ATTENTIVE CUTMIX:
AN ENHANCED DATA AUGMENTATION APPROACH FOR DEEP LEARNING BASED IMAGE CLASSIFICATION

Devesh Walawalkar, Zhiqiang Shen, Zechun Lui, Marios Savvides
Carnegie Mellon University

ICASSP 2020

Presented by Eungyeup Kim

Vision Seminar
21 MAY 2020

Backgrounds

CutMix, ICCV'19 (Oral)

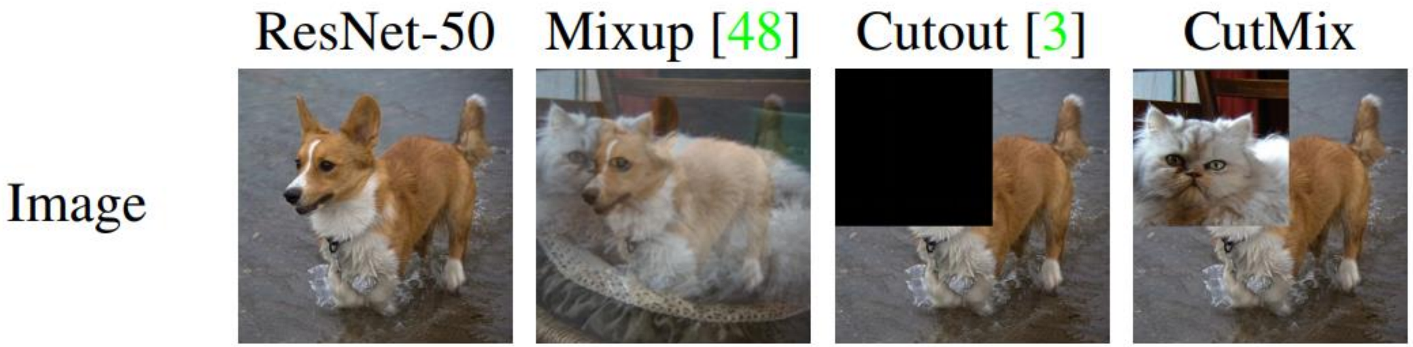
- Conventional drop-out strategies, such as Mixup and Cutout, have enhanced the performance of convolutional neural network classifiers.

- They have guided a model to attend on less discriminative parts of objects, resulting in better generalizability.

- *However, removal of informative pixels lead to information loss and inefficiency during training.*

- So, **CutMix** strategy simply cuts and pastes the patch from image A (cat) to image B (dog).

- Significant improvements in performance of image classification, localization and image captioning tasks.



Backgrounds

CutMix, ICCV'19 (Oral)

Let $x \in R^{W \times H \times C}$ and y denote a training image and its label, respectively. CutMix strategy generates a new sample $(\tilde{x}, \tilde{y})$ by combining two different samples (x_A, y_A) and (x_B, y_B).

$$\tilde{x} = M \odot x_A + (1 - M) \odot x_B$$
$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B,$$

where $M \in \{0,1\}^{W \times H}$ denotes a binary mask, $\odot$ a element-wise multiplication and $\lambda \sim Unif(0, 1)$ the ratio of patches from the first image.

We first sample the bounding box coordinates $B = (r_x, r_y, r_w, r_h)$ indicating the cropping regions on x_A and x_B.

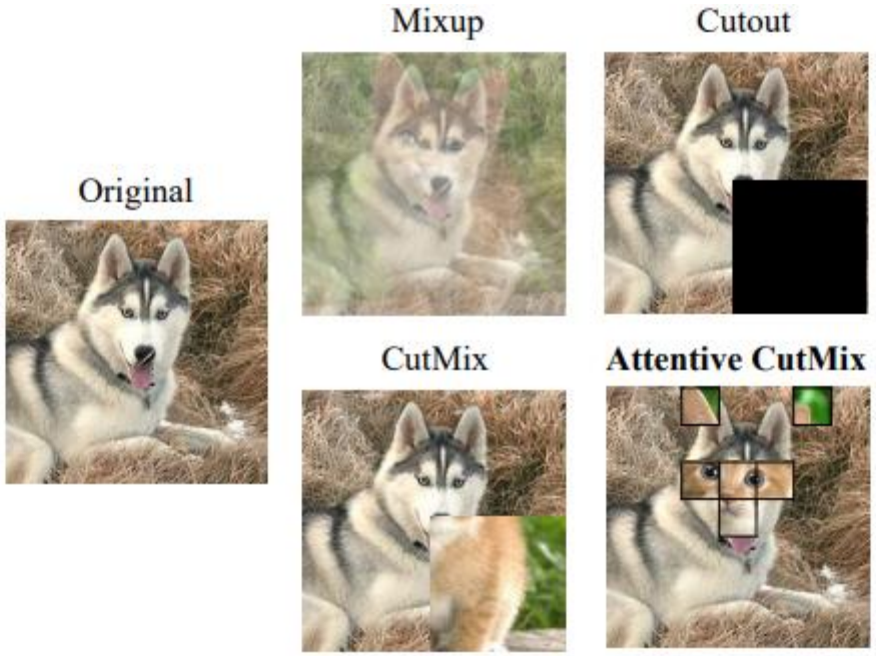
$$r_x \sim Unif(0, W), r_w = W\sqrt{1 - \lambda},$$
$$r_y \sim Unif(0, H), r_h = H\sqrt{1 - \lambda}$$

The binary mask M is then decided by filling with 0 within the bounding box B, otherwise 1.

Motivation

Regional Dropout strategies (Cutout and CutMix) perform the operation randomly.

- The region of dropout (location and size) is randomly selected, without considering its semantical importance.
- *However, it is possible to cutting an unimportant background patch and pasting it into the second image, while the composite label contains a part of first label.*
- This randomness imposes a confusion to the network.



This work

1) Proposes *Attentive CutMix,* cutting the most descriptive regions in an image.
2) Demonstrates that *Attentive CutMix* outperforms the baseline *CutMix* and other methods by a significant margin.

Methods

Attentive CutMix

Let $x \in R^{W \times H \times C}$ and y denote a training image and its label, respectively. Attentive CutMix strategy generates a new sample $(\tilde{x}, \tilde{y})$ by combining two different samples (x_A, y_A) and (x_B, y_B).

$$\tilde{x} = B \odot x_A + (1 - B) \odot x_B$$
$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B,$$

1. We first obtain a heatmap, generally a 7x7 grid map, of the x_A by passing it through an ImageNet pretrained classification model, e.g. Resnet-152.

2. Then we select the top "N" patches from the 7x7 grid.

3. After that, we map the selected attentive patches back to the original image. (Single patch in 7x7 grid would be mapped to 32x32 image on a 224x224 size of x_A .

4. We composite a binary mask B with selected patches on a original image size.

5. Assuming that we select the top 6 patches, λ would be $\frac{6}{7 \times 7}$.

Methods

Attentive CutMix

$$\tilde{x} = B \odot x_A + (1 - B) \odot x_B$$
$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B,$$

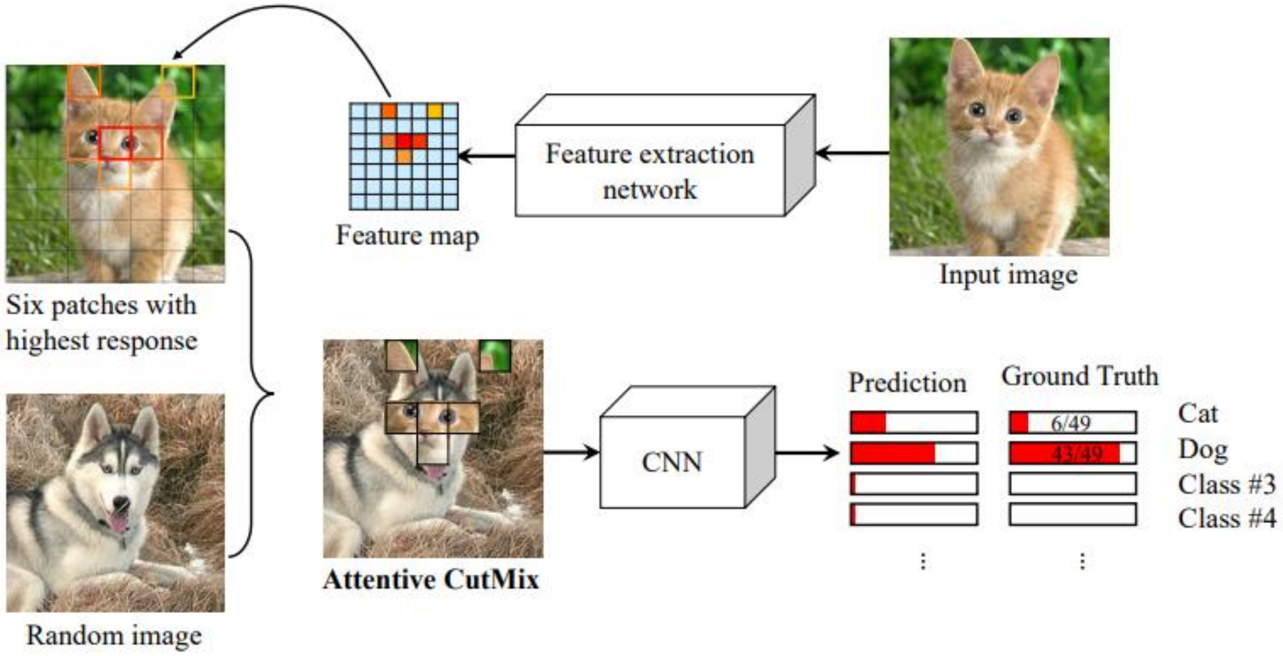


Fig. 2: Framework overview of proposed *Attentive CutMix*.

Experiments

[Ablation Study]

As the number of patches, “N” is a hyperparameter, it needs to be tuned for optimal performance.

$\Rightarrow$ It is found that cutting out top 6 attentive patches results in the best average performance across the experiments.

- **Less than** 6 doesn’t provide enough occlusion to the main subject in the second image .
- **More than** 6 provides excessive occlusion to the subject, which make the label for the image not enough discriminative for the model to learn anything useful.

Experiments

[ImageNet Classification]

	CIFAR-10 (%)			
Method	**Baseline**	**Mixup**	**CutMix**	**Attentive CutMix**
ResNet-18	84.67	88.52	87.92	**88.94**
ResNet-34	87.12	88.70	88.75	**90.40**
ResNet-101	90.47	91.89	92.13	**93.25**
ResNet-152	92.45	94.21	94.35	**94.79**
DenseNet-121	85.65	87.56	87.98	**88.34**
DenseNet-169	87.67	89.12	89.23	**90.45**
DenseNet-201	91.21	93.21	93.45	**94.16**
DenseNet-264	92.78	94.20	94.34	**94.83**
EfficientNet - B0	87.45	88.07	88.67	**88.94**
EfficientNet - B1	90.12	90.99	91.37	**92.10**
EfficientNet - B6	92.74	93.76	93.28	**93.92**
EfficientNet - B7	94.95	95.11	95.25	**95.86**

	ImageNet (Top-1 accuracy %)			
Method	**Baseline**	**Mixup**	**CutMix**	**Attentive CutMix**
ResNet-18	73.54	74.46	75.32	**75.78**
ResNet-34	77.31	79.03	79.22	**80.13**
ResNet-101	78.73	79.42	80.56	**81.16**
ResNet-152	78.98	80.01	80.25	**80.93**
DenseNet-121	75.87	76.89	77.34	**77.98**
DenseNet-169	77.03	79.10	79.32	**79.78**
DenseNet-201	78.67	80.14	80.23	**80.87**
DenseNet-264	79.59	82.11	82.36	**82.79**
EfficientNet - B0	76.12	78.19	78.21	**78.79**
EfficientNet - B1	78.47	79.96	80.17	**81.03**
EfficientNet - B6	83.89	84.43	84.60	**85.29**
EfficientNet - B7	84.34	85.12	85.19	**85.32**

	CIFAR-100 (%)			
Method	**Baseline**	**Mixup**	**CutMix**	**Attentive CutMix**
ResNet-18	63.14	64.40	65.90	**67.16**
ResNet-34	65.54	67.83	68.32	**70.03**
ResNet-101	68.24	70.76	71.32	**72.86**
ResNet-152	71.49	74.81	73.21	**75.37**
DenseNet-121	65.12	66.84	67.62	**69.23**
DenseNet-169	66.42	68.24	69.58	**71.34**
DenseNet-201	70.28	72.89	73.57	**74.65**
DenseNet-264	73.51	76.49	75.23	**77.58**
EfficientNet - B0	64.67	65.78	66.95	**67.48**
EfficientNet - B1	66.89	68.23	68.12	**68.96**
EfficientNet - B6	71.34	73.56	73.75	**74.82**
EfficientNet - B7	75.67	77.21	77.57	**78.52**

- For CIFAR-10, CIFAR-100 and ImageNet, our method provides better results over all tested models compared to CutMix, Mixup and the baseline methods.

Discussions

1. No additional experiments on other downstream tasks such as object detection, image captioning and out-of-distribution detection.

2. Lack of discussion on the effectiveness of covering the semantically important regions of image x_B. Does this aggravate the deterministic ability of the model?

Thank you