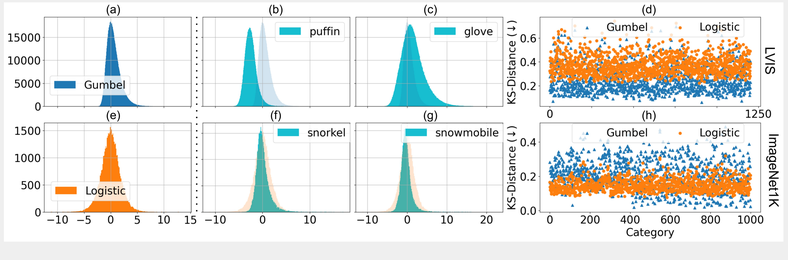
Adaptive Parametric Activation

Konstantinos Panagiotis Alexandridis[1], Jiankang Deng[1], Anh Nguyen[2] and Shan Luo[3]
[1]Huawei Noah's Ark Lab, [2]University of Liverpool, [3]King's College London

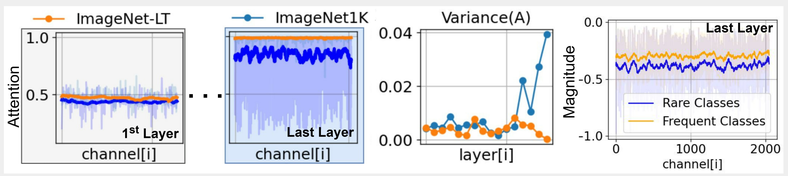
Summary

1. We study the importance of the activation function in balanced and imbalanced classification problems;
2. We propose the novel Adaptive Parametric Activation (APA) function that unifies most common activation functions under a single formula;
3. We have validated the efficacy of APA in balanced and imbalanced benchmarks surpassing the SOTA;

Motivation



The degree of class imbalance affects the shape of the learned logit distributions. When training with the imbalanced LVIS dataset (top), the logit distribution of the classes are closer to the Gumbel distribution, when measured with the Kolmogorov-Smirnov (KS) distance. In contrast, when training with the ImageNet1K (bottom), the logit distributions are closer to the Logistic distribution.



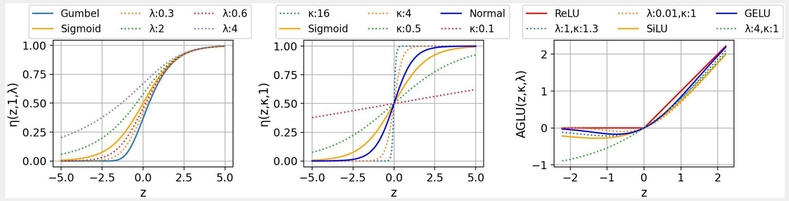
Class imbalance affects the channel attention signals of the model's most semantic layers. When training with ImageNet-LT, the attention signal becomes all-ones and the variance goes to zero. This shows that the semantic attention is biased to the frequent classes, because the activation magnitude is smaller for the rare classes than the frequent classes.

Adaptive Parametric Activation

To boost the activation signals under long-tail learning, we propose the Adaptive Parametric Activation (APA):

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

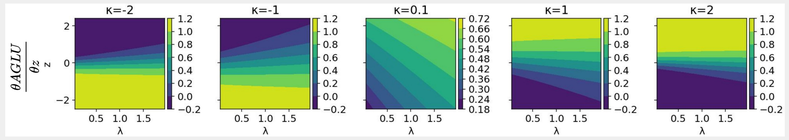
APA adapts its activation rate, according to the input's distribution using two learnable parameters κ and λ.



The Adaptive Generalised Linear Unit (AGLU) is:

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

The κ parameter controls the RELU-ness and the λ parameter controls the leakage, making AGLU flexible and expressive, as shown by it's derivative with respect to z.



AGLU unifies many common activation functions. For example, if $\kappa = \lambda = 1$ then AGLU becomes SiLU, if $\kappa = 1.702$ and $\lambda = 1$, it approximates GELU, if $\kappa \to \infty$, then it approximates ReLU and if $\lambda \to \infty$, then AGLU becomes the identity function.

Name	Formula	Range
RELU	$\eta(z) = \max(0, z)$	$(0, \infty)$
GELU	$\eta(z) = z\sigma(1.702z)$	$(-0.17, \infty)$
SiLU	$\eta(z) = z\sigma(z)$	$(-0.28, \infty)$
Mish	$\eta(z) = z \tanh(\ln(1 + \exp(z))$	$(-0.31, \infty)$
PRELU	$\eta(z,\kappa) = \max(0, z) + \kappa \min(0, z)$	$(-\infty, \infty)$
ELU	$\eta(z,\kappa) = \max(0, z) + \kappa(\exp(\min(0, z)) - 1)$	$(-\kappa, \infty)$
AGLU (ours)	$\eta(z,\kappa,\lambda) = z \cdot (\lambda \exp(-\kappa z) + 1)^{\frac{1}{-\lambda}}$	$(-\infty, \infty)$

Results

Results on long-tailed image classification, using Squeeze and Excite (SE) ResNets (R). $APA*$ shows that both APA and AGLU are used inside the SE model.

Method	Acc
SE-X50	56.7
RIDE(4E)	56.8
ResLT	56.1
$APA*$ (ours)	**59.1**

(a)ImageNet-LT

Method	Acc
SE-R50	71.3
BCL	71.8
ResLT	70.5
$APA*$ (ours)	**74.8**

(b)iNaturalist18

Activations	Acc
ReLU	57.4
GELU	57.5
SiLU	57.1
AGLU	**57.9**

(c)ImageNet-LT

Results on LVISv1 using Mask-RCNN-SE-ResNet50, our $APA*$ increases the rare class performance AP^r.

Method	AP	AP^r	AP^c	AP^f	AP^b
GOL	28.2	20.6	28.9	30.8	28.1
GOL w/ $APA*$ (ours)	**29.1**	**21.6**	**29.6**	**31.7**	**29.0**

$APA*$ generalises well to balanced benchmarks like COCO, V3DET and ImageNet1K, using ResNet50.

Method	AP^b	AP^m
MRCNN	39.2	35.4
w/ SE	40.5	36.9
w/ $APA*$	**41.2**	**37.6**

(a)COCO

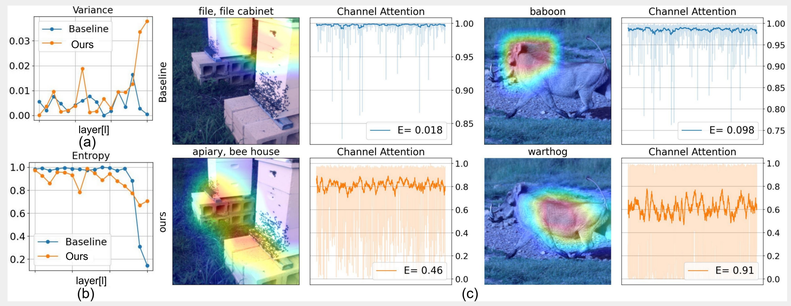
Method	AP^b
CRCNN	31.6
w/ SE	33.3
w/ $APA*$	**35.4**

(b)V3Det

Method	top-1
SE	77.5
w/ APA*	**78.7**

(c)ImageNet1k

$APA*$ increases the variance of the channel attention signals, thus it debiases the model. Also, it increases the channel attention's entropy in the semantic layers of the model, allowing the model to better recognise the rare classes like *apiary* and *warthog*.



Contact: alex.kostas@gmail.com