-0.17, \infty)$
Sigmoid Unit [31]	$\eta(z) = z\sigma(z)$	$(-0.28, \infty)$
Mish [64]	$\eta(z) = z \tanh(\ln(1 + \exp(z)))$	$(-0.31, \infty)$
PRELU [27]	$\eta(z, \kappa) = \max(0, z) + \kappa \min(0, z)$	$(-\infty, \infty)$
ELU [11]	$\eta(z, \kappa) = \max(0, z) + \kappa(\exp(\min(0, z)) - 1)$	$(-\kappa, \infty)$
AGLU (ours)	$\eta(z, \kappa, \lambda) = z \cdot (\lambda \exp(-\kappa z) + 1)^{\frac{1}{-\lambda}}$	$(-\infty, \infty)$



APA behaviour



AGLU derivatives

Results (1)

- APA can be used in both long-tailed classification and long-tailed instance segmentation tasks.

Table 2: Top-1 accuracy (%) on ImageNet-LT test set. E denotes ensemble.

Method	Backbone	Many	Medium	Few	Average
MiSLAS [111]	R50 [28]	61.7	51.3	35.8	52.7
KCL [43]		61.8	49.4	30.9	51.5
TSC [53]		63.5	49.7	30.4	52.4
RIDE (3E)+CMO [68]		66.4	53.9	35.6	56.2
DOC [86]		65.1	52.8	34.2	55.0
CC-SAM [116]		61.4	49.5	37.1	52.4
Our Baseline	SE-R50 [37]	66.2	53.1	37.1	56.0
APA* (ours)		67.5	54.3	39.3	57.4
APA* + AGLU (ours)		**68.3**$^{+1.9}$	**54.8**$^{+0.9}$	**39.4**$^{+2.1}$	**57.9**$^{+1.7}$
RIDE (4E) [95]	X50 [100]	68.2	53.8	36.0	56.8
SSD [52]		66.8	53.1	35.4	56.0
BCL [117]		67.9	54.2	36.6	57.1
CNT [67]		63.2	52.1	36.9	54.2
ALA [110]		64.1	49.9	34.7	53.3
ResLT [13]		63.6	**55.7**	38.9	56.1
ABC-Norm [36]		60.7	49.7	33.1	51.7
RIDE (3E)+CMO+CR [60]		67.3	54.6	38.4	57.4
LWS+ImbSAM [115]		63.2	53.7	38.3	55.3
Our Baseline	SE-X50 [37]	67.9	53.0	37.7	56.7
APA* (ours)		68.9	55.4	39.4	58.4
APA* + AGLU (ours)		**69.8**$^{+1.6}$	55.7$^{0.0}$	**41.1**$^{+2.2}$	**59.1**$^{+1.7}$

Table 5: Comparisons on LVISv1.0 using MaskRCNN-FPN and 2x schedule.

Method	Backbone	AP^m	AP^r	AP^c	AP^f	AP^b
RFS [22]	R50	23.7	13.3	23.0	29.0	24.7
IIF [2]		26.3	18.6	25.2	30.8	25.8
Seesaw [89]		26.4	19.6	26.1	29.8	27.4
LOCE [20]		26.6	18.5	26.2	30.7	27.4
PCB+Seesaw [30]		27.2	19.0	27.1	30.9	28.1
ECM [39]		27.4	19.7	27.0	31.1	27.9
GOL [1]		27.7	21.4	27.7	30.4	27.5
ECM+GAP [107]		26.9	20.1	26.8	30.0	27.2
GOL (baseline)	SE-R50	28.2	20.6	28.9	30.8	28.1
GOL+AGLU(ours)	APA*-R50	**29.1**$^{+0.9}$	**21.6**$^{+0.2}$	**29.6**$^{+0.7}$	**31.7**$^{+0.6}$	**29.0**$^{+0.9}$
RFS [22]	R101 [28]	27.0	16.8	26.5	32.0	27.3
NorCal [65]		27.3	20.8	26.5	31.0	28.1
Seesaw [89]		28.1	20.0	28.0	31.8	28.9
GOL [1]		29.0	22.8	29.0	31.7	29.2
ECM [39]		28.7	21.9	27.9	32.3	29.4
PCB + Seesaw [30]		28.8	22.6	28.3	32.0	29.9
ROG [107]		28.8	21.1	29.1	31.8	28.8
GOL (baseline)	SE-R101	29.7	23.0	29.9	32.5	30.0
GOL+AGLU(ours)	APA*-R101	**30.7**$^{+1.0}$	**23.6**$^{+0.6}$	**31.3**$^{+1.4}$	**33.1**$^{+0.7}$	**31.1**$^{1.1}$

Results (2)

- APA is better than previous activations and can be combined with other attention types in ImageNet-LT:

Activation	Many	Med.	Few	Avg
Sigmoid	66.2	53.1	37.1	56.0
with Temp	65.9	**53.8**	**40.3**	56.6
Gumbel	66.2	53.2	39.3	56.3
with Temp	<u>66.9</u>	<u>53.4</u>	<u>39.7</u>	56.7
APA	**67.1**	**53.8**	39.6	**57.0**

Activations	Avg
ReLU	57.4
PReLU [27]	54.8
ELU [11]	52.6
Mish [64]	57.4
GELU [31]	57.5
SiLU [31]	57.1
AGLU	**57.9**

Attention type	Avg
Spatial [98]	54.8
+APA	55.2
+APA + AGLU	**56.4**
Spatial + Channel [98]	55.6
+APA	56.9
+APA + AGLU	**57.1**

- APA generalizes in balanced classification and detection tasks:

Method	AP^b	AP^m
MaskRCNN-R50	39.2	35.4
w/ SE	40.5	36.9
w/ APA*+AGLU	**41.2**	**37.6**

(a)COCO

Method	AP^b
FasterRCNN-R50	25.4
w/ SE	27.0
w/ APA*+AGLU	**29.9**
CascadeRCNN-R50	31.6
w/ SE	33.3
APA*+AGLU-CRCNN	**35.4**

(b)V3Det

Method	top-1
ResNet50 [37]	76.9
w/ AGLU	**77.5**
SE-ResNet50 [37]	77.5
w/ APA*+AGLU	**78.7**
CBAM-ResNet50 [37]	78.3
w/APA*+AGLU	**78.9**

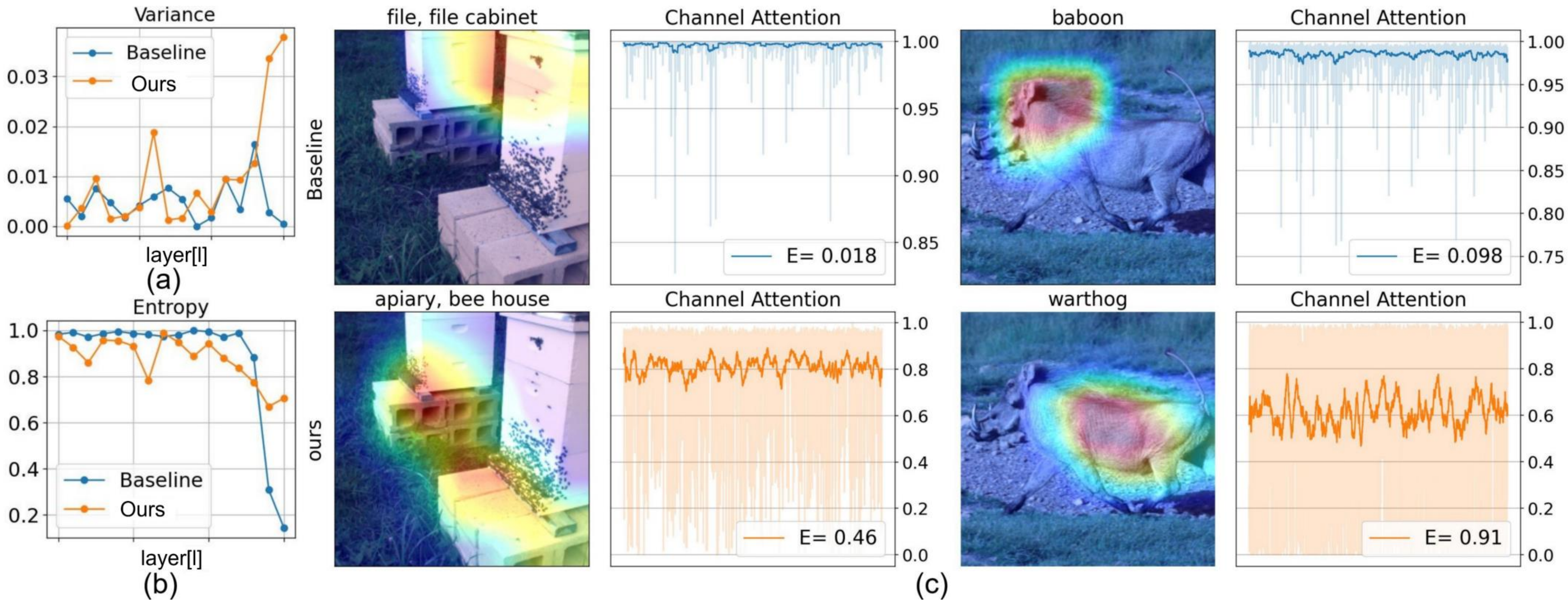
(c)ImageNet1k

Method	top-1
SE-R50	77.5
+APA*	77.9
+APA* + AGLU	**78.7**
SE-R101	79.4
+APA*	79.2
+APA* + AGLU	**80.3**
SE-R152	80.3
+APA*	80.5
+APA* + AGLU	**80.8**

(d) ImageNet1k

Qualitative results

- APA increases the variance (a) and the entropy (b) of the attention signal, allowing the model to correctly classify the rare classes.



Thank you

Poster #24 - Tue 1 Oct 16:30 - 18:30

Paper: https://arxiv.org/abs/2407.08567

Code: https://github.com/kostas1515/AGLU

UK Research and Innovation