TokenMix: Rethinking Image Mixing for Data Augmentation in Vision Transformers

Jihao Liu[1,2], Boxiao Liu[3], Hang Zhou[1], Hongsheng Li[1✉], and Yu Liu[2✉]

[1] CUHK, MMLab
[2] SenseTime Research
[3] SKLP, Institute of Computing Technology, CAS

ECCV 2022

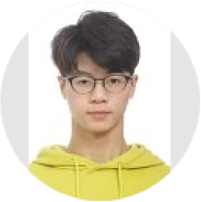
Paper Reading by Zhiying Lu

2022.09.05

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

作者介绍



Jihao Liu

FOLLOW

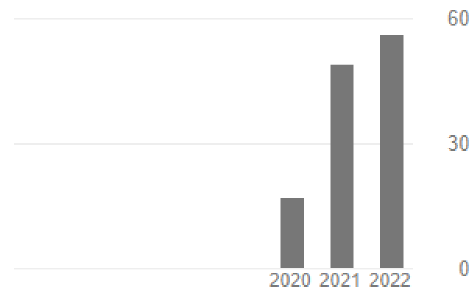
CUHK, MMLab
Verified email at link.cuhk.edu.hk
Computer Vision

TITLE	CITED BY	YEAR
Rotate-and-render: Unsupervised photorealistic face rotation from single-view images H Zhou, J Liu, Z Liu, Y Liu, X Wang Proceedings of the IEEE/CVF conference on computer vision and pattern ...	64	2020
Learning where to focus for efficient video object detection Z Jiang, Y Liu, C Yang, J Liu, P Gao, Q Zhang, S Xiang, C Pan European conference on computer vision, 18-34	28	2020
Towards flops-constrained face recognition Y Liu Proceedings of the IEEE/CVF International Conference on Computer Vision ...	11	2019
Intern: A new learning paradigm towards general vision J Shao, S Chen, Y Li, K Wang, Z Yin, Y He, J Teng, Q Sun, M Gao, J Liu, ... arXiv preprint arXiv:2111.08687	7	2021
Uninet: Unified architecture search with convolution, transformer, and mlp J Liu, X Huang, G Song, Y Liu, H Li arXiv preprint arXiv:2207.05420	4	2022

GET MY OWN PROFILE

Cited by

	All	Since 2017
Citations	122	122
h-index	4	4
i10-index	3	3



Public access VIEW ALL

0 articles 2 articles

作者介绍

Hongsheng Li (李鸿升)

关注

Associate Professor at The Chinese University of Hong Kong
在 ee.cuhk.edu.hk 的电子邮件经过验证 - 首页

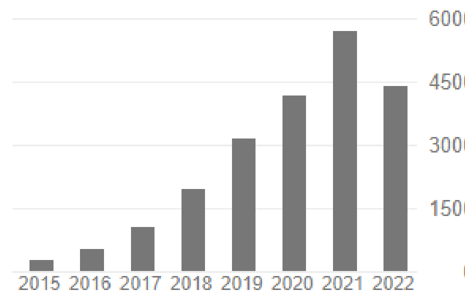
Computer Vision   Machine Learning   Medical Image Analysis

标题	引用次数	年份
StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks H Zhang, T Xu, H Li, S Zhang, X Huang, X Wang, D Metaxas IEEE Int. Conf. Comput. Vision (ICCV), 5907-5915	2342	2017
Pointrcnn: 3d object proposal generation and detection from point cloud S Shi, X Wang, H Li Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern ...	1147	2019
Cross-scene crowd counting via deep convolutional neural networks C Zhang, H Li, X Wang, X Yang Proceedings of the IEEE Conference on Computer Vision and Pattern ...	1119	2015
Learning deep feature representations with domain guided dropout for person re-identification T Xiao, H Li, W Ouyang, X Wang Proceedings of the IEEE Conference on Computer Vision and Pattern ...	1040	2016

创建我的个人资料

引用次数   查看全部

	总计	2017 年至今
引用	22019	20540
h 指数	71	65
i10 指数	135	125

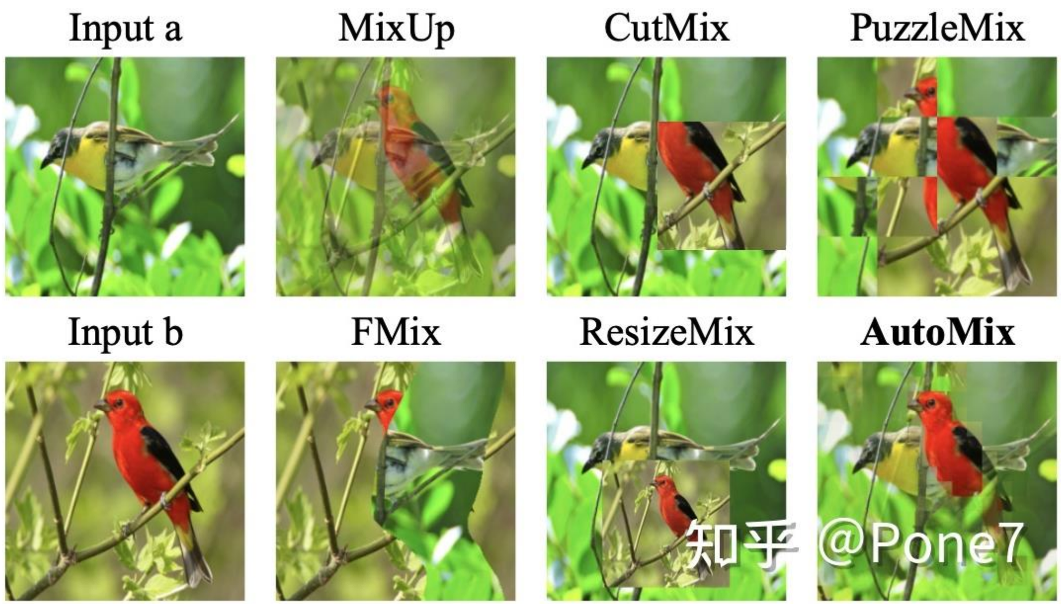


开放获取的出版物数量   查看全部

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

研究背景

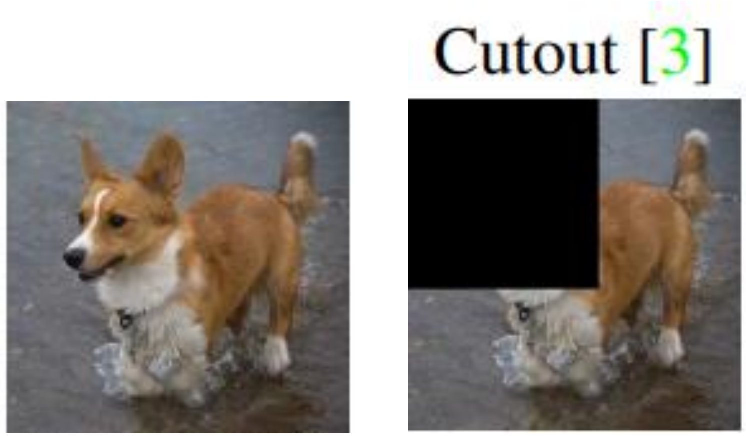
- Data augmentation方法可以分为三类:
- 1. Cutting-based: Cutout, Random-erasing, Hide-and-seek
- 2. Mixing-based: Mixup, Manifold Mixup, Co-mixup, PuzzleMix, AugMix
- 3. Joint of cutting and mixing: CutMix, Attentive CutMix, RICAP, ResizeMix



研究背景

Cutting-based

1. Cutout, Random-erasing: 直接选择一个矩形区域设定为常数值
2. Hide-and-Seek: 将图像划分成grid，随机mask掉一些grid



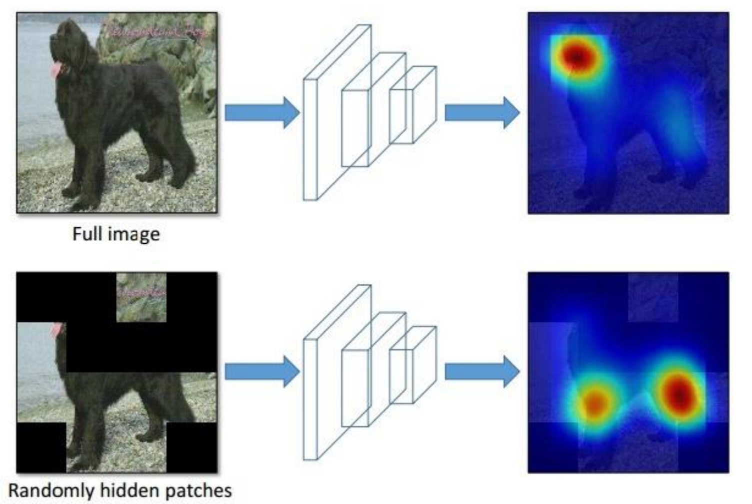
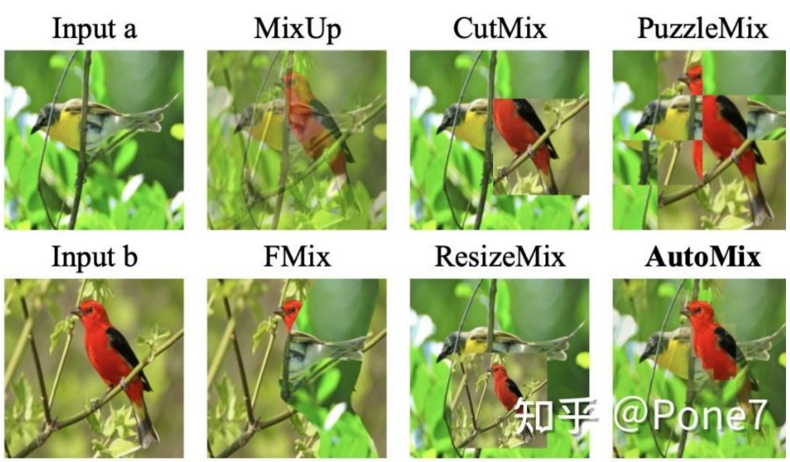


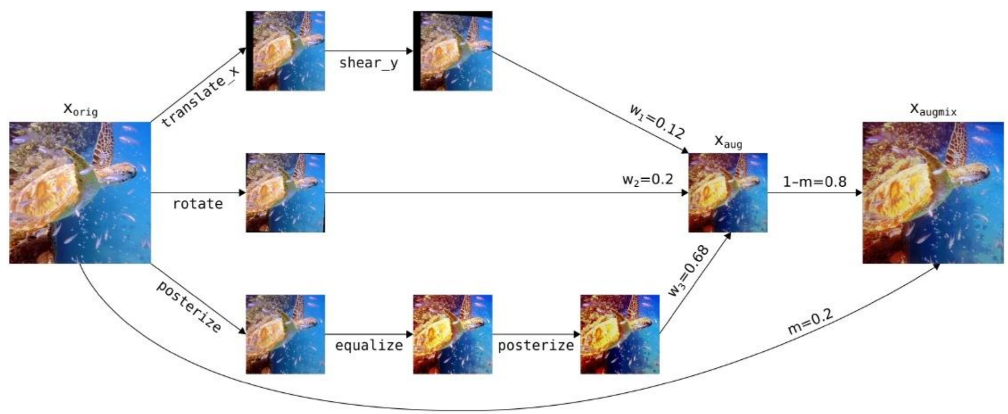
Figure 1. **Main idea.** (Top row) A network tends to focus on the most discriminative parts of an image (e.g., face of the dog) for classification. (Bottom row) By hiding images patches randomly, we can force the network to focus on other relevant object parts in order to correctly classify the image as 'dog'.

研究背景

Mixing-based

1. Mixup: 线性组合两张图像和对应label，组合系数取自beta分布
2. Manifold Mixup: 在feature maps中也进行mixup
3. Co-Mixup, PuzzleMix: 以解优化问题的方式mixup，最大化mixed图片中的saliency
4. Augmix: 从一张图原图和transformed version进行mixup





研究背景

Joint of cutting and mixing

1. CutMix: 随机裁剪拼接两张图的一部分，
2. Attentive CutMix: 根据pretrained网络预测的attentive区域来进行cutmix
3. ResizeMix: 将另一张图整张图进行cutmix

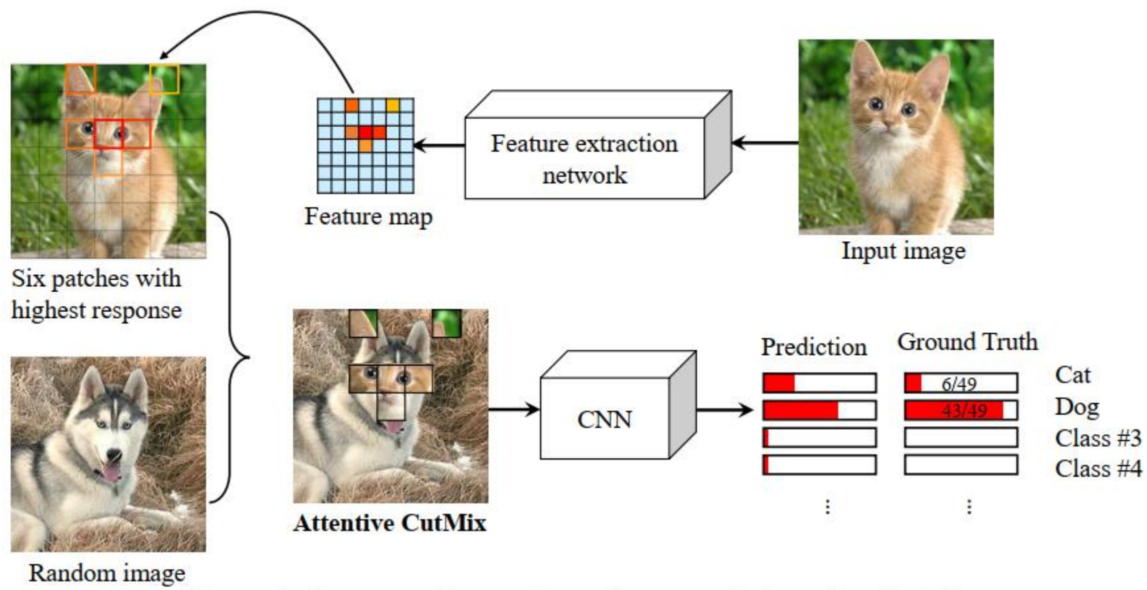
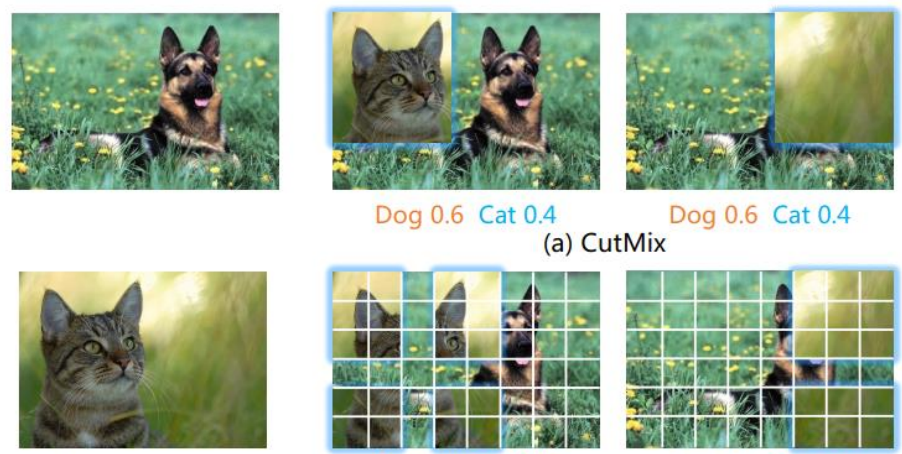


Figure 2: Framework overview of proposed *Attentive CutMix*.

研究背景

- CutMix数据增强方法在训练时可以帮助CNN更好地关注到全局的特征
- 但由于transformer系列方法本身就具有全局的信息捕捉能力，因此CutMix无法充分发挥优势
- 同时CutMix对target label的计算方式也只是线性组合，可能会导致label的不准确，导致网络学习到错误的label (例如当Mix的区域是背景，标签仍然是前景)

	ResNet-50	Mixup [47]	Cutout [3]	CutMix
Image				
Label	Dog 1.0	Dog 0.5 Cat 0.5	Dog 1.0	Dog 0.6 Cat 0.4
ImageNet Cls (%)	76.3 (+0.0)	77.4 (+1.1)	77.1 (+0.8)	**78.6** **(+2.3)**



研究背景

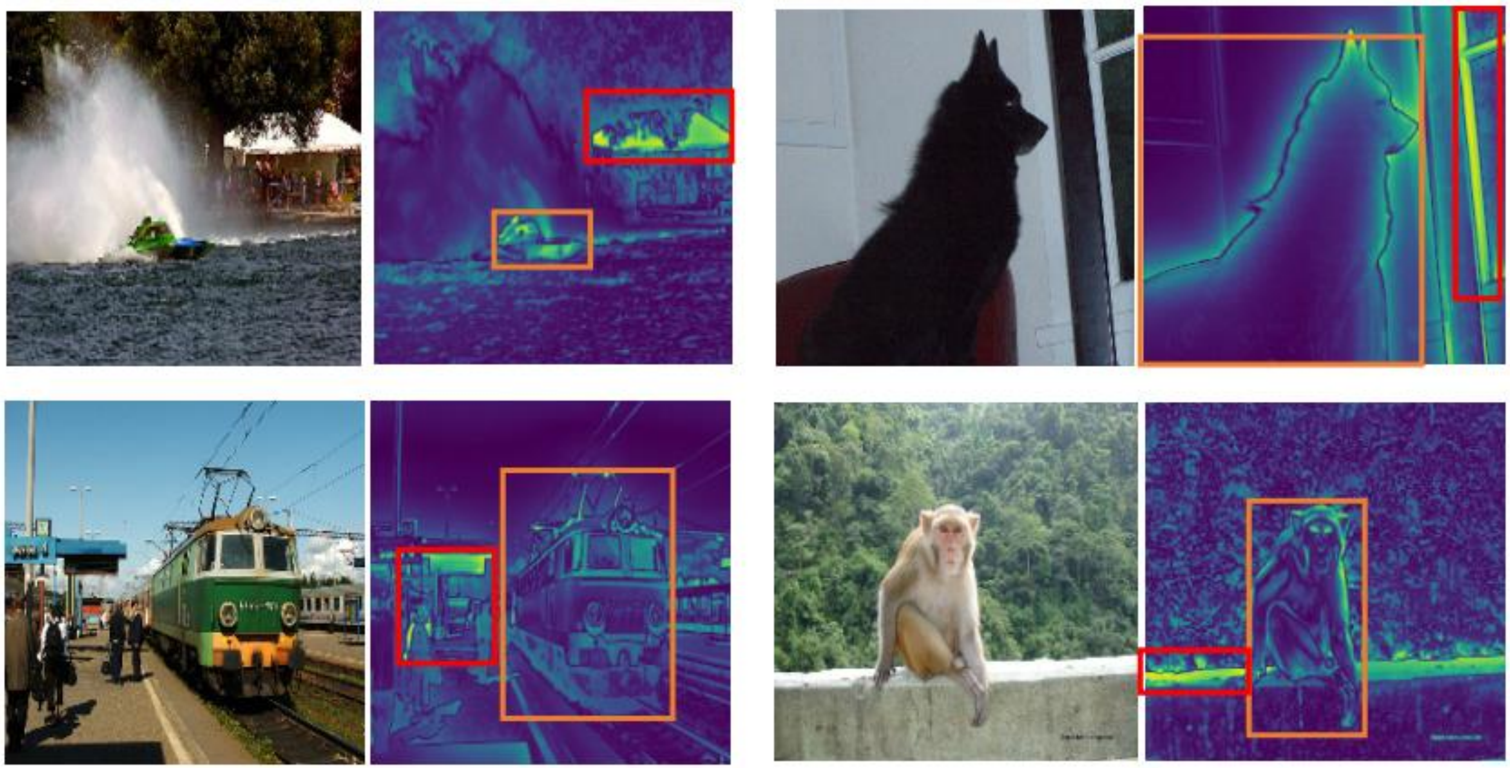
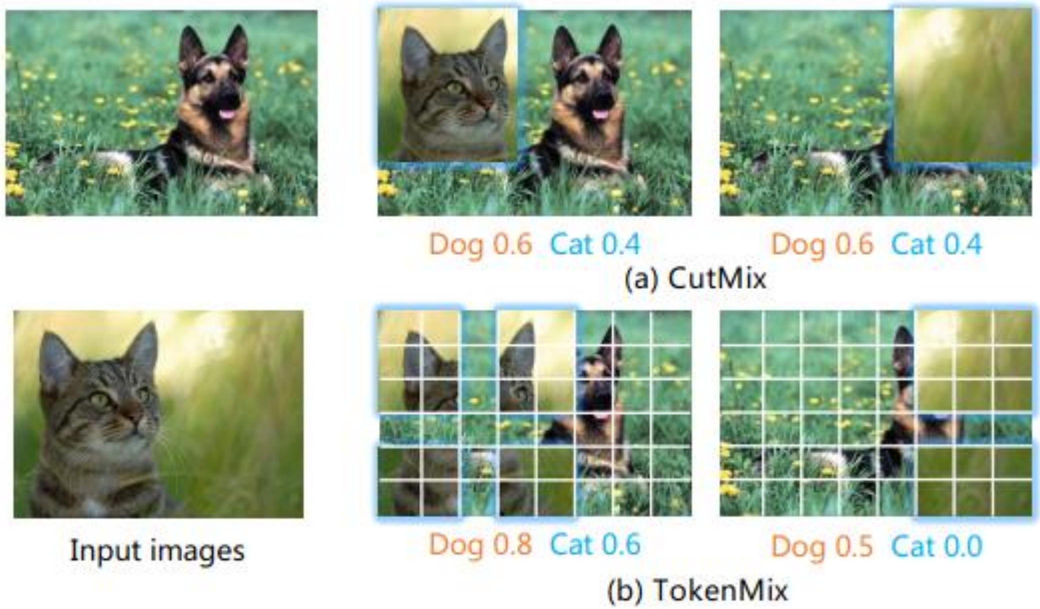


Fig. 8: Examples from ImageNet-1K. Orange boxes indicate foreground regions of the target classes. Red boxes indicate the most salient areas.

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

本文思想



- CutMix类方法的两个核心：mask和label
- 产生类似CutMix的mask，根据mask将两张图片混合在一起
- 产生更合理的mixed label，根据系数将两个label混合在一起

Revisit CutMix

$$\tilde{x} = M \odot x_a + (1 - M) \odot x_b,$$
$$\tilde{y} = \lambda y_a + (1 - \lambda) y_b,$$

$$M \in \{0,1\}^{H \times W}$$

$\mathrm{Beta}(\alpha, \alpha)$ $\quad\quad \frac{\sum M}{HW} = \lambda$

(a) CutMix

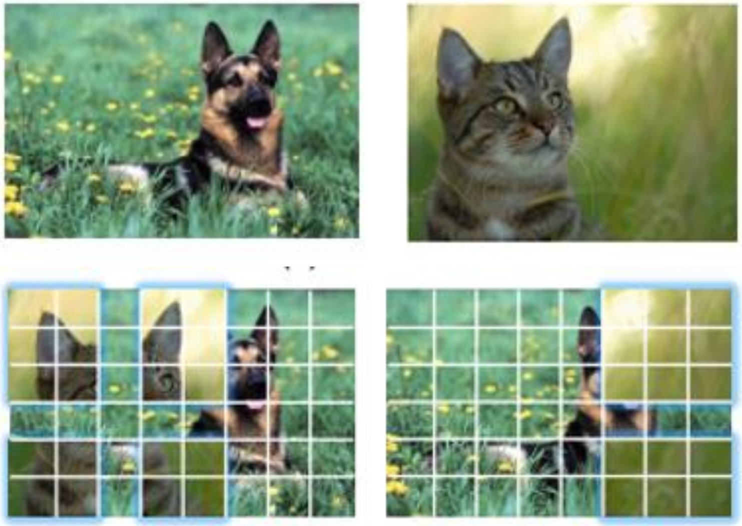
CutMix特点

1、在pixel层面进行mask，mix区域是矩形的，大小和尺度为随机采样得到

2、标签为两个label的线性组合，权重等于mix区域的面积

TokenMix

$$\tilde{x}^p = M_t \odot x_a^p + (1 - M_t) \odot x_b^p,$$
$$\tilde{y} = \sum_{i \in \mathfrak{S}} M_{ti} \odot A_{ai} + \sum_{i \in \mathfrak{S}} (1 - M_{ti}) \odot A_{bi}$$



(b) TokenMix

TokenMix特点

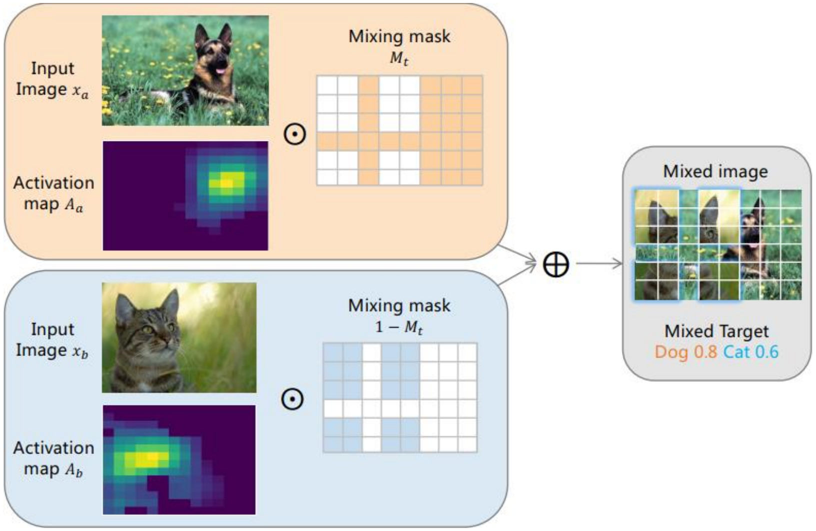
1、在token层面进行mask，mix区域整体是离散的，单块区域是矩形的

2、mix标签根据mask位置激活图的权重对应得到

TokenMix--mask

$$\tilde{x}^p = M_t \odot x_a^p + (1 - M_t) \odot x_b^p,$$
$$\tilde{y} = \sum_{i \in \mathfrak{S}} M_{ti} \odot A_{ai} + \sum_{i \in \mathfrak{S}} (1 - M_{ti}) \odot A_{bi}$$

$$x^p \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times (P^2 \cdot C)}$$
$$M_t \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P}}$$



mask是离散分布的，一次性mask至少14个token，尺度采样自 $[0.3, \frac{1}{0.3}]$

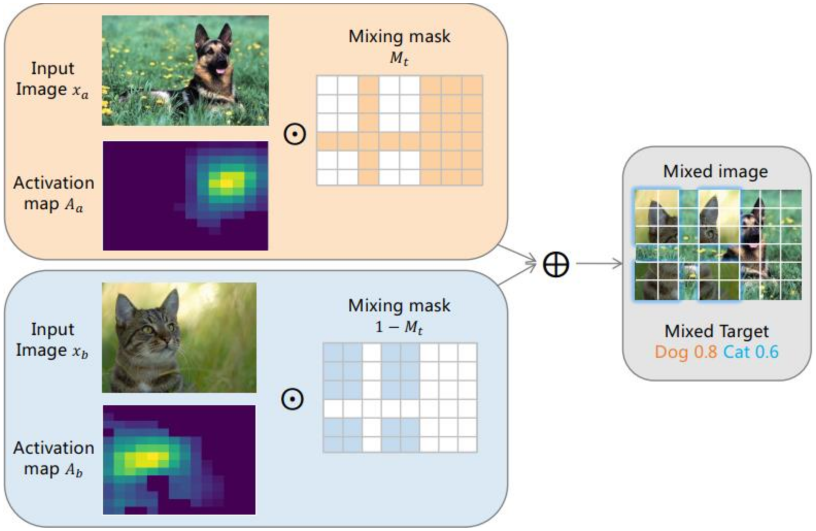
直到全局的mask token数量达到 $\lambda \frac{HW}{P^2}$

λ固定为0.5

TokenMix--label

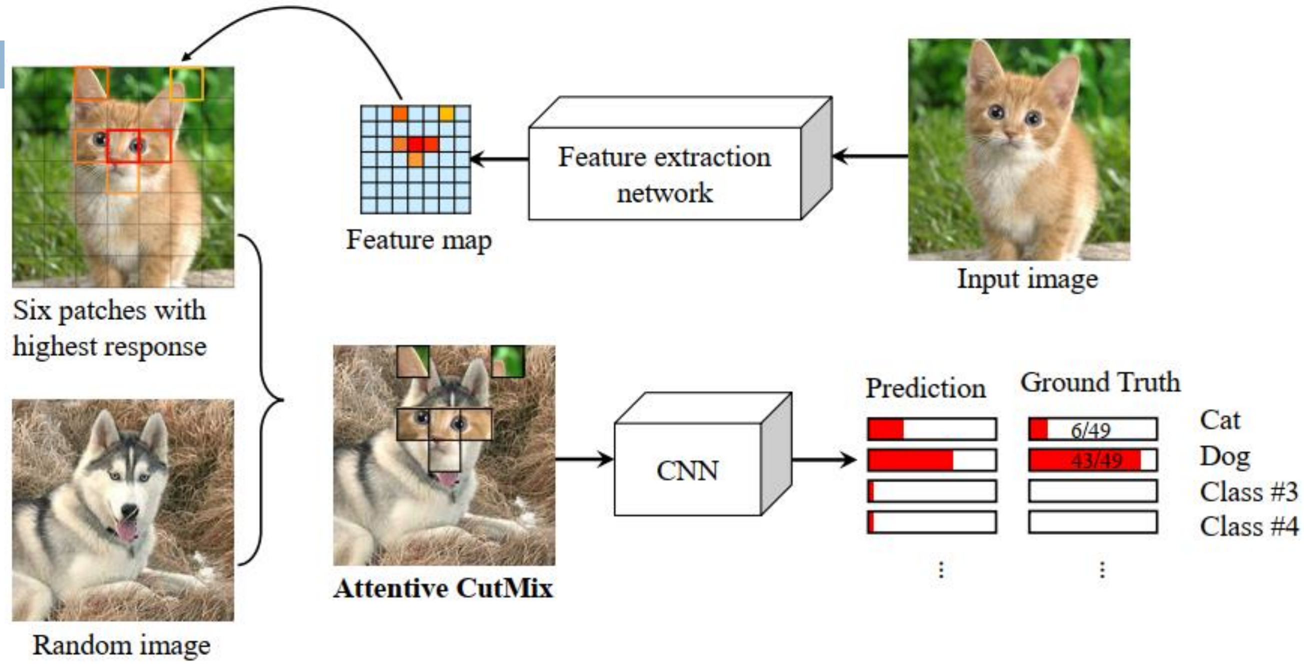
$$\tilde{x}^p = M_t \odot x_a^p + (1 - M_t) \odot x_b^p,$$
$$\tilde{y} = \sum_{i \in \mathfrak{S}} M_{ti} \odot A_{ai} + \sum_{i \in \mathfrak{S}} (1 - M_{ti}) \odot A_{bi}$$

$$x^p \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times (P^2 \cdot C)}$$
$$M_t \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P}}$$



A_{ai} and A_{bi} spatially normalized activation maps

相近方法 Attentive CutMix



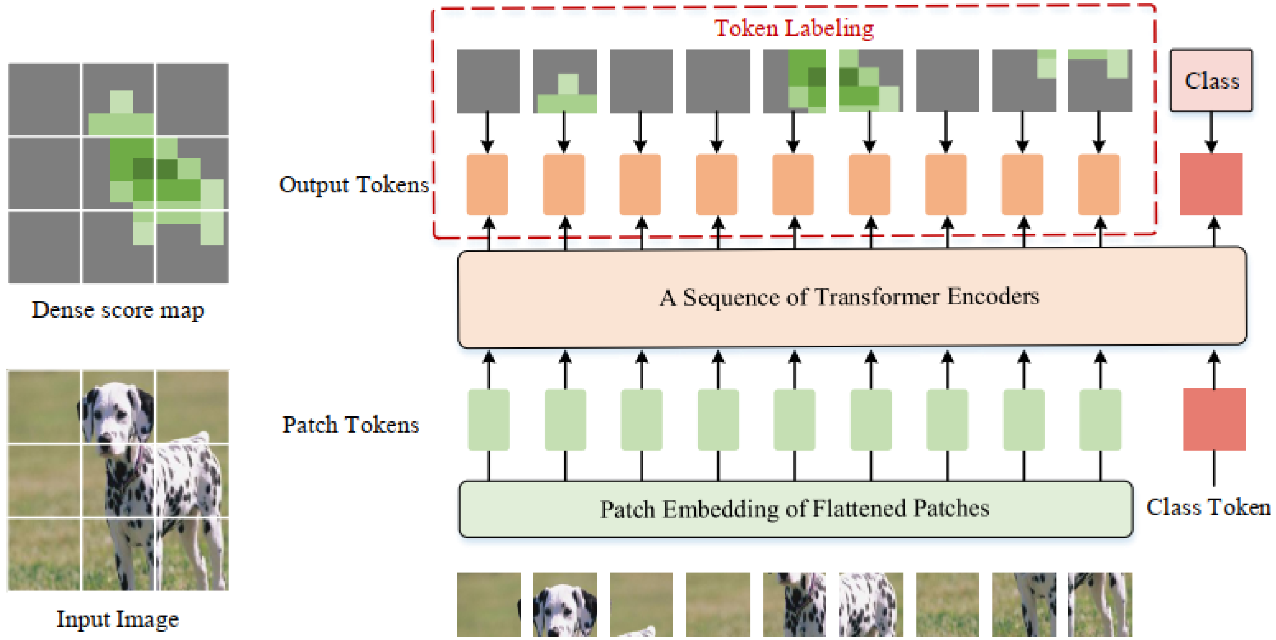
$$\tilde{x} = \mathbf{B} \odot x_1 + (\mathbf{1} - \mathbf{B}) \odot x_2$$

$$\tilde{y} = \lambda y_1 + (1 - \lambda) y_2$$

Attentive CutMix

1. 由网络预测的热力图来得到attentive区域，将该区域与另一张图mix
2. 在grid层面进行操作，但对于标签还是线性组合

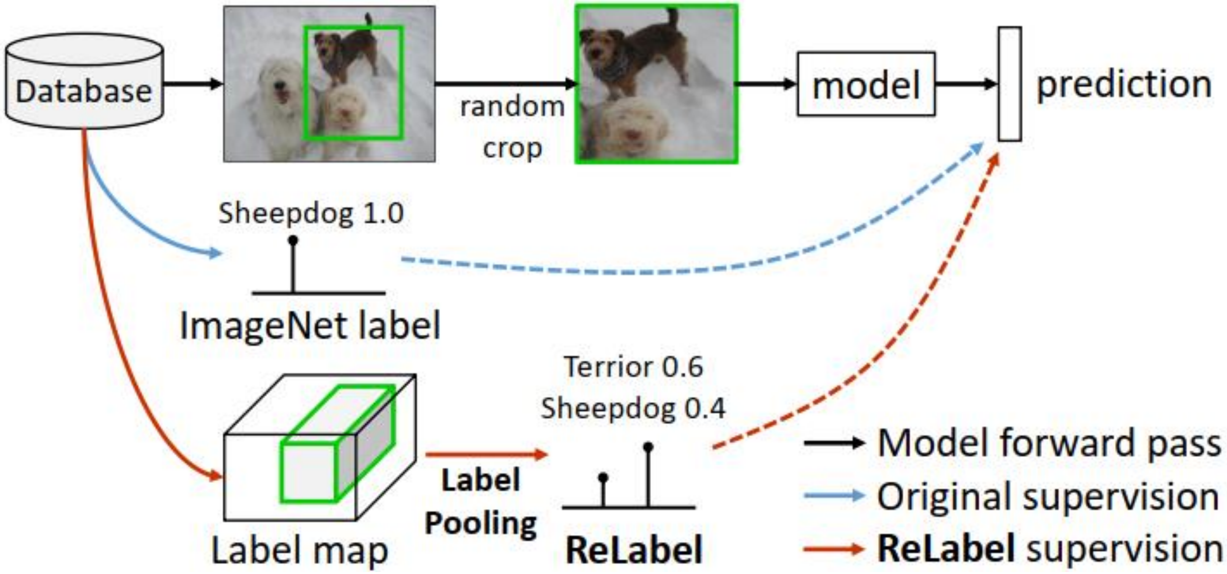
相近方法 Token Labeling



Token Labeling

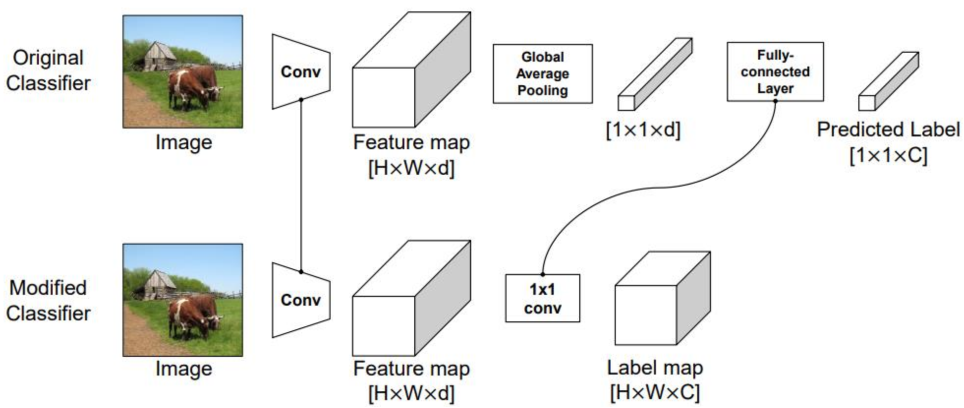
1. 对每个patch token都引入supervision，由activation map得到patch label
2. MixToken：对整个token对应区域mix，保证每个patch里内容的一致性

相近方法 ReLabel



ReLabel

1. 将ImageNet进行多标签labeling，提高数据准确性
2. 对feat map和label map同一位置使用ROIAlign



实验效果:ImageNet/ADE20K

Model	#FLOPs (G)	#Params (M)	CutMix	TokenMix
DeiT-T [29]	1.3	5.7	72.2	**73.2** (+1.0)
PVT-T [34]	1.9	13.2	75.1	**75.6** (+0.5)
CaiT-XXS-24 [30]	2.5	9.5	77.6	**78.0** (+0.4)
DeiT-S [29]	4.6	22.1	79.8	**80.8** (+1.0)
Swin-T [21]	4.5	29	81.2	**81.6** (+0.4)
DeiT-B [29]	17.6	86.6	81.8	**82.9** (+1.1)

Model	TokenMix	mIoU(%)	mAcc(%)	+ms mIoU(%)	+ms mAcc(%)
DeiT-T	✗	36.4	46.7	37.5	47.1
	✓+RL	36.6	47.0	38.1	47.9
	✓+TL	36.9	47.1	38.3	48.1
	✓	**37.1**	**47.5**	**38.6**	**48.2**
DeiT-S	✗	42.3	52.8	43.7	53.8
	✓	**44.5**	**55.0**	**45.9**	**56.1**
DeiT-B	✗	46.3	56.5	47.7	57.6
	✓	**46.8**	**56.9**	**48.2**	**58.1**

消融实验

Augmentation	Supervision	Top-1 Acc.	GPU Time
CutMix	ImageNet	72.2	+0.0%
TokenMix	ImageNet	**72.7**	+0.0%
TokenMix	ReLabel	72.7	+0.8%
TokenMix	TokenLabeling	72.9	+0.8%
TokenMix	TokenMix	**73.2**	+0.8%

Augmentation	Top-1 Acc.
CutMix [36]	72.2
Co-Mix [17]	72.2
SaliencyMix [31]	71.8
Puzzle-Mix [18]	72.3
TokenMix	**72.7**

Teacher	Teacher Top-1 Acc.	Top-1 Acc.
NFNet-F6 [3]	86.1	80.8
ResNet101 [13]	82.3	80.7
ResNet26 [13]	79.8	80.5
Saliency [23]	N/A	80.1

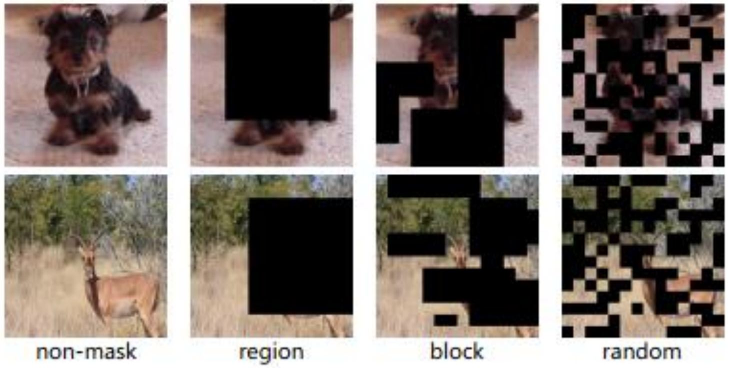
Mixup [38]	CutMix [36]	TokenMix	Top-1 Acc.
✗	✗	✗	75.8
✗	✓	✗	78.7
✓	✗	✗	80.0
✗	✗	✓	**81.5**
✓	✓	✗	81.8
✗	✓	✓	82.0
✓	✗	✓	**82.9**

消融实验

Table 8: Ablation of mask sampling strategy. The *region-based* strategy works best on ResNet50, but degrades on DeiT-S.

Model	region	random	block
DeiT-T [29]	72.2	**72.7**	**72.7**
DeiT-S [29]	79.8	**80.6**	**80.6**
ResNet50 [13]	79.3	78.3	**79.7**

Fig. 7: Illustration of different mask sampling strategies.



Model	Refinement	Top-1 Acc.
DeiT-S [29]	✗	79.8
	✓	**80.5** (+0.7)
Swin-T [21]	✗	81.2
	✓	**81.5** (+0.3)
ResNet50 [13]	✗	79.3
	✓	**79.8** (+0.5)

Model	Mask	Refinement	Top-1 Acc.
DeiT-T	random	✗	72.7
		✓	**72.9** (+0.2)
	block	✗	72.7
		✓	**73.2** (+0.5)
DeiT-S	random	✗	80.6
		✓	80.6
	block	✗	80.6
		✓	**80.8** (+0.2)

消融实验

Table 10: Ablation of training epochs. TokenMix enjoys longer training. The extra 100 epochs of training improve +0.4% accuracy.

Mixing Method	Epoch	Top-1 Acc.
CutMix [36]	300	79.8
	400	79.9 (+0.1)
TokenMix	300	80.8
	400	**81.2** (+0.4)

Table 11: Ablation of the loss function. Binary cross-entropy (BCE) improves TokenMix, compared with multi-class cross-entropy (CE).

Mixing Method	Loss Type	Top-1 Acc.
CutMix [36]	CE	79.8
	BCE	79.8
TokenMix	CE	80.3
	BCE	**80.8** (+0.5)

可视化结果

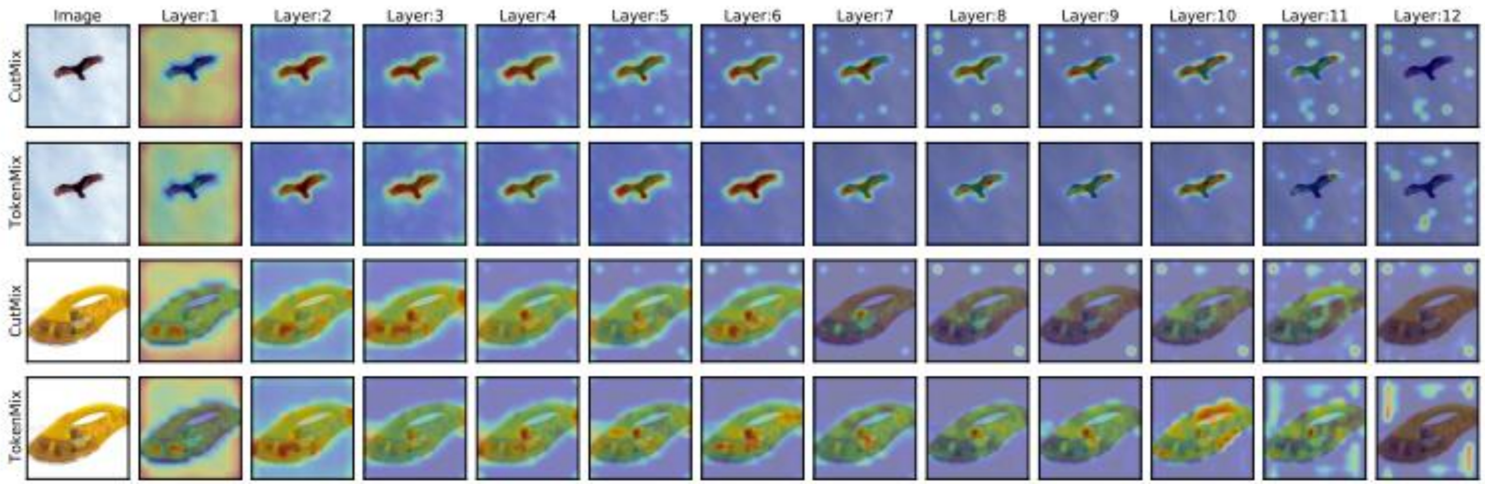


Fig. 3: Visualization of the attention maps of the class token in DeiT-S to attend to patch tokens at different layers. Using CutMix distracts the attention to background areas in the several middle layers. In contrast, the proposed TokenMix helps the class token focus more on foreground objects and leads to consistent performance gain.

可视化结果

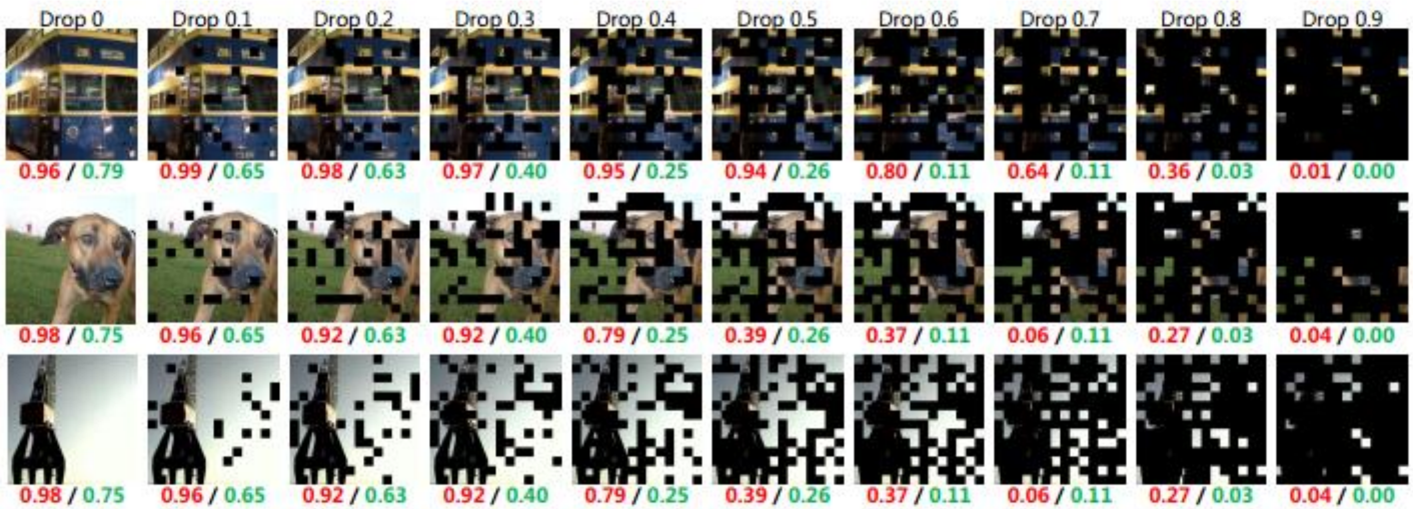


Fig. 4: Example images and the predicted confidences under different occlusion ratios. Red scores under the images are predicted by TokenMix, and green ones by CutMix. The model trained with TokenMix holds high confidence when a large number of patches are dropped, while the model trained with CutMix outputs low confidence.

可视化结果

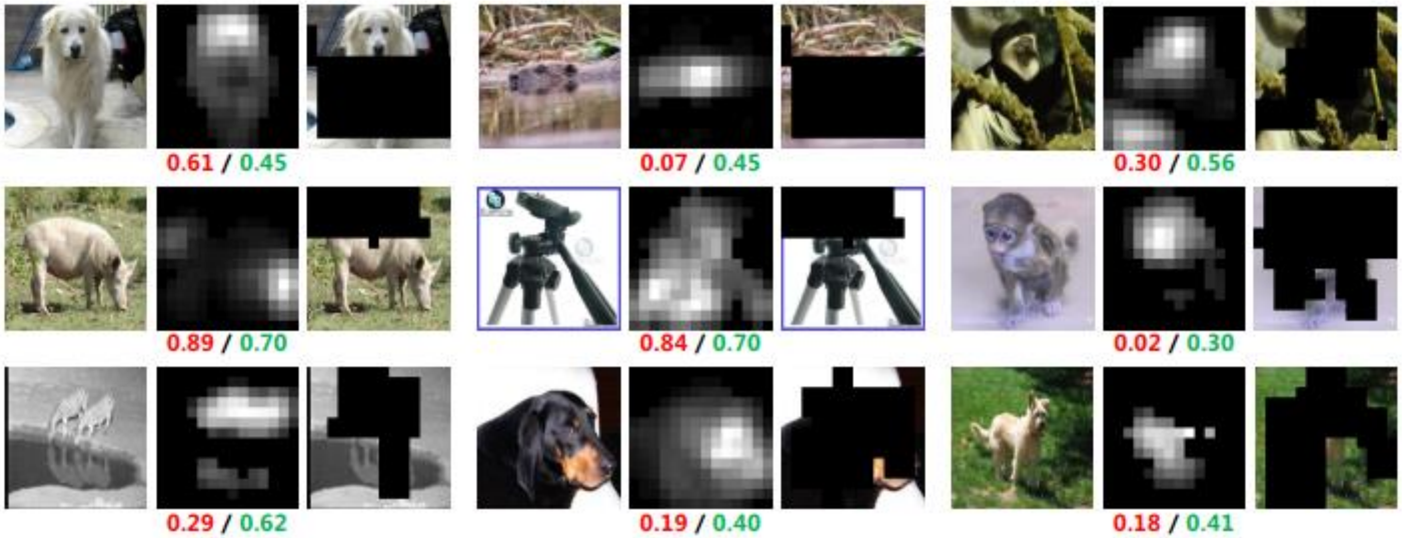


Fig. 6: The target scores generated by TokenMix and CutMix. For each tripled sub-figure, the left is the input image, the middle is the neural activation map, and the right is the masked image. Our approach generates more reasonable target scores, especially when the foreground region is cropped.

- 作者介绍
- 研究背景
- 解决方法
- 实验效果
- 总结反思

总结反思

- 针对Transformer自带的全局性来改进CutMix系列方法，并使用分布式的mask区域来更好地利用全局性
- 根据attentive map来计算混合标签，消除了mix区域与标签无法对应带来的影响
- TokenMix方法使得网络学习能够更加关注前景，对occlusion更加的robust