# ICCV 2021｜“白嫖”性能的MixMo，一种新的数据增强or模型融合方法

原创 CV开发者都爱看的 极市平台 2021-08-04 22:00:00 手机阅读





作者｜小马
编辑｜极市平台

极市导读

本文作者提出了一种新的多输入多输出深度子网学习广义框架MixMo，MixMo可以作为一种集成方法或一种新的混合样本数据增强方法进行分析，同时仍然与两种研究方向的工作保持互补。>>加入极市CV技术交流群，走在计算机视觉的最前沿

## 写在前面

最近的工作提出的不用额外计算的集成方法，大多是在一个网络中同时设置不同的subnet。训练时。每个subnet只学习分类多个输入数据中的其中一个。然而，如何更好地混合这些多个输入的问题迄今尚未被研究。

在本文，作者提出了一种新的多输入多输出深度子网学习广义框架MixMo。作者的Motivation是用一个更合适的混合机制来代替先前方法中求和导致的次优操作。受到混合样本数据增强的启发，作者发现特征的混合可以使subnet更强，使得数据更加多样，进而提高模型performance。

基于MixMo，作者提升了CIFAR-100和Tiny ImageNet数据集上的SOTA性能。

## 1. 论文和代码地址

**MixMo: Mixing Multiple Inputs for Multiple Outputs via Deep Subnetworks**

Alexandre Ramé*†[1], Rémy Sun*[1,2] and Matthieu Cord[1,3]

[1]Sorbonne Université
[2]Optronics & Missile Electronics, Land & Air Systems, Thales
[3]Valeo.ai

论文地址：https://arxiv.org/abs/2103.06132

代码地址：https://github.com/alexrame/mixmo-pytorch

## 2. Motivation

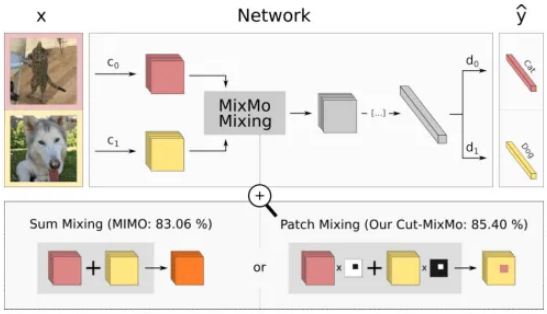
卷积神经网络(cnn)在计算机视觉任务中表现出了出色的性能，尤其是分类任务。为了在真实场景中增加鲁棒性或赢得Kaggle竞赛，cnn通常会采用两种实用策略:**数据增强** 和**模型集成** 。

数据增强可以减少过拟合并提升模型的泛化性。传统的图像增强是保留标签的:例如翻转、裁剪等。然而，最近的混合样本数据增强(MSDA)改变了这种方式：多个输入和它们的标签按比例混合来创建人工样本，代表工作有MixUp，CutMix等等。

模型集成证明了聚合来自多个神经网络的不同预测能够显著提高了泛化能力，尤其是不确定性估计。从经验上讲，几个小网络的集成通常比一个大网络性能更好。然而，在训练和推理方面，集成在时间和显存消耗方面都是昂贵的：这往往限制了模型集成的适用性。

在本文，作者提出了多输入多输出框架MixMo。为了解决传统集成中出现的这些开销，作者将M个独立子网放入一个单一的base网络中。这也是合理的，因为在模型集成时，“最终采纳的网络”其实就和整体的网络表现差不多。

所以，现在最大的问题是如何在没有结构差异的情况下加强subnet之间的多样性。

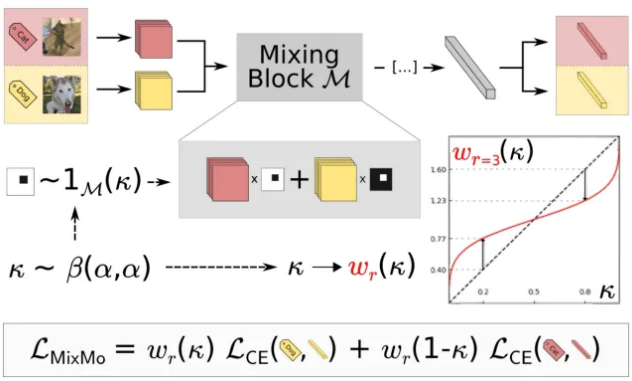



如上图，作者在训练过程中同时考虑了M个输入，M个输入被M个参数不共享的Encoder编码到共享空间中，然后将特征送到核心网络，核心网络最终分成M个分支；这个M个分支用来预测不同输入信息的label。在inference的时候，同一图像重复M次：通过平均M个预测获得“免费”的集成效果。

与现有的MSDA相比，MixMo最大的不同就是multi-input mixing block。如果合并是一个基本的求和，MixMo将变成到MIMO [1]。作者对比了大量的MSDA的工作，设计了更合适的混合块，因此作者采用binary masking的方法来确保子网络的多样性。（**如上图所示，作者对不同样本采用了一个binary masking方法，这一点就类似CutMix，而不是像MIMO那样直接相加**）。

这种不对称的混合也会造成网络特征中的信息不平衡的新问题，因此作者通过一个新的加权函数来解决多个分类训练任务之间的不平衡问题。

## 3. 方法




MixMo的整体结构如上图所示

### 3.1. General overview

如上图所示，多个训练数据$(x_0, x_1), (y_0, y_1)$，通过参数不共享的两个卷积神经网络分别编码到一个共享空间得到$c_0(x_0)$和 $c_1(x_1)$。

为了能够显式地突出混合的信息，作者采用了一个广义多输入混合块$M$(generalized multi-input mixing block)。这种多重混合能够解决模型集成多样性和个体精度权衡的问题，从而达到更高的performance。共享的特征表示$M(c_0(x_0)\ ,\ c_1(x_1))$被送入到下一个卷积层。

核心网络$C$需要同时处理两种输入的特征表示。然后多层的网络D，通过这个mix的特征，再一次把各自样本的类别识别出来。（**个人理解这个网络是一个“分-总-分”的结构，首先，这个网络对不同输入的样本进行分别编码，这是第一个“分”的过程；然后这些被编码的特征通过Mixing Block融合，这是“总”的过程；最后不同的层再根据这个混合的特征，识别出各自样本的类别，这是最后一个“分”的过程**）

训练过程中的损失函数为各自样本的交叉熵损失函数之和（分别乘上各自的权重，权重的计算见下文）：

$$\mathcal{L}_{\text{MixMo}} = w_r(\kappa)\mathcal{L}_{\text{CE}}\left(y_0, \hat{y}_0\right) + w_r(1-\kappa)\mathcal{L}_{\text{CE}}\left(y_1, \hat{y}_1\right).$$

在inference的时候，同一个输入x被输入到不同的分支中，核心网络C的输入为$c_0(x) + c_1(x)$的和，这最大的保留了来自两种编码信息。然后，最终的预测结果为将不同分支的预测平均值$1/2(\hat{y}_0 + \hat{y}_1)$。这使得模型可以在一次前向传播的过程中享受模型融合的结果。

### 3.2. Mixing block$M$

Mixing block 是MixMo的核心，它将两个输入组合成一个共享表示。受MSDA混合方法的启发，MixMo通用框架包含了更广泛的变化。

作者提出的第一个变体是 Linear-MixMo，借鉴了MixUp的思想，直接将两张图片通过一个透明度叠在一起：

$$\mathcal{M}_{\text{Linear-MixMo}}\left(l_0, l_1\right) \ = \ 2\left[\kappa l_0 + (1-\kappa)l_1\right]$$

接着，作者受到MixCut的启发，提出了 Cut-MixMo：

$$\mathcal{M}_{\text{Cut-MixMo}}\left(l_0, l_1\right) = 2\left[\mathbb{1}_{\mathcal{M}} \odot l_0 + (\mathbb{1} - \mathbb{1}_{\mathcal{M}}) \odot l_1\right],$$

与Linear-MixMo不同，这里并不是将整张图片相加，而是像MixCut一样，每次都是加了一个patch。

Cut-MixMo比其他策略表现更好。具体来说，Cut-MixMo中的binary mixing取代了MIMO和 Linear-MixMo中的线性插值，使子网络更加精确和多样化。

**为什么Cut-MixMo会比 Linear-MixMo要更好?**

1）基于CutMix优于Mixup的相同原因，M中的binary mixing训练了更强的单个子网。此外通过binary mixing，模拟了常见的物体遮挡问题。

2）线性插值从根本上不适合诱导多样性，因为两个输入都保留了完整的信息。CutMix通过交替选择的图像patch，显式地增加了数据集的多样性。

### 3.3. Loss weighting$w_r$

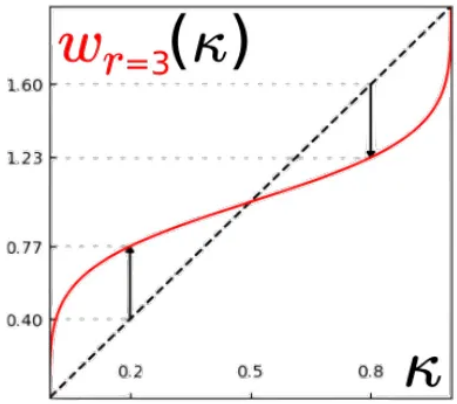
Mixing机制中的不对称可能导致一种输入盖过另一种输入。当$k \neq 0.5$时，权重更大的输入可能更容易预测。因此，作者定义了一个权重函数$w_r$来平衡多个损失函数的重要性。这种加权调整了有效学习率、梯度在网络中的流动方式以及混合信息在特征中表示的方式。

加权函数具体表示如下：

$$\kappa^{1/r}$$

$$w_r(\kappa) = 2\frac{\kappa^{1/r}}{\kappa^{1/r} + (1-\kappa)^{1/r}}.$$

其中$r$是一个超参数，$r = 3$的曲线如下图所示：




### 3.4. From manifold mixing to MixMo

相比于其他MSDA方法，MixMo使用两个独立的编码器(每个编码器输入一个数据)，并且它输出是两个预测而不是一个。

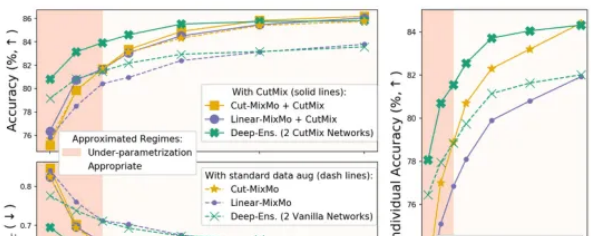
而其他MSDA方法使用一个单一的分类器，该分类器针对一个唯一的软标签，通过线性插值反映不同的类。相反，MixMo选择充分利用混合样本的复合特性，训练分离的dense层，d0和d1，在测试时能够在没有额外计算的情况下，达到模型集成的效果。

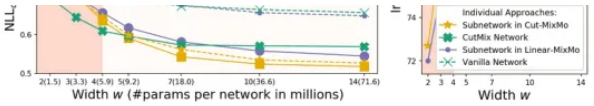
## 4.实验

### 4.1. Main results on CIFAR-100 and CIFAR-10

| Dataset | | CIFAR-100 | | | CIFAR-10 | |
|---|---|---|---|---|---|---|
| Approach | Time Tr./Inf. | Top1 %,↑ | Top5 %,↑ | $NLL_c$ $10^{-2}$,↓ | Top1 %,↑ | $NLL_c$ $10^{-2}$,↓ |
| Vanilla | 1/1 | 81.63 | 95.49 | 73.9 | 96.34 | 12.6 |
| Mixup | | 83.44 | 95.92 | 65.7 | 97.07 | 11.2 |
| Manifold Mixup† | | 81.96 | 95.51 | 73.4 | 97.45 | 12.2 |
| CutMix | | 84.05 | 96.09 | 64.8 | 97.23 | 9.9 |
| ResizeMix† | | 84.31 | - | - | 97.60 | - |
| Puzzle-Mix† | 2/1 | 84.31 | 96.46 | 66.8 | - | - |
| GradAug† | 3/1 | 84.14 | 96.43 | - | - | - |
| + CutMix† | | 85.51 | 96.86 | - | - | - |
| Mixup BA† | 7/1 | 84.30 | - | - | **97.80** | - |
| DE (2 Nets) | 2/2 | 83.17 | 96.37 | 66.4 | 96.67 | 11.1 |
| + CutMix | | 85.74 | 96.82 | 57.1 | 97.52 | 8.6 |
| MIMO | 2/1 | 82.40 | 95.78 | 68.8 | 96.38 | 12.1 |
| Linear-MixMo | | 82.54 | 95.99 | 67.6 | 96.56 | 11.4 |
| + CutMix | | 84.69 | 97.12 | 57.2 | 97.32 | 9.4 |
| Cut-MixMo | | 84.38 | 96.94 | 56.3 | 97.31 | 8.9 |
| + CutMix | | 85.18 | 97.20 | 54.5 | 97.45 | 8.4 |
| MIMO | 4/1 | 83.06 | 96.23 | 66.1 | 96.74 | 11.4 |
| Linear-MixMo | | 83.08 | 96.26 | 65.6 | 96.91 | 10.8 |
| + CutMix | | 85.47 | 97.04 | 55.8 | 97.68 | 8.7 |
| Cut-MixMo | | 85.40 | 97.22 | 53.5 | 97.51 | 8.1 |
| + CutMix | | **85.77** | **97.42** | **52.4** | 97.73 | **7.9** |

上表展示了MixMo在CIFAR10和CIFAR100上的结果，可以看出相比于原始的网络，MixMo对于性能的提升非常明显。

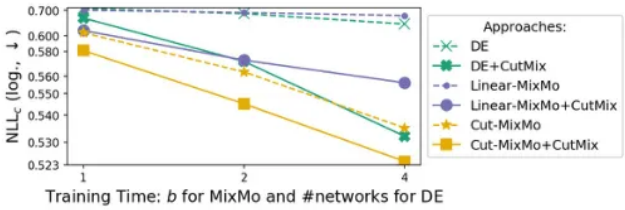





(a) Ensemble Top1 and $\text{NLL}_c$. (b) Individual Top1.

从上图可以看出，随着宽度w的增加，MixMo比DE(绿色曲线)的性能提升更加明显。

### 4.2. Training time



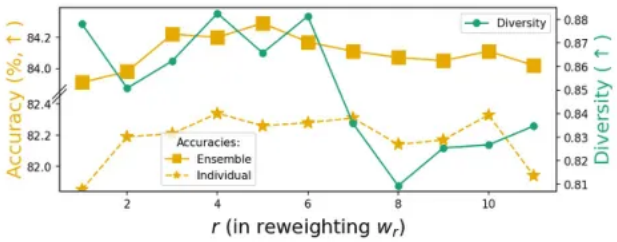

可以看出，在相同的训练时间内，Cut-MixMo的表现优于DE。

### 4.3. The mixing block $M$

| $\mathcal{M}$ approach | Mixup [104] | Horiz. Concat. | Vertical Concat. | PatchUp 2D [20] | FMix [33] | CowMask [23, 24] | CutMix [101] |
|---|---|---|---|---|---|---|---|
| Top1 ↑ | 82.5 | 82.78 | 84.00 | 84.16 | 83.76 | 84.17 | **84.38** |
| $\text{NLL}_c$ ↓ | 0.676 | 0.627 | 0.573 | 0.581 | 0.602 | **0.561** | 0.563 |

上表比较几个mix block的性能，可以看出无论形状如何，binary mixing的性能都优于线性混合。

### 4.4. Weighting function $w_r$




上图比较了加权函数不同r下的性能，r在[3,6]范围内达到了很好的trade-off。

### 4.5. Multiple encoders and classifiers

| # Enc. | # Clas. | $\text{NLL}_c$ ↓ |
|---|---|---|
| 1 | 1 | 0.604 |
| 2 | 1 | 0.666 |
| 1 | $2^{\ominus}$ | 0.687 |
| 1 | $2^{\otimes}$ | 0.598 |
| 2 | 2 | **0.563** |

上表的实验结果表明，2个编码器和2个分类器对于实验结果是比较好的。

### 4.6. Pushing MixMo further: Tiny ImageNet

| Width $w$ (# params) | | $w = 1$ (11.2M) | | $w = 2$ (44.9M) | | $w = 3$ (100.5M) | |
|---|---|---|---|---|---|---|---|
| Approach | Time Tr./Inf. | Top1 %,↑ | $NLL_c$ ↓ | Top1 %,↑ | $NLL_c$ ↓ | Top1 %,↑ | $NLL_c$ ↓ |
| Vanilla | 1/1 | 62.56 | 1.53 | 64.80 | 1.51 | 65.78 | 1.53 |
| Mixup | 1/1 | 63.74 | 1.62 | 66.62 | 1.50 | 67.27 | 1.51 |
| Manifold Mixup† | 1/1 | 58.70 | 1.92 | - | - | - | - |
| Co-Mixup† | 1/1 | 64.15 | - | - | - | - | - |
| CutMix | 1/1 | 65.09 | 1.58 | 67.76 | 1.33 | 68.95 | 1.29 |
| Puzzle-Mix† | 2/1 | 64.48 | 1.65 | - | - | - | - |
| DE (2 Nets) | 2/2 | 65.53 | 1.39 | 68.06 | 1.37 | 68.38 | 1.36 |
| DE (3 Nets) | 3/3 | 66.76 | 1.34 | 69.05 | 1.29 | 69.36 | 1.28 |
| DE (4 Nets) | 4/4 | **67.51** | **1.31** | **69.94** | **1.24** | 69.72 | 1.26 |
| Linear-MixMo | 2/1 | 61.58 | 1.61 | 66.62 | 1.41 | 68.18 | 1.36 |
| Cut-MixMo | 2/1 | 63.78 | 1.48 | 68.30 | 1.30 | 69.89 | 1.26 |
| Linear-MixMo | 4/1 | 62.91 | 1.51 | 67.03 | 1.41 | 68.38 | 1.38 |
| Cut-MixMo | 4/1 | 64.44 | 1.48 | 69.13 | 1.28 | **70.24** | **1.19** |

在更大的规模和更多样的64 × 64图像上，Cut-MixMo在Tiny ImageNet上达到了70.24%的新水平，如上表所示。

## 5. 总结

在本文中，作者提出了MixMo，一个多输入多输出策略的框架。MixMo可以作为一种集成方法或一种新的混合样本数据增强方法进行分析，同时仍然与两种研究方向的工作保持互补。此外，作者引入了一个新的权重函数，以平衡训练时的损失。最终，作者通过实验证明了MixMo的有效性。

**参考文献**

[1]. Marton Havasi, Rodolphe Jenatton, Stanislav Fort,Jeremiah Liu, Jasper Roland Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, and Dustin Tran. Training independent subnetworks for robust prediction. In ICLR,2021.

# 极市平台签约作者#

小马
知乎：努力努力再努力

厦门大学人工智能系20级硕士。
研究领域：多模态内容理解，专注于解决视觉模态和语言模态相结合的任务，促进Vision-Language模型的实地应用。

**作品精选**

CVPR2021最佳学生论文提名：Less is More
Transformer一作又出新作！HaloNet：用Self-Attention的方式进行卷积
超越Swin，Transformer屠榜三大视觉任务！微软推出新作：Focal Self-Attention

**投稿方式：**

添加小编微信Fengcall（微信号：fengcall19），备注：**姓名-投稿**

△长按添加极市平台小编

**觉得有用麻烦给个在看啦~**

阅读原文

喜欢此内容的人还喜欢

15个目标检测开源数据集汇总
极市平台