Adaptive Parametric Activation

Konstantinos Panagiotis Alexandridis[1], Jiankang Deng[1], Anh Nguyen[2], and Shan Luo[3]

[1] Huawei Noah's Ark Lab
{konstantinos.alexandridis,jiankang.deng}@huawei.com
[2] University of Liverpool, Liverpool L69 3BX, United Kingdom
{anguyen}@liverpool.ac.uk
[3] King's College London, London WC2R 2LS, United Kingdom
{shan.luo}@kcl.ac.uk

Abstract. The activation function plays a crucial role in model optimisation, yet the optimal choice remains unclear. For example, the Sigmoid activation is the de-facto activation in balanced classification tasks, however, in imbalanced classification, it proves inappropriate due to bias towards frequent classes. In this work, we delve deeper in this phenomenon by performing a comprehensive statistical analysis in the classification and intermediate layers of both balanced and imbalanced networks and we empirically show that aligning the activation function with the data distribution, enhances the performance in both balanced and imbalanced tasks. To this end, we propose the Adaptive Parametric Activation (APA) function, a novel and versatile activation function that unifies most common activation functions under a single formula. APA can be applied in both intermediate layers and attention layers, significantly outperforming the state-of-the-art on several imbalanced benchmarks such as ImageNet-LT, iNaturalist2018, Places-LT, CIFAR100-LT and LVIS and balanced benchmarks such as ImageNet1K, COCO and V3DET. The code is available at https://github.com/kostas1515/AGLU.

Keywords: Activation function · Long-tailed learning

1 Introduction

Image recognition has witnessed tremendous progress over the last years due to the use of deep learning, large image datasets such as ImageNet1K [17] and advances in model architectures [19, 28], learning algorithms [24, 25, 69], activation layers [21, 27, 31] and normalisation techniques [4, 40]. In this work, we focus on the activation layer of the network.

In balanced image classification works, it was empirically shown that if the activation function is close to the real data distribution then the model converges faster because the learning objective becomes easier [27, 31]. Based on this, the GELU [31] and the PRELU [27] were proposed as alternatives to the commonly used RELU [21] and they were utilised inside the model's layers to

activate the intermediate features of the network. Similarly, in imbalanced image classification, many works have empirically shown that the Sigmoid or the Softmax activation functions, are inappropriate and using another activation function increases the performance [1, 34, 63, 71]. Based on that, the Gumbel activation [1] and the Balanced Softmax [71] activation were proposed and they were used inside the classification layer to predict the classes. In contrast to balanced classification, these works focused only on the classification layer and they disregarded the importance of the intermediate activations. To this end, there is no principled way of choosing the right activation function, and usually, practitioners use the best activation function, according to the task, through cross-validation or parameter-tuning.

In this work, we focus on this problem. First, we theoretically show that the activation function enforces a prior belief of how the data is distributed and therefore it acts as an initialisation point. In practise, a good initialisation point enhances the convergence, therefore, having an appropriate activation function can increase the performance. Second, we study the impact of the activation function on balanced and imbalanced classification from two perspectives, i.e. the classification layer, and the intermediate layers.

Our findings show that the classification logit distribution of a pretrained model heavily depends on the degree of data imbalance. For example, in balanced training, the classification logits align better with the Logistic distribution, while in imbalanced learning they align better with the Gumbel distribution. Regarding the intermediate layers, we study the channel attention layer as an example and we find that it is also affected by the degree of data imbalance. We find that, in balanced learning, the channel attention is robust for all classes, however in imbalanced training the channel attention enhances more the frequent classes than the rare classes.

To this end, we empirically show that the commonly used Sigmoid activation function cannot generalise for both balanced and imbalanced learning, because it is non-parametric and does not align with the imbalanced data distribution. Motivated by this, we develop a novel Adaptive Parametric Activation (APA) function. APA allows the model to align its activations to both the balanced and imbalanced data distributions and reach great performance in both tasks. APA has several benefits, it unifies most previous activations functions such as the Sigmoid, the Gumbel [1], the RELU [21], the SiLU [31] and the GELU [31] under a common formula. Also, it uses two learnable parameters that allow the network to select the best activation function during optimisation enlarging the model's capacity. Moreover, our APA is versatile, it can be used as a direct replacement to RELU, or it can replace the Sigmoid activation function inside the attention mechanism boosting the performance significantly and consistently. Finally, APA can be generalised to both imbalanced and balanced classification and detection tasks and surpass the state-of-the-art (SOTA). Our **contributions** are:

– We demonstrate the importance of the activation function in balanced and imbalanced data distributions, through statistical analysis;

- We propose the novel APA function that unifies most common activation functions under a single formula;
- We have validated the efficacy of APA on a range of long-tailed benchmarks, including ImageNet-LT [58], iNaturalist18 [84], Places-LT [58], CIFAR100-LT [8], LVIS [22] and balanced benchmarks such as ImageNet1k [17], COCO [55] and V3Det [88] largely surpassing the state of the art.

2 Preliminaries

Activation function. First, we show the importance of the activation function in model optimisation, following [1]. Let's consider the example of binary classification, where z is the input, $y \in \{0, 1\}$ is the target class and $f(z) = W^T z + b$ is the classification network. The target class and the input z are related as:

$$y = \begin{cases} 1, & \text{if } f(z) + \epsilon > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where ϵ is the error, that is a random variable and it is distributed according to the real data distribution. In this example, the classification boundary is set to 0, however, it can be adjusted by the network's bias b during optimisation. The probability of the class $P(y = 1)$ is:

$$P(y = 1) = P(f(z) + \epsilon > 0) = P(\epsilon > -f(z)) = 1 - F(-f(z)) \tag{2}$$

where F is the cumulative distribution function. During network optimisation, if one chooses the Sigmoid activation function $\sigma(z) = (1 + \exp(-z))^{-1}$, then the prediction $\bar{y}$ is obtained using $\bar{y} = \sigma(f(z))$ and the probability $P(\bar{y} = 1)$ is:

$$P(\bar{y} = 1) = \frac{1}{1 + \exp(-f(z))} = F_{\text{logistic}}(f(z); 0, 1) = 1 - F_{\text{logistic}}(-(f(z); 0, 1) \tag{3}$$

Comparing Eq. 2 and Eq. 3, it is shown that when we use the Sigmoid, we assume that the error term ϵ follows the standard Logistic distribution. If we use Gumbel activation then we assume that the error follows the Gumbel distribution [1] and, in general, any activation function assumes a different distribution of ϵ.

For this reason, the activation function can be seen as an initialisation point, or a prior belief of how the real data are distributed. If the prior is good, then the learning objective becomes easier and the performance is increased as it was shown empirically by many past studies [1, 21, 27, 77].

2.1 Balanced versus Imbalanced learning

In this subsection, we perform a statistical analysis on the empirical distributions of the logits and the intermediate activations, to understand the importance of the activation function in balanced and imbalanced learning.

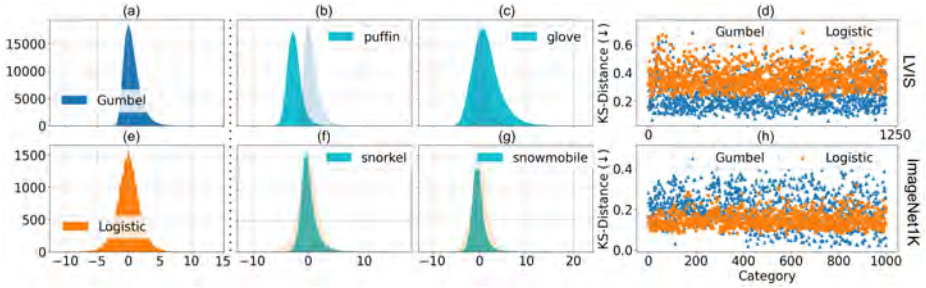

Fig. 1: Top: In imbalanced learning, the logit distributions are more skewed and they have a smaller KS distance to the Gumbel than the Logistic distribution as shown in (d). Bottom: In balanced learning, the logit distributions are less skewed and they align better with the Logistic, than the Gumbel distribution, as shown in (h).

Classification logits. We train a MaskRCNN [26] ResNet50 [28] on LVIS dataset [22], which is a highly imbalanced object detection dataset containing 1203 classes. After the model has converged, we perform inference and store the predicted classification logits, i.e., $f(z_y)$ for every class y. Next, we perform histogram binning on the logits and we visualise the empirical distributions for the rare class *puffin* and the frequent class *glove* in Fig. 1 (b) and (c) respectively.

The rare class has a negative average logit value, because it is dominated by the frequent classes, which have higher average logit values as shown in (c). Regarding its distribution, it resembles the Gumbel distribution, because it has a heavy right tail and it is skewed. To quantify that, we calculate the statistical distance, using the Kolmogorov-Smirnov (KS) test [62], between all empirical class distributions and the Gumbel and Logistic theoretical distributions, shown in (a) and (e). As shown in (d), most classes have smaller distance to the Gumbel distribution. We repeat this test, using ResNet50 trained on the balanced ImageNet1K. As shown in (f) and (g), the classification logits are different this time, for example most average logit values are centered around zero and they are less skewed. In general, the logit distributions are closer to the logistic distribution as shown in (h). This explains why the Sigmoid activation achieves better performance in the balanced classification task and worse performance in imbalanced classification [1, 34, 63, 71]. Next, we show that data imbalance also affects the intermediate layers, by studying the channel attention as an example.

Attention layer. Attention mechanism for input $X \in \mathbb{R}^{H \times W \times C}$ re-weights the input features X by applying an attention function $A(X)$, i.e., $X' = A(X) \otimes X$. For example, in Channel Attention (CA) [37], $X' = \sigma(\text{MLP}(\text{GAP}_c(X))) \otimes X$, where $\text{GAP}_c \in \mathbb{R}^{1 \times 1 \times C}$ is Global Average Pooling and $\otimes$ is the element-wise product. In this case, the attention function is $A_{CA}(X) = \sigma(\text{MLP}(\text{GAP}_c(X)))$.

Balanced vs Imbalanced channel attention. We train SE-ResNet50 [37] models (SE-R5) on balanced ImageNet-1K and imbalanced ImageNet-LT. After training the models, we analyse the average channel attention signals for a random batch of 128 test images. Fig. 2-a shows the output A_{CA}, in the first layer of SE-R50, Fig. 2-b shows the output, in the last layer, and Fig. 2-c shows the variance of A_{CA} across all layers. As shown in Fig. 2-a, the attention is simi-

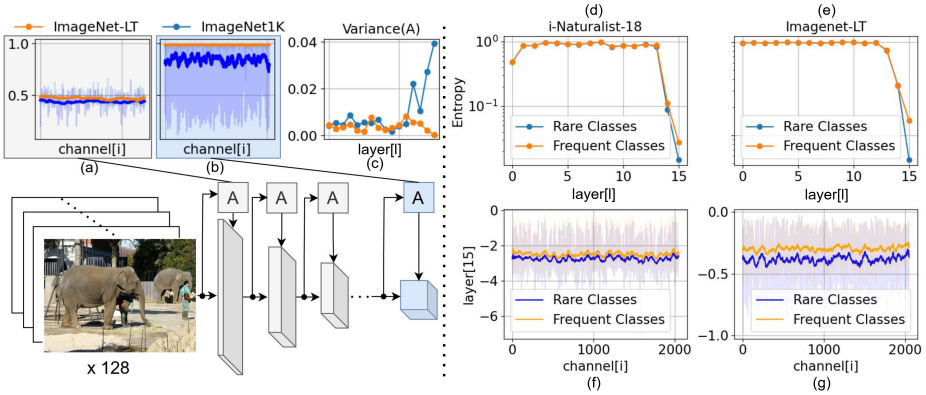

Fig. 2: Visualisations of channel attention (A). In (a) the attention signals when training with imbalanced ImageNet and balanced ImageNet have similar variance in the first layer but completely different in the last most semantic layer in (b). The variance with ImageNet-LT training drops to zero for deeper layers as shown in (c), because the attention promotes only few frequent classes. In (d) and (e) the entropy of channel attention is smaller for the rare classes than the frequent classes in both i-Naturalist18 and ImageNet-LT training. In (f) and (g) the channel responses are smaller for the rare classes than the frequent ones for i-Naturalist18 and ImageNet-LT respectively.

lar in the first layer and it is different in the last layer, as shown in Fig. 2-b. As displayed in Fig. 2-c, the attention variance with ImageNet1K (blue-curve) is larger and the attention reweights all channels and affects all classes. In contrast, for the imbalanced case (orange-curve), the attention signal has small variance, indicating that it is biased to some classes.

Layer-wise analysis. This phenomenon is most prevalent in the last attention layer, which is the most semantic. To quantify which attention layers are affected the most, we use entropy as a measure of signal complexity. Since the channel attention produces a probabilistic weighting vector via the Sigmoid activation, we calculate the total entropy of channel attention, for a layer l, as the sum of the binary channel distribution entropies as follows:

$$E_l = -\frac{1}{C}\sum_{i=1}^{C}[(A_{CA}(X_l)\log(A_{CA}(X_l)) + (1 - A_{CA}(X_l)\log(1 - A_{CA}(X_l))] \quad (4)$$

When the layer's entropy is closer to zero, the channel attention signals are closer to 1 and they do not affect the original features for that layer, i.e., $X' = 1 \otimes X$. If the layer's attention entropy is closer to one, then the channel attention signals are informative, as they affect the signal i.e., $X' = A(X) \otimes X$.

To investigate the complexity of the attention signal, we propagate two batches of 64 test images that contain only frequent and only rare classes respectively, through the pretrained SE-R50 and measure the attention entropy of A_{CA}. In Fig. 2 (d) and (e), the average entropy is similar for both frequent and rare classes for all layers except for the last layer which is the most semantic. The blue curve, that corresponds to rare class channel attention, has lower entropy

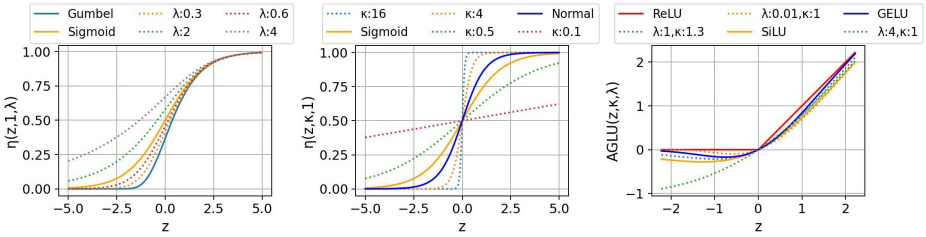

Fig. 3: Our APA unifies most activation functions under the same formula.

than the orange curve that corresponds to frequent class channel attention in the last layer for both i-Naturalist-18 and ImageNet-LT. This indicates that channel attention produces simpler signals for the rare classes and more complex signals for the frequent classes. Finally, in (f) and (g), we show that the average channel responses are smaller when the inputs are rare classes, than frequent classes, which explains why the network cannot model the rare classes effectively. In conclusion, this analysis shows that data imbalance affects the quality of the activations inside the intermediate layers and it highlights the limitation of the Sigmoid activation to model the rare classes.

3 Method

As shown in the previous section, the degree of data imbalance affects both the classification logits and the intermediate layers. While it is possible to perform a statistical analysis and select the appropriate activation for the classification layer, this is difficult to do for all layers of the network, because first, the data distributions inside the layers dynamically change during training [4, 40], and secondly, there is no one-to-one correspondence between classes and intermediate channels, which hinders attribution. For this reason, we propose the Adaptive Parametric Activation (APA):

$$\eta_{ad}(z, \kappa, \lambda) = (\lambda \exp(-\kappa z) + 1)^{\frac{1}{-\lambda}}. \tag{5}$$

APA can adjust its activation rate, dynamically, according to the input's distribution using two parameters κ and λ, that can be learned during optimisation. $\kappa \in \mathbb{R}$ is the gain parameter that controls the function's sensitivity. $\lambda \in (0, \infty)$ is the asymmetrical parameter that controls the function's response rate to positive and negative inputs, allowing the model to have different learning degrees when the input is positive or negative. This function is also known as Richard's curve [73] and it unifies the most common activation functions.

For example, if $\kappa = \lambda = 1$ then APA becomes the Sigmoid activation that has a symmetric response rate for both positive and negative inputs and it is successful for balanced classification tasks. If $\kappa = 1, \lambda \to 0$ then APA becomes the Gumbel activation that has an asymmetric response rate and it is successful for long-tailed instance segmentation tasks [1]. This behaviour is shown in Fig. 3 left and middle. Based on Eq. 5, we also define the Adaptive Generalised Linear

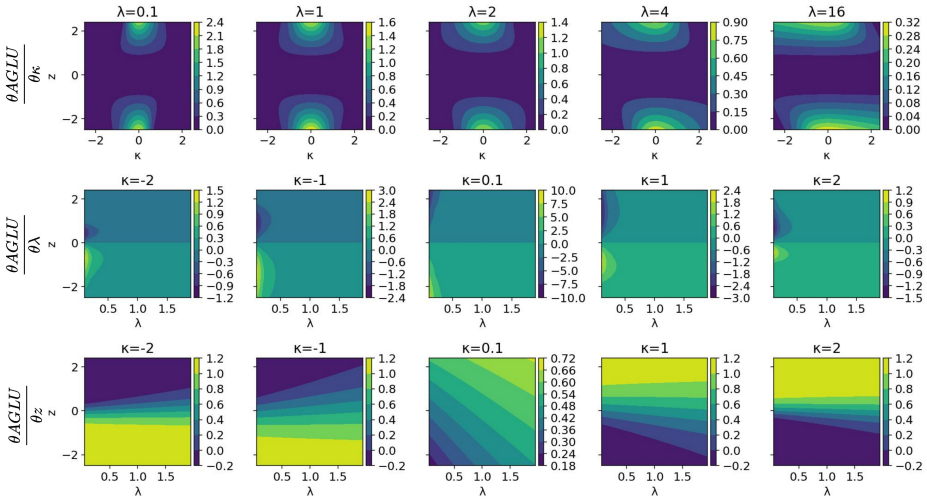

Fig. 4: AGLU derivatives with respect to κ (top), λ (middle) and z (bottom).

Unit (AGLU):

$$AGLU(z, \kappa, \lambda) = z \cdot \eta_{ad}(z, \kappa, \lambda) \tag{6}$$

AGLU has many interesting properties, for example, if $\kappa = \lambda = 1$ then AGLU becomes the Sigmoid Linear Unit (SiLU) [31]. If $\kappa = 1.702, \lambda = 1$, then it becomes the Gaussian Error Linear Unit (GELU) [31]. If $\kappa \to \infty$, then AGLU becomes ReLU [21] and if $\lambda \to \infty$, then AGLU becomes the identity function, as shown in Fig. 3-right.

In other words, the κ parameter controls the RELU-ness and the λ parameter controls the leakage. Also, AGLU could be seen as a smoother version of PRELU, because $AGLU(z, 1, \lambda \to \infty) = PRELU(z, 1)$. We compare $AGLU$ to most common activation functions, in more detail, in Table 1. The derivative of AGLU with respect to κ is:

$$\frac{\partial AGLU(z, \kappa, \lambda)}{\partial \kappa} = z^2 \cdot \frac{\eta_{ad}(z, \kappa, \lambda)}{\lambda + \exp(\kappa z)} \tag{7}$$

the derivative of AGLU with respect to λ is:

$$\frac{\partial AGLU(z, \kappa, \lambda)}{\partial \lambda} = -\frac{z}{\lambda} \cdot \frac{\eta_{ad}(z, \kappa, \lambda)}{\lambda + \exp(\kappa z)} \tag{8}$$

and the derivative of AGLU with respect to z is:

$$\frac{\partial AGLU(z, \kappa, \lambda)}{\partial z} = \kappa z \cdot \frac{\eta_{ad}(z, \kappa, \lambda)}{\lambda + \exp(\kappa z)} + \eta_{ad}(z, \kappa, \lambda) \tag{9}$$

The proofs of the derivatives are shown in the Appendix. The derivatives of AGLU are shown in Fig. 4. Using various κ and λ combinations AGLU has

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)	$\eta(z, \kappa, \lambda) = z \cdot (\lambda \exp(-\kappa z) + 1)^{\frac{1}{-\lambda}}$	$(-\infty, \infty)$

drastically different behaviour and this enhances the capacity of the network and achieves good performance as shown in the experiments.

In the end, our APA is versatile because it can be used not only as an error unit, replacing RELU, but also as an activation function inside the attention mechanism, replacing the Sigmoid activation.

4 Experiment Setup

Datasets. We use CIFAR100-LT [8] with exponential imbalance factor of 100 and 10, ImageNet-LT [58], Places-LT [58] and iNaturalist2018 [84] following the common long-tailed classification protocol. We report our results using top-1 accuracy on the balanced test sets, to fairly evaluate all classes. For ImageNet-LT, we split the categories according to their class frequency in the training set, into *Many* (>100 images), *Medium* (20∼100 images) and *Low* (<20 images) and do per-group evaluation following [58]. Also, we use the LVISv1 [22] instance segmentation dataset, which has 100K training images and 1203 classes, that are grouped according to *Frequent* (>100 images), *Common* (10∼100 images) and *Rare* (<10 images) classes. For this dataset, we report mask average precision AP, bounding box average precision AP^b and AP^r, AP^c and AP^f which is mask average precision for rare, common and frequent classes respectively. Regarding balanced training, we perform experiments on ImageNet1K [17], COCO [55] and the recently proposed V3Det [88], which is a challenging large scale detection dataset with 13K classes and 243K images. We report top-1 accuracy for ImageNet1K, AP^b and AP^m for COCO and AP^b for V3Det.

Implementation Details We primarily use Squeeze and Excite [37] as our baseline with ResNet-32 [28] for CIFAR-LT, ResNet50 for iNaturalist, ResNet50 and ResNext50 [100] for ImageNet-LT and ResNet152 for Places-LT, which has been pretrained on ImageNet1K according to [58]. For LVIS, we use SE-Resnets with MaskRCNN [26], FPN [54], Normalised Mask [89], RFS sampler [22] and GOL loss [1] as a baseline. For V3Det, we use SE-ResNet50 with FasterRCNN [72], FPN and Wrapped Cauchy classifier [23]. All baselines use bag of tricks [109], and strong training techniques that we discuss and ablate in the Appendix. Our motivation for using bag of tricks is two-fold, first, it pushes the performance even further and secondly, it showcases the generalisability of our work.

For our attention models, we replace the Sigmoid with APA, and we further use LayerNorm [4] and attention dropout [32] with $p = 0.1$, for all datasets except

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**$^{+1.7}$

iNaturalist. We denote this configuration as APA* in our Tables and we ablate its components. For our AGLU models, we simply use it in-place of RELU.

5 Results

Long-tailed Classification Benchmark. We compare APA* against ensemble and fusion models [7,13,50,95], margin adjustment [34,108,110], contrastive learning [43,53,91], knowledge transfer [67], knowledge distillation [29,52], decoupled methods [2,36,108,111], sharpness aware minimisation [60,115,116] and data augmentation [68,111].

On ImageNet-LT, as shown in Table 2, our baseline models with bag of tricks reach the state-of-the-art (SOTA) for both SE-ResNet50 (R50) and SE-ResNeXt50 (X50). We want to point out that most performance comes from the bag of tricks and not the SE module, as shown, in detail in the Appendix.

Our APA* outperforms the SE-R50 baseline by 1.4 percentage points (pp) on average, by 1.3pp on frequent categories, 1.2pp on medium and 2.2pp on few classes. Most importantly, it increases the performance of both frequent and rare classes, which is a unique advantage compared to the previous works. Additionally, our APA* exceeds RIDE with 3 Experts (3E) and CMO [68] by 1.2pp on average and by 3.7pp on the rare classes using a single model. When AGLU is combined, it pushes the performance of APA* R50, by 0.5pp on average, by 0.8pp on the frequent, 0.5pp on the medium and 0.1pp on the few classes.

Table 3: (a) Results of AGCA and AGLU for iNaturalist and Places-LT. (b) Results of AGCA and AGLU for CIFAR100-LT with imbalance 10 and 100.

Method	iNat18	PlacesLT
DisAlign [108]	70.6	39.3
MisLAS [111]	71.6	40.4
LADE [34]	70.0	38.8
ALA [110]	70.7	40.1
TSC [53]	69.7	-
CNT [67]	-	39.2
WD+MaxNorm [3]	70.2	-
DOC [86]	71.0	-
BCL [117]	<u>71.8</u>	-
ResLT [13]	70.5	39.8
IIF [2]	-	40.2
ABC Norm [36]	71.4	-
LWS+ImbSAM [115]	71.1	-
CC-SAM [116]	70.9	<u>40.6</u>
AREA [10]	68.4	-
Our Baseline with SE	71.3	40.5
APA (ours)	72.3	41.3
APA +AGLU (ours)	**74.8**$^{+3.0}$	**42.0**$^{+1.4}$

(a)

Method	10	100
BALMS [71]	63.0	50.8
RIDE (4E) [95]	-	49.4
ACE (4E) [7]	-	49.6
DiVE [29]	62.0	45.4
SSD [52]	62.3	46.0
MisLas [111]	63.2	47.0
HSC [91]	63.1	46.7
LADE [34]	61.7	45.4
ResLT (3E) [13]	63.7	49.7
TLC (4E) [50]	-	49.8
TSC [53]	59.0	43.8
RIDE+CMO [68]	60.2	50.0
CC-SAM [116]	-	50.8
RIDE+CMO+CR [60]	61.4	50.7
AREA [10]	60.8	48.9
Our Baseline with SE	<u>65.2</u>	<u>50.9</u>
APA (ours)	65.7	51.9
APA+AGLU(ours)	**66.8**$^{+1.6}$	**52.3**$^{+1.4}$

(b)

APA* with X50 increases the SE-X50 baseline by 1.7pp on average, 1.0pp the many classes, 1.4pp the medium classes and 1.7pp the few classes. Furthermore, it outperforms RIDE(3E)+CMO+CR [60] by 1.0pp on average, 1.6pp on frequent classes, 0.8 on common and 1.0pp on rare classes using a single model. When AGLU is combined to APA* X50, it adds 0.7pp on average, 0.9pp on the frequent categories, 0.3pp on the medium, and 1.7pp on the few classes.

On iNaturalist18, Places-LT and CIFAR100-LT, our SE baselines with bag of tricks again reach the SOTA, as shown in Table 3-a and b.

Regarding, iNaturalist18, APA* improves the SE-baseline by 1.0pp. It also outperforms BCL [117] by 0.5pp, ResLT by 1.8pp and LWS+ImbSAM [115] by 1.2pp. AGLU further enhances the performance by a staggering 3.5pp compared to SE-baseline, which is a significant increase.

On Places-LT, our APA* also increases the performance of the SE baseline by 0.8pp. Moreover, it surpasses MisLAS [111] by 0.9pp and CC-SAM [116] by 0.7pp. When AGLU is combined with APA*, it further increases the performance by 0.7pp reaching 42.0%.

As shown in Table 3-b, on CIFAR100-LT, APA* improves the performance by 0.5pp and 1.0pp compared to SE baseline using an imbalance factor of 10 and 100 respectively. Furthermore, compared to RIDE+CMO [68], CC-SAM [116] and RIDE+CMO+CR [60], APA* achieves 1.9pp, 1.1pp and 1.2pp higher accuracy respectively, under an imbalance factor of 100 using a single model. Finally AGLU, further boosts the performance of APA* by 1.1pp and 0.4pp for imbalance factor of 10 and 100 respectively.

Activation Ablation Study. We compare APA, without the Dropout and Layernorm, against different activation functions such as the Sigmoid and the Gum-

Table 4: For all ablations we use ImageNet-LT. In (a) we compare adaptive activation to other learnable activation functions using SE-ResNet50. In (b) we show that adaptive activation and AGLU generalise to other attention mechanisms. In (c) we show that AGCA and AGLU generalise in ImageNet1K training. In (d) we show that AGCA works effectively with deeper ResNets. In (e) we show the components of AGCA. In (f) we compare AGLU with other activation functions.

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	66.9	53.4	39.7	56.7
APA	**67.1**	**53.8**	39.6	**57.0**

(a) APA comparison.

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**

(b) Attention types w/ R50.

Method	SE	APA*
CE	51.7	**52.9**
PC-Softmax [34]	56.0	**57.4**
cRT [44]	55.6	**56.4**
Decoupled-DRW [8]	55.1	**56.6**
BSCE [71]	56.0	**57.0**

(c) Classifier learning.

Method	Many	Med.	Few	Avg
SE-R101	67.5	53.3	37.6	56.7
APA*-R101	**68.1**	**56.0**	**42.1**	**58.8**
SE-R152	68.0	54.4	39.8	57.6
APA*-R152	**69.0**	**56.8**	**41.2**	**59.4**

(d)Deeper Networks.

SE	APA	Dropout	LayerNorm	Avg
				55.0
✓				56.0
✓	✓			57.0
✓	✓	✓		57.3
✓	✓	✓	✓	**57.4**

(e) APA* ablation.

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**

(f) AGLU comparison.

bel [1] using the SE-ResNet50 baseline. Also, we implement Sigmoid and Gumbel variants with learnable temperature, to further understand their difference to our adaptive activation. As shown in Table 4-a, Sigmoid with temperature achieves slightly better performance for the rare classes, however it over-fits the frequent categories. In contrast, our APA achieves the best performance, increasing the overall performance by 0.3pp, the frequent classes by 0.2pp and the common categories by 0.4pp compared to the second best Gumbel with temperature.

Generalisation to other Attention mechanisms. In Table 4-b, we show that APA can be combined with other attention mechanisms. Specifically, APA improves the performance of Spatial Attention by 0.4pp and the Spatial-Channel Attention by 1.3pp. AGLU further increases the performance of Spatial Attention by 1.2pp and Spatial-Channel Attention by 0.2pp.

Combining APA* with Classifier Learning. APA* is an efficient module that can be easily combined with common classifier learning techniques such as margin adjustment [71], reweighting [8] and resampling [44]. As shown in Table 4-c, APA* consistently boosts the performance of all these methods.

Larger ResNets. We further train deeper models like ResNet-101 and ResNet-152 on ImageNet-LT and compare the SE and APA* attention methods. As Table 4-d shows, the APA* module enlarges the performance of all models consistently for both frequent, common and rare classes. Especially, APA* with ResNet101 promotes overall accuracy by 2.1pp, boosting the rare categories by a significant 4.5pp compared to SE-Resnet101.

Ablation of APA* components. APA* uses Channel attention, APA, Dropout and Layer-Norm and we show their effects in Table 4-e. The plain R50, without channel attention, achieves 55.0% and SE increases its performance by 1.0pp. APA increases the performance of the SE baseline by another 1.0pp. Dropout

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**$^{1.1}$

increases the performance by 0.3pp and LayerNorm enhances the performance by an additional 0.1pp. In the Appendix, we show the full component ablation.

Comparison of AGLU. In Table 4-f, we compare AGLU to other commonly used activations functions, using APA* ResNet50 as backbone and switching the activation function of the intermediate layers. Our AGLU outperforms all the other methods and it is the best choice.

Long-Tailed Instance Segmentation. As the results suggest in Table 5, the SE-R50 backbone increases the overall mask performance compared to plain GOL-R50 by 0.5pp but most improvement comes from the common and frequent categories while the rare categories are significantly reduced by 0.8pp. In contrast, our APA* with AGLU-R50 improves the performance by 0.9pp on average mask and bounding box precision, 1.0pp on AP_r, 0.7pp on AP_c and 0.9pp on AP_f compared to SE-R50-GOL. Compared to GOL with SE-R101, APA* with AGLU also improves the performance by 1.0pp on AP, 0.6pp on AP^r, 1.4pp on AP^c, 0.7pp on AP^f and 1.1pp on AP^b. This highlights that our APA* and AGLU modules are robust for the rare classes and they outperform the previous SOTA in long-tailed instance segmentation.

Qualitative Results. We use APA* and Imagenet-LT for our qualitative analysis. In Fig. 5-(a) we show that APA* increases the variance of channel attention for the most semantic layer by 0.04 compared to the baseline, making it diverse and informative for all channels. In Fig. 5-(b), we show that APA* increases the entropy of the attention signal for the deeper layers by $0.4 \sim 0.6$ compared to the baseline, showing that our module produces informative signals that effectively attend to the rare classes. In Fig.-c 5-top, we show that the baseline channel attention of the rare classes, produces all-pass filters in other words, attention

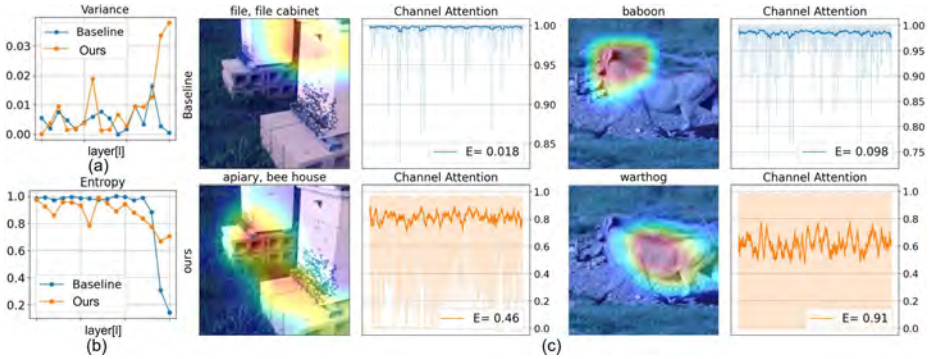

Fig. 5: a) APA* (orange curve) increases the attention variance by 0.04 in the most semantic layer compared to the baseline (blue curve) and removes frequent category attention bias. b) APA* increases the attention entropy in the most semantic layer and retrieves rare class descriptors more efficiently than the baseline. c) Compared to the baseline (top), APA* produces larger entropy attention signals and makes correct rare class predictions. More visualisations are shown in Appendix.

signals that have small entropy, i.e., $E = 0.018$ for the *bee house* and $E = 0.098$ for the *warthog* respectively. This hinders rare class learning and it results in misclassification. In contrast, our method in Fig. 5-bottom produces informative channel attention signals that have larger entropy i.e. $E = 0.46$ and $E = 0.91$ for the *bee house* and *warthog* classes respectively, allowing the model to retrieve rare class features and to make correct predictions. Visualisations of the learned κ and λ parameters are provided in the Appendix.

Generalisation to Balanced tasks. In Table 6-a, we show that APA*+AGLU increases the performance of SE-MaskRCNN-R50 by 0.7pp on AP^b and AP^m on COCO. In Table 6-b, we show that APA*+AGLU increase the performance of SE-FasterRCNN-R50 by 2.9pp and the performance of SE-Cascade-RCNN-R50 by 2.1pp on the challenging V3Det, which are significant increases considering that this dataset has 13K classes. In Table 6-c, we show that our methods increase the performance for R50, SE-R50 and CBAM-R50, by 0.6pp, 1.2pp and 0.6pp respectively. Finally in Table 6-d, we show that our modules increase the performance by 1.2pp on SE-R50, 0.9pp on SE-R101 and 0.5pp on SE-R101, showing that they can generalise in balanced training.

6 Related Work

Long-tailed image recognition. Long-tailed image recognition can be grouped according to representation learning and classifier learning techniques. Representation learning techniques improve the feature quality through rare class features generation [85, 87, 104], contrastive objectives [15, 43, 53, 74, 91, 117], ensemble or fusion models [7, 13, 14, 50, 51, 95, 113], knowledge distillation [29, 41, 51, 52], knowledge transfer [58, 67, 118] and data augmentation [68, 101, 111]. Classifier learning techniques enhance rare class classification through decoupled training [36, 44, 47, 92, 108], margin adjustment [2, 8, 34, 39, 63, 65, 71, 89, 93, 110], cost-

Table 6: Resuls on balanced tasks. In (a) and (b), we combine APA*+AGLU on COCO dataset and V3Det dataset respectively. In (c) we combine AGLU and APA* to ResNet50, SE-ResNet50 and CBAM-ResNet50 in ImageNet1K. In (d) we show that APA*+AGLU generalises to deeper backbones in ImageNet1K.

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

sensitive learning [16,38,46,96], resampling [9,22,33,61,68,76,104,120], dynamic loss adaptation based on batch statistics [35, 80, 94], gradient statistics [49, 79], weight norms [93] and classification scores [20, 30]. These techniques are efficient, however they are applied either during the input phase or during the loss disregarding the intermediate channel activations.

Attention Networks. Attention networks, such as spatial attention [42], channel attention [37] and spatial-channel attention [98], are widely used in image recognition. Self-attention is the core mechanism of the Visual Transformer (ViT) architecture [19], and recently many improvements have been made to the Transformer architecture [56, 81, 83] and its training procedure [6, 78, 82].

In our work, we primarily used APA with channel attention models, but it also works with plain ResNets, Spatial and Spatial-Channel attention models. In the Appendix, we show results with self-attention models.

Activation functions The RELU function [21] dominates the landscape of deep image classification, and it is especially used inside the convolutional networks, while the GELU and SiLU [31] are more commonly used inside the transformer network [19] or the ConNext models [57,97]. The PRELU function [27] is a generalisation of the RELU, because it linearly activates both positive and negative inputs and the ELU [11] is a follow-up work that enforces saturation on the negative inputs after some threshold. Our AGLU activation, generalises the RELU, the GELU and SiLU and it can also be seen as a smoother version of PRELU.

7 Conclusion

Our work highlights the impact of the activation function inside the model's activations for balanced and imbalanced data distributions. We have empirically showed that the degree of data imbalance affects the logit distributions and the intermediate signals and we have showed that the commonly used Sigmoid activation function is unable to model the intermediate features. To this end, we have proposed an novel adaptive parametric activation that unifies most common activation functions under the same formula and we have tested it in several long-tail and balanced classification and detection benchmarks showing great generalisation.

Acknowledgements. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) project "ViTac: Visual-Tactile Synergy for Handling Flexible Materials" (EP/T033517/2).

References

1. Alexandridis, K.P., Deng, J., Nguyen, A., Luo, S.: Long-tailed instance segmentation using gumbel optimized loss. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X. pp. 353–369. Springer (2022)
2. Alexandridis, K.P., Luo, S., Nguyen, A., Deng, J., Zafeiriou, S.: Inverse image frequency for long-tailed image recognition. IEEE Transactions on Image Processing (2023)
3. Alshammari, S., Wang, Y.X., Ramanan, D., Kong, S.: Long-tailed recognition via weight balancing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6897–6907 (2022)
4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
5. Bai, J., Yuan, L., Xia, S.T., Yan, S., Li, Z., Liu, W.: Improving vision transformers by revisiting high-frequency components. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV. pp. 1–18. Springer (2022)
6. Beyer, L., Zhai, X., Kolesnikov, A.: Better plain vit baselines for imagenet-1k. arXiv preprint arXiv:2205.01580 (2022)
7. Cai, J., Wang, Y., Hwang, J.N.: Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 112–121 (2021)
8. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems (2019)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
10. Chen, X., Zhou, Y., Wu, D., Yang, C., Li, B., Hu, Q., Wang, W.: Area: Adaptive reweighting via effective area for long-tailed classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19277–19287 (2023)
11. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015)
12. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 113–123 (2019)
13. Cui, J., Liu, S., Tian, Z., Zhong, Z., Jia, J.: Reslt: Residual learning for long-tailed recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
14. Cui, J., Liu, S., Tian, Z., Zhong, Z., Jia, J.: Reslt: Residual learning for long-tailed recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2022). https://doi.org/10.1109/TPAMI.2022.3174892

15. Cui, J., Zhong, Z., Liu, S., Yu, B., Jia, J.: Parametric contrastive learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 715–724 (2021)
16. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
17. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
18. Dong, Y., Cordonnier, J.B., Loukas, A.: Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In: International Conference on Machine Learning. pp. 2793–2803. PMLR (2021)
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
20. Feng, C., Zhong, Y., Huang, W.: Exploring classification equilibrium in long-tailed object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3417–3426 (2021)
21. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 315–323. JMLR Workshop and Conference Proceedings (2011)
22. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5356–5364 (2019)
23. Han, B.: Wrapped cauchy distributed angular softmax for long-tailed visual recognition. arXiv preprint arXiv:2305.18732 (2023)
24. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
25. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
26. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
27. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
29. He, Y.Y., Wu, J., Wei, X.S.: Distilling virtual examples for long-tailed recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 235–244 (2021)
30. He, Y.Y., Zhang, P., Wei, X.S., Zhang, X., Sun, J.: Relieving long-tailed instance segmentation via pairwise class balance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7000–7009 (2022)
31. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)

32. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
33. Hong, Y., Zhang, J., Sun, Z., Yan, K.: Safa: Sample-adaptive feature augmentation for long-tailed image classification. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV. pp. 587–603. Springer (2022)
34. Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., Chang, B.: Disentangling label distribution for long-tailed visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6626–6636 (2021)
35. Hsieh, T.I., Robb, E., Chen, H.T., Huang, J.B.: Droploss for long-tail instance segmentation. In: Proceedings of the AAAI conference on artificial intelligence. pp. 1549–1557 (2021)
36. Hsu, Y.C., Hong, C.Y., Lee, M.S., Geiger, D., Liu, T.L.: Abc-norm regularization for fine-grained and long-tailed image classification. IEEE Transactions on Image Processing (2023)
37. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
38. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5375–5384 (2016)
39. Hyun Cho, J., Krähenbühl, P.: Long-tail detection with effective class-margins. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII. pp. 698–714. Springer (2022)
40. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. pmlr (2015)
41. Iscen, A., Araujo, A., Gong, B., Schmid, C.: Class-balanced distillation for long-tailed visual recognition (2021)
42. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)
43. Kang, B., Li, Y., Xie, S., Yuan, Z., Feng, J.: Exploring balanced feature spaces for representation learning. In: International Conference on Learning Representations (2021)
44. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: Eighth International Conference on Learning Representations (ICLR) (2020)
45. Karpathy, A.: build-nanogpt. `https://github.com/karpathy/build-nanogpt` (2024)
46. Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R.: Cost-sensitive learning of deep feature representations from imbalanced data. IEEE transactions on neural networks and learning systems **29**(8), 3573–3587 (2017)
47. Kim, B., Kim, J.: Adjusting decision boundary for class imbalanced learning. IEEE Access **8**, 81674–81685 (2020)
48. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
49. Li, B., Yao, Y., Tan, J., Zhang, G., Yu, F., Lu, J., Luo, Y.: Equalized focal loss for dense long-tailed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6990–6999 (2022)

50. Li, B., Han, Z., Li, H., Fu, H., Zhang, C.: Trustworthy long-tailed classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6970–6979 (2022)
51. Li, J., Tan, Z., Wan, J., Lei, Z., Guo, G.: Nested collaborative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6949–6958 (2022)
52. Li, T., Wang, L., Wu, G.: Self supervision to distillation for long-tailed visual recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 630–639 (2021)
53. Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R.S., Indyk, P., Katabi, D.: Targeted supervised contrastive learning for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6918–6928 (2022)
54. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
55. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
56. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
57. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
58. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2537–2546 (2019)
59. Lozhkov, A., Ben Allal, L., von Werra, L., Wolf, T.: Fineweb-edu (May 2024). https://doi.org/10.57967/hf/2497, https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu
60. Ma, Y., Jiao, L., Liu, F., Yang, S., Liu, X., Li, L.: Curvature-balanced feature manifold learning for long-tailed classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15824–15835 (2023)
61. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European conference on computer vision (ECCV). pp. 181–196 (2018)
62. Massey Jr, F.J.: The kolmogorov-smirnov test for goodness of fit. Journal of the American statistical Association **46**(253), 68–78 (1951)
63. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=37nvvqkCo5
64. Misra, D.: Mish: A self regularized non-monotonic activation function. arXiv preprint arXiv:1908.08681 (2019)
65. Pan, T.Y., Zhang, C., Li, Y., Hu, H., Xuan, D., Changpinyo, S., Gong, B., Chao, W.L.: On model calibration for long-tailed object detection and instance segmentation. Advances in Neural Information Processing Systems **34** (2021)

66. Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences **117**(40), 24652–24663 (2020)
67. Parisot, S., Esperança, P.M., McDonagh, S., Madarasz, T.J., Yang, Y., Li, Z.: Long-tail recognition via compositional knowledge transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6939–6948 (2022)
68. Park, S., Hong, Y., Heo, B., Yun, S., Choi, J.Y.: The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6887–6896 (2022)
69. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
70. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
71. Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., Li, H.: Balanced meta-softmax for long-tailed visual recognition. In: Proceedings of Neural Information Processing Systems(NeurIPS) (Dec 2020)
72. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
73. Richards, F.J.: A flexible growth function for empirical use. Journal of experimental Botany **10**(2), 290–301 (1959)
74. Samuel, D., Chechik, G.: Distributional robustness loss for long-tail learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9495–9504 (2021)
75. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
76. Shen, L., Lin, Z., Huang, Q.: Relay backpropagation for effective learning of deep convolutional neural networks. In: European conference on computer vision. pp. 467–482. Springer (2016)
77. Skorski, M., Temperoni, A., Theobald, M.: Revisiting weight initialization of deep neural networks. In: Asian Conference on Machine Learning. pp. 1192–1207. PMLR (2021)
78. Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L.: How to train your vit? data, augmentation, and regularization in vision transformers. arXiv preprint arXiv:2106.10270 (2021)
79. Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q.: Equalization loss v2: A new gradient balance approach for long-tailed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1685–1694 (2021)
80. Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11662–11671 (2020)
81. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)

82. Touvron, H., Cord, M., Jégou, H.: Deit iii: Revenge of the vit. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV. pp. 516–533. Springer (2022)
83. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 32–42 (2021)
84. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
85. Vigneswaran, R., Law, M.T., Balasubramanian, V.N., Tapaswi, M.: Feature generation for long-tail classification. In: Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing. pp. 1–9 (2021)
86. Wang, H., Fu, S., He, X., Fang, H., Liu, Z., Hu, H.: Towards calibrated hypersphere representation via distribution overlap coefficient for long-tailed learning. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV. pp. 179–196. Springer (2022)
87. Wang, J., Lukasiewicz, T., Hu, X., Cai, J., Xu, Z.: Rsg: A simple but effective module for learning imbalanced datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3784–3793 (2021)
88. Wang, J., Zhang, P., Chu, T., Cao, Y., Zhou, Y., Wu, T., Wang, B., He, C., Lin, D.: V3det: Vast vocabulary visual detection dataset. arXiv preprint arXiv:2304.03752 (2023)
89. Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C.C., Lin, D.: Seesaw loss for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9695–9704 (2021)
90. Wang, P., Zheng, W., Chen, T., Wang, Z.: Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=O476oWmiNNp
91. Wang, P., Han, K., Wei, X.S., Zhang, L., Wang, L.: Contrastive learning based hybrid networks for long-tailed image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 943–952 (2021)
92. Wang, T., Li, Y., Kang, B., Li, J., Liew, J., Tang, S., Hoi, S., Feng, J.: The devil is in classification: A simple framework for long-tail instance segmentation. In: European Conference on computer vision. pp. 728–744. Springer (2020)
93. Wang, T., Zhu, Y., Chen, Y., Zhao, C., Yu, B., Wang, J., Tang, M.: C2am loss: Chasing a better decision boundary for long-tail object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6980–6989 (2022)
94. Wang, T., Zhu, Y., Zhao, C., Zeng, W., Wang, J., Tang, M.: Adaptive class suppression loss for long-tail object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3103–3112 (2021)
95. Wang, X., Lian, L., Miao, Z., Liu, Z., Yu, S.: Long-tailed recognition by routing diverse distribution-aware experts. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=D9I3drBz4UC
96. Wang, Y.X., Ramanan, D., Hebert, M.: Learning to model the tail. Advances in neural information processing systems **30** (2017)

97. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16133–16142 (2023)
98. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
99. Xie, L., Yang, Y., Cai, D., He, X.: Neural collapse inspired attraction–repulsion-balanced loss for imbalanced learning. Neurocomputing **527**, 60–70 (2023)
100. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
101. Xu, Z., Chai, Z., Yuan, C.: Towards calibrated model for long-tailed visual recognition from prior perspective. Advances in Neural Information Processing Systems **34**, 7139–7152 (2021)
102. Yang, Y., Chen, S., Li, X., Xie, L., Lin, Z., Tao, D.: Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? Advances in Neural Information Processing Systems **35**, 37991–38002 (2022)
103. Yang, Y., Yuan, H., Li, X., Lin, Z., Torr, P., Tao, D.: Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. arXiv preprint arXiv:2302.03004 (2023)
104. Zang, Y., Huang, C., Loy, C.C.: Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3457–3466 (2021)
105. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y.: Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830 (2019)
106. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
107. Zhang, S., Chen, C., Peng, S.: Reconciling object-level and global-level objectives for long-tail detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18982–18992 (2023)
108. Zhang, S., Li, Z., Yan, S., He, X., Sun, J.: Distribution alignment: A unified framework for long-tail visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2361–2370 (2021)
109. Zhang, Y., Wei, X.S., Zhou, B., Wu, J.: Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. Proceedings of the AAAI Conference on Artificial Intelligence **35**(4), 3447–3455 (May 2021). https://doi.org/10.1609/aaai.v35i4.16458, https://ojs.aaai.org/index.php/AAAI/article/view/16458
110. Zhao, Y., Chen, W., Tan, X., Huang, K., Zhu, J.: Adaptive logit adjustment loss for long-tailed visual recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3472–3480 (2022)
111. Zhong, Z., Cui, J., Liu, S., Jia, J.: Improving calibration for long-tailed recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16489–16498 (2021)
112. Zhong, Z., Cui, J., Yang, Y., Wu, X., Qi, X., Zhang, X., Jia, J.: Understanding imbalanced semantic segmentation through neural collapse. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19550–19560 (June 2023)

113. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9719–9728 (2020)
114. Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., Feng, J.: Deepvit: Towards deeper vision transformer. arXiv preprint arXiv:2103.11886 (2021)
115. Zhou, Y., Qu, Y., Xu, X., Shen, H.: Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11345–11355 (2023)
116. Zhou, Z., Li, L., Zhao, P., Heng, P.A., Gong, W.: Class-conditional sharpness-aware minimization for deep long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3499–3509 (2023)
117. Zhu, J., Wang, Z., Chen, J., Chen, Y.P.P., Jiang, Y.G.: Balanced contrastive learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6908–6917 (2022)
118. Zhu, L., Yang, Y.: Inflated episodic memory with region self-attention for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4344–4353 (2020)
119. Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., Qu, Q.: A geometric analysis of neural collapse with unconstrained features. Advances in Neural Information Processing Systems **34**, 29820–29834 (2021)
120. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European conference on computer vision (ECCV). pp. 289–305 (2018)

A Representations' Quality

We evaluate the quality of the representations learned by the SE and APA* models using the recently proposed Neural Collapse framework [66].

Let $f_{k,j} \in \mathrm{R}^{\mathrm{d}}$ be the features of the penultimate layer, $k = \{1, 2, ..., K\}$ the class, n_k the number of samples in the class k and $n = \sum_{k=1}^{K} n_k$ the total number of samples in the dataset. Then the global feature f_G and class prototype $\bar{f}_k$ are:

$$f_G = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} f_{k,j} \ , \bar{f}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} f_{k,j} \tag{10}$$

The within-class covariance matrix $\Sigma_W \in \mathrm{R}^{\mathrm{d} \times \mathrm{d}}$ and between-class covariance matrix $\Sigma_B \in \mathrm{R}^{\mathrm{d} \times \mathrm{d}}$ are:

$$\begin{aligned} \Sigma_W :=& \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} (f_{k,j} - \bar{f}_k)(f_{k,j} - \bar{f}_k)^\top \\ \Sigma_b :=& \frac{1}{K} \sum_{k=1}^{K} (\bar{f}_k - f_G)(\bar{f}_k - f_G)^\top \end{aligned} \tag{11}$$

The Σ_W matrix shows how distant are individual features $f_{k,j}$ from their class prototype $\bar{f}_k$ and it is an indicator of feature compactness. The Σ_b matrix shows how distant are the class prototypes from the global feature, indicating the class separability. Using these matrices we measure the Neural collapse Variability $NC1$ according to [119] as follows:

$$NC1 := \frac{1}{K} trace(\Sigma_W \Sigma_b^\dagger) \tag{12}$$

where the $\dagger$ symbol denotes the pseudo inverse of Σ_b. $NC1$ measures the magnitude of the within-class covariance Σ_W compared to the magnitude of the between-class covariance Σ_b as explained in [119].

In practise, a low $NC1$ measure shows that the model has more compact features since $\Sigma_W \downarrow$ decreases and more separable class prototypes because the $\Sigma_b \uparrow$ increases. Having more compact features and more separable class prototypes make the representations better and enhance the classification results as shown empirically in previous works [99, 102, 103, 112].

Using Equation 12, we measure the $NC1$ of the deep features of the penultimate layer of SE and APA*, in Table 7 using ImageNet-LT test-set. As the results suggest, our APA* has lower $NC1$ measure for all backbones, showing that APA* produces superior representations that are more compact and seperable than the baseline. This provides another qualitative explanation why our APA* has better performance than SE.

B Implementation Details

The implementation details of APA* and AGLU are shown in Table 8. For balanced ImageNet-1K, the λ parameters are initialised as random variables drawn

Table 7: Neural Collapse NC1 measure, on ImageNet-LT test set. APA* has lower $NC1$ measure than the baseline, which indicates that it has learned superior representations.

Backbone	SE-$NC1\downarrow$	APA*-$NC1\downarrow$
ResNet-50	3.04	**2.71**
ResNeXt-50	3.38	**2.55**
ResNet-101	3.15	**2.69**
ResNet-152	3.24	**2.69**

from a Uniform distribution (U), with low parameter 0, and high parameter 1.0. The APA κ parameters are initialised with $U(0,1)$ and the AGLU κ parameters are with initialised with $U(1,1.3)$. For all other downstream tasks, that use a pretrained model, such as COCO, LVIS, Places-LT and V3Det, we don't re-initialise the κ and λ parameters and we simply load them from the pretained ImageNet1K model.

B.1 Stable APA implementation

During the development of APA, we found that it is more stable to use Softplus $s_f(z,\beta) = \frac{1}{\beta}\ln(1+\exp(\beta z))$, than double exponents, when computing APA. Thus our stable code implementation is:

$$\eta_{ad}(z,\kappa,\lambda) = \exp(\frac{1}{\lambda}s_f(\kappa z - \ln(\lambda), -1)) \tag{13}$$

and it is equivalent to the APA used in the main paper.

B.2 AGLU derivatives

Proof of Eq. 9. Then the gradient of AGLU with respect to κ is:

$$\begin{aligned}\frac{\partial AGLU(x,\kappa,\lambda)}{\partial\kappa} &= \partial\frac{x\cdot(\lambda\exp(-\kappa x)+1)^{\frac{-1}{\lambda}}}{\partial\kappa}\\ &= x\frac{(\lambda\exp(-\kappa x)+1)^{(\frac{-1}{\lambda}-1)}}{-\lambda}\cdot(-\lambda x\exp(-\kappa x))\\ &= x^2\exp(-\kappa x)\frac{(\lambda\exp(-\kappa x)+1)^{(-\frac{-1}{\lambda})}}{\lambda\exp(-\kappa x)+1}\\ &= x^2\frac{\eta_{\text{ad}}(x,\lambda,\kappa)}{\lambda+\exp(\kappa x)}\end{aligned} \tag{14}$$

Method	ImageNet-LT R50/X50	i-Naturalist18 R50	Places-LT R152	C100-LT R32	LVISv1 MRCNN-R50
Batch size	256	1024	256	512	16
Optimiser	SGD	SGD	SGD	SGD	SGD
LR	0.2	0.5	0.1	0.2	0.02
epochs	200	500	40	500	24
Weight Decay	1e-4	1e-4	5e-5	1e-3	1e-4
Norm Weight Decay	1e-4	0.0	5e-5	1e-3	1e-4
Bias Weight Decay	0.0	0.0	0.0	0.0	0.0
Attention Dropout	0.1	0.0	0.1	0.1	0.1
Mixup α	0.2	0.2	0.2	0.2	-
CutMix α	-	1.0	1.0	-	-
Label smoothing ϵ	-	0.1	0.1	-	-
Repeated Aug	-	✓	-	-	-
AutoAugment	✓	-	✓	✓	-
RandAugment	-	✓	-	-	-
Erasing prob	-	0.1	-	-	-
Cutout	-	-	-	✓	-
Cos. Cls. scale	16	16	learnable	learnable	N/A
Norm. Mask scale	N/A	N/A	N/A	N/A	learnable
Sampler	random	random	random	random	RFS
APA κ Init	U(-1,0)	U(0,1)	N/A	U(-1,0)	N/A
APA λ Init	U(0,1)	U(0,1)	N/A	U(0,1)	N/A
AGLU κ Init	U(1,1.3)	U(1,1.3)	N/A	U(1,1.3)	N/A
AGLU λ Init	U(0,1)	U(0,1)	N/A	U(0,1)	N/A

Table 8: Implementation details for Long-tailed Datasets, across various architectures.

Proof of Eq. 10. Then the gradient of AGLU with respect to λ is:

$$\begin{aligned}\frac{\partial AGLU(x,\kappa,\lambda)}{\partial\lambda} &= \partial\frac{x\cdot(\lambda\exp(-\kappa x)+1)^{\frac{1}{-\lambda}}}{\partial\lambda}\\ &= x\frac{(\lambda\exp(-\kappa x)+1)^{(\frac{-1}{\lambda}-1)}}{-\lambda}\cdot(\exp(-\kappa x))\\ &= \frac{-x}{\lambda}\exp(-\kappa x)\frac{(\lambda\exp(-\kappa x)+1)^{(-\frac{1}{\lambda})}}{\lambda\exp(-\kappa x)+1}\\ &= \frac{-x}{\lambda}\frac{\eta_{\text{ad}}(x,\lambda,\kappa)}{\lambda+\exp(\kappa x)}\end{aligned} \tag{15}$$

Proof of Eq. 11. Then the gradient of AGLU with respect to λ is:

$$\begin{aligned}\frac{\partial AGLU(x,\kappa,\lambda)}{\partial x} &= \partial \frac{x \cdot (\lambda \exp(-\kappa x)+1)^{\frac{1}{-\lambda}}}{\partial x} \\ &= \eta_{\text{ad}}(x,\lambda,\kappa) + x \cdot \partial \frac{(\lambda \exp(-\kappa x)+1)^{\frac{1}{-\lambda}}}{\partial x} \\ &= \eta_{\text{ad}}(x,\lambda,\kappa) + x \frac{(\lambda \exp(-\kappa x)+1)^{(\frac{-1}{\lambda}-1)}}{-\lambda} \cdot (-\kappa\lambda \exp(-\kappa x)) \\ &= \eta_{\text{ad}}(x,\lambda,\kappa) + \kappa x \exp(-\kappa x) \frac{(\lambda \exp(-\kappa x)+1)^{(-\frac{1}{\lambda})}}{\lambda \exp(-\kappa x)+1} \\ &= \eta_{\text{ad}}(x,\lambda,\kappa) + \kappa x \frac{\eta_{\text{ad}}(x,\lambda,\kappa)}{\lambda + \exp(\kappa x)}\end{aligned} \tag{16}$$

C Results

C.1 Experiments with AGLU and plain ResNets

In Table 9, we show the result of AGLU when it applied inside plain ResNet50 models, (i.e. without channel attention). As the Table suggests, by simply replancing the RELU with AGLU, our method consistently increases the performance of plain ResNet models.

Table 9: Top-1 accuracy on long-tailed classification datasets using ResNets.

Dataset	CIFAR100-LT		ImageNet-LT		iNaturalist
Imbalance factor	10	100	256		500
Model	ResNet-32		ResNet50	ResNeXt50	ResNet50
RELU	65.7	51.8	55.0	57.0	69.9
AGLU (ours)	**66.8**	**52.0**	**56.0**	**57.6**	**72.4**

C.2 Experiments with AGLU and Vision Transformers on ImageNet1K

We perform a preliminary experiment with Vision Transformer models such as ViT [19] and Swin [56] using ImageNet1K. We replace the GELU activation with AGLU in every feedforward layer and we keep all other settings the same. As shown in Table 10, AGLU performs comparably to GELU. We believe this is because the Self-Attention function makes the features smooth, by removing their harmonising components and it makes them more Gaussian-like [5, 18, 90, 114]. Consequently, the Gaussial linear error unit, GELU, might be a good choice for the ViT network and our AGLU method has comparable performance.

Table 10: Results of AGLU using ViT models on ImageNet1K.

Model	Activation	epochs	top-1
ViT-B [19]	GELU	200	78.3
ViT-B [19]	AGLU	200	**78.5**
Swin-T [56]	GELU	100	78.7
Swin-T [56]	AGLU	100	**78.9**

C.3 Impact of Initialisation

In all of our experiments, we have initialised λ using the Uniform distribution with low parameter 0 and high parameter 1 as a default. Regarding the κ parameter inside AGLU, we initialise it to be close to 1.0, as this works best, as shown in Table 11. Regarding the κ parameter inside the attention layer, we found that initialising it with $Uniform(-1, 0)$ is slightly better than $Uniform(0, 1.0)$ as shown in Table 12.

Table 11: AGLU-κ parameter initialisation, using APA* ResNet50 backbone on ImageNet-LT. The λ is initialised with $Uniform(0, 1)$ by default.

AGLU - κ	top-1
$Uniform(0, 1)$	57.7
$Uniform(-2, 0)$	57.3
$Uniform(-3, 0)$	57.4
$Uniform(-2, 2)$	Failed
$Uniform(1, 1.3)$	**57.9**

Table 12: κ parameter initialisation inside the attention layer, using APA* ResNet50 backbone on ImageNet-LT. The λ is initialised with $Uniform(0, 1)$ by default.

APA - κ	top-1
$Uniform(0, 1)$	57.6
$Uniform(-1, 0)$	**57.9**

C.4 Channel specific $\boldsymbol{\lambda}$ and $\boldsymbol{\kappa}$

We have also tried a variant that uses seperate λ and κ parameters for every channel. As shown in Table 13, this variant performs worse than using shared λ and κ parameters for the channels.

C.5 Baseline enhancements

We show the detailed ablation study for ImageNet-LT in Table 14. First, the vanilla ResNet50 model trained for 100 epochs on ImageNet-LT achieves 44.4%.

Table 13: Results with Channel Specific λ and κ, using APA* ResNet50 backbone on ImageNet-LT.

APA	top-1
Channel Specific	57.5
Channel Shared	**57.9**

Table 14: Detailed Ablation Study, using ResNet50 on ImageNet-LT.

200 epochs	Cosine Classifier	SE-nets [37]	AutoAugment [12]	Mixup [106]	Weight Decay Tuning [3]	PCS [34]	APA (ours)	Dropout [32]	LayerNorm [4]	AGLU (ours)	ImagetNet-LT
											44.4
✓											45.9
✓	✓										46.3
✓	✓	✓									46.8
✓			✓								45.2
✓		✓	✓								45.9
✓	✓	✓	✓								46.6
✓	✓	✓	✓	✓							46.6
✓	✓		✓	✓	✓						49.6
✓	✓		✓	✓	✓	✓					55.0
✓	✓	✓	✓	✓	✓						51.7
✓	✓	✓	✓	✓	✓	✓					56.0
✓	✓	✓	✓	✓	✓	✓	✓				57.0
✓	✓	✓	✓	✓	✓	✓	✓	✓			57.3
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		57.4
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**57.9**

When we train for 200 epochs then it adds 1.5pp and switching from linear classifier to cosine classifier adds another 0.4pp. Stronger training techniques such Mixup [106], Auto-Augment [12] and weight decay tuning further boost the performance by 3.3pp. Post-calibrated Softmax [34] adds an additional 5.4pp and finally the Squeeze and Excite module [37] adds another 1.0pp reaching the final 56.0%. Most baseline performance comes from the PC-Softmax and the weight decay finetuning. On top of this strong baseline, our APA increases the performance by 1.0pp, showing its strong generalisability. Dropout and LayerNorm further increase the performance by 0.4pp and finally AGLU adds a respectable 0.5pp reaching 57.9% accuracy on ImageNet-LT. The absolute improvement of all modules is 13.5pp and our proposed methods, APA and AGLU, contribute by 1.5pp which is a relative 11% of the total absolute improvement.

C.6 Qualitative Results

In Figure 6, we show the learned parameters, when training with the balanced and imbalanced ImageNet. Regarding the λ inside the AGLU layers in (a), we see that both balanced-trained and imbalanced-trained networks prefer an all-pass filter for the early 2-3 layers, possibly, because the networks are uncertain which features to remove. Then in the intermediate layers, we observe smaller λ that corresponds to harder filters and in the final semantic layers we observe larger λ, possibly, because the network prefers smoother semantic features in order to have smoother classification boundaries. In (b), we see a 'down-down-up' κ-pattern in most bottlenecks, for both balanced and imbalanced ImageNet, showing that the networks prefer softer activations, at first, and harder activations before the residual connection. This indicates that the networks, first, keep most information inside the bottleneck's projections, and second, they disregard any redundant information, using harder activation, only before performing addition using the residual connection.

Finally, in the last bottlenecks, i.e. layers 45-50, the κ parameter diminishes, showing that the network prefers overly smooth activations, possibly, to enhance the classification using smoother classification boundaries.

Regarding the attention layers, the κ parameter in (d) dominates over the influence of λ in (c), showing that hard channel attention is more preferable than soft channel attention for all layers.

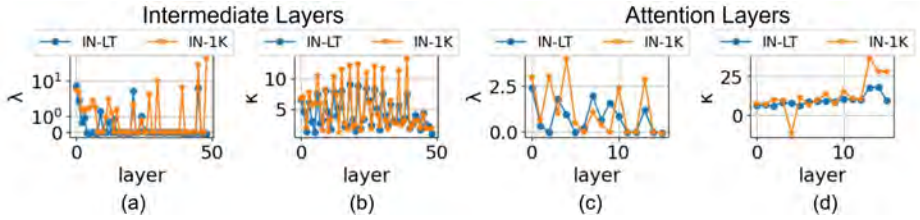

Fig. 6: Visualisations of the learned λ and κ parameters for balanced ImageNet1K (IN-IK) training in blue, and imbalanced ImageNet-LT (IN-LT) training in blue.

Visualisations on Imagenet-LT We further show more qualitative results on ImageNet-LT with ResNet50 backbone. On the left subfigure, we show the model's highest prediction marked with F,C,R that stands for frequent, common and rare class respectively and the Grad-cam activation [75]. On the right subfigure, we show the last layer's channel attention signal and its corresponding entropy denoted with (E). As the Figure shows, APA* produces higher entropy attention signals than the baseline and predicts both frequent and rare classes correctly.

Fig. 7: Comparative Results between the SE-ResNet50 (baseline) and APA*-ResNet50 (ours) with respect to the activations (left) and the attention entropy (right). F,C,R denote frequent, common and rare samples from ImageNet-LT. Our method produces attention signals that have significantly larger entropy than the baseline for both frequent and rare classes.

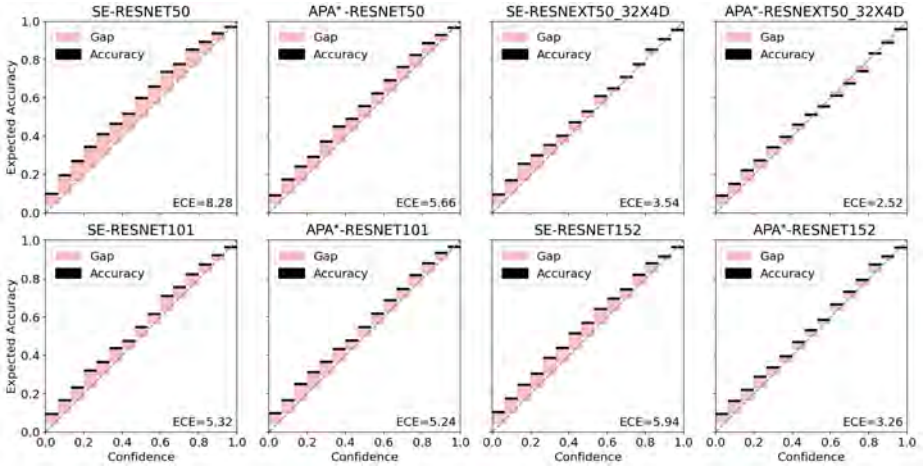

Fig. 8: Calibration results using ResNets on ImageNet-LT. SE (left) is underconfident, i.e., its confidence scores are lower than its actually accuracy due to over-regularisation. Our APA* (right) reduces the ECE and makes more accurate predictions with higher confidence than SE.

C.7 Calibration results

Calibration is an important property of models, since it reassures that the confidence of the prediction matches the actual accuracy. When models are not calibrated, then they give wrong predictions with high confidence score (over-confident models) or make correct predictions with low confidence score (under-confident models). In both situations, the miscalibrated models cannot help in the decision making process because their predictions do not reflect their actual accuracy.

In practice in long-tailed learning, the use of complex augmentations and regularisations like mixup, cutmix, label-smoothing, auto-augment and cosine classifier may improve the accuracy but it also reduces the confidence of the model due to over regularisation. As shown in Figure 8 (left-subfigure), SE-Resnet50 is under-confident due to the usage of complex training that includes heavy augmentations and regularisations. When APA* is applied, it reduces the Expected Calibration Error (ECE) as shown in Figure 8 (right-subfigure) for all backbones.

Table 15: Comparative results using GPT2 smallest model and HellaSwag benchmark.

Method	Accuracy
GELU	31.0
AGLU	**31.4**

C.8 Next textual token prediction experiment

We perform one preliminary next-token prediction experiment using GPT2 [70] and the FineWeb-Edu [59] subset that contains 10 billion GPT2 tokens. The model is based on the GPT2 smallest architecture, which contains 117M parameters, and the code implementation follows [45]. We train the GPT2 model for one epoch, using 8 V100 GPUs, a total batch size of 0.5M tokens, learning rate $6e-4$ and Adam optimizer [48] with momentum. We test the model on the HellaSwag benchmark [105] using zero-shot evaluation. To apply AGLU with GPT2, we simply switch the GELU activation with AGLU inside all MLP layers of the transformer. As the results suggest in Table 15, our AGLU increases the performance of GPT2 by 0.4%, showing that AGLU could be a good alternative for text-classification.