Addressing Discrepancies in Semantic and Visual Alignment in Neural Networks

Natalie Abreu [1] **Nathan Vaska** [1] **Victoria Helus** [1]

Abstract

For the task of image classification, neural networks primarily rely on visual patterns. In robust networks, we would expect visually similar classes to be represented similarly. We consider the problem of when semantically similar classes are visually dissimilar, and when visual similarity is present among non-similar classes. We propose a data augmentation technique with the goal of better aligning semantically similar classes with arbitrary (non-visual) semantic relationships. We leverage recent work in diffusion-based semantic mixing to generate semantic hybrids of two classes, and these hybrids are added to the training set as augmented data. We evaluate whether the method increases semantic alignment by testing model performance on adversarially perturbed data, with the idea that it should be easier for an adversary to switch one class to a similarly represented class. Results demonstrate that there is an increase in alignment of semantically similar classes when using our proposed data augmentation method.

1. Introduction

Within the common task of image classification, neural networks must rely on visual patterns in the images. While semantic relationships often follow from visual alignment, visuals and semantics are not always correlated. For instance, in a system that aims to distinguish child-safe objects from hazardous entities, a harmless object such as a spoon may appear visually similar to a dangerous object such as a knife, and a confusion between the two could have harmful implications. This example highlights the idea of mistake severity in neural networks classification – while most performance measures of classification models treat all errors equally, in reality some errors are much more damaging than others. Despite their visual similarity, any instances of confusion between a knife and a spoon would likely cause extreme distrust in a system that is used to discriminate between harmful and safe objects. To address this concern, we propose a data augmentation method to incorporate prior semantic knowledge into the training process. In particular, we focus on the case of when semantic alignment is at odds with visual similarity, as in this case data-driven learning based solely on visual features may fail due to a lack of crucial information on object semantics and class relationships.

With our method, we aim to increase alignment between semantically similar objects despite a lack of visual similarity. To measure this, we consider the metric of mistake severity over perturbed conditions, with the idea being that a model will be more likely to mistake one class for another if the classes are similarly represented.

The contributions of this work are as follows:

- We propose a method of data augmentation using diffusion-based semantic mixing to increase alignment between semantically similar classes
- We construct a dataset with arbitrary class relationships from CIFAR100 data to evaluate our method when visual similarity is at odds with semantic similarity
- We evaluate our method on mistake severity over adversarially perturbed conditions and find that our data augmentation succeeds in increasing alignment between semantically similar classes

2. Related Work

Previous work has considered methods of incorporating semantic information into training, with methods such as introducing hierarchical loss functions ((Bertinetto et al., 2020), (Zhao et al., 2011), (Verma et al., 2012), (Wu et al., 2016)) and aligning classes using adversarial perturbations ((Ma et al., 2021), (Abreu et al., 2022)). The notion of *mistake severity* arises in many of these works as an alternate measure of model robustness, with the idea being that a mistake between classes that are highly dissimilar is worse than a mistake between semantically similar classes. (Bertinetto et al., 2020) notes that the improvement in the metric of mistake severity has been stagnant in recent years and argues that the metric should be revisited.

Of particular interest in (Bertinetto et al., 2020) is a discussion in which the authors randomize the class relationships such that semantic proximity does not reflect visual similarity. In this setting, the performance of the hierarchical methods considered deteriorates, suggesting that the visual similarity of related classes in one's hierarchy is essential to the success of the proposed methods. The authors note, "while one may wish to enforce application-specific relationships using this approach. . . , the effectiveness of doing so may be constrained by underlying properties of the data" ((Bertinetto et al., 2020)). The work in (Abreu et al., 2022) finds similar behavior when visual relationships no longer support semantic ones. We aim to address this dependency on visual similarity in our data augmentation method.

Additionally, there has been prior work in using diffusion models to generate synthetic training data. (Azizi et al., 2023) uses diffusion models to provide synthetic data for image classification. (He et al., 2022) explores the use of synthetic data generated from the text-to-image generation model GLIDE ((Nichol et al., 2022)) in zero-shot and few-shot settings, as well as for model pre-training. They find that synthetic data can be beneficial in these settings, and further investigate strategies to increase data diversity and reduce data noise for synthetic data generation. Similar to our approach, (Trabucco et al., 2023) proposes a diffusion-based data augmentation method. (Trabucco et al., 2023) uses diffusion models to augment individual images to diversify high-level semantic attributes of images; for instance, modifying the appearance of the face of a truck or the landscape of the background. Our work differs in that we apply our augmentation to create semantic hybrids of images rather than to diversify samples of a given class.

In our method, we utilize semantic perturbations of the training samples as a way of incorporating semantic knowledge. Specifically, we use semantic mixing of training samples, a recent task that aims to blend two different concepts to synthesize a new concept. (Liew et al., 2022) present a method called MagicMix to semantically mix concepts based on pre-trained text-conditioned diffusion models. MagicMix does not require any spatial mask or re-training, making it lightweight enough to be applied over a large dataset.

We use adversarial perturbations in our evaluation to provide insight on how the model aligns the representation of classes. Adversarial perturbations, as introduced in (Szegedy et al., 2014), are small perturbations that can change the model's prediction of an image. (Madry et al., 2018) provides an optimization view of adversarial perturbations that allows us to solve for attacks of an l_2-bounded projected gradient descent (PGD) adversary.

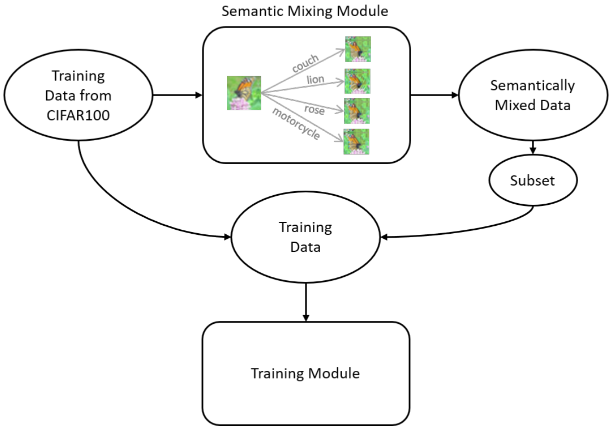

Figure 1. Diagram of method. The semantic mixing module is illustrated with an example used in our experiment, where "Butterfly", "Couch", "Lion", "Rose", and "Motorcycle" are grouped as an arbitrary superclass. In the semantic mixing module, an instance of a butterfly is used to generate new hybrid images by mixing with other concepts in the same superclass. A subset of these mixed images is added to the original clean training data from CIFAR100 to form the final augmented set of training data.

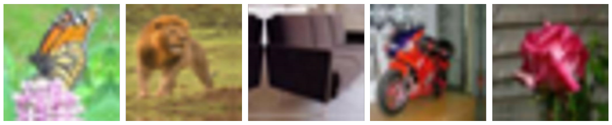

Figure 2. Five images in superclass C, illustrating visual dissimilarity between images in the same "semantic" superclass.

3. Method

We embed semantic knowledge into the training process by incorporating "semantically mixed" data in the training process. Specifically, we propose a data augmentation technique in which the training data is used to generate new training samples which are hybrids of two semantically similar classes. For efficiency, we pre-generate this data using the MagicMix pipeline: for each image in the training set, we generate a new mixed image towards each other class in its superclass (See Figure 1).

We vary the amount of augmented data used in training by specifying a probability of adding an augmented image of the class of any given instance encountered in training. Given the high ratio of augmented data to clean training data, this method allows us to prevent the augmented data from completely dominating the clean data. In our experiments, we refer to "low augmentation" as having a 25% probability of adding an augmented image for any given instance in training, and "high augmentation" as having a 50% probability of adding an augmented image. The augmented image is chosen by randomly selecting a pre-generated image with the same base class as the given instance. Augmented instances are treated 50% as the base class and 50% as the target class when performing optimization.

Superclass	Flowers	Furniture	Insects	Carnivores	Vehicles
A	Orchid	Bed	Bee	Bear	Bicycle
B	Poppy	Chair	Beetle	Leopard	Bus
C	Rose	Couch	Butterfly	Lion	Motorcycle
D	Sunflower	Table	Caterpillar	Tiger	Pickup truck
E	Tulip	Wardrobe	Cockroach	Wolf	Train

Table 1. Chart depicting refactored superclasses. The columns depict the original superclasses (Flowers, Furniture, Insects, Carnivores, Vehicles), and the rows depict the new superclasses (A,B,C,D,E). Note that there is high visual similarity within classes in the same column, but not in the same row.

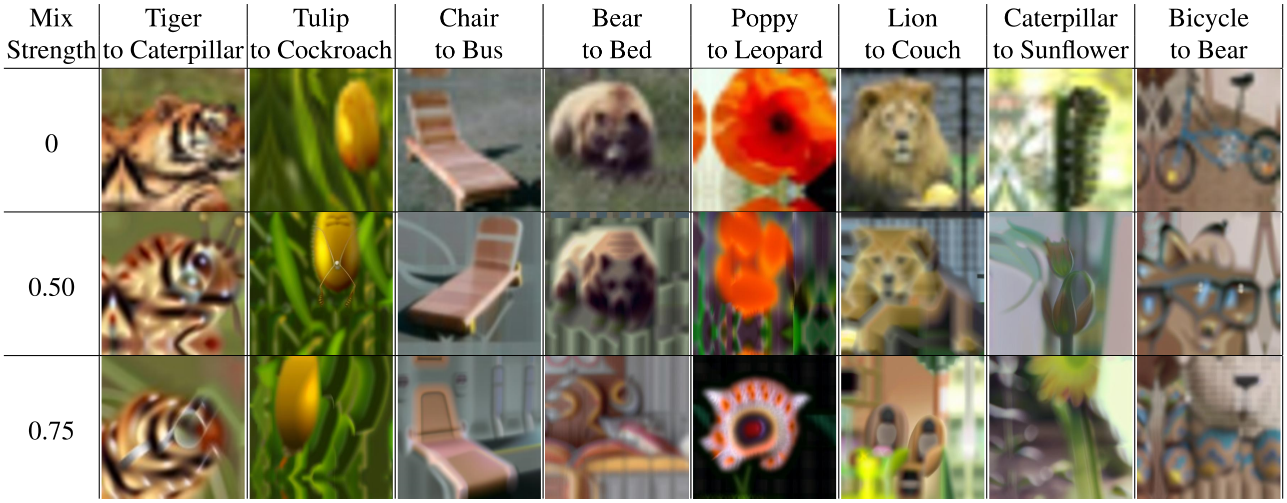

Figure 3. Examples of generated images. The top row, middle row, and bottom row show the original image and images generated with a mix factor of 0.50 and 0.75 respectively.

4. Experiments

To test whether our method could successfully increase alignment with respect to arbitrary class relationships, we formed our training and testing datasets to minimize visual similarity of classes within the same superclass and maxmize visual similarity of classes between different superclasses. We selected five visually dissimilar superclasses from CIFAR100 ((Krizhevsky, 2009)) and redistributed classes such that one class from each original superclass was in each of the new superclasses. Our superclass groupings are shown in Table 1. We will refer to the original superclasses (Flowers, Furniture, Insects, Carnivores, Vehicles) as *visual* superclasses and the new superclasses (A, B, C, D, E) as *semantic* superclasses to avoid confusion.

We pre-generate our hybrid images using the MagicMix pipeline from (Liew et al., 2022) using image-text mixing. For each image in the training set, we create four hybrid images, one hybrid image per each other fine class within the same superclass. The image from the training set is used as the base image and the prompt in the MagicMix module is set to the fine class name in the same superclass. The MagicMix module allows a mix factor in the range $[0, 1]$ to be set to define the strength of the mixing towards the target prompt. We vary the mix factor over models - for low mix strength, we use a mix factor of 0.50 and for high mix strength, we use a mix factor of 0.75. Example images are shown in Figure 3.

All models used a ResNet50 architecture as described in (He et al., 2016) and were trained on the dataset described above. We used a learning rate of 0.1, a batch size of 100, and standard values for remaining training parameters. Additional data augmentation was applied in the form of random cropping and random horizontal flipping. Models were trained for 100 epochs.

We evaluate our method primarily using superclass accuracy on mistakes made by the trained models. Under this metric, models are given credit if their mistakes occur within the correct superclass. In particular, we test models against adversarially-perturbed samples of increasing severity; if a model is not semantically-aligned with respect to the desired semantics, the adversarial perturbation will be able to easily shift the mode's prediction to a class in the incorrect superclass. We also record standard accuracy and semantic superclass accuracy.

Adversarial attacks are modeled with an l_2-bounded pro-

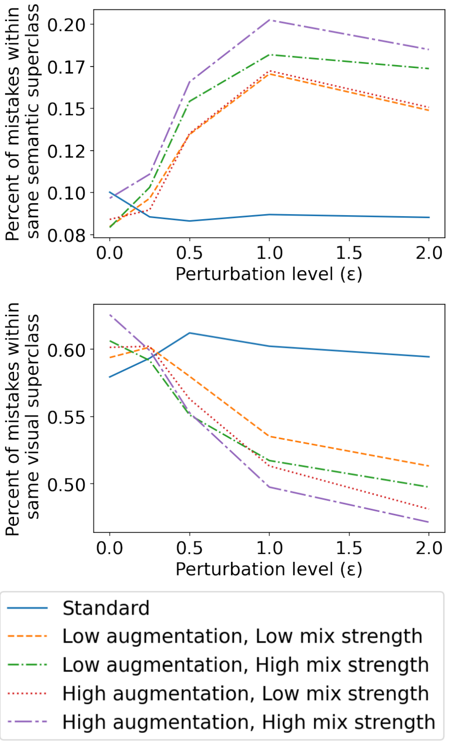

Figure 4. Top: Percent of mistakes that were made within the same semantic superclass. For instance, a tulip getting mistaken for a wardrobe (Class E). Bottom: Percent of mistakes that were made within the same visual superclass. For instance, a tulip being mistaken for a rose (Class Flowers).

jected gradient descent adversary as proposed in (Madry et al., 2018). For a model f with learned parameters θ over a data distribution D and loss function $\mathcal{L}$, we find an adversarial perturbation δ of a given instance x with label y by solving

$$max_{\delta:||\delta||<\epsilon} \mathbb{E}_{(x,y)\sim D}[\mathcal{L}(f_\theta(x+\delta), y)].$$

where ϵ is the l_2 bound of the adversary.

The models we compared were as follows:

- **Standard model**: Model trained with no additional augmented data
- **Low augmentation, low mix strength**: Model trained with 25% additional augmented data and mix strength of 0.5
- **Low augmentation, high mix strength**: Model trained with 25% additional augmented data and mix strength of 0.75
- **High augmentation, low mix strength**: Model trained with 50% additional augmented data and mix strength of 0.5

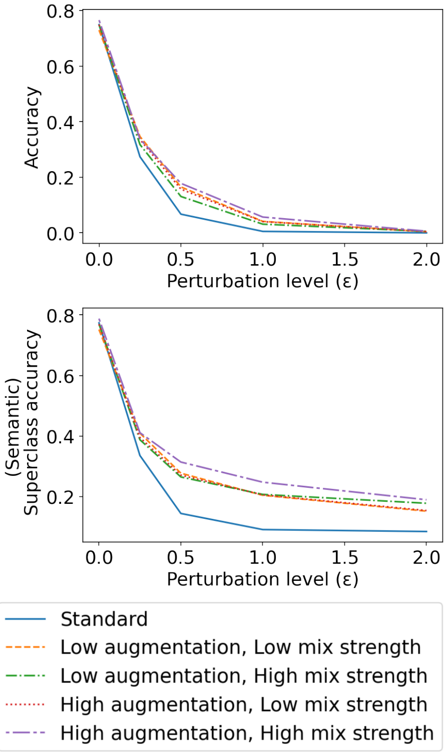

Figure 5. Top: Fine class accuracy over all test instances. Bottom: Superclass accuracy overall (percent of test instances classified as a fine class in the correct semantic superclass).

- **High augmentation, high mix strength**: Model trained with 50% additional augmented data and mix strength of 0.75

5. Results

In this section, we will show results on mistake severity over adversarial perturbations of increasing severity. First, we will demonstrate that the models with our proposed augmentation technique perform better in terms of mistake severity on adversarially perturbed instances. We will additionally demonstrate that our technique decreases mistakes across visually similar classes. These results indicate that our method helps to increase semantic alignment at odds with visual similarities.

The models using our data augmentation technique have considerably higher superclass accuracy on mistakes on perturbed instances than the standard model, as seen in Figure 4. The high data augmentation, high mix strength model performs best overall on this metric, with performance close to the standard model on the clean data and best performance on all nonzero levels of perturbation. To address the similar performance of the standard and data augmentation models on the clean data, we posit that the simplicity of the CIFAR100 dataset causes the models to only makes mistakes on difficult examples (e.g., ones with unique or misleading features) at low levels of perturbation. As the perturbation level increases, the models may start to makes mistakes on examples with more standard features, which offers an explanation to the fact that better performance of the models with data augmentation is only present on more highly perturbed data.

The models using the data augmentation technique additionally have lower percents of mistakes between classes in the same visual superclasses (e.g. ”Flowers”) (shown in Figure 4). This demonstrates that the model learn lower correlations between visually similar classes that are not given as semantically similar.

Finally, we show overall model accuracy and semantic superclass accuracy in Figure 5. The models using the data augmentation technique improve on both metrics over all non-zero levels of adversarial perturbation, and the high augmentation, high mix strength model additionally improves on standard accuracy and superclass accuracy. As the dataset is not very challenging, the improvement of even the best performing model with data augmentation is marginal on clean accuracy and superclass accuracy. As the data set gets more challenging with added perturbation, our method improves on performance as MagicMix distortions help group features of classes in the same semantic superclass. Even at the highest level of perturbation, some semantic alignment is maintained in the models with data augmentation.

6. Discussion and Conclusions

Our findings give promising first results for using data augmentation as a method of increasing semantic alignment between classes with arbitrary visual relationships. More generally, this finding suggests potential for synthetic data to inject prior knowledge into training. As future work, we would like to apply our method to a more complex dataset, where the model is more likely to see ambiguous images or images otherwise more difficult to classify. Additionally, the method could be extended to an application with domain-specific knowledge that needs to be incorporated rather than arbitrary class relationships.

Acknowledgements

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Department of the Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Air Force.

The authors would like to thank Dr. Sanjeev Mohindra, Dr. Lori Layne, and Ms. Paula Donovan for their support and Dr. Rajmonda Caceres for her time and helpful feedback.

References

Abreu, N., Vaska, N., and Helus, V. Addressing mistake severity in neural networks with semantic knowledge. *CoRR*, abs/2211.11880, 2022. doi: 10.48550/arXiv.2211.11880. URL https://doi.org/10.48550/arXiv.2211.11880.

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., and Fleet, D. J. Synthetic data from diffusion models improves imagenet classification. *CoRR*, abs/2304.08466, 2023. doi: 10.48550/arXiv.2304.08466. URL https://doi.org/10.48550/arXiv.2304.08466.

Bertinetto, L., Müller, R., Tertikas, K., Samangooei, S., and Lord, N. A. Making better mistakes: Leveraging class hierarchies with deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 12503–12512. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01252. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Bertinetto_Making_Better_Mistakes_Leveraging_Class_Hierarchies_With_Deep_Networks_CVPR_2020_paper.html.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P. H. S., Bai, S., and Qi, X. Is synthetic data from generative models ready for image recognition? *CoRR*, abs/2210.07574, 2022. doi: 10.48550/arXiv.2210.07574. URL https://doi.org/10.48550/arXiv.2210.07574.

Krizhevsky, A. Learning multiple layers of features from tiny images. pp. 32–33, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Liew, J. H., Yan, H., Zhou, D., and Feng, J. Magicmix: Semantic mixing with diffusion models. *CoRR*, abs/2210.16056, 2022. doi: 10.48550/arXiv.2210.16056. URL https://doi.org/10.48550/arXiv.2210.16056.

Ma, A., Virmaux, A., Scaman, K., and Lu, J. Improving hierarchical adversarial robustness of deep neural networks. *CoRR*, abs/2102.09012, 2021. URL https://arxiv.org/abs/2102.09012.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 2022. URL https://proceedings.mlr.press/v162/nichol22a.html.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Trabucco, B., Doherty, K., Gurinas, M., and Salakhutdinov, R. Effective data augmentation with diffusion models. *CoRR*, abs/2302.07944, 2023. doi: 10.48550/arXiv.2302.07944. URL https://doi.org/10.48550/arXiv.2302.07944.

Verma, N., Mahajan, D., Sellamanickam, S., and Nair, V. Learning hierarchical similarity metrics. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pp. 2280–2287. IEEE Computer Society, 2012. doi: 10.1109/CVPR.2012.6247938. URL https://doi.org/10.1109/CVPR.2012.6247938.

Wu, H., Merler, M., Uceda-Sosa, R., and Smith, J. R. Learning to make better mistakes: Semantics-aware visual food recognition. In Hanjalic, A., Snoek, C., Worring, M., Bulterman, D. C. A., Huet, B., Kelliher, A., Kompatsiaris, Y., and Li, J. (eds.), *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pp. 172–176. ACM, 2016. doi: 10.1145/2964284.2967205. URL https://doi.org/10.1145/2964284.2967205.

Zhao, B., Fei-Fei, L., and Xing, E. P. Large-scale category structure aware image categorization. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 1251–1259, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/d5cfead94f5350c12c322b5b664544c1-Abstract.html.