SAGE: Saliency-Guided Mixup with Optimal Rearrangements

Avery Ma1,2, Nikita Dvornik3, Ran Zhang3, Leila Pishdad4, Konstantinos G. Derpanis3,5, Afsaneh Fazly3
1University of Toronto 2Vector Institute 3Samsung AI Centre Toronto 4Borealis AI 5York University

Motivation

Data augmentation is a key element for training accurate models by reducing overfitting and improving generalization.

- Conventional data augmentation techniques (photometric and geometric transformations) merely create slightly altered copies of the original images and thus introduce limited diversity in the augmented dataset.
- More advanced data augmentation combines multiple training examples into a new image-label pair, leading to increased diversity of the augmented set.
- Nonetheless, these approaches are agnostic to image semantics; they ignore object location cues, and as a result may produce ambiguous scenes with occluded distinctive regions.

To account for such shortcomings, can we explicitly use visual saliency for data augmentation?

SAGE Overview

The main idea behind SAGE is to synthesize novel images (with their labels) by blending pairs of training samples, using spatial saliency information as guidance for optimal blending.

SAGE Overview

We define the saliency of each image pixel as its importance in making the correct prediction, using a given vision model. More formally, we are given

- a training sample, \((x, y)\), where \(x \in \mathbb{R}^{d_x \times 1}\) is an RGB image and \(y \in \mathbb{R}^{d_y}\) is the corresponding one-hot label vector, and
- a classifier, \(f_{\theta}(\cdot)\), that is the current partially trained model, and our task loss, \(K(f_{\theta}(x), y)\), measuring the discrepancy between the classifier’s output and the true label.

We define the saliency, \(s \in \mathbb{R}^{d_x}\), as the magnitude of the gradient with respect to the input image,

\[
s(x) = \|\nabla_x f_{\theta}(f_{\theta}(x), y)\|_1,
\]

where \(\|\cdot\|_1\) denotes the l1-norm along the third (color) dimension. In practice, the saliency map tends to focus on the foreground objects useful for classification and ignores irrelevant background.

Computing Saliency Maps

We propose Saliency-guided Mixup to generate two images, \(x_1^0\) and \(x_2\), and their saliency maps, \(s_1^0\) and \(s_1\), we craft a 2D mixing mask, \(M \in \mathbb{R}^{d_x \times d_x}\), and use it to mix the images:

\[
x' = (1 - M) \odot x_1^0 + M \odot x_2,
\]

where \(x' \in \mathbb{R}^{d_x \times 1}\), \(s_1^0\) and \(s_1\) are spatially-normalized and Gaussian-smoothed saliency maps, \(\zeta\) is a scalar hyperparameter used to avoid division-by-zero and \(\odot\) denotes element-wise product.

Optimal Rearrangements via Saliency Maximization

- **Issue:** When the maximally salient regions in both images spatially overlap, the mask, \(M\), tends to suppress one or both objects, which leads to uninformative new scenes.
- **Solution:** Shift one image relative to the other prior to mixing.

Consider a translation operator that shifts a tensor \(z\) by \(r\) pixels as \(T(z, r)\). To quantify how successfully a given rearrangement is in resolving the saliency overlap, we measure the total saliency \(v(\tau) \in \mathbb{R}\) after the rearrangement:

\[
v(\tau) = \sum_{\tau} [M^* \odot s_1^0 + (1 - M^*) \odot T(s_1^0, \tau)],
\]

where \(T(s_1^0, \tau)\) is the saliency \(s_1^0\) translated by \(r\) and \(M^*\) is the mixing mask (Eq. 2) computed with \(s_1^0\) and \(T(s_1^0, \tau)\). Finally, we find the optimal rearrangement, \(\tau^*\), by solving \(v(\tau^*) = \max_{\tau} v(\tau)\), where \(\Omega\) is the space of all possible offsets.

Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Vanilla Mixup</th>
<th>CutMix</th>
<th>Manifold</th>
<th>SaliencyMix</th>
<th>Puzzle Mix</th>
<th>Co-Mixup</th>
<th>SAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-10</td>
<td>ResNet20</td>
<td>96.07</td>
<td>96.61</td>
<td>96.95</td>
<td>97.23</td>
<td>97.51</td>
<td>78.85</td>
<td>79.65</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>PreActResNet18</td>
<td>78.55</td>
<td>79.83</td>
<td>80.03</td>
<td>79.77</td>
<td>80.16</td>
<td>80.73</td>
<td>80.45</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>ResNet29</td>
<td>80.42</td>
<td>80.45</td>
<td>80.57</td>
<td>80.32</td>
<td>80.44</td>
<td>80.37</td>
<td>80.35</td>
</tr>
</tbody>
</table>

Table 1. Image classification accuracy. CIFAR-10 and CIFAR-100 results are obtained by averaging over three independent training runs.

Figure 4. Possible re-arrangements. In each example, the saliency map corresponding to the rearrangement is shown on the left, the corresponding image is on the right. The rearrangement maximizing the total saliency is shown in red.

Figure 5. Robustness and efficiency analysis of SAGE. (a) Robustness vs. standard accuracy in OOD generalization. The methods in the green area improve both accuracy and robustness relative to vanilla augmentation; while the others in red improve standard test accuracy at the cost of decreased robustness. (b) Runtime comparison of SAGE and other baselines. For SAGE, there is no noticeable overhead besides the additional forward and backward pass to compute the saliency map which approximately doubles the time of Vanilla training.