Dissecting Mixed-Sample Data Augmentation Models via Neural-concept Association

Soyoun Won, Sung-Ho Bae, Seong Tae Kim*
Department of Computer Science and Engineering
Kyung Hee University
Yongin-si, South Korea
{sy.won, shbae, st.kim}@khu.ac.kr

***Abstract*—Data augmentation techniques are widely employed in the training of deep neural networks (DNNs), and recent research verifies their effectiveness across diverse tasks. However, their impact on the model's ability to capture semantic concepts has not been widely investigated. In this paper, we analyze models trained with various mixed-sample data augmentation strategies in terms of neural-concept association. Experimental results suggest that mixed sample data augmentation strategies make the model less reactive to semantic concepts.**

***Index Terms*—mixed sample data augmentation, explainable AI, explainability, concept study**

I. Introduction

Mixed sample data augmentation strategies utilize more than one sample to create augmented (mixed) input [22], [27], [28]. These strategies enhance the generalization ability of deep neural networks (DNNs), achieving higher performance in various fields such as classification [5], [27], object recognition [7], semi-supervised learning [4], and self-supervised learning [11], [18]. Furthermore, some studies reported that models trained with mixed sample data augmentation strategies are more adversarially robust [15], [22], [27].

While augmentation strategies' effect on model performance and robustness is widely studied and verified, their effect on interpretability in terms of neural-concept association has not been widely studied. In this paper, we present semantic concept association with the inner units of the model when the model is trained with various data augmentation strategies.

DNNs achieved high performances in various tasks and mixed sample data augmentation strategies even boosted their success. However, because of their black-box nature, it is challenging to understand their decision-making process. A wide range of efforts has been made to make DNN models reliable and interpretable for humans [1]–[3], [12], [14], [20], [21], [29]. A common approach includes feature attribution methods, where importance scores are assigned to input features (e.g., a heatmap where each pixel value is mapped to the important score). Other lines of research include concept-based studies. It aims to give insight into the internal unit's role in recognizing human-perceptible concepts. In this work, we focus on the latter and study the effect of mixed sample data augmentation by analyzing captured semantic concepts by the model.

This work was supported in part by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-RS-2023-00258649) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation, by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (2022-0-00078, Explainable Logical Reasoning for Medical Knowledge Generation), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2021R1G1A1094990). *Dr. Seong Tae Kim is a corresponding author.

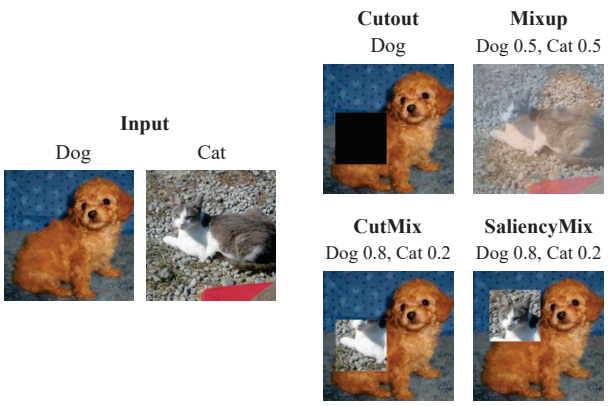

Fig. 1. An overview of the augmented image of Cutout, Mixup, CutMix, and SaliencyMix.

II. Related Work

A. Mixed Sample Data Augmentation

In this study, we explore the effect of mixed sample augmentation on neural-concept association. To this end, we analyze popular mix-based augmentation strategies: Mixup [28] and CutMix [27]. We additionally inspect Cutout [8] and SaliencyMix [22] to understand the effect of CutMix-based augmentation further, as CutMix can be seen as a combination of Cutout and Mixup, and SaliencyMix introduces "saliency guidance" to CutMix. An overview of each method is shown in Fig. 1.

Cutout drops fixed-size square-shaped regions from a randomly chosen position at input space. Dropped regions are filled with zero. Cutout is interpreted as regional dropout, which is a subfield of the regularization techniques. Regional dropout randomly removes continuous regions from an image

or feature space. The main difference between another well-known regional dropout method such as [30] is that Cutout erases a fixed-size box.

Mixup aims to augment both the input image and the label. It mixes two samples by linear interpolation where the beta distribution determines the mix ratio. Various works suggested follow-up Mixup-based augmentation strategies using more elaborate algorithms for mixing [6], [16], [23], [26].

CutMix is inspired by Cutout and Mixup [27]. its algorithm includes removing small square sections from an image, similar to Cutout, and replacing those sections with randomly selected images, like Mixup. The mix ratio is drawn from the beta distribution, and the label is also mixed by the proportion of the augmented image, as in Mixup.

SaliencyMix adds saliency concepts to CutMix. That is, SaliencyMix carefully selects "salient" regions of the source image before attaching them to the target image. Here, the method introduced in [17] is utilized to detect salient regions. Other factors such as mix ratio and label mixing strategies are the same as Mixup and CutMix.

The classification performance (top-1 error) of models trained on architecture ResNet-50 on ImageNet is as follows: SaliencyMix (21.26%), CutMix (21.40 %), Mixup (22.58 %), Cutout (22.93 %), and baseline (23.68 %).

B. Robustness and Interpretability

Other lines of work studied the relationship between adversarial robustness and interpretability of the model. [9] found that adversarially robust models exhibit more interpretable behavior (i.e., their attribution maps are more alike with the input). [24] further explains this observation by considering the decision boundary. They showed that robust models have smoother decision boundaries, and therefore are more interpretable.

However, [9] and [24] defined interpretability as the similarity between an input image and the attribution map. In other words, if the visual pattern of the saliency map resembles the input image, it is considered interpretable. However, this definition of interpretability is an entirely different concept from our work. Our study investigates the interpretability of various models focusing on the semantic representations captured by inner neurons.

III. Concept Study

Human-understandable Concepts are defined using Network Dissection [2], [3]. One can quantify the interpretability of different models by understanding the inner units' roles [13]. Individual units can detect concepts, even though the concepts are not annotated in the training data. For example, units that detect a single concept such as "desk", "computer", or "keyboard" emerge when the input is simply provided as "office". The main idea of Network Dissection is to quantify the disentangled representations learned by individual units of the network. "Concepts" consists of high-leveled ones (e.g., objects) and low-leveled ones (e.g., colors). Network Dissection scans the entire dataset, unit (neuron), and

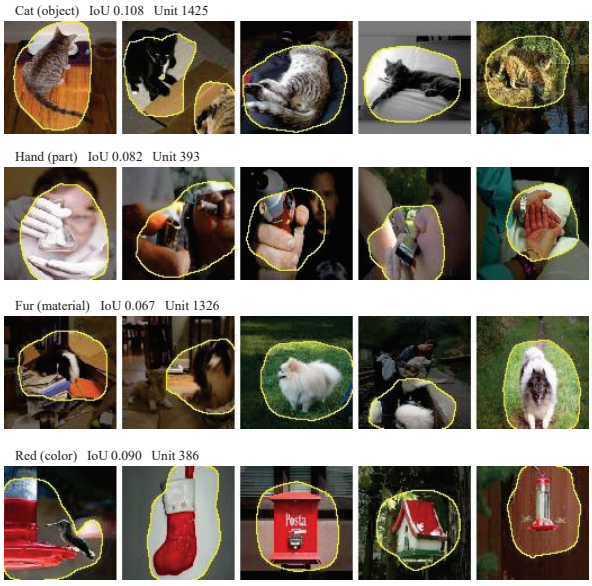

Fig. 2. Detected concepts by network dissection in the baseline (ResNet-50 trained on ImageNet) on each category.

predefined concepts to find detector units. A unit is defined to be a detector of a specific concept if the overlap of the activation of the unit and segmentation annotation of the concept exceeds the threshold. One can understand the role of the inner units by looking at the concepts detected by the units.

Network dissection [2], [3] aims to find disentangled concepts detected by individual units (neurons). It evaluates every unit's activation map for every image in the entire dataset. Firstly, an activation map A_k of unit k is upscaled by S_k to compare it with the input resolution mask for concept c, L_c. Then, a binary mask M_k is created by applying threshold T_k so that only activation above the threshold is set to 1 and others to 0. T_k is determined by the distribution of a unit's activation a_k that satisfies $P(a_k > T_k) = 0.01$. Finally, the intersection over union (IoU) between M_k and L_c is calculated as

$$IoU_{k,c} = \frac{\sum |M_k(\mathbf{x}) \cap L_c(\mathbf{x})|}{\sum |M_k(\mathbf{x}) \cup L_c(\mathbf{x})|}. \quad (1)$$

$IoU_{k,c}$ refers to the accuracy of unit k in detecting concept c. Unit k is defined to be a detector unit of concept c if $IoU_{k,c}$ exceeds a threshold. Examples of detected concepts via individual neurons are shown in Fig. 2.

IV. Experiments

A. Experimental Setup

We introduce the experimental setup used in this study. Utilized models are released by [27] and [22]. All models share the base structure of ResNet-50 [10]. Models are trained on ImageNet [19] with a batch size of 256 for 300 epochs

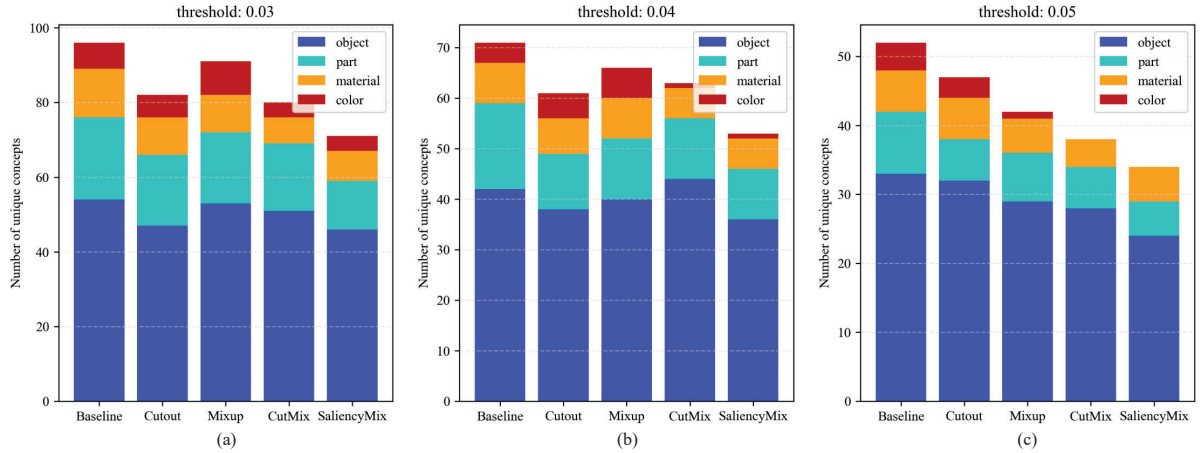

Fig. 3. The number of unique concepts of the baseline and models trained with Cutout, Mixup, CutMix, and SaliencyMix at the last convolutional layer on various IoU thresholds. (a), (b), and (c) shows the experimental results from the IoU threshold 0.03, 0.04, and 0.05, respectively.

with weight decay - the initial learning rate is set to 0.1 and later decayed by the factor of 0.1 at three epoch steps (75, 150, and 225). All models including the baseline are trained with traditional data augmentation strategies such as flipping, cropping, and resizing.

B. Experimental Result

We set the IoU threshold to [0.03, 0.04, 0.05] (i.e., a unit is a detector unit for concept c if $IoU_{k,c} > [0.03, 0.04, 0.05]$). Concepts are divided into four categories - object, part, material, and color. We use the segmentation model UPerNet (Unified Perceptual Parsing Network) [25] trained on the ADE20K dataset [31]. The segmentation model is trained to segment predefined concepts. We evaluate five models trained on ImageNet and probe the final convolutional layer to count the number of unique concepts detected by the detector unit.

The experiment result is shown in Fig. 3. The number of human-recognizable concepts is decreased in models trained with mixed sample augmentation methods on all probed thresholds. When the threshold is set to 0.04, the number of detected unique concepts scored 71, 61, 66, 63, and 53 on the baseline, Cutout, Mixup, Cutmix, and SaliencyMix respectively. Baseline found 42, 17, 8, and 4 detectors on object, part, material, and color respectively. Among the models trained with mixed sample data augmentation, Mixup showed the best results. Cutout and Mixup increase only the number of color detectors of all categories. For CutMix, more object detectors are found than the baseline (42 $\rightarrow$ 44), but fewer detectors are observed with all other categories (12, 6, and 1 on part, material, and color). The smallest number of detectors are found in SaliencyMix in every category (36, 10, 6, and 1 on object, part, material, and color).

A similar tendency is observed for threshold = 0.03, scoring 96, 82, 91, 80, and 71 unique concepts on the baseline, Cutout, Mixup, Cutmix, and SaliencyMix respectively. Mixup increased the number of color detectors than the baseline and SaliencyMix found the smallest number of detectors in all categories except material. Mixup showed the second-best results among probed models and the best results among models that are trained with augmentation methods that utilize more than one sample (i.e., Mixup, CutMix, and SaliencyMix).

However, when the threshold is set higher (0.05), Mixup's ability to capture disentangled concepts dropped to third place. Still, Mixup scored the best among Mixup, CutMix, and SaliencyMix. We hypothesize that CutMix and SaliencyMix only use a small fraction of the source image, thereby less exposed to the whole view of semantic concepts. While Mixup is more successful at detecting disentangled concepts than CutMix and SaliencyMix, it still falls short of the ability compared to the baseline. Mixing algorithms may confuse the classifier to detect semantic concepts.

The interesting thing is that the number of object detectors also decreased on models trained with SaliencyMix (54 $\rightarrow$ 46, 42 $\rightarrow$ 36, and 33 $\rightarrow$ 24 on threshold [0.03, 0.04, 0.05]). It is known that the model classification score is positively correlated with the number of unique object detectors (i.e., the higher the classification score, the more unique object detectors are found). Therefore, the reasonable assumption is that the greatest number of object detectors will be found at SaliencyMix (because SaliencyMix provides the best classification accuracy on ImageNet-trained ResNet-50). However, despite the performance gain, models trained with SaliencyMix have the fewest number of object detectors, presenting the smallest degree of disentanglement.

The decrease of concept detectors in models trained with augmentation strategies despite the performance gain suggests that mixed sample data augmentation strategies are not a panacea - they boost model performance on various tasks but degrade the ability to capture human-perceptible concepts.

V. Conclusion

This paper delves into the overlooked aspects of mixed sample data augmentation strategies: their internal semantic representation. More specifically, models trained with various mixed sample data augmentation are evaluated through Network Dissection, using the number of disentangled concepts. Our experiment revealed that the number of detected disentangled concepts decreased when models were trained with mixed sample augmentation strategies. Data mixing strategy in mix-based methods potentially hinders the emergence of neurons that detect disentangled concepts.

References

[1] Yong Hyun Ahn, Gyeong-Moon Park, and Seong Tae Kim. Line: Out-of-distribution detection by leveraging important neurons. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19852–19862, 2023.

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017.

[3] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020.

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 32, 2019.

[5] Chengtai Cao, Fan Zhou, Yurou Dai, and Jianping Wang. A survey of mix-based data augmentation: Taxonomy, methods, applications, and explainability. *arXiv:2212.10888*, 2022.

[6] Hyeong Kyu Choi, Joonmyung Choi, and Hyunwoo J Kim. Tokenmixup: Efficient attention-guided token-level data augmentation for transformers. *NeurIPS*, 2022.

[7] Ali Dabouei, Sobhan Soleymani, Fariborz Taherkhani, and Nasser M. Nasrabadi. Supermix: Supervising the mixing data augmentation. In *CVPR*, 2021.

[8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[9] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola Schoenlieb. On the connection between adversarial robustness and saliency map interpretability. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1823–1832. PMLR, 2019.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016.

[11] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *NeurIPS*, 33:21798–21809, 2020.

[12] Ashkan Khakzar, Soroosh Baselizadeh, Saurabh Khanduja, Christian Rupprecht, Seong Tae Kim, and Nassir Navab. Neural response interpretation through the lens of critical pathways. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13528–13538, 2021.

[13] Ashkan Khakzar, Sabrina Musatian, Jonas Buchberger, Icxel Valeriano Quiroz, Nikolaus Pinger, Soroosh Baselizadeh, Seong Tae Kim, and Nassir Navab. Towards semantic interpretation of thoracic disease and covid-19 diagnosis models. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 499–508. Springer, 2021.

[14] Ashkan Khakzar, Yang Zhang, Wejdene Mansour, Yuezhi Cai, Yawei Li, Yucheng Zhang, Seong Tae Kim, and Nassir Navab. Explaining covid-19 and thoracic pathology model predictions by identifying informative input features. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 391–401. Springer, 2021.

[15] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. *ICML*, 119:5275–5285, 2020.

[16] Zhijun Mai, Guosheng Hu, Dexiong Chen, Fumin Shen, and Heng Tao Shen. Metamixup: Learning adaptive interpolation policy of mixup with metalearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[17] Sebastian Montabone and Alvaro Soto. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image and Vision Computing*, 2010.

[18] Sucheng Ren, Huiyu Wang, Zhengqi Gao, Shengfeng He, Alan Yuille, Yuyin Zhou, and Cihang Xie. A simple data mixing prior for improving self-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14595–14604, 2022.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[20] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020.

[21] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE international conference on computer vision*, pages 618–626, 2017.

[22] A F M Shahab Uddin, Mst. Sirazam Monira, Wheemyung Shin, TaeChoong Chung, and Sung-Ho Bae. Saliencymix: A saliency guided data augmentation strategy for better regularization. In *International Conference on Learning Representations*, 2021.

[23] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.

[24] Zifan Wang, Matt Fredrikson, and Anupam Datta. Robust models are more interpretable because attributions look normal. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 22625–22651, 2022.

[25] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[26] Wenpeng Yin, Huan Wang, Jin Qu, and Caiming Xiong. Batchmixup: Improving training by interpolating hidden states of the entire minibatch. *Findings of ACL*, pages 4908–4912, 2021.

[27] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[28] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[29] Yang Zhang, Ashkan Khakzar, Yawei Li, Azade Farshad, Seong Tae Kim, and Nassir Navab. Fine-grained neural network explanation by identifying input features with predictive information. *Advances in Neural Information Processing Systems*, 34:20040–20051, 2021.

[30] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[31] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.