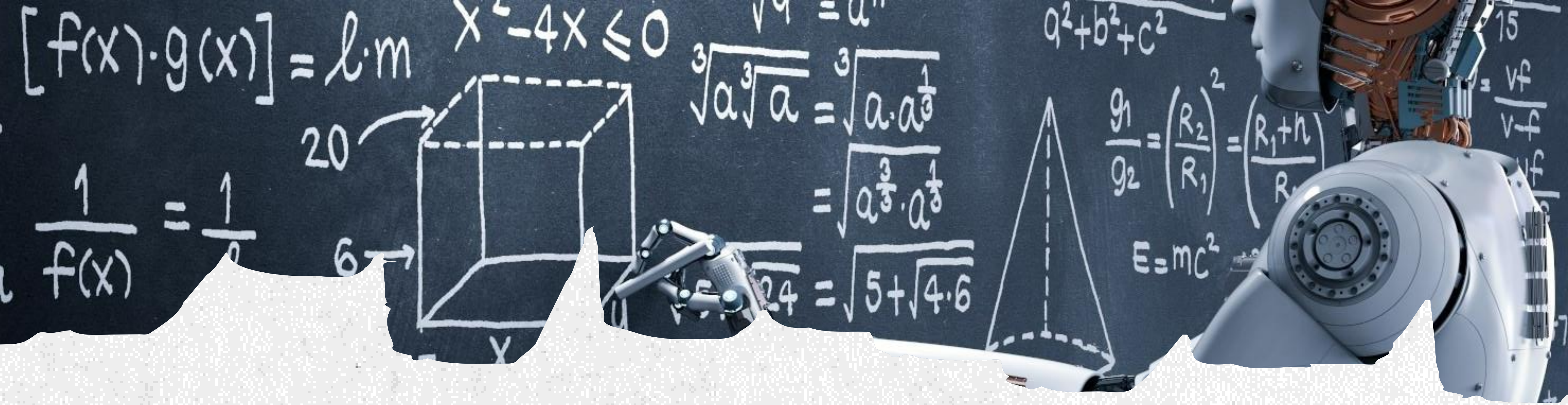


"AI FOR IMAGE" READING GROUP

MOUATH AOUAYEB

31/03/2022

- **Is ViT a PhD Student ?**
  School vs Deep Learning !

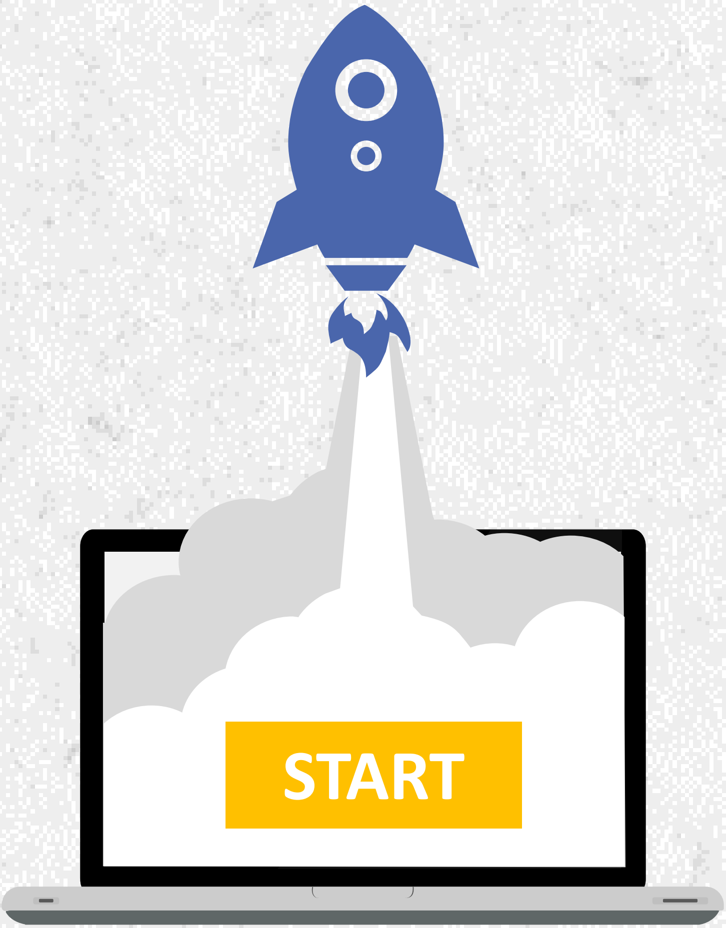
- **Efficient Training of ViT on small DBs : update DL techniques**

What's on the menu Today ?

VAADER
IETR research team

School vs DL

Analogy: School, DL
Common points: ViT, PhD



DL Techniques

ViTAE
DRLOC Loss
SAM Optimizer
AugMix

MY LITTLE BOOK CLUB

Vision Transformer

Student

DATA

Teachers

Internships

Deep Learning

Loss

AI

Courses

Administration

Optimizer

Exams

TD

Fine-tuning

Supervised

Deep Learning

School vs DL

VAADER IETR research team

Student
- Engineering Student
- PhD Student

Teacher
- "Better than a thousand days of diligent study is one day with a great teacher" Japanese Proverb

Lessons
- "Lessons in life will be repeated until they are learned" Frank Sonnenberg

Administration
- "Bad admin, to be sure, can destroy good policy; but good admin. can never save bad policy" Adlai Stevenson

Deep Learning Model
- CNN
- Transformer

Loss
- "Better than a thousand epochs of training is few epochs with a great Loss"

DATA
- "DATA in batch will be repeated until they are learned"

Optimizer
- "Bad optimizer, to be sure, can destroy good Training; but good optimizer can never save bad Training"

School vs DL

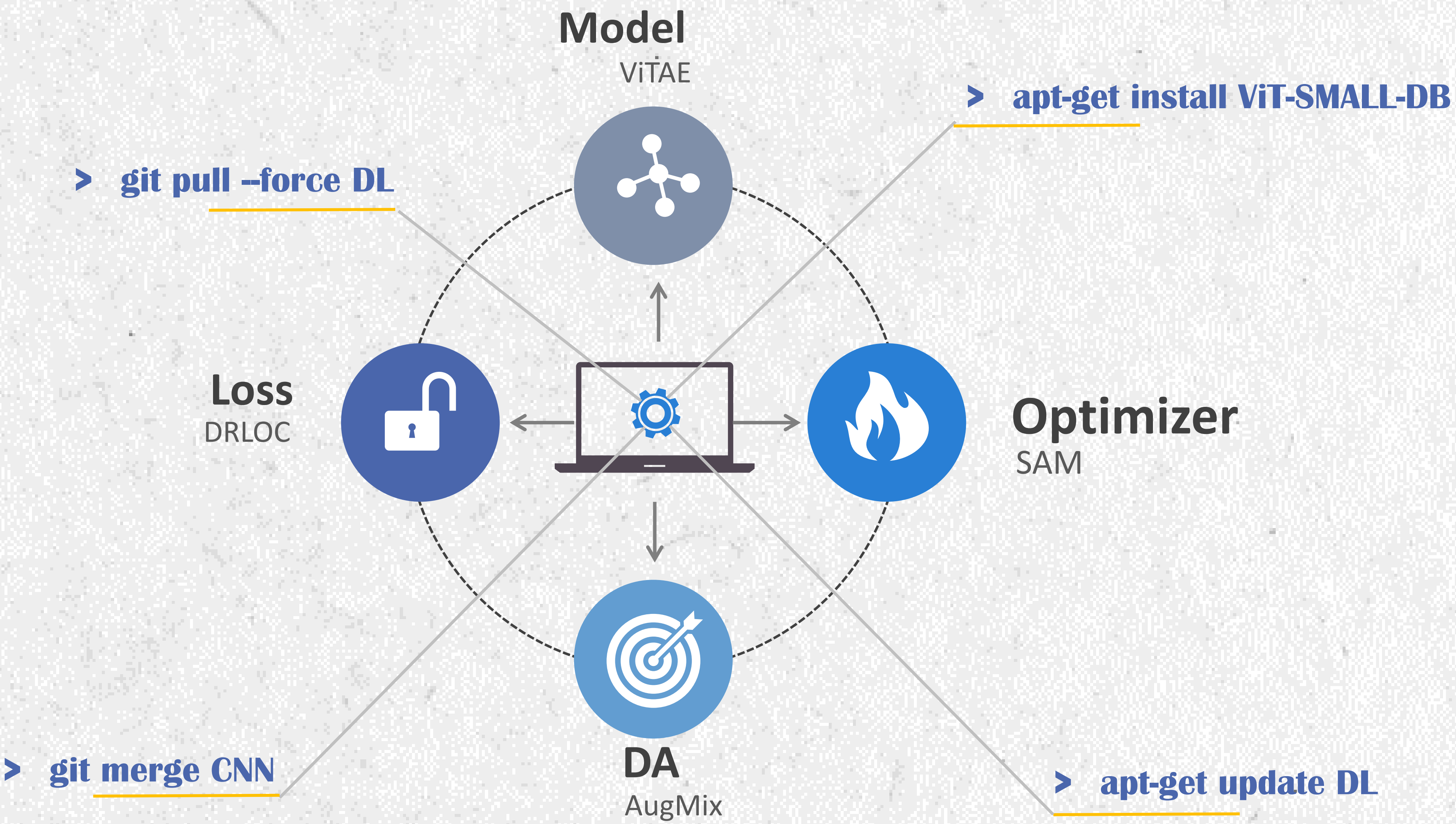
VAADER
IETR research team

Epochs	Years of study
Exams	Supervised Learning
Projects	Unsupervised Learning
Internships	Fine-tuning
Student Community life	Noisy Student Training
Gourp work	GPU Parallelization training
TD: ok , test: not ok	Overfitting

Is ViT a PhD Student ?

A PhD Student has good academic results, has a good spirit of critisim with innovative ideas and reads more corses and papers.

A ViT Model has good performances, has a good robustness with generalisation ability and needs more data for efficient training.


Model
ViTAE
> apt-get install ViT-SMALL-DB
> git pull --force DL
Loss
DRLOC
Optimizer
SAM
> git merge CNN
DA
AugMix
> apt-get update DL

ViTAE Model

ViTAE: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias [1]

ViTAEv2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and Beyond [2]

2020

Lack of Intrinsic Inductive Bias in modeling local visual structures and dealing with scale variance.

CNNs computes local correlation among neighbor pixel and use hierarchly structure to extract multi-scale features.

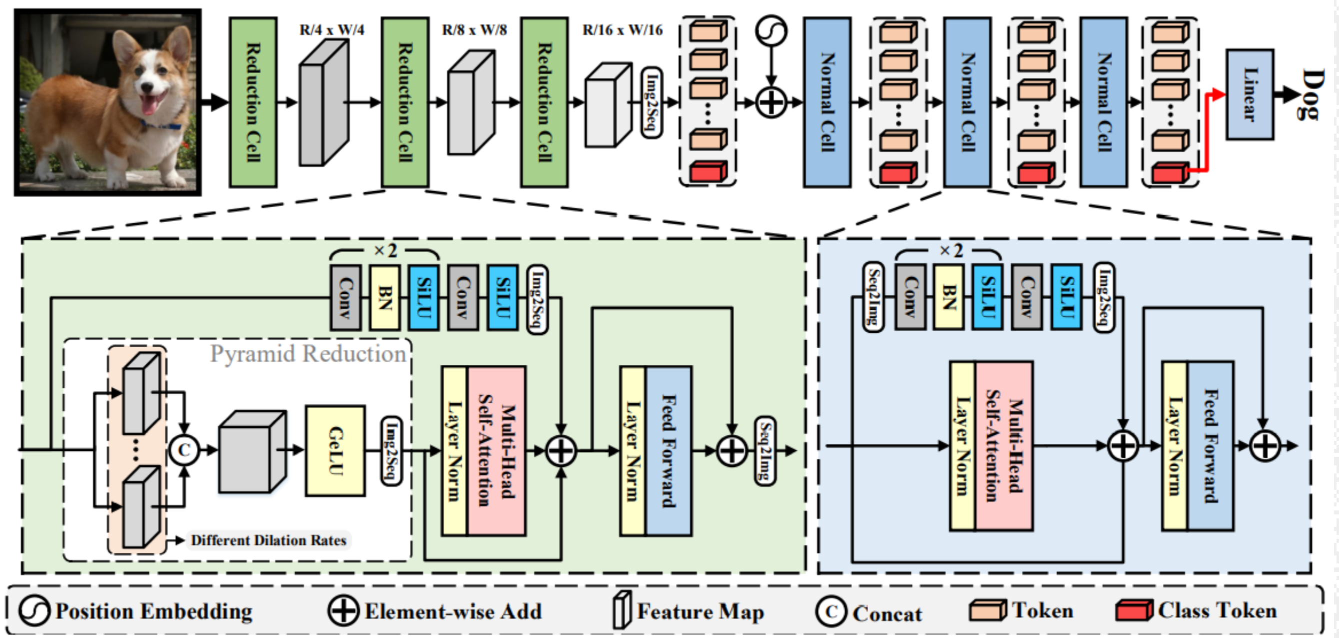
[1] Xu, Y., Zhang, Q., Zhang, J., Tao, D., 2021. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. CoRR abs/2106.03348
[2] Zhang, Q., Xu, Y., Zhang, J., Tao, D., 2022. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond.

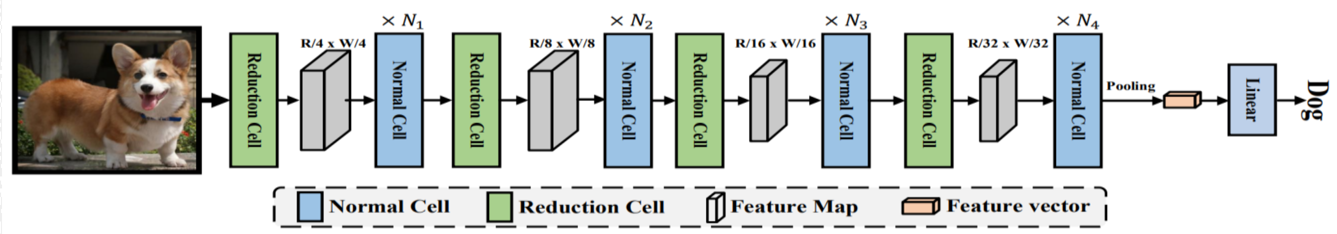
Code: https://github.com/Annbless/ViTAE

ViTAE Model

Multi-scale context:
Hierarchy architecture of spatial token representation using Pyramid Reduction Module (PRM):
Convolution layer to reduce spatial dimension.

Locality context: Parallel Convolution Module (PCM)





[1] Xu, Y., Zhang, Q., Zhang, J., Tao, D., 2021. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. CoRR abs/2106.03348
[2] Zhang, Q., Xu, Y., Zhang, J., Tao, D., 2022. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond.

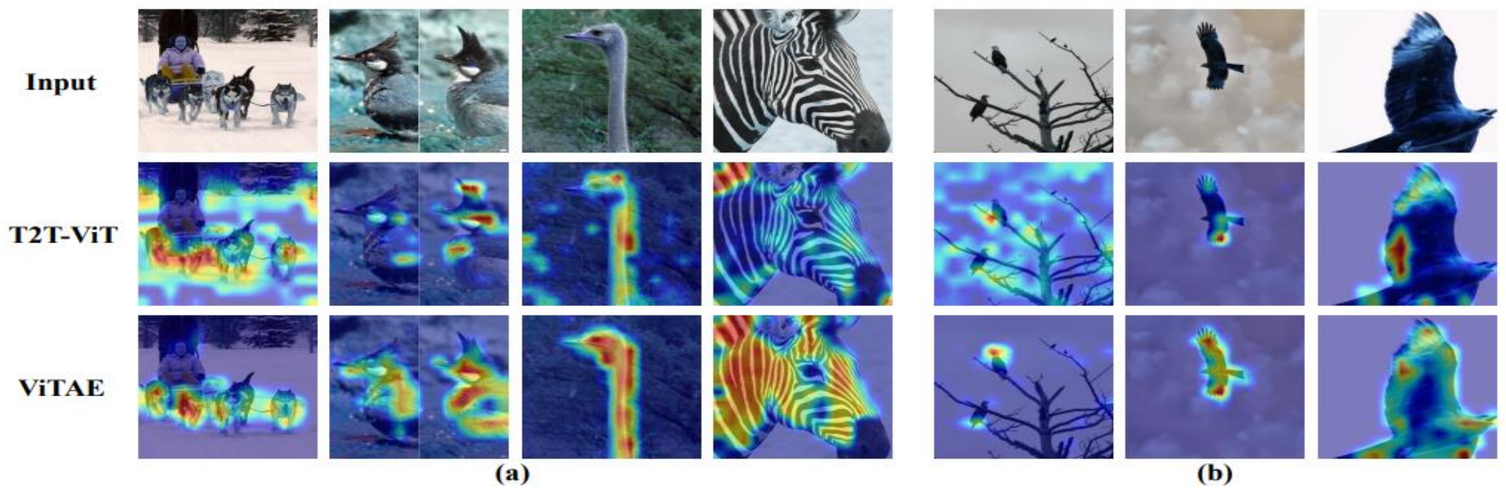
Code: https://github.com/Annbless/ViTAE

ViTAE Model

By scaling up the ViTAE to 644M parameters, they optain the state-of-the-art classification performance, i.e., 88.5% Top-1 classification accuracy on ImageNet validation set and the best 91.2% Top-1 classification accuracy on ImageNet real validation set, without using extra private data.

Table 4 Generalization of ViTAE and SOTA methods on different downstream image classification tasks.

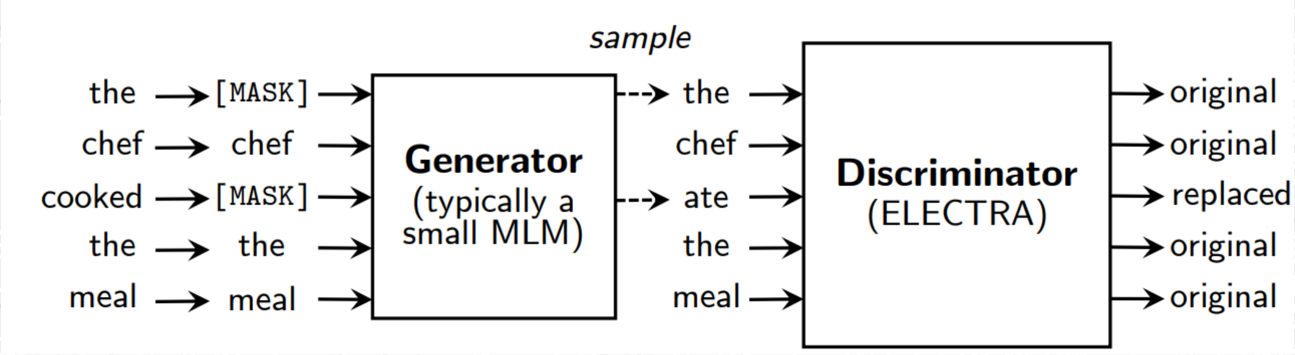
Model	Params (M)	Cifar10	Cifar100	iNat19	Cars	Flowers	Pets
Grafit ResNet-50 [73]	25.6	-	-	75.9	92.5	98.2	-
EfficientNet-B5 [70]	30	98.1	91.1	-	-	98.5	-
ViT-B/16 [22]	86.5	98.1	87.1	-	-	89.5	93.8
ViT-L/16 [22]	304.3	97.9	86.4	-	-	89.7	93.6
DeiT-B [72]	86.6	99.1	90.8	77.7	92.1	98.4	-
T2T-ViT-14 [92]	21.5	98.3	88.4	-	-	-	-
ViTAE-T	4.8	97.3	86.0	73.3	89.5	97.5	92.6
ViTAE-S	23.6	98.8	90.8	76.0	91.4	97.8	94.2



[2] Wen Y., Zhang K., Li Z., Qiao Y. (2016) A Discriminative Feature Learning Approach for Deep Face Recognition. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016.
Code: https://github.com/KaiyangZhou/pytorch-center-loss

Loss DRLOC

Efficient Training of Visual Transformers with Small Datasets [3]



Motivation

NLP: ELECTRA [4]

[3] Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., Nadai, M.D., 2021.Efficient training of visual transformers with small-size datasets. CoRR abs/2106.03746.
[4] Clark, K., Luong, M.T., Le, Q.V., Manning, C.D., 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators, in: ICLR.

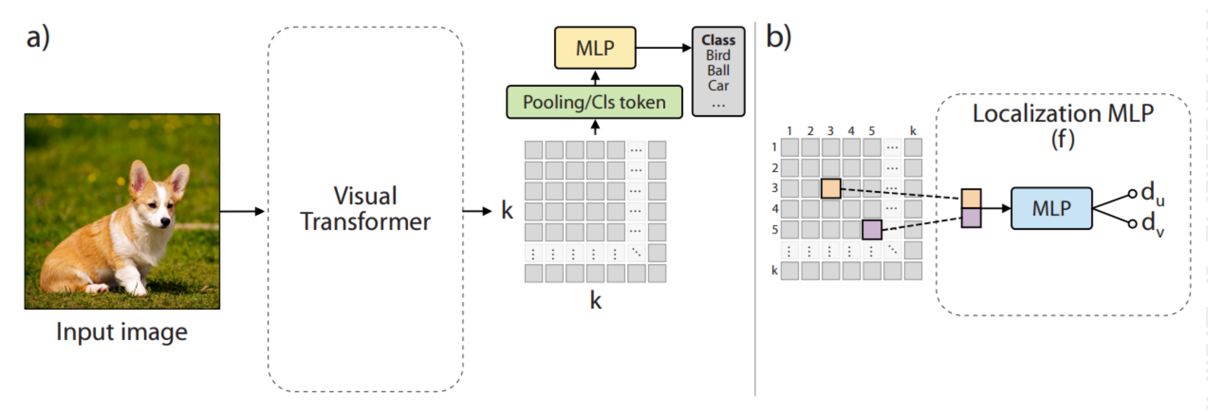
Code: https://github.com/yhlleo/VTs-Drloc

Loss DRLOC

Add an unsupervisd task to address the lack of locality inductive bias of ViT.

Add an unsupervised task based on the position of the tokens

Lack of convolutional inductive bias => data hungry than CNNs



[3] Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., Nadai, M.D., 2021.Efficient training of visual transformers with small-size datasets. CoRR abs/2106.03746.
[4] Clark, K., Luong, M.T., Le, Q.V., Manning, C.D., 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators, in: ICLR.

Code: https://github.com/yhlleo/VTs-Drloc

Loss DRLOC

Table 1: The size of the datasets used in our empirical analysis.

Dataset		Train size	Test size	Classes
ImageNet-1K [48]		1,281,167	100,000	1000
ImageNet-100 [52]		126,689	5,000	100
CIFAR-10 [31]		50,000	10,000	10
CIFAR-100 [31]		50,000	10,000	100
Oxford Flowers102 [41]		2,040	6,149	102
SVHN [40]		73,257	26,032	10
DomainNet	ClipArt	33,525	14,604	345
	Infograph	36,023	15,582	
	Painting	50,416	21,850	
	Quickdraw	120,750	51,750	
	Real	120,906	52,041	
	Sketch	48,212	20,916	

Table 4: Top-1 accuracy of VTs and ResNets, trained from scratch on different datasets (100 epochs).

		CIFAR-10	CIFAR-100	Flowers102	SVHN	ClipArt	Infograph	Painting	Quickdraw	Real	Sketch
CvT	CvT-13	89.02	73.50	54.29	91.47	60.34	19.39	54.79	70.10	76.33	56.98
	CvT-13+$\mathcal{L}_{drloc}$	**90.30**	**74.51**	**56.29**	**95.36**	**60.64**	**20.05**	**55.26**	**70.36**	**77.05**	**57.56**
		(+1.28)	(+1.01)	(+2.00)	(+3.89)	(+0.30)	(+0.67)	(+0.47)	(+0.26)	(+0.68)	(+0.58)
Swin	Swin-T	59.47	53.28	34.51	71.60	38.05	8.20	35.92	24.08	73.47	11.97
	Swin-T+$\mathcal{L}_{drloc}$	**83.89**	**66.23**	**39.37**	**94.23**	**47.47**	**10.16**	**41.86**	**69.41**	**75.59**	**38.55**
		(+24.42)	(+12.95)	(+4.86)	(+22.63)	(+9.42)	(+1.96)	(+5.94)	(+45.33)	(+2.12)	(+26.58)
T2T	T2T-ViT-14	84.19	65.16	31.73	95.36	43.55	6.89	34.24	69.83	73.93	31.51
	T2T-ViT-14+$\mathcal{L}_{drloc}$	**87.56**	**68.03**	**34.35**	**96.49**	**52.36**	**9.51**	**42.78**	**70.16**	**74.63**	**51.95**
		(+3.37)	(+2.87)	(+2.62)	(+1.13)	(+8.81)	(+2.62)	(+8.54)	(+0.33)	(+0.70)	(+20.44)
ResNet	ResNet-50	91.78	72.80	46.92	96.45	63.73	19.81	53.22	71.38	75.28	**60.08**
	ResNet-50+$\mathcal{L}_{drloc}$	**92.03**	**72.94**	**47.65**	**96.53**	**63.93**	**20.79**	**53.52**	**71.57**	**75.56**	59.62
		(+0.25)	(+0.14)	(+0.73)	(+0.08)	(+0.20)	(+0.98)	(+0.30)	(+0.19)	(+0.28)	(-0.46)

[3] Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., Nadai, M.D., 2021.Efficient training of visual transformers with small-size datasets. CoRR abs/2106.03746.
[4] Clark, K., Luong, M.T., Le, Q.V., Manning, C.D., 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators, in: ICLR.

Code: https://github.com/yhlleo/VTs-Drloc

Optimizer SAM

Sharpness-Aware Minimization for Efficiently Improving Generalization [5]
When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations [6]

Figure 1: (left) Error rate reduction obtained by switching to SAM. Each point is a different dataset / model / data augmentation. (middle) A sharp minimum to which a ResNet trained with SGD converged. (right) A wide minimum to which the same ResNet trained with SAM converged.

[5] Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B., 2020. Sharpness-aware minimization for efficiently improving generalization. CoRR abs/2010.01412.
[6] Chen, X., Hsieh, C., Gong, B., 2021. When vision transformers outperform resnets without pretraining or strong data augmentations. CoRR abs/2106.01548.

Code: https://github.com/google-research/sam

Optimizer SAM

$$S \triangleq \bigcup_{i=1}^{n} \{(x_i, y_i)\}$$ Training dataset from distibution D

$w \in W$ Model parameters

$l : W * X * Y \to R_+$ Loss function per data point

$$L_S \triangleq \frac{1}{n} \sum_{i=1}^{n} l(w, x_i, y_i)$$ Training Loss

$$L_D \triangleq E_{(x,y) \sim D}\,[l(w, x, y)]$$ Population Loss

The goal of model training is to select w having low population loss L_D(w), having observed only S.

- L_S is not convex in w (modern models) ,
- Low training loss for the neighbohrs of w and not only for w,

[5] Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B., 2020. Sharpness-aware minimization for efficiently improving generalization. CoRR abs/2010.01412.
[6] Chen, X., Hsieh, C., Gong, B., 2021. When vision transformers outperform resnets without pretraining or strong data augmentations. CoRR abs/2106.01548.

Code: https://github.com/google-research/sam

Optimizer SAM

Theorem 1: For any $\boldsymbol{\rho} > \mathbf{0}$, with high probability over training set S generated from distribution D.

$$L_D(w) \leq \max_{||\epsilon||_2 \leq \rho} L_S(w+\epsilon) + h(||\epsilon||_2^2/\rho^2),$$

Where $\boldsymbol{h}$: $\boldsymbol{R}_+ \rightarrow \boldsymbol{R}_+$ is a strictly increasing function

$$\left[\max_{||\epsilon||_2 \leq \rho} L_S(w+\epsilon) - L_S(w)\right] + L_S(w) + h\left(||\epsilon||_2^2/\rho^2\right)$$

Sharpness + Training Loss + Regularizer

- h a standard L2 Regularization

$$\min_w L_S^{SAM}(w) + \lambda||w||_2^2 \text{ , where}$$

$$L_S^{SAM}(w) \triangleq \max_{||\epsilon||_2 \leq \rho} L_S(w+\epsilon)$$

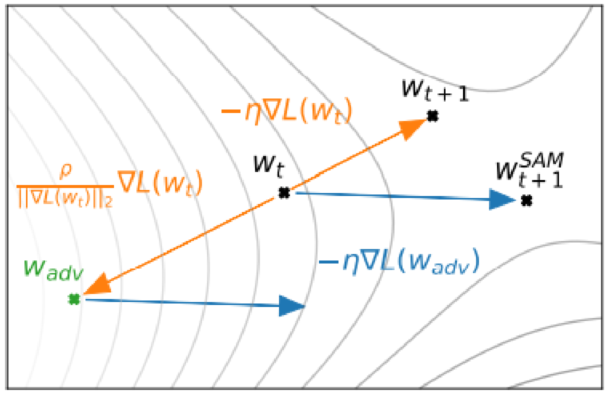
- To make a more efficient and effective more approximations have been used on the estimation of $\boldsymbol{\epsilon}$ and on the $\boldsymbol{\nabla_w L_S^{SAM}(w)}$.
- 2 forwards

[5] Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B., 2020. Sharpness-aware minimization for efficiently improving generalization. CoRR abs/2010.01412.
[6] Chen, X., Hsieh, C., Gong, B., 2021. When vision transformers outperform resnets without pretraining or strong data augmentations. CoRR abs/2106.01548.

Code: https://github.com/google-research/sam

Optimizer SAM

- Improves the ImageNet top-1 error-rate of ResNet-152 with 2%.
- More accuracy with more epochs without overfitting.
- Improvesperformance relative to finetuning
- More robust: with 80% Noise rate, 79,9% accuracy is obtained with SAM instead of the 26,2% accuracy with SGD.
- +7,6% improvement on accuracy with SGD+SAM vs SGD on IN
- ViT + SAM outperforms ResNet and ResNet + SAM on ImageNet



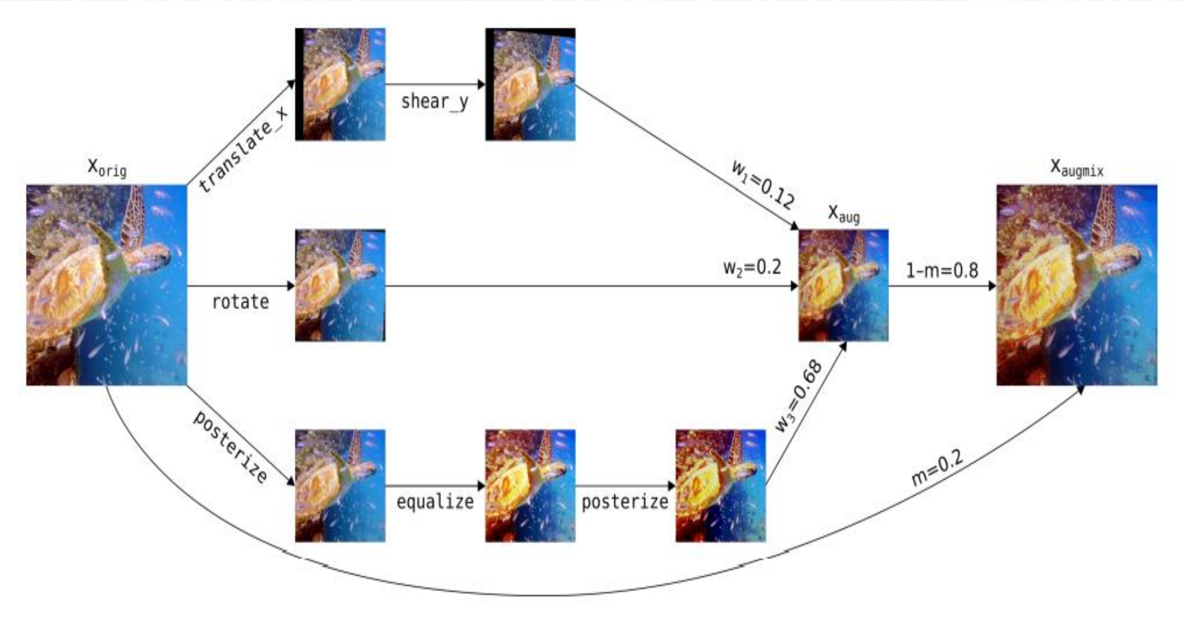
Model	#params	Throughput (img/sec/core)	ImageNet	ReaL	V2	ImageNet-R	ImageNet-C
ResNet							
ResNet-50-SAM	25M	2161	76.7 (+0.7)	83.1 (+0.7)	64.6 (+1.0)	23.3 (+1.1)	46.5 (+1.9)
ResNet-101-SAM	44M	1334	78.6 (+0.8)	84.8 (+0.9)	66.7 (+1.4)	25.9 (+1.5)	51.3 (+2.8)
ResNet-152-SAM	60M	935	79.3 (+0.8)	84.9 (+0.7)	67.3 (+1.0)	25.7 (+0.4)	52.2 (+2.2)
ResNet-50x2-SAM	98M	891	79.6 (+1.5)	85.3 (+1.6)	67.5 (+1.7)	26.0 (+2.9)	50.7 (+3.9)
ResNet-101x2-SAM	173M	519	80.9 (+2.4)	86.4 (+2.4)	69.1 (+2.8)	27.8 (+3.2)	54.0 (+4.7)
ResNet-152x2-SAM	236M	356	81.1 (+1.8)	86.4 (+1.9)	69.6 (+2.3)	28.1 (+2.8)	55.0 (+4.2)
Vision Transformer							
ViT-S/32-SAM	23M	6888	70.5 (+2.1)	77.5 (+2.3)	56.9 (+2.6)	21.4 (+2.4)	46.2 (+2.9)
ViT-S/16-SAM	22M	2043	78.1 (+3.7)	84.1 (+3.7)	65.6 (+3.9)	24.7 (+4.7)	53.0 (+6.5)
ViT-S/14-SAM	22M	1234	78.8 (+4.0)	84.8 (+4.5)	67.2 (+5.2)	24.4 (+4.7)	54.2 (+7.0)
ViT-S/8-SAM	22M	333	81.3 (+5.3)	86.7 (+5.5)	70.4 (+6.2)	25.3 (+6.1)	55.6 (+8.5)
ViT-B/32-SAM	88M	2805	73.6 (+4.1)	80.3 (+5.1)	60.0 (+4.7)	24.0 (+4.1)	50.7 (+6.7)
ViT-B/16-SAM	87M	863	79.9 (+5.3)	85.2 (+5.4)	67.5 (+6.2)	26.4 (+6.3)	56.5 (+9.9)

[5] Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B., 2020. Sharpness-aware minimization for efficiently improving generalization. CoRR abs/2010.01412.
[6] Chen, X., Hsieh, C., Gong, B., 2021. When vision transformers outperform resnets without pretraining or strong data augmentations. CoRR abs/2106.01548.

Code: https://github.com/google-research/sam

Data Aumentation

AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty [7]



couple with this augmentation scheme a loss that enforces smoother neural network responses. Since the semantic content of an image is approximately preserved with AUGMIX, we should like the model to embed xorig, xaugmix1, xaugmix2 similarly.

$$L(p_{orig,y}) + \lambda JS(p_{orig}; p_{augmix1}; p_{augmix2})$$

$$JS(p_{orig}; p_{augmix1}; p_{augmix2}) =$$

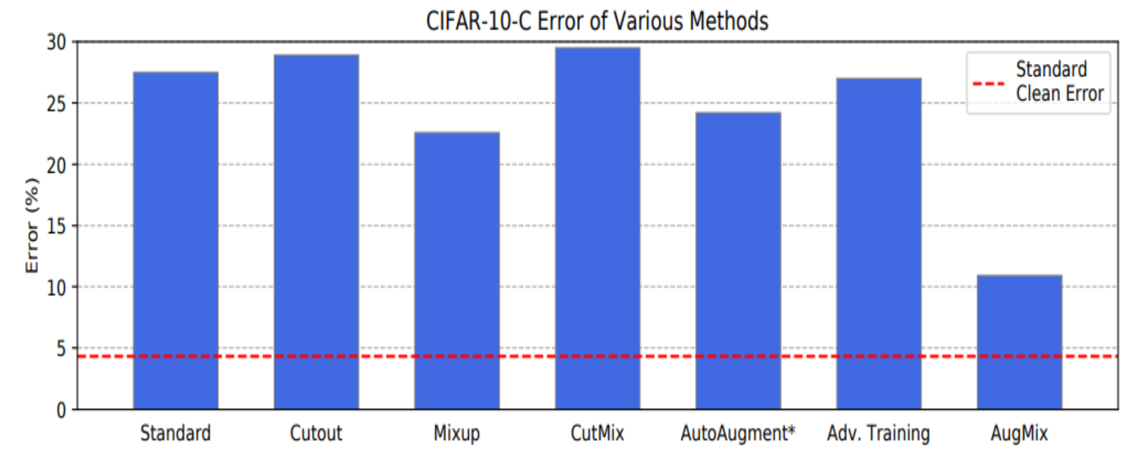
$$\frac{1}{3}\left(KL[p_{orig}||M] + KL[p_{augmix1}||M] + KL[p_{augmix1}||M]\right)$$

[7] Hendrycks*, D., Mu*, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshmi-narayanan, B., 2020. Augmix: A simple method to improve robustness and uncertainty under data shift, in: International Conference on Learning Representations.

Code: https://github.com/google-research/augmix

Data Aumentation

AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty [7]



		Standard	Cutout	Mixup	CutMix	AutoAugment*	Adv Training	AugMix
CIFAR-10-C	AllConvNet	30.8	32.9	24.6	31.3	29.2	28.1	**15.0**
	DenseNet	30.7	32.1	24.6	33.5	26.6	27.6	**12.7**
	WideResNet	26.9	26.8	22.3	27.1	23.9	26.2	**11.2**
	ResNeXt	27.5	28.9	22.6	29.5	24.2	27.0	**10.9**
	Mean	29.0	30.2	23.5	30.3	26.0	27.2	**12.5**
CIFAR-100-C	AllConvNet	56.4	56.8	53.4	56.0	55.1	56.0	**42.7**
	DenseNet	59.3	59.6	55.4	59.2	53.9	55.2	**39.6**
	WideResNet	53.3	53.5	50.4	52.9	49.6	55.1	**35.9**
	ResNeXt	53.4	54.6	51.4	54.1	51.3	54.4	**34.9**
	Mean	55.6	56.1	52.6	55.5	52.5	55.2	**38.3**

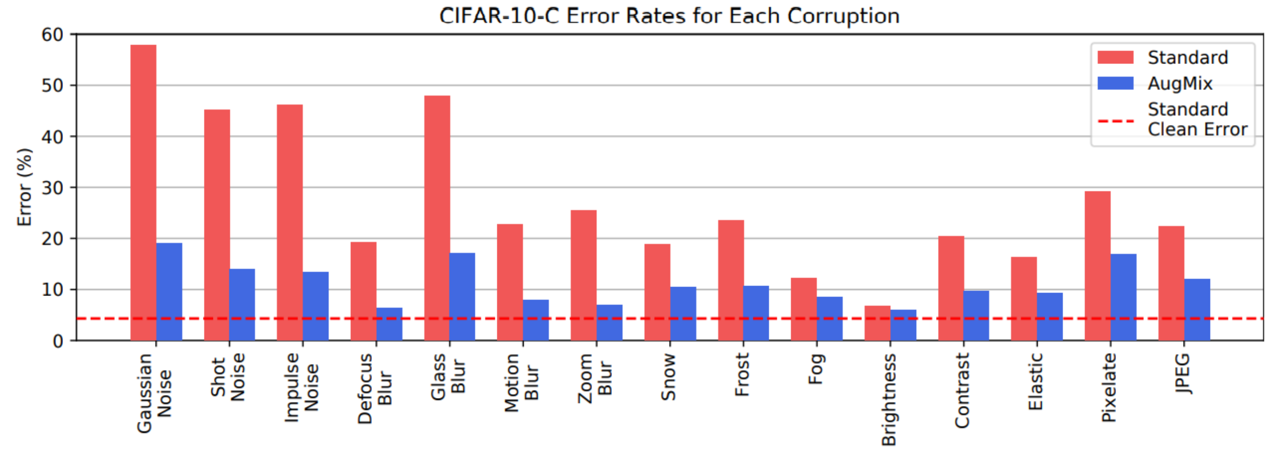
[7] Hendrycks*, D., Mu*, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshmi-narayanan, B., 2020. Augmix: A simple method to improve robustness and uncertainty under data shift, in: International Conference on Learning Representations.

Code: https://github.com/google-research/augmix

Data Aumentation

AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty [7]

Method	CIFAR-10-C Error Rate	CIFAR-100-C Error Rate
Standard	26.9	53.3
AutoAugment*	23.9	49.6
Random AutoAugment*	17.0	43.6
Random AutoAugment* + JSD Loss	14.7	40.8
AugmentAndMix (No JSD Loss)	13.1	39.8
AugMix (Mixing + JSD Loss)	11.2	35.9



[7] Hendrycks*, D., Mu*, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshmi-narayanan, B., 2020. Augmix: A simple method to improve robustness and uncertainty under data shift, in: International Conference on Learning Representations.

Code: https://github.com/google-research/augmix

THANK YOU

IS IT THE TIME TO LEARN FROM MACHINES HOW TO IMPROVE THE HUMAN EDUCATION SYSTEM?

MOUATH