Multitask Learning With Self-Supervision to Improve Robustness

Zixi Wang zw2774, Sixuan Wang sw3513

Dec 20, 2021

Abstract

While deep neural networks achieves good accuracy in many applications, they remain vulnerable under adversarial attacks. In our paper, we investigate the architecture of multitask learning, specifically multitask learning combined with self-supervised learning, and its affect on model robustness. We use Contrastive Learning as the self-supervised task that is trained in parallel with the main supervised task. Our experiment shows that training with auxiliary Contrastive Learning task yields a robust accuracy of 60% in the main classification task. Moreover, we compare the effects of a Supervised auxiliary task and a Contrasitve auxiliary task, and the result shows that Contrastive auxiliary task provides much larger robustness improvement than the Supervised auxiliary task. These two experiments together show that Contrastive Learning provides a strong regularization when trained together with a Classification task, forcing the model to learn a more robust image representation.

1 Introduction

Deep learning models have shown great success at many tasks[1] such as image classification, NLP, time series forecasting. While deep learning models achieve high accuracy in these tasks, they are still brittle under adversarial attacks. As deep learning models become more and more prevalent in real-world applications, especially high-stake applications such as Automatic Driving, Medicines, Laws, etc., it becomes increasingly important to ensure the security and the robustness of the model.

A growing body of research has been dedicated to answering the cause and mitigation of adversarial attacks on deep neural networks. Some common approaches towards model robustness include Robust Optimization[2, 3, 4], Robust Regularization[5, 6], and Online Defense [7].

Another line of investigation[8] on multitask learning sheds light on a new perspective of model robustness: multitask learning, besides improving the performance of specific tasks[9], also increases model robustness. Mao et al [8] theoretically showed that increasing output dimensionality improves the robustness of the entire model. Moreover, they also empirically show that multitask learning improves the model robustness both when a single task is attacked or several tasks are simultaneously attacked.

In our paper, we take this idea further and investigate the role of self-supervised learning when trained in conjunction with a supervised task. Recent work shows that unlabeled data can improve adversarial robustnes[1]. Another work[7] shows that contrastive learning

representations can be used to detect adversarial examples at inference time. Self-supervised learning[10, 11, 12, 13, 14] trains models on unlabeled data in a supervised manner by utilizing self-generated labels from the data itself. Without using labels, self-supervised learning has the potential to learn more robust representations from the images themselves.

Our work verifies this potential by training Self-supervised learning tasks and Supervised-learning tasks in parallel. We will formulate our model set up in Section 3. Our experiments in Section 4.1 shows that Self-supervised learning tasks serve as a strong regularization when trained in parallel with the traditional classification tasks, forcing the model to learn a more robust representation of the input images. By combining Self-supervised tasks and supervised tasks in a Multi-task model, we are able to achieve a 60% marginal improvement on the model's accuracy against single-task PGD attack.

Moreover, in Section 4.2, we also compare the robust accuracy of Supervised exclusive multitask model and Self-supervised + Supervised multitask model. This comparison allow us to further verify that Self-supervised learning can improve robustness more than a Supervised learning task. Self-supervised learning task provides the robustness boost not only from an additional task, but also from a task of its noise-insensitive nature.

2 Related Work

2.1 Adversarial Attacks

The adversarial attack can be categorized into black-box, white-box and grey-box[1] according to the threat model's knowledge of the adversaries. In this paper, we will focus on white-box attack, which means the adversaries will have full knowledge of the target model, including the model structure and model parameters, which the adversaries can rely on to generate adversarial samples. Some famous white-box attack like the fast gradient sign method(FGSM)[3], the basic iterative method(BIM)[15] and the projected gradient descent(PGD)[2]. In this paper, we will use PGD attack to evaluate the adversarial robustness of our model.

2.2 Adversarial Robustness

Adversarial training improves models' robustness against attacks, where the training data is augmented using adversarial samples[2]. In combination with adversarial training, later works[5] achieve improved robustness by regularizing the feature representations with additional loss, which can be viewed as adding additional tasks. Despite the improvement of robustness, adversarially trained models lose significant accuracy on clean (unperturbed) examples[2]. Moreover, generating adversarial samples slows down training several-fold, which makes it hard to scale adversarial training to large datasets. Our method enhances robustness by naturally training a multitask model with a main Image Classification task and an auxiliary Self-Supervised task. While reaching a high robust accuracy, our method does not involve adversarial training. Instead, through multitask learning, the auxiliary tasks serve as an "regularization" that forces the models to learn robust representations, although not explicitly regularizing against adversarial examples as in [5, 6].

2.3 Multi-task Learning

Multi-task learning is a training paradigm where machine learning models are trained simultaneously, using shared represenations to learn the common ideas between a collection of related tasks[16]. These shared representations increase data efficiency and can potentially yield faster learning speed for related and downstream tasks. Multi-task learning has been used successfully across all applications of machine learning, from PLT to computer vision. MTL comes in many guises: learning jointly, learning with auxiliary tasks, learning to learn. Recent work shows input reconstruction as a self-supervised task improves robustness. It connects the vulnerability of adversarial robustness under attacks with multitask learning and hints towards a new direction of research. Chengzhi Mao and Amogh Gupta[8] show that deep networks are vulnerable partly because they are trained on too few tasks.

While previous work incorporates a self-supervised learning task into adversarial training[9], or leverages adversarial defense at inference[17], we added stronger self-supervised learning tasks into a shared-weights multitask learning architecture, and investigated how do self-supervised algorithms improve robustness for the architecture.

2.4 Self-supervised Learning

Self-supervised learning allows us to learn high quality representations from images without annotations. It can learn generalizable representations from unlabeled data without annotation, and then will be able to self-generate a supervisory signal exploiting implicit information. The contribution of self-supervised learning is two-fold: on the one hand, it can utilize high quality unlabeled data[13]. On the other hand, it can potentially extract representations that capture the underlying structure of such data, which helps to improve model performance and the convergence speed of the downstream tasks. Recent work also shows that these self-supervised learned representations can be used to detect if a image is under adversarial attack[7]. This sheds a new light on self-supervised learning's role in Robust Machine Learning. We take this notion to a different track and use the representations learned from self-supervised task to regularize Image Classification tasks and enhance adversarial robustness.

3 Formulation

3.1 Multitask Learning Objective

Notations In this work, we utilize multitask learning with shared parameters[18, 19, 20, 21, 22], where all the tasks share the same back-bone network $F(\cdot)$ as a feature extractor. Then on top of the backbone feature extractor, for every task c, we add a task-specific prediction head $H_c(\cdot)$. Let x denote an input example.

Main Task, Base Model The base model consists of only the main task, which is a single classification task that uses the backbone feature extractor $F(\cdot)$ and a single prediction head $H(\cdot)$. Let y denote the ground-truth label for the task. The objective of the single-task model is formulated as:

$$L(x, y) = L(H(F(x)), y) \tag{1}$$

Supervised Multitask Model The supervised multitask model consists of the Main Task and other auxiliary supervised tasks. The model uses one backbone feature extractor $F(\cdot)$ and a prediction head $H_s(\cdot)$ for every task s. Let y_s denote the corresponding ground truth label for task s. Then for every task, the task-specific loss is formulated as:

$$L_s(x, y_s) = l(H_s(F(x)), y_s) \tag{2}$$

where l is the appropriate loss function. For simplicity, we denote $(y_1, ..., y_M)$ as $\bar{y}$, where M is the number of tasks. The total loss for the multitask model is the weighted sum of all the individual losses:

$$L_{all}(x, \bar{y}) = \sum_{s=1}^{M} \lambda_s L_s(x, y_s) \tag{3}$$

Contrastive Multitask Model The Contrastive Multitask Model consists of the Main Task and a Contrastive Learning Task[12]. Specifically, we add a classification prediction head $H_s(\cdot)$ and a Contrastive Learning prediction head $H_c(\cdot)$ on top of the base encoder $F(\cdot)$.
In Contrastive Learning, we create two views of each image x, denoted as x_i and x_j. Let $sim(u, v) = u^T v/\|u\|\|v\|$ denote the dot product between l_2 normalized u and v (i.e.cosine similarity). Then the loss function for a positive pair of examples (x_i, x_j) is defined as

$$l_{c,(x_i,x_j)} = -\log \frac{\exp\left(sim(z_i, z_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp\left(sim(z_i, z_k)/\tau\right)} \tag{4}$$

where $z_i = H_c(F(x_i))$, $\mathbb{1}_{[k\neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $k \neq i$, and τ denotes a temperature parameter.
Since every input image x is augmented into two views x_i and x_j, the main classification task need to predict the correct labels y_i, y_j for both of the views x_i, x_j, where $y_i = y_j = y$, y being the ground-truth label for image x. The total loss for the Contrastive Combined Multitask model is:

$$L_{comb}(x, y) = \lambda_c\Big(l_{c,(x_i,x_j)} + l_{c,(x_i,x_j)}\Big) + \lambda_s\Big(L(x_i, y_i) + L(x_j, y_j)\Big) \tag{5}$$

3.2 Adversarial Single-task Attack Objective

In this work, we focus on single-task attacks as formulated in [8]. We focus on attacking the main classification task in order to investigate the effects of different auxiliary tasks on the main classification task's robustness. In general, given an input example x, the objective function for single-task attacks against models with multiple outputs is the following:

$$\underset{x_{adv}}{\operatorname{argmax}}\, L_s(x_{adv}, y_s), \;\; \textbf{s.t.} \|x_{adv} - x\|_p \leq r. \tag{6}$$

4 Experiments

We analyze the effects that the auxiliary tasks have on the main tasks' robustness. We do this by constructing two case studies: one with CIFAR10 and one with CIFAR100. In the first experiment, we look at how contrastive learning task affects the main task's robustness. In the second experiment, we compare the two Multitask models: the Supervised Multitask

Model and the Contrastive Multitask Model, and compare which auxiliary task can bring larger marginal increase in the main task's robustness.

4.1 Implementation Details

For both of the experiments, We use a ResNet18[12, 23] as the backbone encoder $F(\cdot)$, a batch size = 256, an Adam optimizer with learning rate = 0.0003 and weight decay = 1e-4, a cosine annealing scheduler after the first 10 epochs, and train for 100 epochs. For adversarial attack, we use a 7-step PGD attack with $\epsilon = 8/225, \alpha = 2/225$ on the Main Classification task.

4.2 Experiment 1: CIFAR-10

4.2.1 Architecture

In this experiment, we compare the robustness against single-task adversarial attack of two architecture: Single-Task Base Model (Figure 1) and Contrastive Multitask Model (Figure 2). In the Sintle-Task Base Model, we only learn one main task, that is the Image Classifiacation of CIFAR10[24].

In the Contrastive Multitask Model, we learn the Image Classification task and the Contrastive Learning task in parallel. The two task have a shared-weight encoder and two separate fully connected prediction head. The shared-weight encoder is the same base encoder as the Single-Task model, ResNet18.

In Contrastive Learning, we generate two views of the original Image and minimize the similarity between the two views. Therefore in order to train Image Classification and Contrastive Learning in conjuncture, we also need to generate two identical labels for the two transformed images, and learn to classify both transformed images correctly. Another thing worth noting here is that the output dimension of Contrastive Learning is 128, according to the original SimCLR paper[12], whereas the output of the Image Classification task is 10, since the CIFAR-10 dataset has 10 classes.

4.2.2 Experiment Result

For both models, we trained for 100 epochs and test the natural accuracy and the robust accuracy of the Image Classification Task against PGD attack. The full result is shown in Table 1. We can see that while natural accuracy of the second model drops by a margin of 10%, the robust accuracy reaches a very high level of 60%. This results even reaches the state-of-the-art robust defenses against CIFAR10 adversarial examples. This improvement in Image Classification robustness in the Contrastive Multitask Model shows a preliminary support for our hypothesis: Contrastive Learning provides a strong regularization for the learned representation. Since Contrastive Learning aims to learn representations of the images that hold across different transformations without label, it is very likely that, when trained in parallel with Image Classification, it forces the model to learn a more general and robust representation of the images, one that is less sensitive to the original images' non-robust noises correlated with the labels, and one that the Image Classification head uses to make more robust predictions.

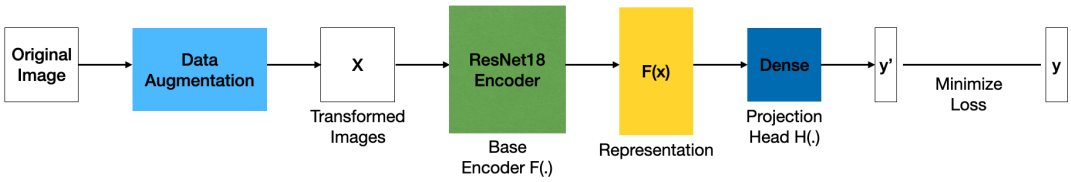

Figure 1: Single-Task Base Model

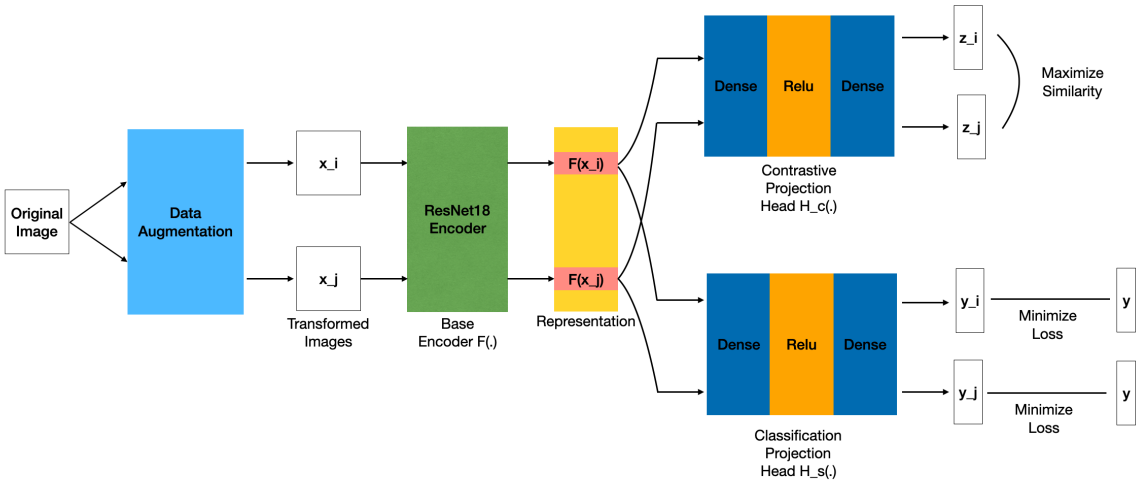

Figure 2: Contrastive Multitask Model

Image Classification	Single-Task	Contrastive Multitask
Natural Accuracy	82.46 %	73.16 %
Robust Accuracy (FGSM)	2.89 %	58.61 %
Robust Accuracy (PGD-7)	0.04 %	60.62 %
Robust Accuracy (PGD-20)	0.00 %	59.95 %
Robust Accuracy (PGD-50)	0.00 %	59.89 %

Table 1: Natural and Robust Accuracy of Image Classification Tasks for Single-Task Model and Contrasitve Multitask Model

4.3 Experiment 2: CIFAR-100

In this experiment, we have two objectives. First, we want to re-do the analysis above on a larger dataset CIFAR100, and verify if similar result still holds. More importantly, we want to disentangle the robustness improvement of Contrastive Multitask learning from Supervised Multitask Learning. That is, we want to verify that the robustness improvement we observe in Experiment 1 is largely coming from the Contrasitively Learned representations, rather than the sheer result of adding additional tasks, as shown in[8].

4.3.1 Architecture

In additional to the Single-Task Model and the Contrastive Multitask Model, we propose a third architecture, called the Supervised Multitask Model, shown in Figure 3. In this model, instead of using a Contrastive Learning task as the auxiliary task, we add another Image Classification task to parallelize with the main task.

The auxiliary Image Classification task is to predict the "super-class" that each image in CIFAR-100 belongs to, namely, Coarse Classification. The CIFAR-100 dataset has 100 classes containing 600 images each. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. Each image comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs)[25]. The classes and their corresponding superclasses are shown in Figure 4. The Coarse Classification task predicts the "Coarse" labels and the main Image Classification task predicts the "Fine" labels, or the original labels. Each of the two classification tasks has its own fully connected prediction head, with Coarse Classification an output dimension of 20 and Fine Classification an output dimension of 100. The two prediction head share a ResNet18 Encoder, like in the previous experiment.

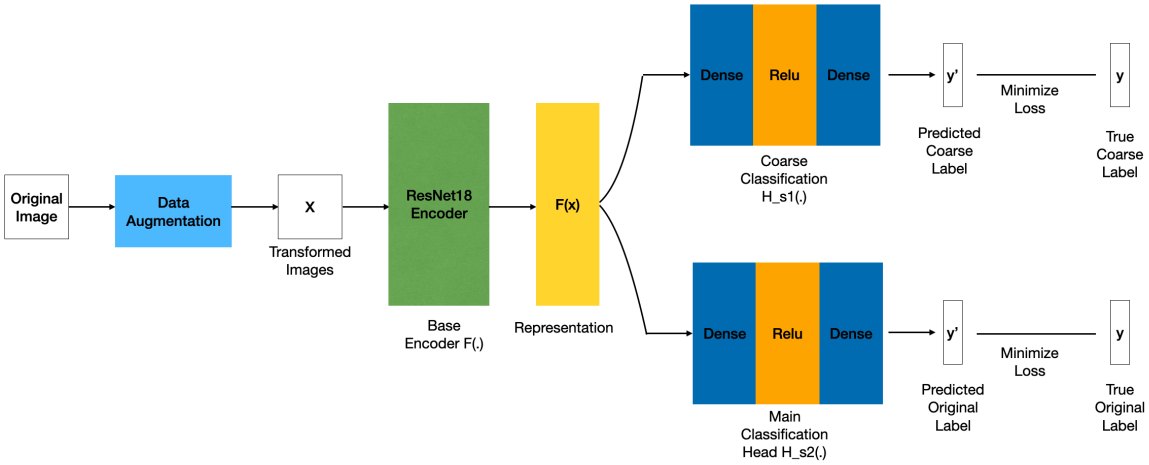

Figure 3: Supervised Multitask Model

4.3.2 Experiment Result

For each model, we again trained for 100 epochs and test the natural accuracy and robust accuracy of the Main Image Classification test (in the third model it's the "Fine" Classification). The result is shown in Table 2. Note that while the Supervised Multitask model yields a robustness increase of 25%, the Contrastive Multitask model yields a larger increase of 42%. This result further shows that Contrastive Learning can provide extra robustness boost other than simply being an auxiliary task.

Superclass	Classes
aquatic mammals	beaver, dolphin, otter, seal, whale
fish	aquarium fish, flatfish, ray, shark, trout
flowers	orchids, poppies, roses, sunflowers, tulips
food containers	bottles, bowls, cans, cups, plates
fruit and vegetables	apples, mushrooms, oranges, pears, sweet peppers
household electrical devices	clock, computer keyboard, lamp, telephone, television
household furniture	bed, chair, couch, table, wardrobe
insects	bee, beetle, butterfly, caterpillar, cockroach
large carnivores	bear, leopard, lion, tiger, wolf
large man-made outdoor things	bridge, castle, house, road, skyscraper
large natural outdoor scenes	cloud, forest, mountain, plain, sea
large omnivores and herbivores	camel, cattle, chimpanzee, elephant, kangaroo
medium-sized mammals	fox, porcupine, possum, raccoon, skunk
non-insect invertebrates	crab, lobster, snail, spider, worm
people	baby, boy, girl, man, woman
reptiles	crocodile, dinosaur, lizard, snake, turtle
small mammals	hamster, mouse, rabbit, shrew, squirrel
trees	maple, oak, palm, pine, willow
vehicles 1	bicycle, bus, motorcycle, pickup truck, train
vehicles 2	lawn-mower, rocket, streetcar, tank, tractor

Figure 4: CIFAR-100 Classes and Superclasses

Image Classification	Single-Task	Supervised Multitask	Contrastive Multitask
Natural Accuracy	48.99 %	50.47%	44.37 %
Robust Accuracy (FGSM)	2.77 %	23.20%	32.49 %
Robust Accuracy (PGD-7)	0.37 %	25.64%	42.62 %
Robust Accuracy (PGD-20)	0.18 %	23.59%	33.92 %
Robust Accuracy (PGD-50)	0.13 %	23.49%	33.53 %

Table 2: Natural and Robust Accuracy of Image Classification Tasks for the Three Architectures

5 Conclusions

Through our experiments, we have shown that the Contrastive Learning, while trained in parallel with the main Supervised Task, can provide strong regularization to the model, forcing the model to learn a robust representation of images. Moreover, we have also disentangle the robustness improvement brought by Contrastive Learning from the improvement brought by adding another supervised task. This comparison shows that training Contrastive Learning in parallel with Supervised Learning can provide not only robustness improvement by adding another task, but also additional improvement from the noise-insensitive nature of Contrastive Learning.

References

[1] Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020.

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2016.

[5] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[6] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy, 2019.

[7] Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, and Carl Vondrick. Adversarial attacks are reversible with natural supervision, 2021.

[8] Chengzhi Mao, Amogh Gupta, Vikram Nitin, Baishakhi Ray, Shuran Song, Junfeng Yang, and Carl Vondrick. Multitask learning strengthens adversarial robustness. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 158–174. Springer, 2020.

[9] Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, and Carl Vondrick. Adversarial attacks are reversible with natural supervision. *CoRR*, abs/2103.14222, 2021.

[10] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization, 2016.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[13] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[15] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2016.

[16] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017.

[17] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 699–708, 2020.

[18] Iasonas Kokkinos. Ubernet: Training a 'universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory, 2016.

[19] Trevor Standley, Amir Roshan Zamir, Dawn Chen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? *CoRR*, abs/1905.07553, 2019.

[20] Tyler Lee and Anthony Ndirango. Generalization in multitask deep neural classifiers: a statistical physics approach, 2019.

[21] Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning, 2015.

[22] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding, 2019.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[24] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

[25] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research).