

---

# On Adversarial Training without Perturbing all Examples

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Adversarial Training (AT) is the de-facto standard for improving robustness against  
2 adversarial examples. This usually involves a multi-step adversarial attack applied  
3 on each example during training. In this paper, we explore only constructing  
4 Adversarial Examples (AEs) on a subset of the training examples. That is, we  
5 split the training set in two subsets  $A$  and  $B$ , train models on both  $(A \cup B)$  but  
6 construct AEs only for examples in  $A$ . Starting with  $A$  containing only a single  
7 class, we systematically increase the size of  $A$  and consider splitting by class and by  
8 examples. We observe that: (i) adv. robustness transfers by difficulty and to classes  
9 in  $B$  that have never been adv. attacked during training, (ii) we observe a tendency  
10 for hard examples to provide better robustness transfer than easy examples, yet find  
11 this tendency to diminish with increasing complexity of datasets (iii) generating  
12 AEs on only 50% of training data is sufficient to recover most of the baseline AT  
13 performance even on ImageNet. We observe similar transfer properties across tasks,  
14 where generating AEs on only 30% of data can recover baseline robustness on the  
15 target task. We evaluate our subset analysis on a wide variety of image datasets  
16 like CIFAR-10, CIFAR-100, ImageNet-200 and show transfer to SVHN, Oxford-  
17 Flowers-102 and Caltech-256. In contrast to conventional practice, our experiments  
18 indicate that the utility of computing AEs varies by class and examples and that  
19 weighting examples from  $A$  higher than  $B$  provides high transfer performance.

## 20 1 Introduction

21 Imperceptible changes in the input can change the output of a well performing model dramatically.  
22 These so-called Adversarial Examples (AEs) have been the focus of a large body on deep learning  
23 vulnerabilities of works since its discovery [1]. To date, Adversarial Training (AT) [2, 3] and its  
24 variants [4–6] is the de-facto state-of-the-art in improving the robustness against AEs. Essentially, AT  
25 generates adversarial perturbations for all examples seen during training. While adversarial training  
26 is known to transfer robustness to downstream tasks [7–9] and that robustness is distributed unevenly  
27 across classes [10, 11], common practice dictates that AT “sees” adversarial examples corresponding  
28 to the whole training data, including all classes and concepts therein. This is independent of whether  
29 only adversarial robustness is optimized or a trade-off between robustness and clean performance  
30 is desired [12]. This also holds for variants that treat individual examples differently [13–15] or  
31 adaptively select subsets to attack during training to reduce computational overhead [16, 17]. It is  
32 largely unclear how adversarial robustness is affected when training is limited to seeing adversarial  
33 examples only on specific subsets of the training data.

34 To shed light on this issue, we consider the adversarial training setup depicted in figure 1, called  
35 Subset Adversarial Training (SAT), where we split the training data into two subsets  $A$  and  $B$ , train  
36 the model conventionally on the union  $(A \cup B)$ , but generate AEs only on examples from  $A$  (indicated

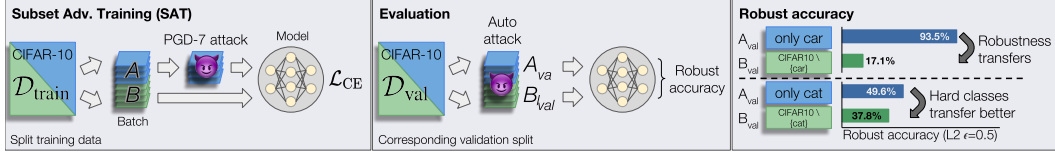


Figure 1: Adversarial robustness transfers among classes. Using Subset Adversarial Training (SAT), during which only a subset of all training examples ( $A$ ) are attacked, we show that robust training even on a single class provides robustness transfer to all other, non adv. trained, classes ( $B$ ). E.g., SAT for  $A=cat$ , we observe an robust accuracy of 37.8% on  $B$ . Noteworthy is the difference of transfer utility between classes. I.e.  $A=car$  provides very little transfer to  $B$  (17.1%). We investigate this transfer among classes and provide new insights for robustness transfer to downstream tasks.

37 by the emoji). For example, we can split training data by class, with  $A = \{car\}$  or  $A = \{cat\}$  and  
 38  $B = A^c$ , and investigate how adversarial robustness transfers. Surprisingly, we observe significant  
 39 adversarial robustness on  $B_{val}$  at test time, the degree of which depends on the class(es) in  $A$ . Of  
 40 course,  $A$  and  $B$  can be arbitrary partitions of the training data. For example, we could put only  
 41 “difficult” examples in  $A$  during training. At test time, we evaluate overall adversarial robustness  
 42 (since there is no natural split into  $A_{val}$  or  $B_{val}$ ). These experiments reveal a rather complex interaction  
 43 of adversarial robustness between classes and examples.

44 Our analysis provides a set of **contributions** revealing a surprising generalizability of robustness  
 45 towards non-adv. trained classes and examples even under scarce training data setups. **First**, selecting  
 46 subsets of whole classes, we find that SAT provides transfer of adversarial robustness to classes which  
 47 have never been attacked during training. E.g. only generating adversarial examples for class *car* on  
 48 CIFAR-10, achieves a non-trivial robust accuracy of 17.1% on all remaining CIFAR-10 classes (see  
 49 figure 1, right). **Secondly**, we observe classes and examples that are hard to classify do generally  
 50 provide better robustness transfer than easier ones. I.e. class *cat* achieves more than twice the robust  
 51 accuracy on the remaining classes (37.8%) over class *car* (17.1%). **Thirdly**, SAT with 50% of  
 52 training data is sufficient to recover the baseline performance with vanilla AT even on hard datasets  
 53 like ImageNet. Lastly, we observe similar transfer properties of SATed models to downstream tasks.  
 54 In this setting, exposing the model to only 30% of AEs during training, can recover baseline AT  
 55 performance on the target task.

## 56 2 Related Work

57 Since their discovery [1], robustness against adversarial examples has mainly been tackled using  
 58 adversarial training [18, 2, 4]. Among many others, prior work proposed adversarial training variants  
 59 working with example-dependent threat models [19, 13–15], acknowledging that examples can have  
 60 different difficulties. Some works also mine hard examples [16] or progressively prune a portion of  
 61 the training examples throughout training [17, 20]. However, all of these methods generally assume  
 62 access to adversarial examples on the whole training set. That is, while individual examples can  
 63 be dropped during training or are treated depending on difficulty, the model can see adversarial  
 64 perturbations for these examples if deemed necessary. Adversarial training is also known to transfer  
 65 robustness to downstream tasks [8, 9, 7] and adversarially robust representations can be learned  
 66 in a self-supervised fashion [21]. Here, a robust backbone is often adapted to the target task by  
 67 re-training a shallow classifier – sometimes in an adversarial fashion. It is generally not studied  
 68 whether seeing adversarial examples on the whole training set is required for good transfer. This is  
 69 despite evidence that achieving adversarial robustness is easier for some classes/concepts than for  
 70 others [22, 23, 11, 10], also for robustness transfer [24]. Complementing these works, we consider  
 71 only constructing adversarial examples on a pre-defined subset of the training set, not informed by  
 72 the model or training procedure, and study how robustness transfers across examples and tasks.

### 73 3 Background and Method

#### 74 3.1 Adversarial Training (AT)

75 It is a well known fact that conventional deep networks are vulnerable to small, often imperceptible,  
76 changes in the input. As mitigation, AT is a common approach to extend the empirical risk minimiza-  
77 tion framework [2]. Let  $(x, y) \in \mathcal{D}_{\text{train}}$  be a training set of example and label pairs and  $\theta$  be trainable  
78 parameters, then AT is defined as:

$$\min_{\theta} \mathbb{E}_{(x,y) \in \mathcal{D}_{\text{train}}} \left[ \max_{\|\delta\|_2 \leq \epsilon} \mathcal{L}(x + \delta, y; \theta) \right], \quad (1)$$

79 where  $\delta$  is a perturbation that maximizes the training loss  $\mathcal{L}$  and thus training error. The idea being  
80 that, simultaneously to minimizing the training loss, the loss is also optimized to be stable within a  
81 small space  $\epsilon$  around each training example  $\|\delta\|_2 \leq \epsilon$  (we consider the  $L_2$  norm). This additional  
82 inner maximization is solved by an iterative loop; conventionally consisting of 7 or more steps. In  
83 some settings [18, 12, 4], the robust loss is combined with the corresponding loss on clean examples  
84 in a weighted fashion to control the trade-off between adversarial robustness and clean performance.

#### 85 3.2 AT without Perturbing all Training Examples

86 Most proposed AT methodologies generate AEs on the whole training set. This being also valid for  
87 methods which adaptively select subsets [16, 17] during training or more traditional AT in which  
88 only a subset per batch is adversarially attacked. These methods do not guarantee the exclusion of  
89 examples, that is, the model is likely to see an AE for every example in the training set. From a broader  
90 perspective, the necessity to generate AEs exhaustively for all classes appears unfortunate though.  
91 Ideally, we desire robust models to be scalable, i.e. transfer flexibly from few examples and across  
92 classes to unseen ones [25]. We propose SAT to investigate to what extent AT provides this utility.  
93 To formalize, let  $A$  be a training subset and  $B$  contain the complement:  $A \subset \mathcal{D}_{\text{train}}, B = \mathcal{D}_{\text{train}} \setminus A$ .  
94 Then SAT applies the inner maximization loop of AT on the subset  $A$  only; on  $B$  the conventional  
95 empirical risk is minimized:

$$\min_{\theta} \mathbb{E}_{(x,y) \in \mathcal{D}_{\text{train}}} \left[ w_A \mathbb{1}_{(x,y) \in A} \max_{\|\delta\|_2 \leq \epsilon} \mathcal{L}(x + \delta, y; \theta) + w_B \mathbb{1}_{(x,y) \in B} \mathcal{L}(x, y; \theta) \right], \quad (2)$$

96 where  $\mathbb{1}_{(x,y) \in A}$  is 1 when the training example is in  $A$  and 0 otherwise.  $w_A$  and  $w_B$  define optional  
97 weights, which are by default both set to 1. Note that this is different from balancing robust and clean  
98 loss as discussed in [18, 12, 4], where the model still encounters adversarial examples on the whole  
99 training set.

100 **Loss balancing.** The formulation in equation 2 implies an imbalance between left and right loss as  
101 soon as the training split is not even ( $|A| \neq |B|$ ). To counteract, we assign different values to  $w_A$  and  
102  $w_B$  based on their subset size. E.g., to equalize the loss between both subsets, we assign  $w_B = 1$   
103 and  $w_A = |B|/|A|$ . We will utilize this loss balancing to improve robustness for transfer learning in  
104 section 4.3.

#### 105 3.3 Training and evaluation recipes

106 Consider the depiction of SAT in figure 1. Prior to training, the training set is split into  $A$  and  $B$   
107 (left). For evaluation (middle), we split the validation set into a corresponding split of  $A_{\text{val}}$  and  $B_{\text{val}}$ ,  
108 if possible. For **Class-subset Adversarial Training (CSAT)**, this split aligns with the classes on the  
109 dataset:  $A$  and  $B$  are all training examples corresponding to two disjoint sets of classes while  $A_{\text{val}}$   
110 and  $B_{\text{val}}$  are the corresponding test examples of these classes. As experimenting with all possible  
111 splits of classes is infeasible, we motivate splits by class difficulty where we measure difficulty by  
112 the average entropy of predictions per class – introduced as  $\mathcal{H}_C$  in the next paragraph. In contrast,  
113 we can also split based on individual example difficulty. We provide empirical support for this  
114 approach in the experimental section 4. Additionally, example difficulty has been frequently linked to  
115 proximity between decision boundary and example [26, 13, 15, 16, 27]. The closer the example is  
116 to the boundary, the harder it is likely to classify. The hypothesis: hard examples provide a larger

117 contribution to training robust models, since they optimize for large margins [13, 14]. We refer to  
 118 this experiment as **Example-subset Adversarial Training (ESAT)**. In contrast to CSAT, however,  
 119 there is no natural split of the test examples into  $A_{\text{val}}$  and  $B_{\text{val}}$  such that we evaluate robustness on  
 120 the whole test set (i.e.,  $\mathcal{D}_{\text{val}}$ ).

121 As difficulty metric, we utilize entropy over softmax, which we empirically find to be as suitable as  
 122 alternative metrics (discussed in the supplement). Consider a training set example  $x \in \mathcal{D}_{\text{train}}$  and a  
 123 classifier  $f$  mapping from input space to logit space with  $N$  logits. Then the entropy of example  $x$  is  
 124 determined by  $\mathcal{H}(f(x))$  and of a whole class  $C \subset \mathcal{D}_{\text{train}}$  is determined by  $\mathcal{H}_C(f)$  – the average over  
 125 all examples in  $C$ :

$$\mathcal{H}(f(x)) = - \sum_{i=1}^N \sigma_i(f(x)) \cdot \log \sigma_i(f(x)), \quad \mathcal{H}_C(f) = \frac{1}{|C|} \sum_{x \in C} \mathcal{H}(f(x)),$$

126 where  $\sigma$  denotes the softmax function. For our SAT setting, we rank examples prior to adversarial  
 127 training. This requires a classifier pretrained on  $\mathcal{D}_{\text{train}}$  enabling the calculation of the entropy. To  
 128 strictly separate the effects between entropy and AT, we determine the entropy using a non-robust  
 129 classifier trained without AT. Similar to [27], we aggregate the classifier states at multiple epochs  
 130 during training and average the entropies. Let  $f_1, f_2, \dots, f_M$  be snapshots of the classifier from multiple  
 131 epochs during training, where  $M$  denotes the number of training epochs. Then the average entropy  
 132 for an example is given by  $\bar{\mathcal{H}}(x)$  and for a class by  $\bar{\mathcal{H}}_C(f)$ :

$$\bar{\mathcal{H}}(x) = \frac{1}{M} \sum_{e=1}^M \mathcal{H}(f_e(x)), \quad \bar{\mathcal{H}}_C = \frac{1}{M} \sum_{e=1}^M \mathcal{H}_C(f_e). \quad (3)$$

## 133 4 Experiments

134 As aforementioned, common practice performs AT for the whole training set. In the following, we  
 135 explore CSAT and ESAT, which splits the training set in two subsets  $A$  and  $B$  and only constructs AEs  
 136 for  $A$  such that the model never sees AEs for  $B$ . We start with single-class CSAT –  $A$  contains only  
 137 examples of a single class – and increase the size of  $A$  (section 4.1) by utilizing the entropy ranking  
 138 of classes  $\mathcal{H}_C$  (equation 3). ESAT, which splits into example subsets is discussed in section 4.2.  
 139 Both SAT variants reveal complex interactions between classes and examples while indicating that  
 140 few AEs can provide high transfer performance to downstream tasks when weighted appropriately  
 141 (section 4.3).

142 **Training and evaluation details.** Since AT is prone to overfitting [28], it is common practice to stop  
 143 training when robust accuracy on a hold-out set is at its peak. This typically happens after a learning  
 144 rate decay. We adopt this “early stopping” for all our experiments by following the methodology  
 145 in [28] but utilize Auto Attack (AA) to evaluate robust accuracy. Throughout the course of the training,  
 146 we evaluate AA on 10% of the validation data  $\mathcal{D}_{\text{val}}$  after each learning rate decay and perform final  
 147 evaluation with the model providing the highest robust accuracy. This final evaluation is performed  
 148 on the remaining 90% of validation data. This AA split is fixed throughout experiments to provide  
 149 consistency. If not specified otherwise, we generate adversarial examples during training with PGD-7  
 150 within an  $L_2$  epsilon ball of  $\epsilon = 0.5$  (all CIFAR variants) or  $\epsilon = 3.0$  (all ImageNet variants) – typical  
 151 configurations found in related work. We train all models from scratch and use ResNet-18 [29] for all  
 152 CIFAR-10 and CIFAR-100 [30] experiments and ResNet-50 for all ImageNet-200 experiments. Here,  
 153 ImageNet-200 corresponds to the ImageNet-A subset [31] to render random baseline experiments  
 154 tractable (to reduce training time). This ImageNet-200 dataset, contains 200 classes that retain the  
 155 class variety and breadth of regular ImageNet, but remove classes that are similar to each other  
 156 (e.g. fine-grained dog types). We use all training and validation examples from ImageNet [32] that  
 157 correspond to this subset classes. All training details can be found in the supplement.

### 158 4.1 Class subset splits

159 We start by investigating the interactions between individual classes in  $A$  using CSAT on CIFAR-10,  
 160 followed by an investigation on increasing the number of classes. **Single-class subsets (CSAT)**. We  
 161 train all possible, single class CSAT runs (10) and evaluate robust accuracies on the **adv. trained**  
 162 **class (A)** and the **non-adv. trained classes (B)**. The results are shown in figure 2, left. Each rows

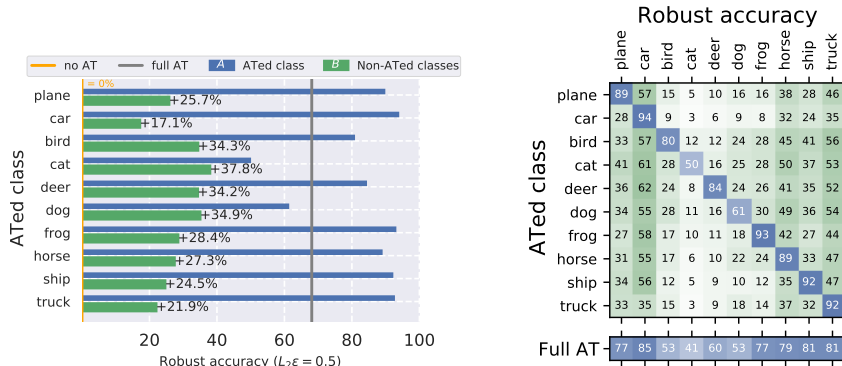


Figure 2: CSAT on a single CIFAR-10 class  $A$  (blue), we observe non-trivial transfer to the non-adv. trained classes  $B$  (green). Classes considered hard in CIFAR-10 (cat) offer best generalization (+37.8% gain on non-adv. trained), while easy classes offer the worst (car, +17.1% gained). Note that without AT, robust accuracy is close to 0% for all classes (orange). Right: same as left, but robust accuracy is evaluated per class (along columns). Here, we observe an unexpected transfer property: hard classes provide better transfer to seemingly unrelated classes (cat  $\rightarrow$  truck: 53%) than related classes (car  $\rightarrow$  truck: 35%).

163 represents a different training run. Note that the baseline robust accuracy, trained without AT achieves  
 164 practically 0% (indicated by red line). Most importantly, we observe non-trivial robustness gains for  
 165 all classes that have never been attack during training ( $B$ -sets). That is, irrespective of the chosen  
 166 class, we gain at least 17.1% robust accuracy ( $A=car$ ) on the remaining classes and can gain up to  
 167 37.8% robust accuracy when  $A=cat$ . These robustness gains are unexpectedly good, given many  
 168 features of the non-adv. trained classes can be assumed to not be trained robustly.

169 To investigate this phenomenon further, we analyze robust gains for  
 170 each individual class and present robust accuracies in the matrix in  
 171 figure 2, right, where training runs are listed in rows and robust accuracies  
 172 per class are listed in columns. Blue cells denote the adv. trained class and green cells denote non-adv.  
 173 trained classes. While we see some expected transfer properties, e.g. CSAT on  $car$  provides greater  
 174 robust accuracy on the related class  $truck$  (46%) than unrelated animal  
 175 classes  $bird, cat, deer, dog$  (between 5% and 16%), the reverse is not  
 176 straight-forward. CSAT on  $bird$  provides 56% robust accuracy on the  
 177 seemingly unrelated class  $truck$ , 10%-points more than CSAT on  $car$ .  
 178 More generally, animal classes provide stronger robustness throughout  
 179 all classes than inanimate classes. We observe, that these classes are  
 180 also harder to classify and have a higher entropy  $\mathcal{H}_C$  as shown in figure 3.

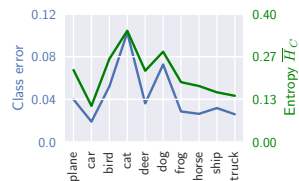


Figure 3: The hardest classes (blue) have the highest entropy (green).

182 **Many-class subsets (CSAT).** To increase the number of classes in  $A$  while maintaining a minimal  
 183 computational complexity, we utilize the average class entropy  $\overline{\mathcal{H}_C}$  proposed in equation 3 to inform  
 184 us which ranking to select from. To improve clarity, we begin with a reduced set of experiments  
 185 on CIFAR-10 before transitioning to larger datasets. We utilize the observed correlation between  
 186 class difficulty, average class entropy and robustness transfer  $\overline{H}_C$  to rank classes and construct 4  
 187 adv. trained subsets. Ranked by class entropy  $\mathcal{H}_C$ , we select 4 subsets showing in figure 4, left. As  
 188 observed before,  $cat$  and  $dog$  are hardest and thus first chosen to be in subset  $A$ .  $Truck$  and  $car$  on  
 189 the other hand are easiest and thus last. To gauge the utility of this ranking, we provide a robust and  
 190 clean accuracy comparison with a random baseline in figure 4, center and right. I.e., for each subset  
 191  $A$  we select 10 random subsets and report mean and std. deviation (red line and shaded area). Similar  
 192 to the single-class setup, we observe subsets of the hardest classes to consistently outperform the  
 193 random baseline (upper middle plot), up until a subset size of  $|A| = 8$ , when it draws even. Also  
 194 note that the robust accuracy on  $B_{val}$  is improved across all splits, thus providing support that harder  
 195 classes – as initially observed on animate vs inanimate classes – offer greater robustness transfer.



Figure 4: Ranking CIFAR10 classes by difficulty (using entropy as proxy), we perform CSAT with an increasing size of *adv. trained classes* in *A*. Class splits used for training (*A* and *B*) are stated on the left. The resulting robust and clean accuracies on the validation set is shown on the right, separated into performance on  $B_{\text{val}}$  and *all*. Compared with a random baseline of random class ranking (red), we find the ranking by difficulty to have consistently better transfer to *non-adv. trained classes* (*B*). Overall, this results in an improved robust accuracy on average over all classes.

196 For our experiments on larger datasets like CIFAR-100 and ImageNet-200, we additionally evaluate  
 197 a third ranking strategy. Beside selecting at random and selecting the hardest first, we additionally  
 198 compare with selecting the easiest (inverting the entropy ranking). We construct 9 subsets per type of  
 199 ranking (instead of 4) and report robust accuracies for selecting the easiest classes as well. Results are  
 200 presented in three columns in figure 5; one dataset per column. As before, we show robust accuracies  
 201 on the tested dataset (upper row) and robust accuracies on  $B_{\text{val}}$  (lower row). For CIFAR-10, we  
 202 calculate mean and std. dev. over 10 runs, for CIFAR-100 over 5 runs and for ImageNet-200 over  
 203 3 runs. Selecting hardest first (highest entropy) is marked as a solid line and easiest first (lowest  
 204 entropy) as a dashed line. First and foremost, we observe that irrespective of the dataset and the  
 205 size of *A*, we see robustness transfer to  $B_{\text{val}}$ . This transfer remains greatest with classes we consider  
 206 hard, while easy classes provide the least. Nonetheless, we see diminishing returns of such an  
 207 informed ranking when dataset complexity is increased. E.g. the gap between dashed and solid line  
 208 on ImageNet-200 is small and random class selection is on-par with the best. The results are similar  
 209 on CIFAR-100, as shown in figure 5, middle). Based on these results, entropy ranking and selecting  
 210 classes provides only slight improvements in general. Importantly though, we continue to see the  
 211 tendency of increased robustness transfer to  $B_{\text{val}}$ , which we will come back to in section 4.3.

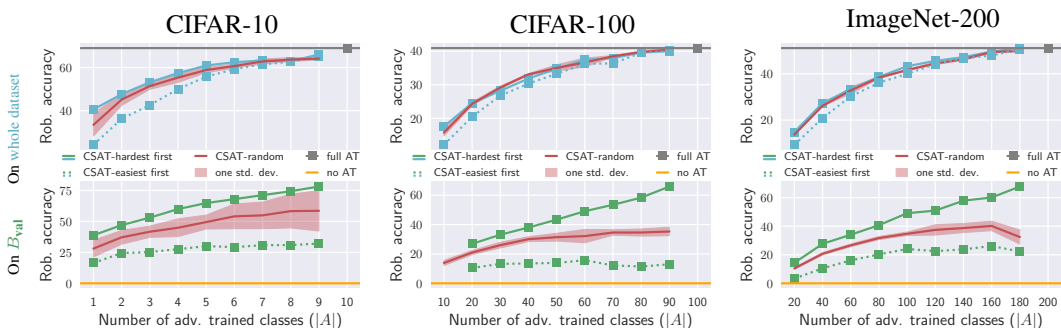


Figure 5: Class-subset Adversarial Training (CSAT) produces non-trivial robustness on classes that have never been attacked during training ( $B_{\text{val}}$ ). Along the x-axes we increase the class subset size of *A* on which AEs are constructed and compare three different class-selection strategies: select hardest first (solid lines), select easiest first (dashed line) and select at random (red). On average, random selection performs as well as informed ranking (upper row), while the robustness transfer to  $B_{\text{val}}$  is best for the hardest classes (lower row). AT on a single class provides already much greater robust accuracies than without AT (orange).



212 **4.2 Example subset splits (ESAT)**

213 Considering that splits along classes are inefficient in terms of reaching the full potential of adversarial  
 214 robustness, we investigate ranking examples across the whole dataset (ESAT). We follow with the  
 215 same setup as before but rank examples – and not classes – by entropy  $\overline{\mathcal{H}}$ . Since it is not feasible to  
 216 construct corresponding rankings on the validation set, we cannot gauge robustness transfer to  $B_{\text{val}}$ .  
 217 Instead, we will test transfer performance to downstream tasks in section 4.3. We consequently report  
 218 robust accuracy and clean accuracy on the whole validation set in figure 6.

219 Firstly, note that the increase in robust accuracy is more rapid than with CSAT w.r.t. the size of  
 220  $A$ . AT only on 50% of training data (25k examples on CIFAR and 112k on ImageNet-200) and the  
 221 resulting average robust accuracy is very close to the baseline AT performance (gray line). Secondly,  
 222 note that gap between hard (solid line) and easy example selection (dashed line) has substantially  
 223 widened. In practice, it is therefore possible to accidentally select poor performing subsets, although  
 224 the chance appears to be low given the narrow variance of random rankings (red). To some extent,  
 225 this observation supports the hypothesis that examples far from the decision border (the easiest to  
 226 classify) provide the least contribution to robustness gains. This is also supported by the reverse  
 227 gap in clean accuracy (bottom row in figure 6). That is, easiest-first-selection results in higher clean  
 228 accuracies than hardest-first, while robust accuracies are much lower. In contrast however, we observe  
 229 random rankings (red) to achieve similar performances to hard rankings (solid lines) on all datasets  
 230 and subset sizes. This is somewhat unexpected, especially on small sizes of  $A$  (e.g. 5k). Given  
 231 the results, we conjecture that the proximity to the decision boundary plays a subordinate role to  
 232 increasing robustness. Instead, it is plausible to assume that diversity in the training data has a large  
 233 impact on learning robust features, also indicated by [33].

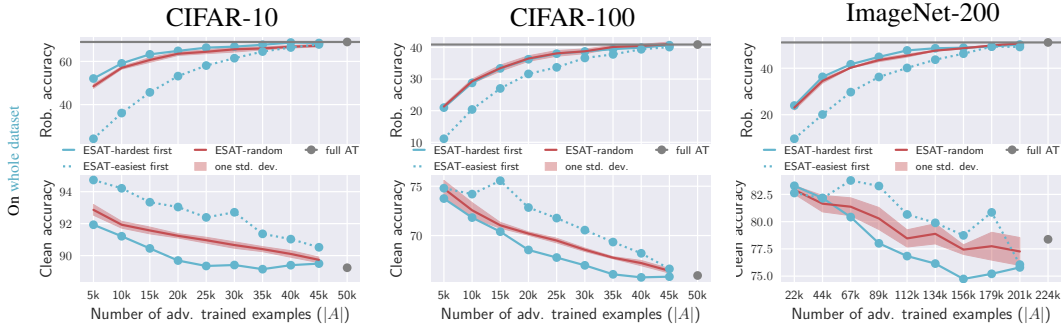


Figure 6: Example-subset Adversarial Training (ESAT) on CIFAR datasets and ImageNet-200, provide quick convergence to a full AT baseline (gray line and dot) with increasing size of  $A$ . We report robust accuracy (upper row) and clean accuracy (lower row) and observe similar characteristics as with CSAT (figure 5). I.e., selecting the hardest examples first (solid line) provide higher rob. accuracy than easy ones (dashed line), although the gap substantially widens. Random example selection (red) provides competitive performance on average. Across all datasets, we see the common clean accuracy decrease while robust accuracy increases [34].

234 **4.3 Transfer to downstream tasks**

235 Previous experiments on ESAT could not provide explicit robust accuracies on the non-adv. trained  
 236 subset  $B_{\text{val}}$  since training and testing splits do not align naturally – recall the evaluation recipe outlined  
 237 in section 3.3. In order to test transfer performance regardless, we make use of the fixed-feature  
 238 task transfer setting proposed in [7]. The recipe just slightly changes: split the data into  $A$  and  $B$  as  
 239 usual and perform SAT. Fix all features, replace the last classification layer with a 1-hidden layered  
 240 classifier and finetune only the new classifier on the target task. Importantly, neither training nor  
 241 validation set for the target task are split. We consider CIFAR-100 and ImageNet-200 and transfer  
 242 to CIFAR-10, SVHN, Caltech-256 [35] and Flowers102 [36]. We call SAT trained for transfer  
 243 Source-task Subset Adversarial Training (S-SAT), to emphasize that the subset training is performed  
 244 on the source-task dataset.

245 In this section, we consider models that have “seen” only a fraction of AEs on the source task and  
 246 investigate the robustness transfer capabilities to tasks on which they have not explicitly adversarially



Figure 7: Impact of cross-entropy weighting on robustness transfer. For subset AT, we test different weighting strategies for sets A and B given they are of unequal size. We observe that vanilla cross-entropy (*circle*) offers the worst robustness transfer to CIFAR-10 (right). The best transfer (*plus*) is provided when loss weights are chosen such that training is overemphasized on A, indicated by dropping robust accuracies on B (compare left and center).

247 trained on. We find unexpectedly strong transfer performances for models that have both low clean  
 248 and robust accuracy, only by putting more weight on the AEs.

249 **Loss balancing improves robustness transfer.** In contrast to the previously explored setting, we  
 250 observe the transfer setting to benefit from loss balancing. Recall equation 2 in section 3.2 in which  
 251  $w_A$  and  $w_B$  can be assigned different values to balance the loss when  $|A| \neq |B|$ . We show that the  
 252 vanilla configuration  $w_A = w_B = 1$  transfers robustness to downstream tasks poorly, that balancing  
 253 the loss with  $w_B = 1, w_A = |B|/|A|$  lacks transfer performance for small  $|B|$  and that weighting  
 254 examples from A higher results in improved robustness transfer. We present results for all three  
 255 weightings in figure 7. The figure is organized in three columns, all reporting robust accuracy. The  
 256 first column reports the robust accuracy on subset  $A_{val}$ , the second on subset  $B_{val}$  and the third reports  
 257 the robust accuracy on the downstream task. Here, we train on CIFAR-100 and transfer to CIFAR-10.  
 258 The vanilla loss is indicated by circles and a solid line, the balanced loss  $w_A = |B|/|A|$  by squares  
 259 and a dotted line and the loss overemphasizing A by a plus and a dashed line.

260 First and foremost, note that the robustness transfer for the vanilla configuration is substantially worse  
 261 than both alternatives (robust accuracy in top right). Transfer improves with use of loss balancing, e.g.  
 262 for  $|A| = 10$ , robust accuracy improves from 8% to 30%, but does not converge to the baseline AT  
 263 performance (gray line). This is an unwanted side effect of equalizing the weight between A and B.  
 264 When A is much smaller than B, less weight is assigned to the AEs constructed for A and robustness  
 265 reduces. Note, this effect can also be seen on  $A_{val}$  (top left in figure). Instead, we find it beneficial  
 266 to overemphasize on the AEs (plus with dashed line). This configuration assigns  $w_A = 2|B|/|A|$  for  
 267  $|A| = 10$  and increases the weight to  $w_A = 10|B|/|A|$  for  $|A| = 90$ . This results in improved robust  
 268 accuracy on  $A_{val}$ , but low robust and clean accuracy on  $B_{val}$ . Interestingly, while the generalization to  
 269  $B_{val}$  is low, robustness transfer to CIFAR-10 is very high. We use this loss weighting for all following  
 270 task transfer experiments.

271 **Robustness transfer from example subsets.** Using the weighted loss, we focus in the following  
 272 on S-ESAT on two source tasks: CIFAR-100 and ImageNet-200, and train on three downstream  
 273 tasks. Similar results for S-CSAT and SVHN as additional downstream task can be found in the  
 274 supplement. Figure 8 presents results for three settings: CIFAR-100  $\rightarrow$  CIFAR-10 and ImageNet-200  
 275  $\rightarrow$  Caltech-256, Oxford-Flowers-102. The first and second row show robust and clean accuracy on  
 276 the downstream task respectively. As before, we compare with a random (red) and a full AT baseline  
 277 (gray line). Selecting A to contain the hardest examples first (highest entropy) is marked by a solid  
 278 line; selecting easiest is marked by a dashed line.

279 In line with the improvements seen using the appropriate loss weighting, we see similarly fast  
 280 recovery of baseline AT performance across all dataset. In fact,  $|A|$  containing only 30% of training  
 281 data (15k and 70k) is sufficient to reach near baseline performance. On CIFAR-100  $\rightarrow$  CIFAR-10  
 282 and ImageNet-200  $\rightarrow$  Flowers-102 even slightly outperforming the same with a further increase  
 283 in size. Similar to the non-transfer settings tested before, we also see similar interactions between



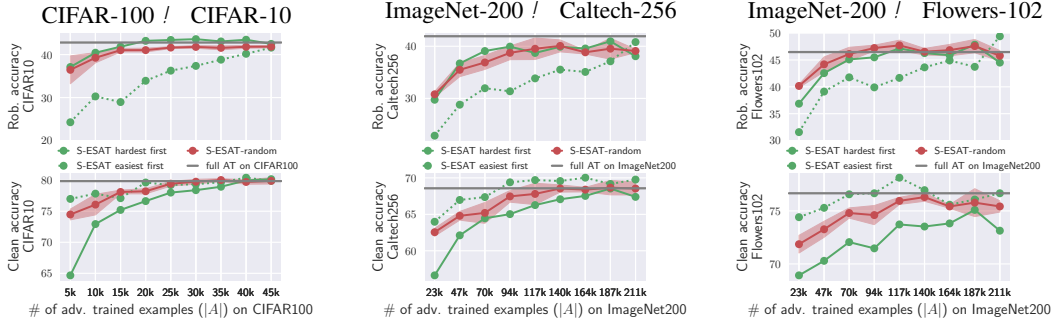


Figure 8: Transfer from S-ESAT to three different downstream tasks. S-ESAT is trained on source dataset CIFAR-100 (left) and ImageNet-200 (middle and right). We report robust (top row) and clean (bottom) accuracies for increasing size of  $A$ . Similar to our investigation on transfer from  $A$  to  $B$ , we find that hard examples provide better robustness transfer than easy ones, but random selections (red) achieve competitive performances. Most importantly, “seeing” only few AEs (here 30% of source data) recovers baseline AT performance (gray line).

284 subset selection strategies. I.e. hardest examples (solid line) provide greater robustness transfer  
 285 than easiest (dashed line) while a random baseline (red) achieves competitive performances. The  
 286 latter consistently outperforming entropy selection on ImageNet-200  $\rightarrow$  Flowers-102, supporting our  
 287 observation in section 4.2: with increasing dataset complexity, informed subset selection provides  
 288 diminishing returns. Note that all robust accuracy increases proportionally correlate to an increase in  
 289 clean accuracy as well. This is in stark contrast to the inverse relationship in previous settings. C.f.  
 290 figure 5 and 6, for which clean accuracy decreases. This interaction during transfer is similar to what  
 291 is reported in [8]: increased robustness of the source model results in increased clean accuracy on the  
 292 target task (over a non-robust model). Intriguingly though, with appropriate weighting, the biggest  
 293 robustness gains on the downstream task happen under fairly small  $A$ . This is a promising outlook  
 294 for introducing robustness in the foundational setting [37], where models are generally trained on  
 295 very large datasets, for which AT is multiple factors more expensive to train. Note that our results  
 296 generalize to single-step attacks like fast gradient sign method (FGSM) [18, 38] as well. We provide  
 297 evaluations in the supplement. While we consider the fixed-feature transfer only, recent work has  
 298 shown this to be a reliable indicator for utility on full-network transfer [8, 39].

## 299 5 Conclusion

300 In this paper, we presented an analysis of how adversarial robustness transfers between classes,  
 301 examples and tasks. To this end, we proposed the use of Subset Adversarial Training (SAT), which  
 302 splits the training data into  $A$  and  $B$  and constructs AEs on  $A$  only. Trained on CIFAR-10, CIFAR-  
 303 100 and ImageNet-200, SAT revealed a surprising generalizability of robustness between subsets,  
 304 which we found to be based on the following observations: (i) adv. robustness transfers among  
 305 classes even if some or most classes have never been attacked during training and (ii) hard classes  
 306 and examples provide better robustness transfer than easy ones. These observations remained largely  
 307 valid in the transfer to downstream tasks like Flowers-102 and Caltech-256 for which we found that  
 308 overemphasizing loss minimization of AEs in  $A$  provided fast convergence to baseline AT robust  
 309 accuracies, even though transfer to  $B$  was severely reduced. Specifically, it appears that only few AEs  
 310 ( $A$  containing 30% of the training set) learn all of the robust features which generalize to downstream  
 311 tasks. This finding could be particularly interesting for AT in the foundational setting, in which very  
 312 large datasets render training computationally demanding.

313 More broadly, improving adversarial robustness remains one of the most important problems to  
 314 solve in deep learning, especially in high-stake decision making like autonomous driving or medical  
 315 diagnostics. Our findings shed new light onto the properties of adversarial training and may lead to  
 316 more efficient robustness transfer approaches which would allow easier deployment of robust models.  
 317 We provided an account on a broad variety of datasets and used models commonly evaluated in  
 318 related work. It needs to be seen whether our findings generalize to other threat models [40] as well.

319 **References**

- 320 [1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Good-  
321 fellow, and Rob Fergus. Intriguing properties of neural networks. *ICLR*, 2014.
- 322 [2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.  
323 Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.
- 324 [3] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris  
325 Tsipras, Ian J Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial  
326 robustness. *ICLR*, 2019.
- 327 [4] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan.  
328 Theoretically principled trade-off between robustness and accuracy. *ICML*, 2019.
- 329 [5] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled  
330 data improves adversarial robustness. *NeurIPS*, 2019.
- 331 [6] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust  
332 generalization. *NeurIPS*, 2020.
- 333 [7] Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and  
334 Tom Goldstein. Adversarially robust transfer learning. *ICLR*, 2020.
- 335 [8] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do  
336 adversarially robust imagenet models transfer better? *NeurIPS*, 2020.
- 337 [9] Yutaro Yamada and Mayu Otani. Does robustness on imagenet transfer to downstream tasks?  
338 *CVPR*, 2022.
- 339 [10] Qi Tian, Kun Kuang, Kelu Jiang, Fei Wu, and Yisen Wang. Analysis and applications of  
340 class-wise robustness in adversarial training. *SIGKDD*, 2021.
- 341 [11] Zhikang Xia, Bin Chen, Tao Dai, and Shu-Tao Xia. Class aware robust training. *ICASSP*, 2021.
- 342 [12] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and  
343 generalization. *CVPR*, 2019.
- 344 [13] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training:  
345 Direct input space margin maximization through adversarial training. *ICLR*, 2020.
- 346 [14] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving  
347 adversarial robustness requires revisiting misclassified examples. *ICLR*, 2020.
- 348 [15] Minseon Kim, Jihoon Tack, Jinwoo Shin, and Sung Ju Hwang. Entropy weighted adversarial  
349 training. *ICML Workshop on Adversarial Machine Learning*, 2021.
- 350 [16] Weizhe Hua, Yichi Zhang, Chuan Guo, Zhiru Zhang, and G Edward Suh. Bullettrain: Acceler-  
351 ating robust neural network training via boundary example mining. *NeurIPS*, 2021.
- 352 [17] Hadi M Dolatabadi, Sarah Erfani, and Christopher Leckie. 1-inf-robustness and beyond: Un-  
353 leashing efficient adversarial training. *ECCV*, 2022.
- 354 [18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adver-  
355 sarial examples. *ICLR*, 2015.
- 356 [19] Yogesh Balaji, Tom Goldstein, and Judy Hoffman. Instance adaptive adversarial training:  
357 Improved accuracy tradeoffs in neural nets. *arXiv preprint*, 2019.
- 358 [20] Maximilian Kaufmann, Yiren Zhao, Iliia Shumailov, Robert Mullins, and Nicolas Papernot.  
359 Efficient adversarial training with data pruning. *arXiv preprint*, 2022.
- 360 [21] Sven Gowal, Po-Sen Huang, Aäron van den Oord, Timothy A. Mann, and Pushmeet Kohli.  
361 Self-supervised adversarial robustness for the low-label, high-data regime. In *ICLR*, 2021.

- 362 [22] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Robustness may be at odds  
363 with fairness: An empirical study on class-wise accuracy. *NeurIPS Workshop on Pre-registration*  
364 *in Machine Learning*, 2020.
- 365 [23] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness  
366 through robustness: Investigating robustness disparity in deep learning. *ACM FAccT*, 2021.
- 367 [24] Saachi Jain, Hadi Salman, Alaa Khaddaj, Eric Wong, Sung Min Park, and Aleksander Madry.  
368 A data-based perspective on transfer learning. *CVPR*, 2023.
- 369 [25] Mohamed Omran and Bernt Schiele. Towards systematic robustness for scalable visual recogni-  
370 tion. *ICML Shift Happens Workshop*, 2022.
- 371 [26] Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of  
372 example difficulty. *NeurIPS*, 2021.
- 373 [27] Chirag Agarwal, Daniel D’souza, and Sara Hooker. Estimating example difficulty using variance  
374 of gradients. *CVPR*, 2022.
- 375 [28] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning.  
376 *ICML*, 2020.
- 377 [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image  
378 recognition. *CVPR*, 2016.
- 379 [30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.  
380 2009.
- 381 [31] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural  
382 adversarial examples. *CVPR*, 2021.
- 383 [32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale  
384 hierarchical image database. *CVPR*, 2009.
- 385 [33] Paul Gavrnikov and Janis Keuper. Adversarial robustness through the lens of convolutional filters.  
386 *CVPR*, 2022.
- 387 [34] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry.  
388 Robustness may be at odds with accuracy. *ICLR*, 2019.
- 389 [35] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.
- 390 [36] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large  
391 number of classes. *ICVGIP*, 2008.
- 392 [37] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von  
393 Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the  
394 opportunities and risks of foundation models. *arXiv preprint*, 2021.
- 395 [38] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial  
396 training. *ICLR*, 2020.
- 397 [39] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better?  
398 *CVPR*, 2019.
- 399 [40] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense  
400 against unseen threat models. *ICLR*, 2021.