
Subclass Distillation

Rafael Müller¹ Simon Kornblith¹ Geoffrey Hinton¹

Abstract

After a large “teacher” neural network has been trained on labeled data, the probabilities that the teacher assigns to incorrect classes reveal a lot of information about the way in which the teacher generalizes. By training a small “student” model to match these probabilities, it is possible to transfer most of the generalization ability of the teacher to the student, often producing a much better small model than directly training the student on the training data. The transfer works best when there are many possible classes because more is then revealed about the function learned by the teacher, but in cases where there are only a few possible classes we show that we can improve the transfer by forcing the teacher to divide each class into many subclasses that it invents during the supervised training. The student is then trained to match the subclass probabilities. For datasets where there are known, natural subclasses we demonstrate that the teacher learns similar subclasses and these improve distillation. For clickthrough datasets where the subclasses are unknown we demonstrate that subclass distillation allows the student to learn faster and better.

1. Introduction

The idea of compressing a teacher model into a smaller student model one by matching the predictions of the teacher was introduced by Bucila et al. (2006). After training the teacher, they performed the transfer on new, unlabelled data by minimizing the squared difference between the logits of the final softmax of the teacher and student models. A related technique, called “distillation”, was introduced by Hinton et al. (2014). That paper performed the transfer on the labelled training data rather than on new, unlabelled data. The student is trained to minimize a weighted sum of two different cross entropies. The first is the cross entropy with the correct answer using a standard softmax. The second is

¹Google Brain, Toronto, Canada. Correspondence to: Rafael Müller <rafaelmuller@google.com>.

the cross entropy with the probability distribution produced by the teacher when using a temperature higher than 1 in the softmax of both models. The point of using a higher temperature is to emphasize the differences between the probabilities of wrong answers that would all be very close to zero at a temperature of 1.

There have since been some interesting theoretical developments of distillation (Lopez-Paz et al., 2016) and it is now being widely used to produce small models that generalize well. These are needed for resource constrained applications of neural networks such as text-to-speech (Oord et al., 2018a) and mobile on-device convolutional neural networks (Howard et al., 2017).

In this work, we focus on distillation for datasets where there are only a few possible classes, resulting in limited information to be transferred (e.g. binary classification). We show that we can improve the transfer by forcing the teacher to divide each class into many subclasses that it invents during the supervised training. We propose an auxiliary loss that encourages each subclass to be used equally while ensuring that each prediction is “peaky”. We show experimentally that the subclasses learned have semantic meaning and help distillation. The subclasses can also be used to interpret the models predictions by clustering them in discrete bins.

The paper is organized as follows. We start with a description of subclass distillation and a comparison to a related method, penultimate layer distillation. First, we train models on a binary split of CIFAR-10 (Krizhevsky, 2009) that we call CIFAR-2x5, where we group sets of 5 classes together to create a binary classification task. We show that a teacher trained to produce subclasses is able to discover the original CIFAR-10 classes, despite receiving only binary supervision. We also show that distilling from this teacher using these learned subclasses leads to better results as compared to conventional distillation and penultimate layer distillation. We next move to the CelebA dataset (Liu et al., 2015), in which each example has 40 binary labels. We show that when predicting a single one of these binary labels, the subclasses produced by the teacher are highly correlated with the other binary labels it has never been trained on, which helps subsequent subclass distillation.

We conclude the experimental section with two additional results. First, on the Criteo click prediction dataset (Criteo-

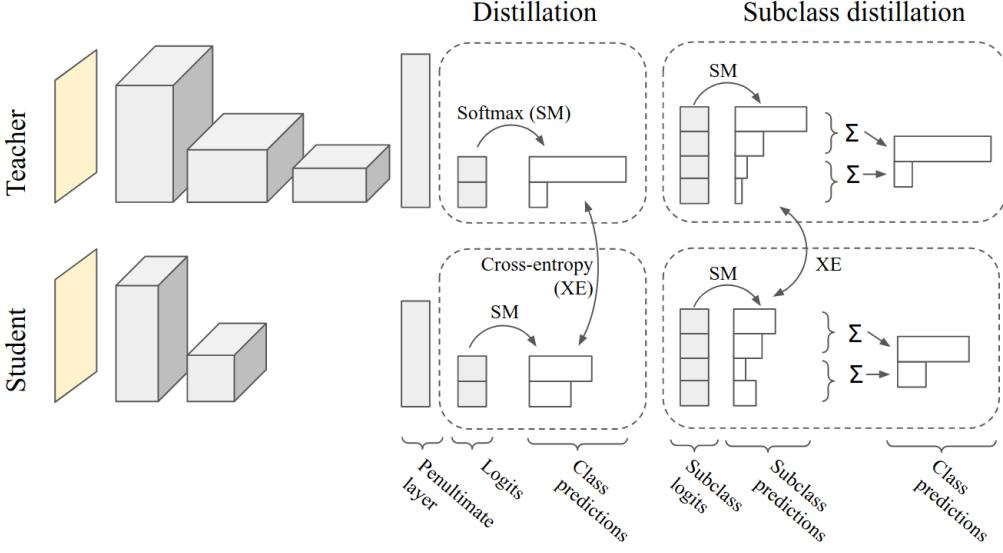


Figure 1. Comparison between distillation and subclass distillation using 2 classes and 2 subclasses per class. The teacher is usually deeper and/or wider than the student. For distillation, the student mimics (using temperature-scaled cross-entropy) the teacher’s class predictions while in subclass distillation the student mimics the subclasses predictions that were invented by the teacher. The class predictions are derived by summing the subclass predictions and the only ground-truth supervision for both cases are binary class labels.

Labs, 2017), we show that subclass distillation outperforms conventional distillation in terms of training speed. We also show that when the student does not see the full dataset, subclass distillation provides significant generalization gains. Second, using MNIST-2x5 (LeCun et al., 1998), we show that the student can learn to predict the binary label by learning to predict the relative subclass probabilities (intra-class), without having ever seen the binary labels or receiving class relative probabilities from the teacher.

2. Subclass distillation

During distillation, the amount of information that the student network receives about the generalization tendencies of the teacher network depends on the number of classes. The information provided by the hard target labels is logarithmic in the number of classes, but the information about how the teacher generalizes is linear in the number of classes provided we distill using the logits or using cross-entropy at a high temperature. This means that distillation is considerably less efficient for models with few classes.

Binary classifiers are important in many applications, and the aim of this paper is to make distillation more efficient for such models by forcing the teacher to invent s subclasses for each of the c classes in the dataset, as shown in Fig. 1. The teacher computes $c \times s$ logits and puts these through a softmax to get $c \times s$ probabilities that sum to 1. The probabilities of all the subclasses of a class are then added to get the teacher’s predicted probability for that class. The

teacher is trained by minimizing the cross-entropy with the class probabilities:

$$\mathcal{L}_{\text{xent}} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^c \mathbf{Y}_{i,j} \log \left(\sum_{k=1}^s P_{i,j,k} \right) \quad (1)$$

where $\mathbf{Y}_{i,j} \in \{0, 1\}$ are the correct targets for the j^{th} class of the i^{th} example as by the dataset and $P_{i,j,k}$ is the output probability for the k^{th} subclass of that example. Given logits Z , the output probabilities P are computed in the usual fashion by performing a softmax operation over all logits belonging to the same example:

$$P_{i,j,k} = \frac{\exp(Z_{i,j,k}/T)}{\sum_{l=1}^c \sum_{m=1}^s \exp(Z_{i,l,m}/T)}. \quad (2)$$

The temperature parameter T controls the entropy of the output distribution. When training the teacher, it is set to 1. When distilling knowledge from the teacher to the student, it is often beneficial to increase the temperature.

In subclass distillation, as in conventional distillation, the student is trained to match the teacher. However, rather than use only the c classes in the original dataset, the student learns to mimic the teacher’s output for $c \times s$ subclasses. Like the teacher, the student produces $c \times s$ output probabilities $\tilde{P}_{i,:,:}$ for each example i , resulting in the subclass distillation loss:

$$\mathcal{L}_{\text{distill}} = -T^2 \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^c \sum_{k=1}^s P_{i,j,k} \log (\tilde{P}_{i,j,k}), \quad (3)$$

where we scale the loss by T^2 in order to keep gradient magnitudes approximately constant when changing the temperature (Hinton et al., 2014). Thus, with this loss, knowledge is transferred from the teacher to the student not merely through the probabilities the teacher assigns to the classes in the original dataset, but also through the probabilities assigned to the subclasses.¹ When training the student, we typically use a combination of the distillation loss $\mathcal{L}_{\text{distill}}$ and the standard cross-entropy loss $\mathcal{L}_{\text{xent}}$:

$$\mathcal{L}_{\text{student}} = \alpha \mathcal{L}_{\text{distill}} + (1 - \alpha) \mathcal{L}_{\text{xent}} \quad (4)$$

where $\alpha \in [0, 1]$ controls the balance between hard and soft targets, which we call “task balance”.

2.1. Penultimate layer distillation

An alternative to subclass distillation that also incorporates more information into distillation is to distill not from the logits, but from the penultimate layer’s activations (or from other layers as in Romero et al. (2014)). In this case:

$$\mathcal{L}_{\text{distill}} = \frac{1}{n} \sum_{i=1}^n \|\mathbf{a}_i - \mathbf{W}\tilde{\mathbf{a}}_i\|^2. \quad (5)$$

where $\tilde{\mathbf{a}}_i$ are the penultimate layer’s activations of the student for the i^{th} example in the minibatch, \mathbf{a}_i are the respective activations in the teacher and \mathbf{W} is a projection matrix to match the dimensions of teacher/student *learned in the distillation phase*. Note that, the student will use its capacity to match the teacher’s representations even for directions that may not be relevant for predicting the classes.

In subclass distillation, the teacher’s subclass logits are a projection of the teacher’s penultimate layer activations into a lower dimension which is *learned during the teacher’s training phase*. Therefore, the projection into subclasses can remove irrelevant information present in the penultimate layer while retaining more information compared to the “class” logits.

Note that Hinton et al. (2014) shows that minimizing the squared difference between the zero-meaned logits of the teacher and student is the limit of distillation as the temperature goes to infinity, provided that the learning rate is scaled as the squared temperature. Therefore, subclass distillation, as the temperature goes to infinity, is equivalent to penultimate layer distillation applied not on the full penultimate layer, but on a low-dimensional projection of that layer.

¹In conventional distillation the cross-entropy loss is $\mathcal{L}_{\text{distill}} = -T^2 \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^c P_{i,j} \log(\tilde{P}_{i,j})$ since the teacher only produces class probabilities.

3. Auxiliary loss

In subclass distillation, the cross-entropy loss (Eq. 1) constrains only the class probabilities and not the subclass probabilities. Without an additional loss encouraging the network to use all subclasses, it may consistently assign high probability to a single subclass of each class and assign extremely low probability to the others. In this case, the subclasses would provide almost no additional signal for distillation. We thus propose an auxiliary loss that encourages the network to assign different examples to different subclasses, even when they belong to the same class. Given a minibatch of n logit vectors $\mathbf{v}_i = \text{vec}(\mathbf{Z}_{i,:,:})$, we compute:

$$\mathcal{L}_{\text{aux}} = -\frac{1}{n} \sum_{i=1}^n \log \frac{e^{\hat{\mathbf{v}}_i^\top \hat{\mathbf{v}}_i / T}}{\frac{1}{n} \sum_{j=1}^n e^{\hat{\mathbf{v}}_i^\top \hat{\mathbf{v}}_j / T}} \quad (6)$$

$$= \frac{1}{n} \sum_{i=1}^n \log \left(\sum_{j=1}^n e^{\hat{\mathbf{v}}_i^\top \hat{\mathbf{v}}_j / T} \right) - \frac{1}{T} - \log(n), \quad (7)$$

where $\hat{\mathbf{v}}_i$ is a normalized version of \mathbf{v}_i (zero-mean, unit-variance) to prevent easy solution of the minimization by making the logits large. As above, T is a temperature hyper-parameter, although its value need not correspond to the temperature used for distillation. This auxiliary loss encourages the normalized logit vector corresponding to each example to have a low dot product with other normalized logit vectors. In practice, the network accomplishes this by distributing examples across subclasses.

The total loss for the teacher is:

$$\mathcal{L}_{\text{teacher}} = \mathcal{L}_{\text{xent}} + \beta \mathcal{L}_{\text{aux}} \quad (8)$$

where β controls the strength of the auxiliary loss.

4. Experimental results

4.1. CIFAR-10

In this section, we experimentally test the ideas presented in the previous sections. We start by providing a visual demonstration that the hidden representations of neural networks contain semantically meaningful information that is not present in the class logits. In Fig. 2 (top), we show the nearest neighbors using Euclidean distance in the class logits layer of a network trained on CIFAR-10 classification. We observe that the nearest neighbors are examples of the same class (horse) as we expected. However, if instead of using the logits layer, we find the nearest neighbors in the penultimate layer, we notice that not only the closest examples are from the same class, but they are also semantically similar to the query image (horse head). This is the sort of information that is present in the penultimate layer but not in the logits that we want to use to improve distillation.

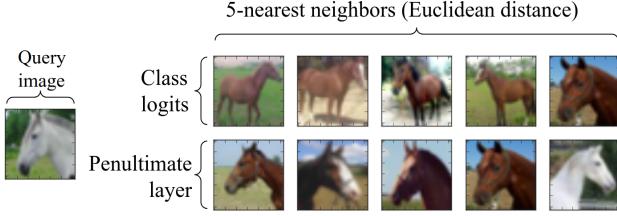


Figure 2. Finding the nearest neighbor in a network trained on CIFAR-10. Query is a close-up on a horse’s head. If the nearest neighbor is calculated in the “class” logits layer, we find examples from the same class (horse), but the semantically similar image with a close-up head is only the 5th nearest-neighbor. If distance is calculated in the penultimate layer, all nearest neighbors are semantically similar to the query. This shows that some semantic information is lost in the “class” logits and distillation can benefit from using more information.

Next, we move to the quantitative results. We use the CIFAR-10 dataset to construct an artificial binary classification task where we group together examples from the classes airplane, automobile, bird, cat and deer to construct the first class and dog, frog, horse, ship and truck to construct the second one. We call this task CIFAR-2x5 and by using this artificial construction we have natural semantic subclasses corresponding to the original CIFAR-10 classes.

4.1.1. UNSUPERVISED SUBCLASS CLASSIFICATION

We train a ResNet (He et al., 2016) network with 20 layers to be used as a teacher (see results in Table 1 and training details including hyperparameters in Appendix A). We first train this network on CIFAR-10 as a baseline and obtain 93.5% accuracy (averaged over 3 runs as all the results in this section). We use the same network with frozen weights to evaluate how well it does on the binary classification task and we obtain 95.6% (+2.1%). If we train this network directly on the binary classification task (CIFAR-2x5), we get 94.3%. Note that although it is evaluated on the same task, the first network is trained with 3.32 ($\log_2 10$) label bits per example compared to only 1 label bit per example in the second network. This difference in the number of bits of label information explains the 1.3% accuracy gap between them in the binary classification task and the benefit of using “subclass” information even when the evaluation is done at the “class” level.

Next, we investigate how making the teacher “invent” subclasses affects the network performance. The subclass head enables the network to output 10 logits (5 subclasses per class) which are marginalized (after softmax) over the subclasses before binary cross-entropy loss. Simply adding the head produces no improvement in binary classification despite the increase in the number of parameters in the last layer by a factor of 5. We also measure the accuracy of

Table 1. Teacher/ResNet results over 3 runs trained on CIFAR-10 or CIFAR-2x5 and top-1 accuracy evaluation on both tasks. Additionally, we evaluate the effect of adding a subclass head and auxiliary loss on unsupervised subclass classification. For reference we include the state-of-the art result on fully unsupervised CIFAR-10 using the invariant information clustering method (IIC) in last line.

CIFAR-	HEAD	AUX. LOSS	ACC. (2)	ACC. (10)
10			95.6 ± 0.1	93.5 ± 0.2
2x5			94.3 ± 0.2	
2x5	✓		94.2 ± 0.2	39.3 ± 4.0
2x5	✓	✓	94.2 ± 0.0	64.6 ± 4.8
UNSUPERVISED (JI ET AL., 2018)				57.6 ± 5.0

this network on all the 10 classes by directly taking the arg max of the subclass layer and picking the permutation that maximizes the accuracy. Although the result of 39.3% is better than chance (20%)², we observed that since there is nothing encouraging the network to use all subclasses, they can “die” during training. The subclass accuracy can significantly be improved by adding the auxiliary loss which increases the accuracy to 64.6%. Note that this network has only seen binary labels, but is able to separate the classes in meaningful subclasses without extra supervision.

Fig. 3 shows how the best network out of 3 runs (70.2%) splits a subset of examples in the validation set into subclasses. Most errors arise in distinguishing among cats, birds and deer, while other subclasses correspond to the original dataset classes. For comparison, the state-of-the-art (Ji et al., 2018) on fully unsupervised classification on CIFAR-10 is 57.6% using the invariant information clustering method (IIC). Here, we show that, with little extra supervision, (binary labels) we can outperform this result with a very simple approach.

In the analysis above, we use the accuracy on 10-class classification as a measure of how well the network separates the examples into meaningful subclasses. The idea is that this subclass information will help the student generalize better through subclass distillation. We can use a very simple model to measure how much extra label information the subclass teacher can provide. In the ideal case where the teacher perfectly learns the subclasses, it provides $1 + 2.32$ label bits ($\log_2 2 + \log_2 5$) per example, where the first bit comes from the binary class and the remaining ones from the subclass. In the case where the teacher can “relabel” $P \times 100\%$ of the subclasses correctly and the remaining errors are distributed equally over the remaining 4 subclasses, the effective number of label bits is given by the q -ary sym-

²Corresponding to perfect knowledge of the class and random choice of the subclass.

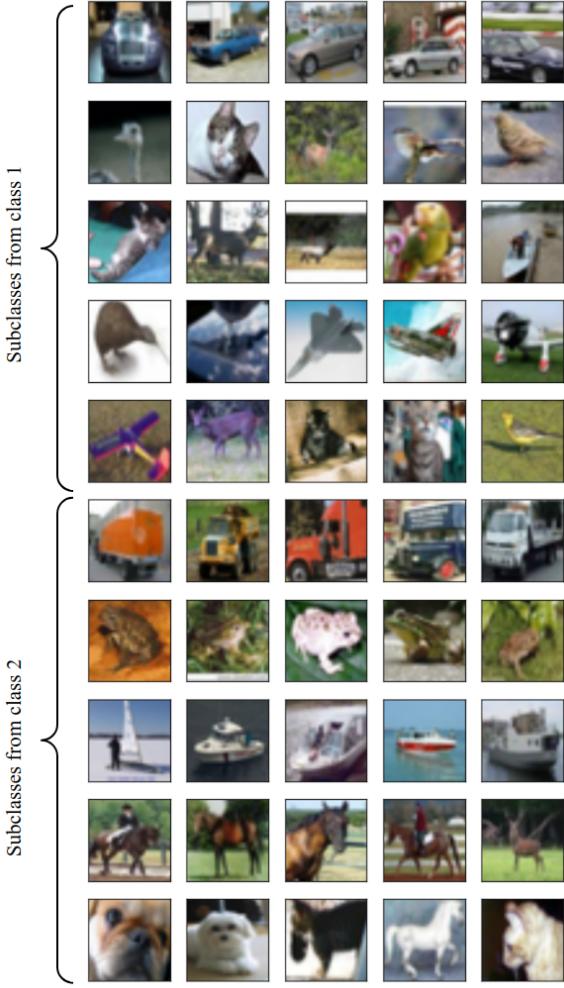


Figure 3. Unsupervised subclass discovery. Examples of the validation set grouped by the subclass logit they activate most (one row per subclass). Using the validation set, we find the 1-to-1 assignment that maximizes accuracy, resulting in the following permutation: automobile, cat, bird, airplane, deer (first class), truck, frog, boat, horse and dog (second class).

metrical channel (Cover & Thomas, 2012) and is equal to $\log_2 5 + P \log_2 P + (1 - P) \log_2 (1 - P)/4$. The teacher trained with binary classification + the subclass head + the auxiliary loss gets on average $67.7 \pm 4.5\%$ subclass accuracy on the training set. This result is slightly better than results from Table 1 from validation set, but they are relevant for the analysis since with distillation we reuse the training set in the transferring phase. The best of the 3 runs gets 73.0%, which results in 0.94 effective extra label bits per example given by the teacher compared to a student that only sees the binary labels. This assumes that the teacher provides noisy one-hot encoded subclass labels (“hard information”) to the student, while distillation can also benefit from “soft” information (small differences in relative proba-

bilities) which can increase the effective number of subclass bits per example, but with the simple model our subclass teacher can already provide roughly the double amount of label information per example.

Additionally, we would like the subclass predictions for each example to be “peaky”, resulting in probability mass concentrated mostly in a single subclass. This can be translated to having low-entropy predictions. For the network trained without the auxiliary loss the average entropy is 0.13 ± 0.02 bits while it increases to 0.42 ± 0.05 bits using the auxiliary loss, which is still far away from 3.32 bits for the uniform distribution. However, just having low-entropy predictions is not enough, since, for all examples belonging to a given dataset class, the network may assign a confident prediction to the same subclass. Therefore, we would like to ensure that after making a hard decision (arg max), the distribution of subclass utilization is close to the uniform distribution (high entropy). The subclass utilization entropy is 1.87 ± 0.11 bits (without) and 3.19 ± 0.02 bits (with) the auxiliary loss. This shows that the auxiliary loss helps the subclass predictions to be confident and diverse at the same time, resulting in discovery of the original subclasses for the CIFAR-2x5 example.

4.1.2. SUBCLASS DISTILLATION

In this section, we investigate how to transfer the teacher’s knowledge to a low capacity student. We pick the AlexNet architecture as the student (Krizhevsky et al., 2012). Results are shown in Table 2. We start by training the network on the two tasks without distillation, as a baseline. We observe a gap of 2.2% between a network trained with subclass labels (CIFAR-10) and a network without access to this extra information (CIFAR-2x5). Next, we train the student in two different situations. First we use conventional distillation. We observe a 1.0% accuracy gain compared to the baseline student. Then we train the same student with penultimate layer distillation and we get similar gain to conventional distillation: 1.0% accuracy gain. Finally, we test subclass distillation, where we distill from a teacher that was trained to perform binary classification, but with the subclass head and auxiliary loss. With subclass distillation, we observe a 2.3% accuracy improvement compared to the baseline student. The subclass distillation student can also classify the examples over 10 classes with 68.3% accuracy which is slightly below the teacher (70.2% which was the best of 3 runs). Note that the student trained with subclass distillation can completely recover the 2.2% gap between the models trained with hard targets on CIFAR-10 and CIFAR-2x5 without ever seeing the “true” subclass labels.

Table 2. Student/AlexNet results over 3 runs. The baselines (two first rows) correspond to training the network with only the labels in the dataset. Then the distillation results correspond to training the student to match the teacher’s class predictions (**Distillation**), to match the penultimate layer’s activations (**Penultimate Layer Distillation**) or the teacher’s subclass predictions (**SubClass Distillation**).

CIFAR-	D	PL-D	SC-D	ACC. (2)	ACC. (10)
10				91.3 ± 0.1	86.7 ± 0.2
2x5				89.1 ± 0.2	
2x5	✓			90.1 ± 0.1	
2x5		✓		90.1 ± 0.1	
2x5			✓	91.4 ± 0.2	68.3 ± 0.2

4.1.3. TRAINING SPEED

In addition to improving performance, subclass distillation also makes training faster. Figure 4 shows the evolution of accuracy on the validation set through training. First we train a baseline network using only the dataset’s “hard” labels represented by the blue curve and the second row in Table 2. We observe a large variation of performance early in training and performance increase is slow. When we train the student with conventional distillation (D), shown in green, training progresses much faster, and the final performance is better than the baseline. Since the teacher provides only a single real number per training example, there is not much information to enable the student to significantly outperform the baseline. Subclass distillation (SC-D), shown in red, addresses this issue. This results in faster training, more stable performance and higher final accuracy, matching a student trained directly on the “true hidden” subclasses (blue dashed line). Note that both the subclass teacher and student have only seen binary labels. Finally, we show the results of penultimate layer distillation (PL-D). Although the performance is similar to distillation, training is slower, as the student tries to match the 128-dimensional teacher’s activations, which may have directions that are not important for final classification.

4.2. CelebA

Although CIFAR-2x5 is suitable to demonstrate the subclass distillation concept and we can show significant gains in performance and training speed, the fact that the true subclass structure matches our choice of the number of subclasses makes the task easier. Therefore, we decided to test our approach on CelebA, a more realistic and challenging dataset.

CelebA comprises 202,599 images of celebrity faces, annotated with 40 binary attributes that are highly correlated

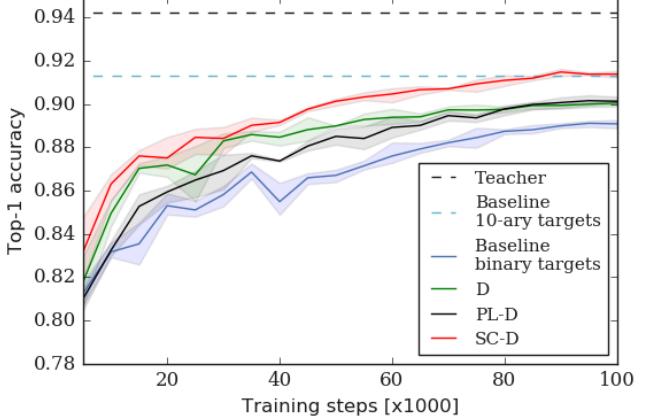


Figure 4. CIFAR-2x5: Evolution of validation accuracy of a student (AlexNet) during training and comparison between: training only with dataset labels (baseline binary targets), distillation (D), penultimate layer distillation (PL-D) and our proposed solution, subclass distillation (SC-D). For reference, we add the performance of the teacher (ResNet-20) trained on binary labels and a student trained on 10-ary labels but evaluated on binary classification (baseline 10-ary targets).

and unbalanced. We pick the male/female classification task and we use 10 subclasses per class, which does not match the number of features. We obtain 1.51% error rate using a ResNet-20 network (averaged over 3 runs). For some of the annotated labels, we can find a corresponding subclass that is activated by said feature. For example, in Fig. 5, we show the proportion of examples in the validation set labeled “blond” in each subclass, where the first 10 subclasses represent the “female” class and the remaining the “male” one. Dashed lines represent the average of the class (more female than male blonds in the dataset). We highlight examples that activate the first and ninth subclass and we observe that indeed the teacher has split the predictions into semantic subclasses and we speculate that this helps distillation.

Next, we transferred knowledge from the teacher (ResNet-20) to a student (AlexNet). Results are shown in Table 3 in terms of error rate for the male/female prediction. The teacher achieves 1.51% error rate while a student trained only with the hard labels achieves 2.05%. Using conventional distillation, the error drops to 1.83% while with subclass distillation we achieve the best performance of 1.70%. This shows that the learned subclass factorization is useful for distillation and helps the student generalize better.

4.3. Criteo

In our CIFAR-2x5 and CelebA experiments, we ignored some of the available supervision during training time and instead used it for evaluation, in order to verify that our

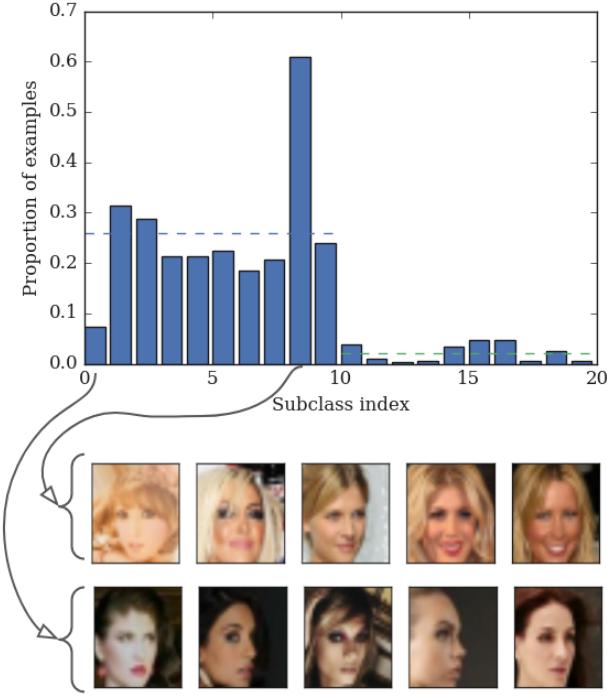


Figure 5. CelebA: proportion of examples per subclass that have the “blond-hair” feature. We highlight some examples of subclass “0” and “8”, where we observe that our teacher network splits the dataset in semantic meaningful subclasses. Dashed lines represent the class-average (female/male).

approach learns meaningful subclasses. In a real-world scenario, we would use all the available information for training. Therefore, we also tested our approach on a binary dataset without a known subclass structure, the Criteo click prediction dataset (CriteoLabs, 2017). This dataset consists of anonymized real-valued and categorical features. The target is a binary label indicating whether the ad was clicked.

Subclass distillation accelerates training on the Criteo dataset and leads to accuracy improvements when limited data is used for distillation. We use the large version of this dataset and we downsample the non-click examples to create a balanced dataset. The teacher is a 5-layer fully-connected network achieving 71.5% accuracy, while the student is a 1-hidden layer network achieving 71.4%. Note that a tiny accuracy improvement is significant in click prediction tasks since it results in large revenue increase for large user bases (Wang et al., 2017). We then compare distillation to subclass distillation. Both achieve 71.6% accuracy, which is better than the teacher. More important, subclass distillation again trains faster, as it provides more information about teacher generalization per example, but the dataset is so big this does not affect final performance. If we artificially reduce the amount of data that the student is trained on (10% of

Table 3. CelebA: results over 3 runs. Both the teacher (ResNet-20) and student (AlexNet) are trained with to predict the binary male/female label. Distillation results correspond to training the student to match the teacher’s class predictions (**D**) or the teacher’s subclass predictions (**SubClass Distillation**).

NET	D	SC-D	ERROR RATE
TEACHER			1.51 ± 0.04
STUDENT			2.05 ± 0.07
STUDENT	✓		1.83 ± 0.05
STUDENT		✓	1.70 ± 0.12

the total) to exaggerate the performance difference, then we observe accuracy gains by using subclass distillation. The ability to perform distillation with limited data is attractive for large datasets such as Criteo (over 1 terabyte in size).

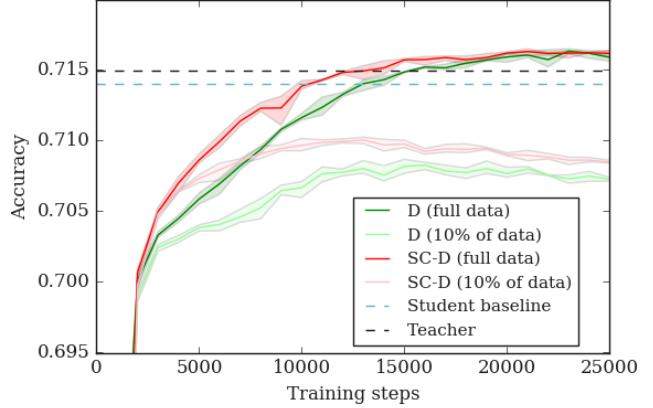


Figure 6. Criteo click prediction: Evolution of validation accuracy of a student during training and comparison between: distillation (D) and subclass distillation (SC-D). When the transfer set contains all the training data, SC-D trains faster but final performance is comparable. By reducing the transfer set by a factor of 10, we exaggerate the performance gap and SC-D outperforms D as it provides the student more bits per training example.

4.4. MNIST

As our final experiment, we split the MNIST dataset into a binary classification (MNIST-2x5), by grouping digits 0 to 4 in one class and digits 5 to 9 in the other. We train a convolutional teacher to produce 10 subclasses. Fig. 7 shows how the network groups the examples into subclasses (each column represents one subclass). This network achieves $0.73\% \pm 0.09$ error rate in the binary classification task. A fully connected 2 hidden layer student achieves $1.57\% \pm 0.06$. We then distill the teacher using distillation ($1.23\% \pm 0.04$) while subclass distillation achieves $0.93\% \pm 0.06$.

More interestingly, we can train the student without the

hard targets by encouraging the student to mimic the intra-class relative probabilities provided by the teacher. We apply a separate softmax to each group of subclass logits to keep relative intra-class probabilities and erase relative class probability. Then we train the student with two cross-entropy losses over 5 subclasses, one per class. This way, the student never sees the binary label, but surprisingly learns it indirectly, obtaining $2.06\% \pm 0.18$ error rate. This is analogous to the experiment in Section 3 of Hinton et al. (2014), where the authors omit the digit “3” in the transfer set and the networks learned to correctly classify them just by observing the soft prediction of the digit “3” for the remaining digits it has seen.



Figure 7. MNIST unsupervised subclass discovery. Examples of the validation set and which subclass logit they activate most (one column per subclass).

5. Related work

Several distillation methods have been proposed in the last few years (Zagoruyko & Komodakis, 2016; Tung & Mori, 2019; Peng et al., 2019; Ahn et al., 2019; Park et al., 2019; Passalis & Tefas, 2018; Heo et al., 2018; Kim et al., 2018; Yim et al., 2017; Huang & Wang, 2017). Some methods focus on teachers and students with the same architecture which can be trained sequentially (Furlanello et al., 2018; Xie et al., 2019) (using unlabeled data and noisy student), (Bagherinezhad et al., 2018) (using extensive data augmentation) or in parallel (Anil et al., 2018) (ensemble). Other methods distill from earlier layers using L_2 loss (Romero et al., 2014; Sun et al., 2019). The relationship between our method and these methods is described in section 2.1. Recently, Tian et al. (2019) proposed to distill from the penultimate layer using a contrastive loss. The relationship between our approach contrastive distillation is more vague; we use a contrastive loss during the teacher training phase to learn the subclasses while in their method it is used during distillation phase.

Our method also bears some resemblance to clustering meth-

ods. Ji et al. (2018) use a contrastive loss similar to our auxiliary loss (they use pairs of data augmented examples to create an anchor, whereas our loss effectively pairs the example with itself) to obtain state-of-the art results on CIFAR-10 in unsupervised and semi-supervised settings. A similar loss has been used for representation learning in (Hjelm et al., 2018; Tian et al., 2019; He et al., 2019; Oord et al., 2018b). In these works, the loss is applied either in an unsupervised setting, or in a semi-supervised setting where only part of the dataset has labels. By contrast, in our case, all examples have a binary label, and we want to learn the hidden subclass labels. Moreover, these methods learn a high dimension representation of the data, whereas we learn exactly the number of subclasses with no need for a linear layer on top. An alternative method for unsupervised clustering with deep neural networks that is not based on the contrastive loss can be found in Kosiorek et al. (2019), where they use capsule networks to directly learn MNIST and CIFAR-10 classes. The closest method to ours is that of Krause et al. (2010), which also uses a probabilistic classifier for clustering by optimizing for class balance and class separation, although the authors use a different loss for this purpose, and perform experiments with kernel methods rather than deep neural networks.

6. Conclusion

We propose subclass distillation, a distillation method where the teacher divides each class into many subclasses that it invents, and the student matches these subclass probabilities. We show that we can improve learning compared to conventional distillation and penultimate layer distillation in terms of generalization and/or training speed. We showed that with a simple auxiliary loss, our teacher divides examples of the dataset into semantically meaningful subclasses. The loss encourages the subclass predictions to be confident and diverse.

We showed that when the underlying subclass structure is known and matches the choice of number subclasses (CIFAR-2x5 and MNIST-2x5), we can discover the original subclasses with high accuracy, and subclass distillation outperforms other distillation methods. When there is a subclass structure in the dataset which does not match the number of subclasses chosen (CelebA), our method can still discover semantic subclasses which help subclass distillation. Finally, when there is no known subclass structure (Criteo), subclasses can provide faster transfer and more bits per example when the data available is limited. We further validated that subclass distillation provides additional bits per example by showing on MNIST that we can learn to predict the binary label without any binary supervision, just by mimicking the (intra-class) teacher subclass relative probabilities.

References

- Ahn, S., Hu, S. X., Damianou, A., Lawrence, N. D., and Dai, Z. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9163–9171, 2019.
- Anil, R., Pereyra, G., Passos, A., Ormandi, R., Dahl, G. E., and Hinton, G. E. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235*, 2018.
- Bagherinezhad, H., Horton, M., Rastegari, M., and Farhadi, A. Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*, 2018.
- Bucila, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541. ACM, 2006.
- Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.
- CriteoLabs. Download terabyte click logs, Jan 2017. URL <http://labs.criteo.com/2013/12/download-terabyte-click-logs/>.
- Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., and Anandkumar, A. Born again neural networks. In *International Conference on Machine Learning*, pp. 1607–1616, 2018.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.
- Heo, B., Lee, M., Yun, S., and Choi, J. Y. Knowledge transfer via distillation of activation boundaries formed by hidden neurons, 2018.
- Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *NIPS 2014 Deep Learning Workshop*, 2014.
- Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- Huang, Z. and Wang, N. Like what you like: Knowledge distill via neuron selectivity transfer, 2017.
- Ji, X., Henriques, J. F., and Vedaldi, A. Invariant information distillation for unsupervised image segmentation and clustering. *arXiv preprint arXiv:1807.06653*, 2018.
- Kim, J., Park, S., and Kwak, N. Paraphrasing complex network: Network compression via factor transfer. In *Advances in Neural Information Processing Systems*, pp. 2760–2769, 2018.
- Kosiorek, A. R., Sabour, S., Teh, Y. W., and Hinton, G. E. Stacked capsule autoencoders. *arXiv preprint arXiv:1906.06818*, 2019.
- Krause, A., Perona, P., and Gomes, R. G. Discriminative clustering by regularized information maximization. In *Advances in neural information processing systems*, pp. 775–783, 2010.
- Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- LeCun, Y., Cortes, C., and Burges, C. J. The mnist database of handwritten digits, 1998. URL <http://yann.lecun.com/exdb/mnist>, 10:34, 1998.
- Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- Lopez-Paz, D., Bottou, L., Schölkopf, B., and Vapnik, V. Unifying distillation and privileged information. *International Conference on Learning Representations*, 2016.
- Oord, A. v. d., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., et al. Parallel WaveNet: Fast high-fidelity speech synthesis. In *International Conference on Machine Learning*, pp. 3915–3923, 2018a.
- Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018b.

Park, W., Kim, D., Lu, Y., and Cho, M. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Passalis, N. and Tefas, A. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 268–284, 2018.

Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., and Zhang, Z. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5007–5016, 2019.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Sun, S., Cheng, Y., Gan, Z., and Liu, J. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4314–4323, 2019.

Tian, Y., Krishnan, D., and Isola, P. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Tung, F. and Mori, G. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1365–1374, 2019.

Wang, R., Fu, B., Fu, G., and Wang, M. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD’17*, pp. 1–7. 2017.

Xie, Q., Hovy, E., Luong, M.-T., and Le, Q. V. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*, 2019.

Yim, J., Joo, D., Bae, J., and Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4133–4141, 2017.

Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

A. Experimental setup

CIFAR-2x5 The teacher is a ResNet-20 trained for 64k steps using a minibatch size of 128. We used a cosine learning rate schedule that drops to 0 at the end of training and Nesterov momentum (0.9). The starting learning rate was 0.1, weight decay 0.0003, temperature in the auxiliary loss 2.0. For the student, we switch the architecture to AlexNet, and increased the number of training steps to 100k. For the baseline student trained from scratch without distillation, we performed a grid search over weight decay (0.0001, 0.0003, 0.001, **0.003**, 0.01, 0.03) and learning rate (0.03, 0.1, **0.3**, 1.0). For every point, we averaged the validation accuracy over 3 runs. Values achieving the maximum accuracy are highlighted in bold. We use these same basic hyperparameters for all distillation results, but distillation adds additional hyperparameters, which we again tune by grid search. For conventional distillation, we sweep temperature T (1.0, 2.0, 4.0, 8.0, 16.0, **32.0**, 64.0) and task balance α (0.0, 0.25, 0.5, **0.75**). For subclass distillation, we also sweep temperature (1.0, 2.0, 4.0, **8.0**, 16.0) and task balance (0.0, 0.25, 0.5, **0.75**). For penultimate layer distillation we sweep the weight we give to the L_2 distillation loss (0.3, 1.0, 3.0, **10.0**, 30.0).

CelebA For CelebA experiments, we follow a similar procedure to the CIFAR-2x5 experiments. We optimize the ResNet-20 teacher with 10 subclasses per class, picking values of weight decay, learning rate and auxiliary loss and temperature (0.001, 0.1, 2.0). We optimize the AlexNet student trained without distillation by performing grid search over weight decay (0.00003, 0.0001, **0.0003**, 0.001, 0.003) and learning rate (0.003, 0.01, **0.03**, 0.1). We pick these values for the distillation results and tune temperature and task balance. For conventional distillation, temperature (1.0, 2.0, **4.0**, 8.0, 16.0) and task balance (0.0, 0.25, 0.5, **0.75**). For subclass distillation, temperature (1.0, 2.0, 4.0, 8.0, **16.0**) and task balance (0.0, 0.25, **0.5**, 0.75).

Criteo The teacher is a 4 layer fully connected network with ReLU nonlinearity and the following number of neurons per layer: 2048, 1024, 512, 256. We have 13 integer-valued features and 26 categorical which are embedded with dimension 32 after being hashed to 1e6 buckets. The teacher has a total number of 10 subclasses, the auxiliary loss has temperature 2.0 and is multiplied by 0.1 before being added to the hard targets cross-entropy. The student has a single hidden layer of size 256. Both are trained using minibatch size of 8192, momentum of 0.9, learning rate starting at 0 and increasing quadratically up 0.1 at 10k steps then staying constant. Teacher and student baseline results represent early stopping at 27k steps. For the distillation results we use task balance of 0.5 and temperature of 2.0.

MNIST The teacher is a deep convolutional network with ReLU nonlinearity and the following layers in sequence: convolutional with 32 output channels and kernel size of 3x3, max-pooling with kernel size of 2x2, convolutional with 64 output channels and kernel size of 2x2, max-pooling with kernel size of 2x2 and strides of 2, dropout with rate 0.5, fully connected with 128 output neurons and dropout with rate of 0.5. The teacher uses a total of 10 subclasses and auxiliary loss temperature of 1.0. The student is a 2 hidden layer fully connected network (784 neurons per layer) and relu activation. For conventional distillation and subclass distillation, the temperature is 4.0 and task balance 0.5. Every network is trained for 12 epochs using a batch size of 256 and Adam optimizer with following parameters: learning rate of 0.1, β_1 of 0.9, β_2 of 0.999 and ϵ of 10^{-7} .