From Convolution to Attention: a Survey on OOD Detection

Matheus V. Bernat (347152), Pawel Mlyniec (344872), Sepideh Mamooler (273704)
CS-503 Final Project Report

***Abstract*—In this study, we compare how the architectural differences of attention-based and convolution-based models affect the ability of the models to identify out-of-distribution (OoD) data when performing image classification. The approach was to select models with different levels of usage of attention and convolution and compare their performances on identifying OoD data. The key contributions of this work are a semantic OoD dataset based on ImageNet-21k, ImageNet-1K, and Wordnet; a performance comparison of several models on detecting OoD data using softmax, entropy of softmax, and temperature scaling.**

I. Introduction

The presence of false positives in image classification is a reoccurring problem in computer vision applications. This problem makes itself even more present in an open-world setting, where the test images might not belong to the finite set of classes that the model has been trained to recognize. It is, thus, crucial to implement robust errors that avoid the error of classifying an unknown object as if it belonged to a known class.

The particular attention-based architectures, such as Vision Transformers (ViTs) [1], have recently attracted attention in computer vision and achieved comparable results to convolutional neural networks (CNNs) for tasks such as image classification. However, unlike convolution, the role of attention in the performance of classifying OoD data is poorly studied. In this work, we study the role of the architectural differences of attention and convolution in successfully detecting semantic and non-semantic shifts in OoD data. Our experiments show that, in OoD detection tasks, hybrid vision transformers are comparable to ResNet and EcaResNet with around four times as many parameters.

Note that the terminology found in ODIN [2] to separate OoD data into semantic and non-semantic shifts will be used; semantic shifts are synonymous with different classes and non-semantic shifts are synonymous with shifts within the same class, e.g. texture shifts, sketches, etc. Figure 1 shows a clear view of the difference between semantic and non-semantic.

II. Related Work

A. Out-of-distribution detection

Out-of-distribution detection is a well-known yet poorly solved problem when it comes to machine learning models. In [3], Zhang et. al show that ViTs outperform CNNs in generalizability under different groups of distribution shifts,

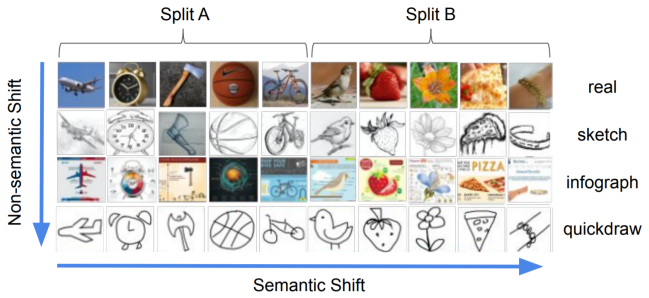

Figure 1: Semantic vs non-semantic shifts in OoD data. Taken from [2].

all of which are *non-semantic*, but none of their experiments takes into account *semantic* shifts. In [4], Fort et. al compare three OoD detection methods including maximum over softmax probabilities (MSP), Mahalanobis distance, and outlier exposure over different models from MLPMixers to ViTs. They categorize OoD data in two groups: far-OoD (CIFAR10 and SHVN) and near-OoD (CIFAR10 and CIFAR100). However, they do not distinguish between semantic and non-semantic distribution shifts. Generalized ODIN [2] uses temperature scaling and input pre-processing to improve softmax probabilities' reliability for OoD detection. However, it does not evaluate attention-based architectures' performance in its experiments. This paper classifies OoD shifts in two groups: non-semantic shifts resulting in samples with the same object as ID data but different styles, and semantic shifts consisting of samples belonging to categories unseen during training. We used the same terminology in this work.

B. Out-of-distribution dataset

ImageNet-O [5] consists of semantically shifted samples collected from ImageNet-21K images classified by ResNet-50 as an ImageNet-1K class with high confidence. However, this data curation strategy makes this dataset biased in favor of models other than ResNet-50. Unlike ImageNet-O, our OoD dataset only depends on the semantic distance of chosen classes from ImageNet-1K classes.

Regarding OoD samples with non-semantic shifts, we found ImageNet-R [6] which is a collection of images from 200 categories of ImageNet-21K but with different styles such as cartoon, graffiti, painting, line drawing, etc. We used this dataset for our experiments around non-semantic shifts.

C. Model comparison

In [7], the authors compare ViTs' robustness to input and model perturbations taking ResNet as a baseline. Despite extensive studies, they only consider non-semantic shifts.

III. Method

The solution that was chosen to approach the problem of understanding how attention and convolution affect a model's ability to detect OoD data was composed of five steps, which will be presented in the sections to follow.

A. Construction of a semantic OoD dataset

To evaluate the models' ability to detect non-semantic OoD data, we needed to construct a non-semantic OoD dataset ourselves. As named in the introduction, ImageNet-O is not an appropriate non-semantic dataset, as it consists of semantically shifted samples collected from ImageNet-21K and classified by Resnet as ImageNet-1K classes with high probability. The classes in ImageNet-O are those that Resnet performs the worst in them, making comparisons between Resnet's and other models' ability to detect OoD detection biased if using ImageNet-O.

The approach used to construct a non-semantic OoD dataset was to choose classes from ImageNet-21K that were not present in ImageNet-1k, and that, in addition, were semantically distant to all the 1000 classes in ImageNet-1K. The semantical distance was measured using the lexical database for the English language, Wordnet [8]. The curated dataset contains roughly 16000 images belonging to 98 different classes.

B. Finding appropriate models

The requirements for the models we wanted were that they were all trained on ImageNet-1K, that they had a similar amount of parameters, and that their architectures belonged to diverse architectures from linear, to convolution, and attention.

To have a full spectrum of model types, we decided to compare models from fully feed-forward through convolutional, ending on the attention-based vision transformer. The first model is MLPMixer-B16 [9], which has only multi-layer perceptrons, that are repeatedly applied across either spatial locations or feature channels. It consists of per-patch linear embeddings, mixer layers, and a classifier head. Our fully convolutional model is ResNet-50 [10]. The main idea behind this model is to introduce a shortcut connection that turns the network into its counterpart residual version. Another model that we considered is ECAResNet, which is a convolution and attention-based architecture [11]. It introduces the efficient channel attention (ECA) module with no channel dimensionality reduction, cross-channel interaction, and having fewer parameters than other attention modules, consisting of 1x1 convolutions layers generating channel weights. This attention mechanism differs from the purely attention-based model tested by us: DeiT-S [12]. DeiT is a convolution-free transformer that was created by distilling information from a convolutional network with the teacher-student approach. It obtains better results than ResNet while having the same amount of parameters. We have chosen DeiT over ViT because ViT's true capability is revealed only when trained on large datasets like ImageNet-21k. Another model candidate was the hybrid vision transformer which consists of ViT on top of ResNet backbone as explained in [13]. However, we were unable to find any version of this model pre-trained on ImageNet-1k with a number of parameters comparable to other candidates. Thus, we kept the four previous models as the *core* models in our work but also considered two versions of the hybrid model: R+Ti/16 (Hybrid_Tiny_ViT), and R26+S/32 (Hybrid_Small_ViT).

To ensure the *core* models are comparable concerning their size and capacity, we choose model versions with the same amount of parameters, as suggested in the robustness study by Bhojanapalli et. al [7]. Table I summarizes the sizes of the models used in our experiments.

C. Image classification on ID and non-semantic OoD data

The chosen models were evaluated on both ID data and non-semantic OoD data, and compared with respect to the precision, recall, and f1-score performance metrics. The ID data is made of the 50000 images of the validation set of ImageNet-1K, while the non-semantic OoD dataset is the ImageNet-R dataset containing 30000 images of 200 different ImageNet-1k classes.

The goal of this experiment is to get a better understanding of how the models perform on ID data, as the performance of a model on ID data is believed to be a strong predictor of the performance on OoD data [14].

D. OoD detection on semantic and non-semantic OoD data

We compared the reliability of the models' softmax probabilities in presence of semantic and non-semantic OoD data based on three different strategies: maximum of raw softmax probabilities, the entropy of softmax probabilities, and maximum of softmax probabilities with temperature scaling [15]. We chose these methods due to their simple and intuitive approach and the fact that they don't require further parameter optimization. In each of the settings, the performance metrics used to compare the models are the area under the ROC curve and false positive rate when the true positive rate is 90%.

In the first setting, we use the maximum of raw softmax probabilities. Ideally, the maximum probability obtained by the model should be on average high for ID samples and low for OoD ones, i.e the model should be more certain of its prediction for ID samples than OoD samples. As a result by thresholding this probability, we can evaluate the reliability of the model's certainty of its prediction in presence of OoD data.

Model	MLPMixer-B16	ResNet-50	ECAResNet-50D	DeiT-S	R+Ti/16	R26+S/32
# Params	22M	25M	25M	22M	6.4M	36.6M

Table I: Parameters by models

In a second setting, we used temperature scaling as done in ODIN [15]. ODIN combines temperature scaling with input pre-processing which requires further optimization of pre-processing parameters. Here we only use temperature scaling to calibrate the softmax results. Ideally, the temperature used for calibration should be optimized for each model. However, due to our limited computational resources we used the same temperature for all models as done in ODIN experiments [15]. Similar to the first setting the prediction is based on maximum softmax probability. Equation 1 shows how the softmax probability is computed using temperature scaling for the i^{th} class among N classes, given input $\mathbf{x}$ and temperature T.

$$S_i(\mathbf{x}; T) = \frac{exp(f_i(\mathbf{x})/T)}{\sum_{j=1}^{N} exp(f_j(\mathbf{x})/T)} \quad (1)$$

Finally, inspired by Tent [16], we used entropy of softmax probabilities. Higher entropy indicates higher uncertainty. So we expect the entropy of softmax probabilities to be higher for OoD data. Tent minimizes entropy using input pre-processing which requires optimization. Here we only use the entropy of the resulting probability distribution.

IV. Experiments

In this section, the experiments performed in this study are presented. We used pre-trained weights provided in the timm library [17] for all models. In all experiments with temperature scaling, we followed the same setting as ODIN [15] and used $T = 1000$.

A. Image classification on ID data

Table II shows the performance metrics precision, recall, and f1-score for the four *core* and the hybrid models trained on ImageNet-1k when performing image classification on the 50000 images of the validation set of ImageNet-1k.

In this experiment, ECAResNet has the highest overall precision, recall, and f1-score among *core* models. Remember that ECAResNet has both convolution and channel attention layers as parts of its architecture.

B. Image classification on non-semantic OoD data

Table III shows the performance metrics of the four models when running image classification on ImageNet-R.

Observe that the precision of all models is roughly three times larger than the corresponding recall. This is explained by the fact that the models are generally reticent in correctly classifying (non-semantically) shifted images since the distribution shifts have not been seen during training. The precision metrics are high since the model has seen similar objects during training, and some of the shifts might still be close to the non-shifted objects. When it comes to the recall metrics, its low values are explainable since the model misses to correctly classify non-semantically shifted images since their modalities (sketch, graffiti, etc.) have not been seen during training.

Models \ Metrics	Precision ↑	Recall ↑	f1-score ↑
MLPMixer	0.761	0.757	0.754
ResNet	0.762	0.749	0.747
ECAResNet	0.800	0.794	0.792
DeiT	0.777	0.770	0.766
Hybrid_Tiny_ViT	0.672	0.650	0.649
Hybrid_Small_ViT	0.810	0.806	0.804

Table II: Performance metrics for models trained on ImageNet-1k when classifying the 50000 validation images of ImageNet-1k. ↑ indicates higher values are better. ↓ indicates lower values are better.

Another observation to draw from this experiment is that, similarly as in the evaluation of ID data, ECAResNet delivers the highest overall metrics compared to the three other *core* models.

Models \ Metrics	Precision ↑	Recall ↑	f1-score ↑
MLPMixer	0.734	0.249	0.348
ResNet	0.803	0.283	0.396
ECAResNet	0.804	0.309	0.423
DeiT	0.783	0.299	0.409
Hybrid_Tiny_ViT	0.694	0.148	0.227
Hybrid_Small_ViT	0.831	0.344	0.461

Table III: Performance metrics for the models when fed with the non-semantic OoD data of ImageNet-R.

C. OoD detection using semantic OoD data

In this experiment, we perform a binary classification, where a *positive* prediction means the image is *in-distribution* and a *negative* means the object is *out-of-distribution*. We fed the models with 66000 images, of which 16000 are semantic OoD data and 50000 are ID data from the validation set of ImageNet-1K.

The prediction values for each image – which will be then be compared to a threshold to get positive or negative predictions – were calculated as one of the following three approaches:

- the max probability of the raw softmax probabilities;
- the entropy of the softmax probabilities;
- the max probability of the temperature scaled softmax probabilities.

See Figure 2 for the comparison of the models' ROC curve when fed with both ID and semantic OoD data using the three approaches listed above. Comparing the area under the curve (AUC) for the *core* models, one can easily conclude that the ResNet and ECAResNet models perform the best and that DeiT performs the worst no matter which approach is chosen. Moreover, despite having fewer number of parameters, the tiny hybrid ViT obtains similar performance to ECAResNet when entropy of softmax is used.

In addition to the ROC curves, we plotted the box plots for models' predictions for ID and OoD samples separately. Ideally, we expect the probability distribution over all classes to have higher maximum and lower entropy for ID samples. Table VII shows the difference between the mean of the prediction values generated by the three methods when the models are fed with ID and semantic OoD data.

As shown in Table VII, among the four *core* models, ResNet has the best performance while DeiT the worst. In addition, the tiny hybrid ViT outperforms EcaResNet with four times as many parameters. These results are very close to what was obtained by comparing models based on their FPR when $TPR = 90\%$. The box plots can be found in the appendix VII-A.

	Difference of means		
Models \ Methods	Raw softmax↑	Entropy↓	Temp. scaling↑
MLPMixer	0.209	-1.092	1.030e-06
ResNet	0.299	-1.792	1.774e-06
ECAResNet	0.226	-1.024	7.799e-07
DeiT	0.149	-0.652	7.777e-07
Hybrid_Tiny_ViT	0.228	-1.092	2.445e-06
Hybrid_Small_ViT	0.325	-1.579	3.722e-06

Table IV: Difference between the mean of the prediction values of ID data and the mean of the prediction values of *semantic* OoD data. The greater the *absolute difference*, the clearer is the separation between ID and semantic OoD.

	Difference of means		
Models \ Methods	Raw softmax↑	Entropy↓	Temp. scaling↑
MLPMixer	0.164	-0.661	5.284e-07
ResNet	0.308	-2.046	2.045e-06
ECAResNet	0.237	-1.252	9.759e-07
DeiT	0.140	-0.617	7.455e-07
Hybrid_Tiny_ViT	0.280	-1.516	3.429e-06
Hybrid_Small_ViT	0.330	-1.805	4.287e-06

Table V: Difference between the mean of the prediction values of ID data and the mean of the prediction values of *non-semantic* OoD data. The greater the *absolute difference*, the clearer is the separation between ID and non-semantic OoD.

D. OoD detection using non-semantic OoD data

Similarly to the experiment described in IV-C, we perform binary classification of the images where a *positive* prediction means the image is *in-distribution* and a *negative* means the object is *out-of-distribution*. The only difference compared to the experiment in IV-C is that the used OoD dataset is ImageNet-R.

	FPR at 90% TPR ↓		
Models \ Metrics	Raw softmax	Entropy	Temp. scaling
MLPMixer	0.609	0.687	0.618
ResNet	0.617	0.572	0.627
ECAResNet	0.578	0.691	0.606
DeiT	0.683	0.755	0.712
Hybrid_Tiny_ViT	0.718	0.578	0.670
Hybrid_Small_ViT	0.494	0.474	0.392

Table VI: For semantic OoD data: false positive rates (FPR) at 90% true positive rate (TPR). The smaller the FPR at $TPR = 90\%$ the better.

	FPR at 90% TPR ↓		
Models \ Metrics	Raw softmax	Entropy	Temp. scaling
MLPMixer	0.624	0.931	0.653
ResNet	0.552	0.596	0.529
ECAResNet	0.548	0.670	0.537
DeiT	0.668	0.803	0.694
Hybrid_Tiny_ViT	0.615	0.497	0.468
Hybrid_Small_ViT	0.488	0.522	0.326

Table VII: Non-semantic shift, FPR at $TPR = 90\%$. The smaller the FPR at $TPR = 90\%$ the better.

Figure 3 shows the comparisons of the models' ROC curves and the respective areas. Similar to the results obtained in IV-C, among *core* models, ResNet and ECAResNet have the greatest AUC and thus perform best in this setting where we wish to maximize the true positive rate and minimize the false positive rate. Also, DeiT performs the worst together with MLPMixer. In addition, the tiny hybrid ViT outperforms EcaResNet and has similar performance to ResNet with four times as many parameters.

V. Conclusion and Limitations

To conclude, we curated a dataset of images with ImageNet-21K categories that are not present in ImageNet-1K and that are within a chosen semantical distance from all classes in ImageNet-1K using the tree hierarchy of Wordnet. This dataset can be used as a semantic OoD dataset for models trained on ImageNet-1k. In addition, over a range of models from purely linear to purely convolutional and purely attention-based, we used softmax probabilities in three different ways to detect OoD samples: maximum of raw softmax probabilities, entropy of softmax probability distribution, and maximum of temperature-scaled softmax probabilities. Our results show that ResNet and ECAResNet outperform MLPMixer and DeiT in presence of both semantic and non-semantic OoD samples. We also found that, apart from MLPMixer, models detect non-semantic shifts easier than semantic shifts. Finally, our experiments show

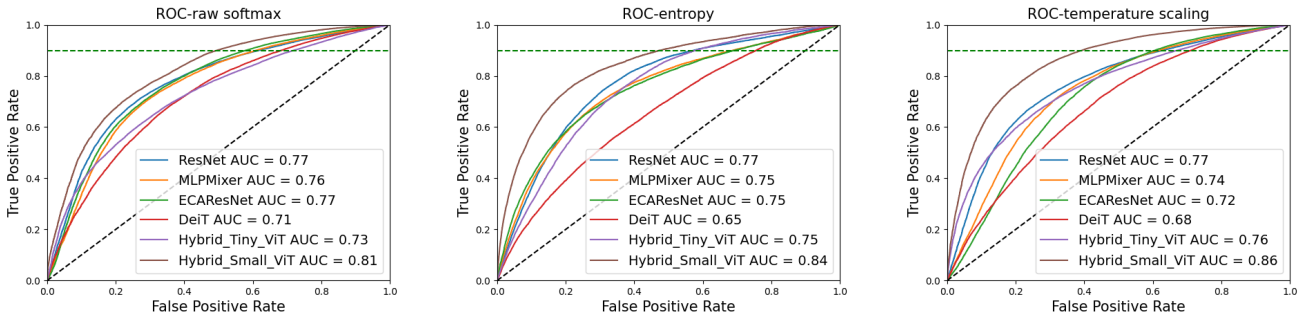

Figure 2: ROC curves of the models when fed with ID and *semantic* OoD data. The black dash line represents random guessing, and the green horizontal line illustrates $TPR = 90\%$.

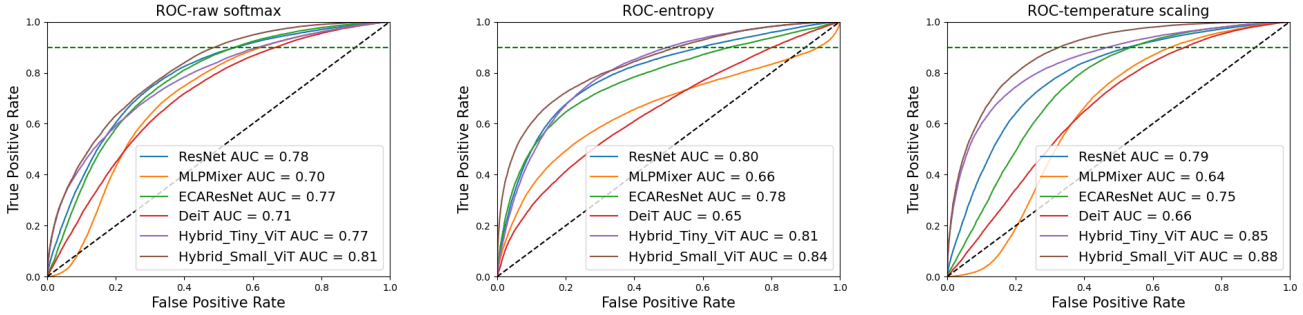

Figure 3: ROC curves of the models when fed with ID and *non-semantic* OoD data. The black dashed line represents random guessing, and the green horizontal line illustrates the true positive rate 90%.

that hybrid vision transformers perform as good as ResNet and EcaResNet with four times as many parameters in OoD detection tasks. Nevertheless, based on our results, no conclusion can be made about the superiority of convolution or attention-based architectures in presence of OoD data at inference time. Further studies on a more diverse set of models should be conducted. In addition, evaluating the models on ImageNet-1k test set instead of the validation set would result in a more realistic understanding of models' performance as the stopping criteria of the training process might depend on the validation set. However, this should not affect how models compare to one another.

Due to our limited computational resources, we did not use OoD detection methods that require optimization. Future work can take into account such methods. Moreover, A comparison between uni-modal (image only) and multi-modal (image and text) models' performance in presence of OoD data would be a valuable contribution to this work.

VI. Individual contributions

Bernat constructed the semantic OoD dataset using Wordnet, assisted Mamooler with the code for the evaluation of the models for ID data, and implemented the performance metrics of the semantic OoD evaluation. Mamooler found the hybrid models, implemented the evaluations on the semantic and non-semantic OoD data, ran the OoD detection experiments, and studied how the predictions using temperature scaling and entropy perform compare to softmax. Mlyniec found the the *core* models, set up the clusters, and led the work related to parallelizing the evaluation code. All the authors contributed equally to the conception of this report.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[2] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 948–10 957, 2020.

[3] C. Zhang, M. Zhang, S. Zhang, D. Jin, Q. Zhou, Z. Cai, H. Zhao, S. Yi, X. Liu, and Z. Liu, "Delving deep into the generalization of vision transformers under distribution shifts," 2021.

[4] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," *ArXiv*, vol. abs/2106.03004, 2021.

[5] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," *CVPR*, 2021.

[6] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," *ICCV*, 2021.

[7] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," *ArXiv*, vol. abs/2103.14586, 2021.

[8] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, p. 39–41, nov 1995. [Online]. Available: https://doi.org/10.1145/219717.219748

[9] A. K. L. B. X. Z. T. U. J. Y. A. S. D. K. J. U. M. L. A. D. Ilya Tolstikhin, Neil Houlsby, "Mlp-mixer: An all-mlp architecture for vision," *ArXiv*, 2021.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[11] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," 2020.

[12] M. D. F. M. A. S. H. J. Hugo Touvron, Matthieu Cord, "Training data-efficient image transformers distillation through attention," 2020.

[13] A. Steiner, A. Kolesnikov, , X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.

[14] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," 2021.

[15] R. Sjögren and J. Trygg, "Odin: Outlier detection in neural networks," 2018.

[16] D. Wang, E. Shelhamer, S. Liu, B. A. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *ICLR*, 2021.

[17] R. Wightman, "Pytorch image models," https://github.com/rwightman/pytorch-image-models, 2019.

VII. Appendix

A. Prediction Distribution

In this section, we provide the box plots of prediction distributions for different models and OoD detection methods. They were used to evaluate the assumption that, in general, maximum and entropy of softmax probabilities are respectively higher and lower for ID samples.

Figures 4, 5, and 6 illustrate the box plots of model prediction when using raw softmax probabilities, entropy, temperature scaling respectively for semantic distribution shifts. Figures 7, 8, and 9 illustrate the box plots of model prediction when using raw softmax probabilities, entropy, temperature scaling respectively for non-semantic distribution shifts.

The further apart the means of the prediction values of the ID and OoD data are, the fewer errors the model will commit when the prediction values are compared to a threshold. Ideally, there would be no overlap between the prediction values of ID and OoD data. Comparing the medians of the prediction values given by the models when evaluating ID and OoD data would be more robust to outliers, but such a comparison has not been made.

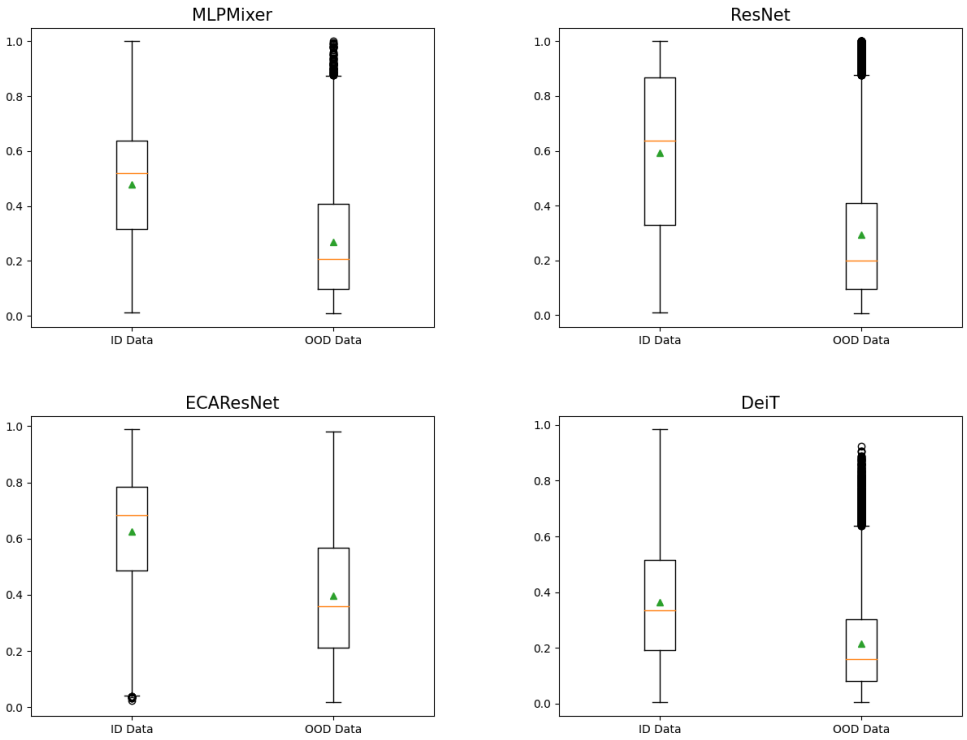

Figure 4: Maximum probability distribution of the models when fed with ID and *semantic* OoD data.

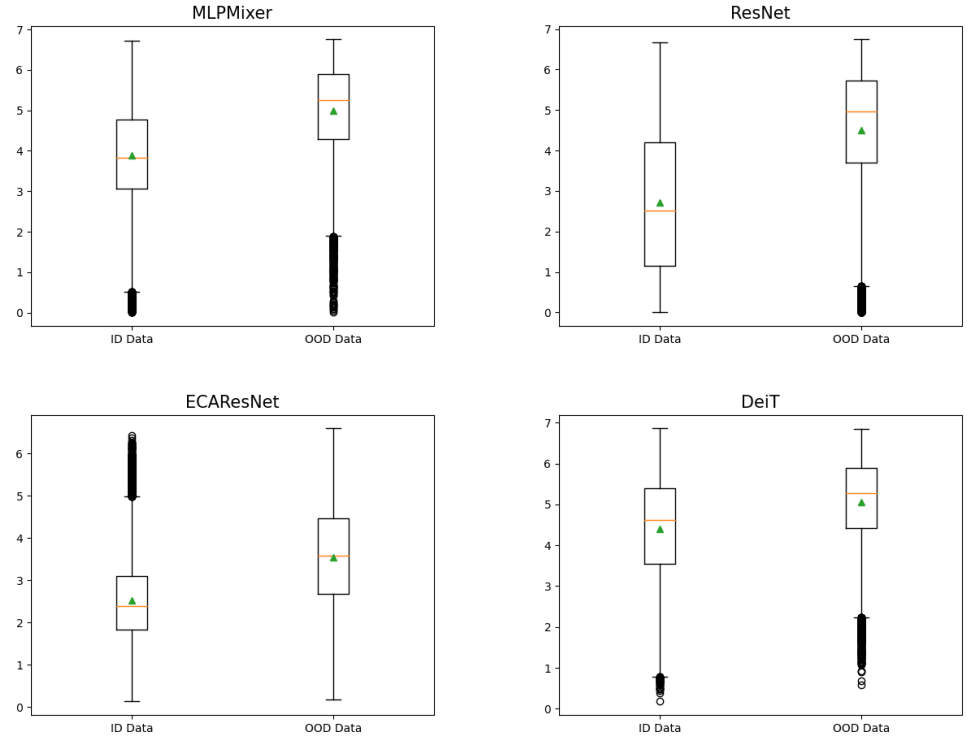

Figure 5: Entropy distribution of the models when fed with ID and *semantic* OoD data.

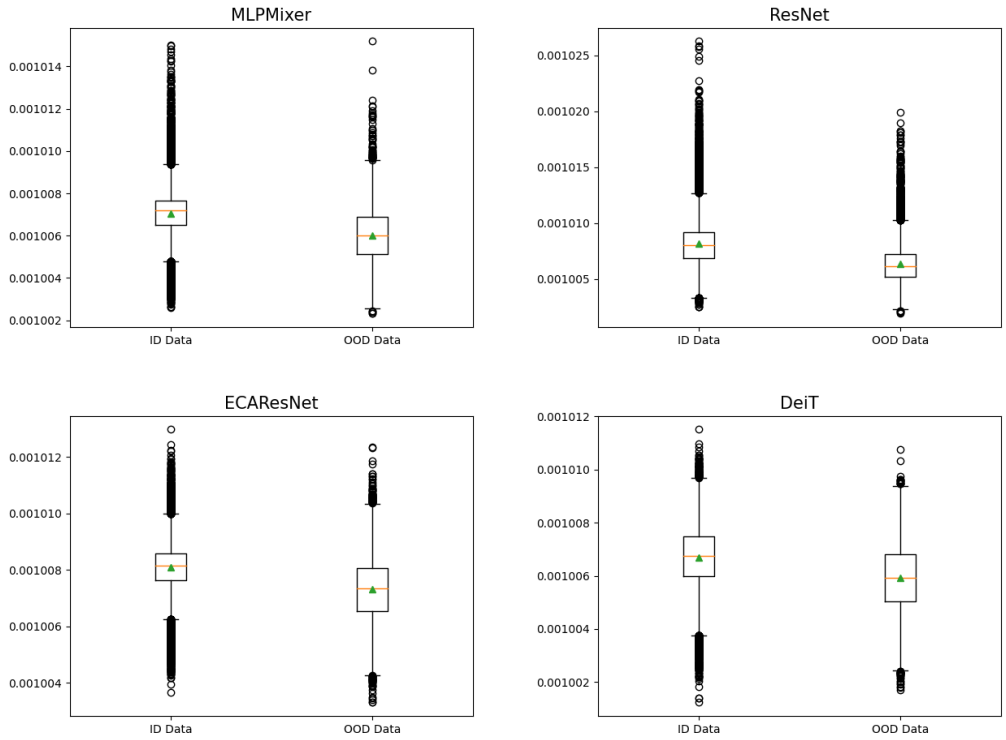

Figure 6: Maximum temperature scaled probability distribution of the models when fed with ID and *semantic* OoD data.

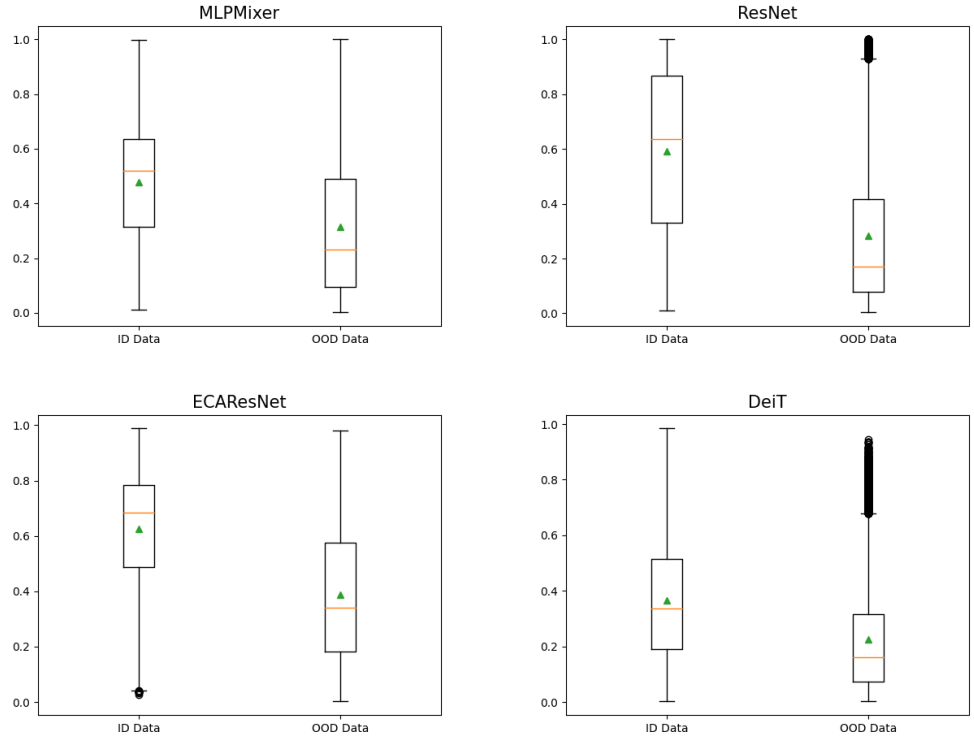

Figure 7: Maximum probability distribution of the models when fed with ID and *non-semantic* OoD data.

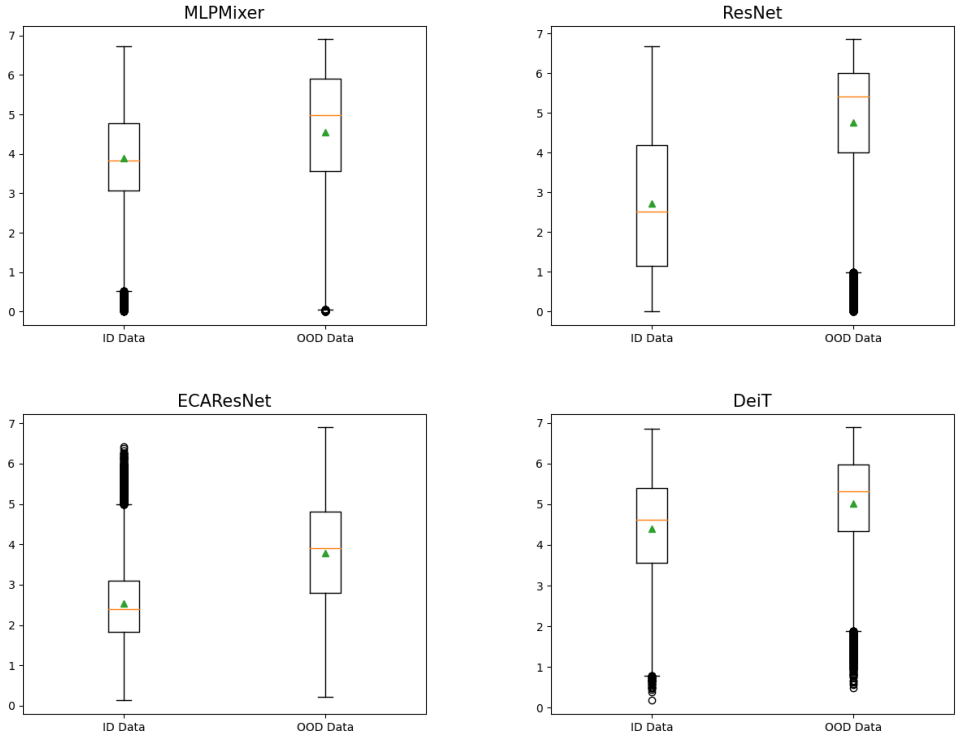

Figure 8: Entropy distribution of the models when fed with ID and *non-semantic* OoD data.

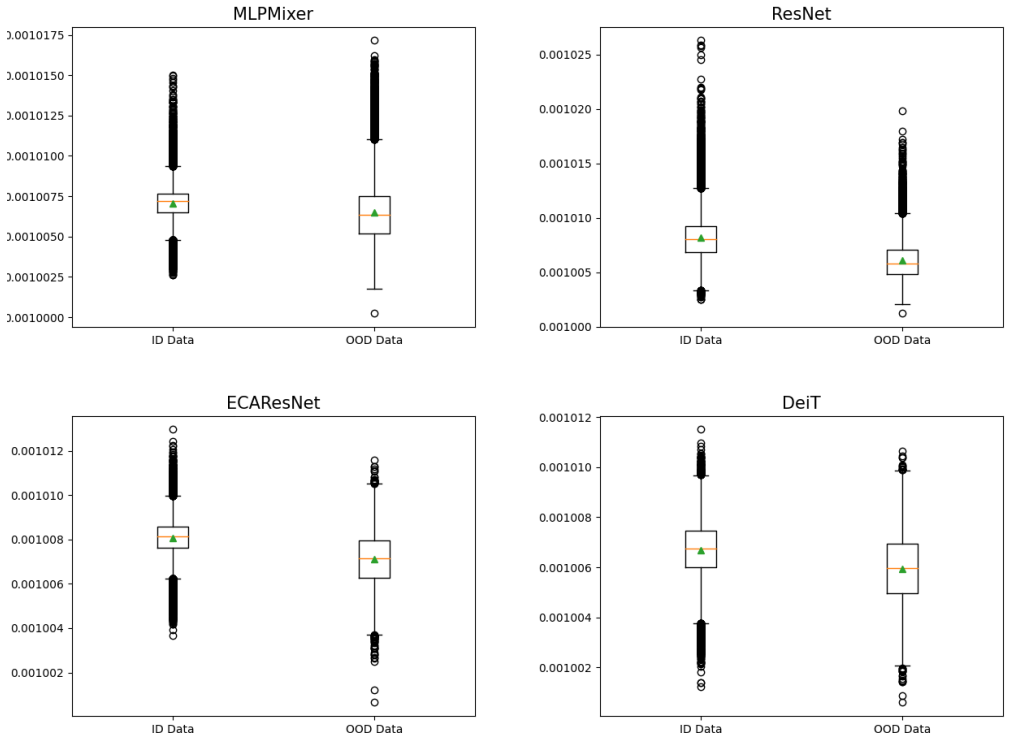

Figure 9: Maximum temperature scaled probability distribution of the models when fed with ID and *non-semantic* OoD data.