Sub-Task Imputation via Self-Labelling to Train Image Moderation Models on Sparse Noisy Data

Indraneil Paul
Amazon.com Inc.
Bangalore, India
indranep@amazon.com

Sumit Negi
Amazon.com Inc.
Bangalore, India
suminegi@amazon.com

Abstract

E-commerce marketplaces protect shopper experience and trust at scale by deploying deep learning models trained on human annotated moderation data, for the identification and removal of advert imagery that does not comply with moderation policies (a.k.a. *defective images*). However, human moderation labels can be hard to source for smaller advert programs that target specific device types [1] with separate formats or for recently launched locales with unique moderation policies. Additionally, the sourced labels can be noisy due to annotator biases or policy rules clubbing multiple types of transgressions into a single category. Therefore, training advert image moderation models necessitates an approach that can effectively improve the sample efficiency of training, weed out noise and discover latent moderation sub-labels in one go.

Our work demonstrates the merits of automated sub-label discovery using self-labelling. We show that self-labelling approaches can be used to decompose an image moderation task into its hidden sub-tasks (corresponding to intercepting a single sub-label) in an unsupervised manner, thus helping with cases where the granularity of labels is inadequate. This enables us to bootstrap useful representations quickly, via low-capacity but fast-learning teacher models that each specialize in a single distinct sub-task of the main classification task. These *sub-task specialists* then distil their logits to a high-capacity but slow-learning *generalist* student model, thus allowing it to perform well on complex moderation tasks with relatively fewer labels than vanilla supervised training. We conduct all our experiments on the moderation of sexually explicit advert images (though this method can be utilized for any defect type) and show a sizeable improvement in NPV (+30.2% absolute gain) viz-a-viz regular supervised baselines at a 1% FPR level. A long-term A/B test of our deployed model shows a significant relative reduction (-45.6%) in the prevalence of such advertisements compared to the previously deployed model.

CCS Concepts

• **Computing methodologies** → *Learning from implicit feedback*; **Computer vision**.

Keywords

self-labelling, knowledge distillation, computer vision, neural networks

ACM Reference Format:
Indraneil Paul and Sumit Negi. 2022. Sub-Task Imputation via Self-Labelling to Train Image Moderation Models on Sparse Noisy Data. In *Proceedings of 31st ACM International Conference on Information and Knowledge Management (CIKM '22).* ACM, Atlanta, GA, USA, 11 pages. https://doi.org/10.1145/3511808.3557149

1 Introduction and Motivation

E-commerce marketplaces enable sellers and vendors to promote their products and foster brand awareness via self-service advertisement creation workflows. In the interest of ensuring shopper experience and safety, the advert creation workflows usually entail stringent moderation checks. These checks usually look out for transgressions in advert images that render an advert unfit for public viewing (such as nudity and obscenity) or engender societal harm (such as the sale of weapons or narcotics). We term these transgressions as *defects* in the moderation domain. E-commerce marketplaces lay out the locale specific defect categories to advertisers via publicly available moderation policies[2][3].

1.1 Constraints and Contributions

With ever-increasing advert creation volumes, keeping up with the growing public expectations of platform safety and accessibility requires e-commerce marketplaces to automate the interception and takedown of defective advert images. This is usually enforced using deep learning models that are trained on past expert moderator judgements. This presents a three-fold problem.

Sample Scarcity: High-quality defect classification audits, conducted by subject-matter experts on previously created ads, are rare in smaller advertising programs[4][5], with models often trained on a few thousand samples. Third-party crowdsourcing of labels presents its own challenges, as the costs and difficulty associated with training uninitiated annotators in the nuances of the moderation policy are prohibitive. This means that high inductive bias sample-efficient architectures such as CNNs are better suited to the learning task. However, results on popular image classification benchmarks [40] suggest that these architectures have a lower performance ceiling compared to self-attention based approaches [12, 14, 33, 46]. Self-attention based models represent a more flexible and generic function approximator than CNNs (hence their higher performance ceiling), which makes them an attractive

[1] advertising.amazon.com/resources/ad-specs/kindle

[2] advertising.amazon.com/resources/ad-policy/creative-acceptance
[3] www.ebayads.com/advertising-policies
[4] advertising.amazon.com/solutions/products/posts
[5] advertising.amazon.com/library/guides/advertising-books-on-amazon-authors

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557149

option for image moderation tasks. Their weaker inductive bias, however, also makes these models more data-hungry and liable to overfit to noisy data. Thus, we need an approach to train high capacity transformer based models in a sample efficient and generalizable manner. Existing work [48] shows increases in transformer training sample-efficiency by distilling knowledge from lower capacity but faster learning teacher model. We adopt this paradigm and extend it by automatically inferring latent sub-tasks and training a specialist teacher for each sub-task.

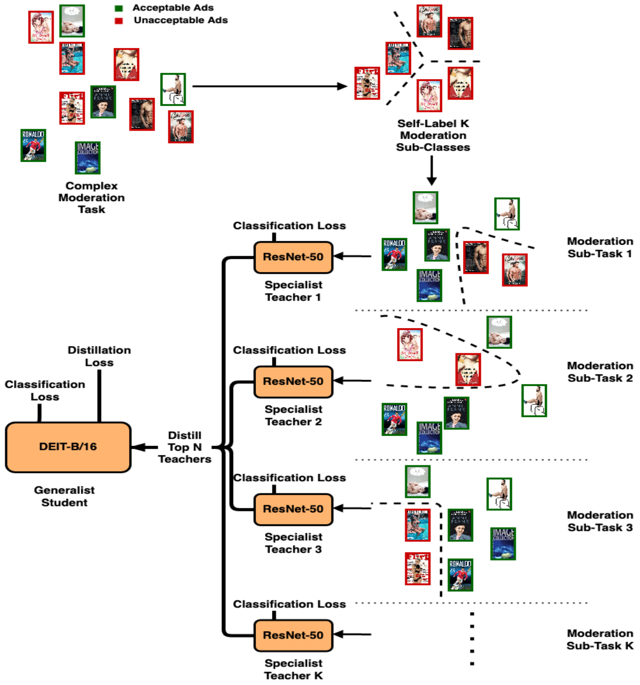


Figure 1: High-level overview of out approach

Latent Sub-Tasks: In established high-traffic advert programs[6] with a longer history of online advertising (and thus more expert audits), learning the defect classification task in its entirety from expert moderator labels may sometimes still not be possible. One prominent example is when the task consists of multiple latent sub-tasks. In these cases, models are better off learning the sub-tasks separately. For example, the task of intercepting sexually explicit images may consist of the separate sub-tasks of intercepting images containing romantic embraces as well intercepting images containing nudity. In such a setting, vanilla supervised training using moderation labels places an unreasonable burden of imputing class substructures on the model. Soliciting crowdsourced labels for these semantically simpler sub-tasks has its limitations. Human inspection for sub-task discovery is an imperfect process that is dependent on the individual's biases and does not respect the inter-sample distance in the deployed model's embedding space. Hence, we look for means to automatically impute sub-tasks from data.

Approaches such as Subclass Distillation [35] deal with this problem by training a teacher model on the whole task and distilling the inferred subclasses of candidate classes from its penultimate layer to a student model. This process can be compromised in situations where only a few semantic subclasses exist, as it limits the amount of knowledge that can be transferred. DeepCluster [6] uses alternating rounds of representation learning and `K-Means` clustering to discover subclasses, but is prone to producing trivial cluster assignments. Additionally, it does not provide an obvious way to leverage subclass assignments for downstream classification tasks. We leverage optimal transport via the Sinkhorn algorithm [11] to reliably impute non-trivial defect subclasses from data. These subclass assignments break the moderation task into constituent sub-tasks (discriminating the defect subclass from the whole set of acceptable advert images).

Label Noise: Limiting ourselves to only expert labelled data is not always enough to avoid noise. Moderation policies may be subjectively interpreted by individuals and are also subject to change in response to real world events.

Existing methods to circumvent this, either leverage contrastive objectives along with involved data augmentation strategies[23] or employ self-training based student-teacher mechanisms [51, 53] to compensate for data scarcity and noisy labels. We show that our paradigm lends itself nicely to noise removal by automatically discarding out-of-domain or uninformative sub-tasks.

Our technical contributions (refer to Figure 1) can be summarized in order as follows:

1. Sub-Task Discovery via Self-Labelling: We uncover sub-defects latent in a complex defect type using cluster-based self-labelling. We leverage optimal transport via the Sinkhorn algorithm to reliably avoid trivial and uninformative sub-tasks. All defective images are assigned to one of the disjoint clusters. Discriminating each of these clusters from the whole set of acceptable adverts represents a moderation sub-task. Details in subsection 4.1.

2. Sub-Task Distillation with Specialist Teachers: We break our `val` set into separate `teacher val`, `teacher test` and `student val` sets (See section 3 for details). We train sample-efficient teacher models that each specialize in a distinct sub-task (discriminating one unique cluster of defects [class 0] from acceptable advert images [class 1]), with model selection done on the `teacher val` set. The individual specialist teachers are then benchmarked for their competency on the whole task using the `teacher test` set. We subsequently select the most competent teachers (using a performance threshold on the `teacher test` set) and their corresponding unique cluster of defective images. This allows us to discard noisy, non-informative and out-of-domain data points. The selected teachers finally distill their logits (for only data points they specialize in) to a generic, high-capacity student model which learns the task space in its entirety (discriminating all defective advert images [class 0] from acceptable advert images [class 1]). The student model selection is done on the `student val` set, and the best model is deployed for production. See subsection 4.2 for details.

3. Framework Extensions: We explore the performance gains afforded to this paradigm by different distillation curricula, i.e., the order in which the selected specialist teachers distil their knowledge to the generalist student model (Refer to subsection 4.3 for details). We discuss how the self-labelled clusters can be leveraged to discard outlier samples, that otherwise hurt final performance,

[6] https://advertising.amazon.com/solutions/products/sponsored-products

from annotated data as well as unlabelled near-domain data (See subsection 5.4 for details).

Our sub-task imputation and selective knowledge distillation framework is able to tackle the domain-specific challenges of uncovering latent sub-tasks, increasing training sample-efficiency and omitting noisy data points in one go. We detail the infrastructure stack serving our final model and the resultant user-facing production impact of our work (via a long-term A/B test) in section 6.

2 Related Work

Adversarial advert Classification: Initial attempts at adversarial advert detection [42] hand-engineered relevant features, learnt representations using ranking objectives and employed a cascade of classifiers in order to ensure high precision. However, such methods demand many labelled samples. Owing to the skewed long-tail distribution of adversarial adverts over defect types, it is important to be able to effectively learn from a few labelled samples. Subsequent work [19] has leveraged a few thousand text moderation labels to train models, which then utilize unlabelled near-domain data for self-training more general representations [53]. While these methods are not adjacent to ours, we leverage the ideas of self-training extensively in our work.

Discrete Prototypical Self-Supervised Feature Learning: Methods such as BEiT [3] have recreated the discrete autoencoding objective used in text transformers. Image patches from all unlabelled images are tokenized using a discrete VAE [39]. Pre-training requires the model to reconstruct the discrete token for deleted image patches. Alternatively, DeepCluster [6], quickly bootstraps clustered representations using alternating steps of `K-Means` clustering followed by learning to classify the dataset samples using the obtained assignments. However, the method was sensitive to hyperparameter choices and employed undesirable workarounds to avoid trivial cluster assignments. SeLa [2] circumvented this problem by formulating the cluster assignment problem as an optimal transport problem and solving it using the *Sinkhorn-Knopp* algorithm. Subsequently, Swav [7] extended this idea to allow for the cluster prototypes to themselves be represented using a learnable vector in the embedding space which can be trained by back-propagation using a contrastive loss based on swapping the cluster assignments of two views of the same image.

To the best of our knowledge, we are the first to leverage cluster-based self-labelling in the context of disjoint sub-task discovery. We chose to utilize the Sinkhorn-Knopp formulation due to its speed, owing to its lack of learnable cluster prototypes. Additionally, the optimal transport formulation prevents trivial clusters out of the box.

Uncovering Learning Sub-Tasks: The idea of deconstructing a complex task into simpler constituent sub-tasks has been best explored in reinforcement learning. Early work [13] has sought to decompose an MDP into a hierarchy of sub-MDPs, re-composing the overall value function using a combination of the value functions of the sub-MDPs. Later work [16] has shown that the task basis matrix of Multi-task LMDPs [41] can be inferred using NMF [27], effectively uncovering sub-tasks. Existing work [45] shows that given a sub-task dependency graph, a general controller can be used to determine the optimal sub-task in a situation, solving which is delegated to a low-level controller. In the multiagent setting, existing results [49] demonstrate the benefits of devolving large joint action spaces into constituent *role* spaces, via clustering actions by their effects. Additionally, substantial gains in imitation learning and behavioural cloning performance can be observed by decomposing complex state-action sequences into meaningful subsequences. CompILE [24] modifies behavioural cloning using an auto-encoder to map disjoint state-action subsequences to a fixed-size codebook, thus allowing the model to learn more general sequence transitions, which helps it generalize more easily. Such a formulation was later extended [28] to imitation learning using self-supervision from a learnt sub-task transition model.

Hierarchical decompositions of task spaces have also shown to improve the quality of self-supervised representation learning. PCL [29] modifies the InfoNCE loss with a contrastive loss against multi-level cluster prototypes, which are inferred using the EM algorithm. More recently, methods [56] have leveraged the internal knowledge of language models coupled with external documents to provide textual descriptions of sub-tasks, given complex tasks in a supervised setting. We show, in the context of our image moderation task, that decomposing supervised image datasets into conceptual clusters and selectively learning from relevant ones can improve the final performance and sample-efficiency.

Multi-Task Multi-Teacher Distillation: Distilling knowledge from multiple teachers has mainly been explored in the single task setting. In such situations, one can average the teacher predictions [18, 54] before computing the loss, weigh the teachers by entropy [26], weigh the structural losses of the teachers based on alignment with student features [32] or jointly optimize gradient update directions [15].

Our setting, however, more closely resembles multi-teacher multitask distillation [30], where each teacher specializes in a single task. While methods exist to uncover helpful task groupings [17, 55] in multitask learning, selecting useful task-specialist teachers has been rather unexplored. We show the benefits to student performance allowed by selecting good teachers and sequentially distilling their logits using a curriculum.

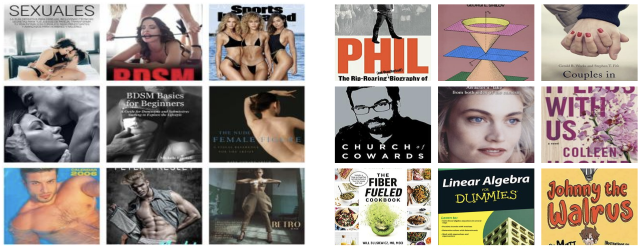
3 Dataset

In this work, we seek to demonstrate the value added by our methods through benchmarking them on the task of moderating sexually explicit advert images submitted in the Sponsored Books[7] ad program. The domain provides a good opportunity to demonstrate our technical contributions as it is a smaller advertising space with few labelled examples as well as on a complex defect type that hides many subtypes (e.g., book cover images with bare chested models, sexual health guides, etc). The defect category also represents a good opportunity to demonstrate production impact, as it is one of the most common type of transgressions committed by advertisers. To that end, we detail the dataset construction process below.

We retrieve all prior adverts with human moderation verdicts over a one-year span (going back further risks introducing label noise due to policy changes). These adverts are first de-duplicated by headline. Further deduplication is performed over images using perceptual hashing [50] to prevent visible duplicates from inflating

[7] advertising.amazon.com/library/videos/sponsored-products-for-books

recall numbers. The images labelled as sexually explicit are considered negative ($\mathcal{D}_-$ with general task label 0) while the rest are considered positive ($\mathcal{D}_+$ with general task label 1). We obtain 3,764 negative samples and 22,156 positive samples for a total of 25,920 samples. Sample images from each class are shown in Figure 2.



(a) Sexually Explicit Images **(b) Acceptable Images**

Figure 2: Retrieved Advert Image Samples

This set is then temporally split into the `train`, `val` and `test` sets with the same proportion of positive and negative samples, to simulate production settings. We throw away the last 15 days of data from the train and val sets to avoid result inflation caused due to bursts of submissions by a single advertiser (which are usually quite similar). The resulting `train`, `val` and `test` splits are detailed in Table 1 under the section *regular split*. We use this split to train all the baseline models we compare against.

Table 1: Dataset split details

Split Name	Set Name	$\lvert\mathcal{D}_-\rvert$	$\lvert\mathcal{D}_+\rvert$
Regular	`train`	2500	15000
	`val`	800	4800
	`test`	776	4656
Student-Teacher	`train`	2500	15000
	`teacher val`	160	960
	`student val`	480	2880
	`teacher test`	160	960
	`test`	776	4656

Our approach, however, uses a two-stage procedure where the teachers are trained and evaluated, following which the student learns from the most competent teachers. Using the same val set for model selection in both the stages would expose us to circular analysis [25] bias that can lead to inflated performance numbers, overfitting and poor generalization. Our approach demands separate validation and holdout sets for the specialist teacher models. Thus, following existing recommendations of sample splitting [21], we subsample 20% of the aforementioned regular split `val` samples in a stratified manner, as the `teacher val` set and a further 20% of the regular split `val` samples as the `teacher test` set. The model selection of the student set is done of the remaining 60% samples of the regular split `val` samples, which we refer to as the `student val` set. The `train` and `test` splits are exactly the same as in the regular split, which ensures fair comparisons in section 5. We deem this split the *student-teacher split*. We compare our models trained on this split against baselines trained on the regular split in section 5. The exact set sizes and compositions are summarized in Table 1 under the student teacher split section.

4 Methods

4.1 Self-Labelling to Reveal Latent Clusters

In the context of defect moderation, we observe that each image annotated as a certain defect class (e.g., sexually explicit) is defective in its own way, contravening its own unique combination of clauses present in the moderation policy. However, the non-defective advert images are all alike in their compliance to all the clauses. Hence, we seek to uncover latent subclasses among defective images in the `train` set, in an unsupervised manner. We do so by framing subclass assignment as a self-labelling problem, as illustrated below.

Consider a learner $\mathcal{F}_\theta(\mathbf{x})$ parametrized by learnable weights θ, that maps images $\mathbf{x}$ to feature representations r as part of classifying the image as y. In the general supervised setting with N labelled image-label pairs and K classes, the parameters θ would be inferred by minimizing the average cross-entropy loss between the label distribution $q(y|\mathbf{x})$ and the model predicted distribution $p(y|\mathbf{x})$ produced by $\texttt{softmax}(hf_\theta(x))$ (where h is the activation function) as follows:

$$\underset{\theta}{\operatorname{argmin}} \; -\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{K} q(y|\mathbf{x}_i)\log p(y|\mathbf{x}_i) \qquad (1)$$
$$\text{where } q(y|\mathbf{x}_i) = \delta(y - y_i)$$

In the semi-supervised case, some images lack labels and thus their corresponding $q(y|\mathbf{x})$ can be jointly estimated along with network parameters θ. In the unsupervised case, however, the absence of labels y_i implies that the objective in Equation 1 turns into a joint feature learning and label assignment problem. This formulation for uncovering latent subclasses in unlabelled defective images would suffer from one obvious degenerate solution where all the images $\mathbf{x}_i$ are assigned to one label, thus providing no analytical value. Another common degenerate solution is a large majority of data points being assigned to a single label, with the remaining labels being assigned very few data points. These problems are commonly observed in deep unsupervised clustering approaches [6] that use off the shelf algorithms like `K-Means++` [1] or `PIC` [31]. In addition to preventing the aforementioned trivial assignments, when we wish to impose partition constraints when assigning the defective images into K subclass categories in a way that ensures that each subclass possesses at least a certain pre-specified fraction of the images. Our revised formulation with partitioning constraints reads as follows:

$$\underset{q,\theta}{\operatorname{argmin}} \; -\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{K} q(y|\mathbf{x}_i)\log p(y|\mathbf{x}_i)$$
$$\text{subject to } \forall y : q(y|\mathbf{x}_i) \in \{0,1\} \text{ and } \sum_{i=1}^{N} q(y|\mathbf{x}_i) = \frac{N}{K} \qquad (2)$$

The integer program in Equation 2 is combinatorial in complexity. One could relax the hard partition constraints and formulate an *optimal transport* problem, minimizing the cost (determined from network outputs as $-\log p(y|\mathbf{x})$) of a soft assignment of images to clusters. Consider the cost matrix P of size $K \times N$ where, $P_{yi} = -\log p(y|\mathbf{x}_i)$ and an assignment matrix Q of size $K \times N$ where, $Q_{yi} = q(y|\mathbf{x}_i)$. The transport objective can thus be defined as:

$$\begin{gathered}\underset{Q}{\operatorname{argmin}}\ \langle Q, P\rangle \\ \text{subject to } Q \in \mathbb{R}_+^{K\times N},\ Q\mathbb{1}^{N\times 1} = \frac{N}{K}\mathbb{1}^{K\times 1},\ Q^T\mathbb{1}^{N\times 1} = \mathbb{1}^{N\times 1}\end{gathered} \tag{3}$$

where $\langle . \rangle$ represents the Frobenius dot product. However, the Network Simplex [37] algorithm needed to solve this problem defined by Equation 3 possesses polynomial time complexity. While better than the original integer program, this un-regularized optimal transport formulation can prohibitively slow for even moderately sized unlabelled datasets. Thus, we formulate the unsupervised clustering problem as an entropy regularized [11] optimal transport problem. The regularized objective can be written as:

$$\begin{gathered}\underset{Q}{\operatorname{argmin}}\ \langle Q, P\rangle - \frac{1}{\lambda}H(Q) \\ \text{subject to } Q \in \mathbb{R}_+^{K\times N},\ Q\mathbb{1}^{N\times 1} = \frac{N}{K}\mathbb{1}^{K\times 1},\ Q^T\mathbb{1}^{N\times 1} = \mathbb{1}^{N\times 1} \\ \text{where } H(Q) = \sum_{y,i} q_{yi}\log q_{yi}\end{gathered} \tag{4}$$

The Lagrangian formulation in Equation 4 has the desirable property that the minimizing solution assumes the form:

$$Q_{\mathsf{OPT}} = Ue^{-\lambda P}V \tag{5}$$

Since $P \in \mathbb{R}_+^{N\times K}$, the Sinkhorn theorem [43] states that there exist α and β such that:

$$\begin{gathered}U = \operatorname{diag}(\alpha), V = \operatorname{diag}(\beta) \\ \text{and } Q_{\mathsf{OPT}} \in \mathbb{R}_+^{K\times N},\ Q_{\mathsf{OPT}}\mathbb{1}^{N\times 1} = \frac{N}{K}\mathbb{1}^{K\times 1},\ Q_{\mathsf{OPT}}^T\mathbb{1}^{N\times 1} = \mathbb{1}^{N\times 1}\end{gathered} \tag{6}$$

Furthermore, the matrices α and β can be converged to in a number of steps that scales linearly with the number of entries in Q, using the Sinkhorn-Knopp [44] iterative approximation scheme as follows:

$$\begin{gathered}\beta_{t+1} \leftarrow \left[\left(e^{-\lambda P}\right)^T {\alpha_{\mathsf{t}}}^{-1}\right] \\ \alpha_{t+1} \leftarrow \frac{K}{N}\left[\left(e^{-\lambda P}\right)^T {\beta_{\mathsf{t+1}}}^{-1}\right]\end{gathered} \tag{7}$$

The Lagrangian formulation in Equation 4 also constrains the assignment matrix w.r.t the cost matrix. It can be shown that Equation 4 minimizes the KL divergence between Q and $e^{-\lambda P}$. This leads to diffuse cluster assignments, which go against our requirement of a hard partition. In practice, however, even moderately large values of λ can lead to very sharp assignments. In these cases, one can be reasonably certain in hard-assigning an image to the cluster with the highest probability. Thus, λ provides a useful lever for controlling the trade-off between the sharpness of assignments and convergence time. As a practical detail, the sub-defect assignment procedure must be robust to common image perturbations. Thus, the aforementioned self-labelling procedure is always performed on randomly transformed images $\psi(\mathbf{x})$ (detailed in section 5).

The self-labelling procedure detailed thus far considers the probabilities output by a feature learner $\mathcal{F}_\theta$ is trained to discern the subclass assignments as a fixed cost in the cluster-assignment problem. However, in the absence of an expert model, the features θ must themselves be learnt alongside the cluster assignments. This is tackled by alternating phases of representation learning and optimal transport. At each representation learning phase, the learner $\mathcal{F}_\theta$ is trained to discern the existing subclass assignments. This leads to a virtuous cycle where better representations lead to improved subdefect discovery and the improved knowledge of latent subclasses improves the learnt representations. In all future discussions, we refer to this model as the *latent discriminator* $\mathcal{F}_{\mathsf{LD}}$.

Our final sub-defect discovery procedure is summarized as:

1. We extract class probabilities over a pre-specified K classes using a ResNet-50 [20] $\mathcal{F}_{\mathsf{LD}}$ model.
2. We obtain equipartitioned hard cluster assignments over the aforementioned K clusters by solving an entropy regularized optimal transport problem using $\mathcal{F}_{\mathsf{LD}}$ predictions as the cost.
3. Subsequent iterations of the latent discriminator are trained using the cluster assignments obtained from the previous iteration as class labels.
4. Steps 1-3 are repeated up to an epoch budget B or the convergence of subclass assignments.

4.2 Training Sub-Task Specialists and Distilling Knowledge to a Generalist

While self-labelling helps us uncover K latent subclasses amongst the defective images, employing a $K+1$ class multi-class classification at this stage is suboptimal for two reasons:

- Individual defect subclasses may have a few hundred data points at most. Training high capacity models such as ViT or Swin transformers with such data sizes leaves room for improvement in performance.
- Some defect subclasses may be majorly composed of acceptable ads that were misclassified by moderator noise. Other subclasses may contain a mixture of unrelated concepts as a result of the partitioning constraints in subsection 4.1. Yet others may contain out-of-domain data that may be counterproductive to the moderation task.

Sample-Efficient Training: The architecture of any learner $\mathcal{F}_\theta$ can encode a set of useful inductive biases that can affect sample-efficiency. Restrictive inductive biases, that are relevant to the downstream task, facilitate the quick bootstrapping of salient representations with minimal training but also impose a lower performance ceiling. For image tasks, convolutions constitute a strong inductive bias in the form of translation equivariance and invariance. The local receptive field enforced by convolutional layers has been shown to minimize Minimum Description Length (MDL) over image datasets, with fully connected networks trained on such data implicitly learning convolutional filters [36]. Meanwhile, self-attention constitutes a relatively weaker inductive bias, able to learn a larger space of functions [9] (and hence having a higher performance ceiling [8]) but being less sample-efficient. Thus, training CNN teacher models which then distill their knowledge to a transformer based student model, offers a principled approach to getting the best of both worlds. This idea has been explored [47] for small classification datasets using a single CNN teacher. However, in the presence of subclass labels (obtained via self-labelling), we demonstrate that casting the overall task into K binary classification tasks (discriminating one subclass of defective advert images [class 0] from acceptable advert images [class 1]) allows us to push performance in low data scenarios to the extreme. As we show in Table 2, training K CNN *specialist teachers* (in our case ResNet-50 models) $\mathcal{F}_{\mathsf{S}j}$ where $j \in \{1, \ldots, K\}$ to discern acceptable adverts $\mathcal{D}_{+}$ from their respective assigned subclass of defective adverts $\mathcal{D}_{-j}$ and distilling that knowledge to a transformer based *generalist student* model leads to large performance gains. The student model is a modified version of ViT-base detailed in [47], rigged with a distillation token (`[DIST]`) in the input in addition to the usual classification token (`[CLS]`). Analogous to the `[CLS]` token, the `[DIST]` token has a distillation head at the corresponding output location where the teacher models distill their logits.

We train the generalist model $\mathcal{F}_{\mathsf{G}}$ on all the selected sub-tasks $\{(\mathcal{D}_{-c}, \mathcal{D}_{+})\}$, sequentially, cycling through the tasks for B epochs. For each sub-task in an epoch, the `[CLS]` head is trained conventionally against hard human moderation labels using a cross-entropy loss. In addition, the soft logits of the respective specialist teacher $\mathcal{F}_{\mathsf{S}c}$ are distilled to $\mathcal{F}_{\mathsf{G}}$ using the `[DIST]` head, again with a cross-entropy loss. The final loss is summarized as:

$$\mathcal{L}_{\mathsf{GLOBAL}} = \mathcal{L}_{\mathsf{CE}}\left(\mathcal{F}_{\mathsf{G}}(\mathbf{x}), y\right) + \mathcal{L}_{\mathsf{CE}}\left(\mathcal{F}_{\mathsf{G}}(\mathbf{x}), \mathcal{F}_{\mathsf{S}c}(\mathbf{x})\right) \tag{8}$$

The `[CLS]` and `[DIST]` heads interact via self-attention, which allows for a principled way to contextualize sub-task probabilities in the backdrop of the general task label. We observe low cosine similarities between the representations produced by the `[DIST]` head and the `[CLS]` head, thus verifying the value added by the introduction of this modification. In the interest of retaining this relative task sub-task context during inference, the predictions are made by averaging the logits of the `[CLS]` and `[DIST]` heads.

Data Denoising: As we later practically demonstrate in Table 2, directly distilling the knowledge from all teacher models to the student transformer model is far from ideal. Due to the presence of clusters composed entirely of noise, out-of-domain data or a mixture of concepts, some teachers' learnt knowledge is antithetical to the overall moderation task. We needed to select teachers based on their competence, and let the general model imbibe knowledge from only the selected teachers (and their corresponding sub-defects). To that end, each of these teachers is evaluated for their competency on the original defect moderation task, using a small `teacher test` set (a subset of the usual `val` set separate from the `test` set of the overall task, refer to section 3). This acts as an implicit marker of cluster importance. Teachers that do not clear an NPV threshold of $\mathcal{T}$ (a pre-specified hyperparameter) at the 1% FPR level are discarded along with their associated cluster data points. This step allows us to discard the vast majority of uncertain or mislabelled data points. All the remaining specialists $\mathcal{F}_{\mathsf{S}c}$ where, $c \in C$ and $C \subseteq \{1, \ldots, K\}$ then distill their knowledge to the student model. Hence, we are able to perform dataset denoising with a small extension to our method.

Algorithm 1 Sample-Efficient Image Moderation Training

```
Require: Labelled moderation data $[\mathcal{D}_{+}, \mathcal{D}_{-}]$, Number of sub-classes in defects $K$, Iteration budget $B$, Teacher selection NPV threshold $\mathcal{T}$
for $\mathcal{I} = 1, \ldots, B$ using $\mathcal{D}_{-}$ do
    Estimate $\mathcal{F}_{\mathsf{LD}}$ parameters mirroring assignments $Q$
    Estimate assignments $Q$ using Sinkhorn-Knopp with $\mathcal{F}_{\mathsf{LD}}$ probabilities as cost
end for
Initiate the set of selected sub-task indices $C \leftarrow \varnothing$
for $\mathcal{I} = 1, \ldots, K$ using $[\mathcal{D}_{+}, \mathcal{D}_{-1}, \ldots, \mathcal{D}_{-K}]$ do
    Estimate $\mathcal{F}_{\mathsf{S}\mathcal{I}}$ parameters on $[\mathcal{D}_{+}, \mathcal{D}_{-\mathcal{I}}]$ over $B$ epochs using teacher val set for model selection
    if NPV at 1% FPR of $\mathcal{F}_{\mathsf{S}\mathcal{I}}$ on teacher test $\geq \mathcal{T}$ then
        $C \leftarrow C \cup \mathcal{I}$
    end if
end for
Initiate $O = \mathcal{M}(C)$ where $\mathcal{M}$ denotes the set permuation function representing a distillation curriculum e.g. for performance curriculum $\mathcal{M}(C)$ is the selected task indices $C$ in descending order of NPV of $\mathcal{F}_{\mathsf{S}c}$ for $c \in C$
for $\mathcal{I} = 1, \ldots, B$ using $[\mathcal{D}_{+}, \mathcal{D}_{-1}, \ldots, \mathcal{D}_{-K}]$ do
    for $\mathcal{J} = 1, \ldots, |O|$ using $[\mathcal{D}_{+}, \mathcal{D}_{-1}, \ldots, \mathcal{D}_{-K}]$ do
        Estimate $\mathcal{F}_{\mathsf{G}}$ parameters on $\left[\mathcal{D}_{+}, \mathcal{D}_{-O[\mathcal{J}]}\right]$ by back-propagating classification loss on moderation labels at the [CLS] head and the distillation loss on the logits of $\mathcal{F}_{\mathsf{S}O[\mathcal{J}]}$ at the [DIST] head using the student val set for model selection
    end for
end for
Ensure: Production ready generalist model $\mathcal{F}_{\mathsf{G}}$
```

4.3 Curriculum-Aware Sub-Task Distillation

The next design choice is to select the order in which the sub-task knowledge is distilled into the generalist model. Existing work [22, 57] shows that a teacher model's order of knowledge delivery over both intermediate feature and predicted label changes can have a significant impact on final student performance. Our final motivation for formulating a sub-task curriculum comes from findings[38] showing that a competency curriculum (based on how predictive of the general task each sub-task is) in a multitask distillation setting (similar to ours) provides measurable benefits over jointly learning the tasks. In the interest of verifying the impact of a multitask curriculum over distillation performance, we experiment with random (samples intermixed), curricular (increasing confidence) and anti-curricular (decreasing confidence) approaches while also trying out different approaches to score sub-task confidence.

The pseudocode for the complete training procedure of the model we launch in production is detailed in algorithm 1.

5 Experiments

We benchmark all experiments using relative improvements to defect interception precision (NPV) at a 1% defect misclassification rate (FPR). Though our methods are equally applicable to any complex defect category, we report our NPV improvements on the interception of sexually explicit advert images (dataset construction is detailed in section 3).

Default Hyperparameter Choices: To control for the effects of hyperparameter choice, all $\mathcal{F}_{\mathtt{G}}$ and $\mathcal{F}_{\mathtt{S}}$ models discussed below are, unless otherwise mentioned, trained with the same configuration. We experimented with a range of learning rates $\in [8e-7, 1e-6, 5e-5, 3e-4]$ and settled upon a learning rate of $8e-7$ for the encoder layers and a learning rate of $3e-4$ for the classification and distillation heads. We employ an ADAM optimizer with decoupled weight decay [34], setting the weight decay to 0.01 for all parameters except the bias weights and clip the gradient norms to 2.0. We also use a dropout of 0.1. For all model training runs, we use a cross-entropy loss with a label smoothing coefficient of 0.05 and sample image augmentations using RandAugment [10] with 4 operations per image and a max magnitude coefficient of 8. Finally, we set an epoch budget B of 20 in all training runs as it seemed like a safe cut-off point with our all training runs converging well before this point.

5.1 Regular Fine-Tuning Baselines

We commence our evaluations by verifying the advantages of leveraging the best of convolution and self-attention, in the image moderation domain. First, we train a ViT model on the regular split of the general task labels, using conventional fine-tuning. We choose the ViT-Base model, which has roughly 86M parameters, for all our experimentation. Specifically, we use a version that works on image patches of size 16×16. This model is the baseline to which all other variations are compared to, as it represents the previously deployed moderation model. In the interest of fairness, we compare this baseline to a CNN model having roughly the same number of parameters. A ResNeXt-101 [52] model, with 84M parameters, is trained in a manner identical to our baseline. We also evaluate two other baselines that combine the best attributes of convolution and self-attention. Firstly, we fine-tune a CoAtNet-2 model with 75M parameters. This architecture explicitly combines convolution and self-attention blocks. Secondly, we evaluate the Swin transformer, which modifies the attention mechanism to partly incorporate the inductive biases of convolution. We use the Swin-Base model with a patch size of 4×4 and a shift window of 7, having 88M parameters. As shown under the section Regular Fine-Tuning Baselines in Table 2, models combining the helpful properties of convolution and self-attention perform the best.

Next, we justify our choice of distilling knowledge from CNN teacher models to improve the sample efficiency of image transformers. We run an ablation where a DeiT model learns from the moderation labels as well as the logits of a ResNet-50 teacher model (which itself has been trained on all the data). As shown in the Single Teacher Distillation section of Table 2, this approach outperforms architectural interventions that seek into combine the properties of convolution and self-attention.

5.2 Sub-Task Discovery

The negative training split samples sourced in section 3 are clustered for subclass discovery via the self-labelling approach described in subsection 4.1. We run the iterative procedure using an ImageNet pre-trained ResNet-50 learner $\mathcal{F}_{\mathtt{LD}}$ over varying settings of epoch count $\in [100, 300, 1000]$ and defect sub-label count $K \in [20, 100, 500]$. Empirically, we settled on a high learning rate of 0.03 (multiplied by a factor of 0.1 after every quartile of the training is complete) and an aggressive weight decay of 0.1 as the combination of hyperparameters that reliably provided us with salient cluster assignments. We experiment with $\lambda = \{10, 50, 125, 250, 1000\}$ as our entropy-regularization coefficient. We find $\lambda = 250$ to be the sweet spot, as it is sufficiently large to prevent diffuse cluster assignments and not large enough to make self-labelling prohibitively slow. As noted in Table 2, the 300 epoch and 100 cluster configuration consistently performs the best.

Subsequently, we train separate ResNet-50 specialist teachers $\mathcal{F}_{\mathtt{S}j}$ for $j \in \{1, \ldots, K\}$ to distinguish the cluster j of negative samples (special task label 0) from all the positive samples (special task label 1), with checkpoint selection done on the `teacher val` set of the student-teacher split. All the teacher models are then evaluated on the `teacher test` set of the student-teacher split. We retain only the teachers that clear the NPV threshold $\mathcal{T} = 0.1$ at our prespecified 1% FPR level[8]. Subsequent knowledge transfer to the generalist model is only conducted using images in the clusters associated with the selected models.

5.3 Sub-Task Distillation Ablations

Before we delve into the performance gains afforded by the combination of uncovering latent sub-tasks and denoising via omitting irrelevant sub-tasks, we wish to investigate the value of each in its own right. To that end, we investigate the performance of a Swin-Base model when trained on the $K + 1$ class classification task cast by the K defect subclasses uncovered by self-labelling under the section Self-Label Multi-Class in Table 2. Additionally, we evaluate the performance improvements on the baseline brought about by ditching the teacher selection phase and distilling knowledge from all teachers (AT) under the All Teacher Distillation section of Table 2. These represent value added by uncovering latent sub-defects types in the absence of denoising. We also evaluate the performance of a Swin-base model on the binary classification task cast by discriminating the sub-defects corresponding to selected teachers (class label 0) from the set of acceptable ads (class label 1) under the section Filtered Fine-Tuning in Table 2. This represents the value added by denoising without utilizing the fine-grained sub-defect information uncovered during self-labelling. While each of these interventions of ours has a positive effect, the later sections of Table 2 show that combining them can really push the moderation performance significantly.

In the Filtered Teacher Ablations section of Table 2, we detail the sizeable performance gains afforded by filtering teachers (FT) that clear the NPV threshold and only learning from them. We explore a range of self-labelling configurations (number of epochs and pre-specified clusters). Additionally, we benchmark the comparative effects of a soft and hard label transfer during knowledge distillation. While soft label transfer is more efficient at transferring the inductive biases of the teacher model, it can also pass on more of the learnt noise. We compare both these approaches using a random training order where the image samples, general task labels and teacher sub-task logits from all samples across all sub-tasks are randomly interspersed. In our experiments, we find that soft label transfer (SD) consistently outperforms hard label transfer (HD)

[8] While we experimented with multiple teacher selection NPV thresholds, we find that a threshold of 0.1 was by far the best

Table 2: Benchmarks and ablations over `test` set (Relative metrics owing to confidentiality constraints)

Model Category	Method	NPV Absolute % Gain	
		@5% FPR	@1% FPR
Regular Fine-Tuning Baselines	ViT-Base/16 [14]	–	–
	ResNeXt-101 [52]	-3.3%	-3.9%
	CoAtNet-2 [12]	+1.2%	+1.3%
	Swin-B/P4W7 [33]	+3.7%	+4.1%
Single Teacher Distillation	DeiT-Base/16 + Teacher ResNet-50	+8.4%	+10.8%
Self-Label Multi-Class	100 Epoch SL K20 + Swin-B/P4W7	+0.9%	+1.4%
	300 Epoch SL K20 + Swin-B/P4W7	+1.5%	+2.2%
	1000 Epoch SL K20 + Swin-B/P4W7	+0.7%	+1.6%
	100 Epoch SL K100 + Swin-B/P4W7	+1.4%	+1.9%
	300 Epoch SL K100 + Swin-B/P4W7	+8.6%	+10.8%
	1000 Epoch SL K100 + Swin-B/P4W7	+8.3%	+10.1%
	100 Epoch SL K500 + Swin-B/P4W7	+1.2%	+1.8%
	300 Epoch SL K500 + Swin-B/P4W7	+4.4%	+6.3%
	1000 Epoch SL K500 + Swin-B/P4W7	+4.6%	+6.3%
All Teacher Distillation	300 Epoch SL K100 + DeiT-Base/16 + AT	+4.0%	+6.7%
Filtered Data Fine-Tuning	300 Epoch SL K100 + Swin-B/P4W7	+4.5%	+7.1%
Filtered Teacher Ablations	100 Epoch SL K20 + DeiT-Base/16 + FT + HD	+8.5%	+9.1%
	300 Epoch SL K20 + DeiT-Base/16 + FT + HD	+9.7%	+9.6%
	1000 Epoch SL K20 DeiT-Base/16 + FT + HD	+8.9%	+9.2%
	100 Epoch SL K100 + DeiT-Base/16 + FT + HD	+9.7%	+12.6%
	300 Epoch SL K100 + DeiT-Base/16 + FT + HD	+10.4%	+15.1%
	1000 Epoch SL K100 + DeiT-Base/16 + FT + HD	+10.2%	+13.0%
	100 Epoch SL K500 + DeiT-Base/16 + FT + HD	+9.9%	+12.5%
	300 Epoch SL K500 + DeiT-Base/16 + FT + HD	+10.4%	+14.7%
	1000 Epoch SL K500 + DeiT-Base/16 + FT + HD	+10.5%	+15.0%
	300 Epoch SL K100 + DeiT-Base/16 + FT + SD	+13.4%	+15.6%
Distillation Curriculum Ablations	300 Epoch SL K100 + DeiT-Base/16 + FT + SD + APC	+13.8%	+14.2%
	300 Epoch SL K100 + DeiT-Base/16 + FT + SD + PC	+17.7%	+20.3%
	300 Epoch SL K100 + DeiT-Base/16 + FT + SD + ACC	+11.9%	+14.5%
	300 Epoch SL K100 + DeiT-Base/16 + FT + SD + CC	+22.4%	+27.1%
Semi-Supervised Benchmarks	SWSL [53]	+19.7%	+23.9%
	NoisyStudent-L2 [51]	+21.8%	+23.5%
	300 Epoch SL K100 + DeiT-Base/16 + FT + SD + CC + UL	**+28.8%**	**+30.2%**

w.r.t NPV. This is an interesting finding as it contradicts existing findings [48].

So far, all the multi-teacher distillation evaluations have allowed the teachers to distill their knowledge to the student in random order. We now benchmark the effects of distilling using a task curriculum (see Distillation Curriculum Variations in Table 2). Herein, samples from sub-tasks are grouped and fed to the student sequentially in increasing order of confidence. This reduces the occurrence of divergent samples over time as the sub-task data distribution increasingly approximates the general task distribution. We compare two measures of confidence. First, we arrange sub-tasks in increasing order of average cluster confidence (CC) as given by the $\mathcal{F}_{\text{LD}}$ ResNet-50 in subsection 4.1. Secondly, we order the selected sub-tasks $c \in C$ in order of the performance (PC) of their respective learners $\mathcal{F}_{\text{Sc}}$ over the `teacher val` set of the student-teacher split. We find that while both the curricula present improvements in final downstream performance, CC does yield the greatest improvements on the NPV front. This effect is further verified by testing the anti-curriculum order (APC and ACC respectively) i.e., distilling the samples in decreasing order of confidence, and observing the obvious weaker effect it has viz-a-viz the random order distillation.

5.4 Unlabelled Data Sourcing

When faced with a classification task that offers only a few labelled samples, sourcing unlabelled near-domain data is a common workaround. Such data may be provided with *pseudo-labels* using label propagation mechanisms [4, 5]. However, the label propagation paradigm does not avail us with means to selectively discard uninformative or counterproductive samples from the unlabelled data and hence usually only work well with in-domain rather than near-domain unlabelled data. Alternatively, for near-domain data, one might attempt self-training based learning approaches [51, 53] to bootstrap some measure of classifier confidence w.r.t. data points. This formulation suffers from the lack of an idiomatic means to specify partitioning constraints among semantic sub-concepts. Our framework, thus, presents an easy way to solve both the challenges. In the specialist teacher selection process, the set of classes C corresponding to selected teachers, can be seen as concepts useful in solving the overall task. Thus, one could retain the unlabelled samples classified as $c \in C$ by our latent discriminator $\mathcal{F}_{\text{LD}}$.

To explore the competitiveness of our approach, in such a setting, we obtain $50,000$ unlabelled generic product catalogue images. The distribution of product types and defect types in this set differs significantly from our moderation task, thus presenting a pertinent real-world test of handling near-domain data. The Semi-Supervised section of Table 2 demonstrates how our method is able to positively leverage near-domain unlabelled data (UL) for even greater performance improvements, affording a 30.2% gain in NPV over the previously deployed baseline. This is the model we finally deploy to production. Our method compares favourably to two strong self-training semi-supervised learning baselines, owing to the fact that it does not simply rely on imputed class probabilities, but also factors in subclass structure.

6 Production Details and Impact

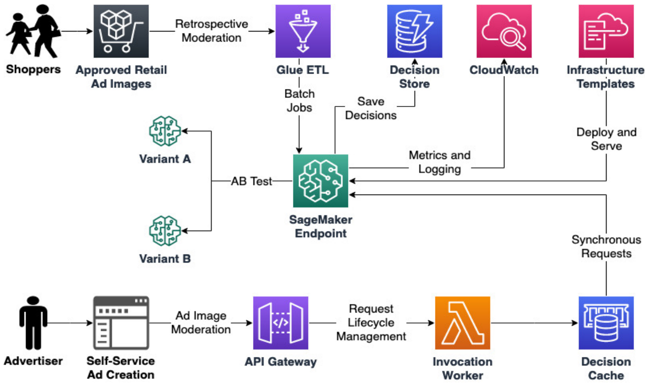


Figure 3: Components and interactions in our model serving architecture

Our described solution, with the best configuration in Table 2, has been deployed in production on an AWS Sagemaker endpoint. The model scores newly created adverts as part of the self-service advert creation workflow. Additionally, we use our new model to score and retrospectively moderate previously approved ads. To ensure provenance and idempotency, we maintain our infrastructure using the Infrastructure as Code templates offered by AWS Cloud Development Framework[9]. Our model invocation infrastructure is completely serverless, as evidenced by Figure 3.

To gauge the effectiveness of our interventions, we carry out periodic audits of advertisements that have made it past the moderation phase. These are sampled in a weighted manner, with the sampling weight being biased by the impressions that an advert gains in the marketplace. These adverts are then moderated in accordance with the aforementioned moderation policy, and the weighted proportion of adverts exhibiting our defect of interest are calculated and tracked as Weighted False Positive Rate (WFPR).

Concurrently, new models are launched via a long-term A/B Test. The aforementioned WFPR is then broken down and attributed to specific model variants, based on the model logs. This enables us to track the reduction in WFPR for the new model on its subset w.r.t the incumbent model on its subset of the scored adverts. On performing such an analysis, we observe that the model has had a demonstrable contribution in reducing the WFPR attributable to our defect of concern. In the course of a few weeks post deployment, a 48.7% relative reduction in WFPR has been observed (refer to Table 3) compared to the incumbent ViT model (trained via vanilla fine-tuning). Given that sexually explicit adverts are the most prevalent defect category, this reduction of WFPR will lead to a safer and more accessible customer experience when retail shopping on Amazon.

Table 3: Percentage WFPR reduction w.r.t 4-week average one week before deployment (Relative metrics owing to confidentiality constraints)

Weeks Since Deployment	4	8	12	16
% WFPR Reduction	-15.1%	-20.7%	-29.3%	-48.7%

7 Conclusion

We delve into efficiently training vision transformers for complex image moderation tasks and outline an effective technique to achieve this by decomposing the task into several sub-tasks via optimal transport based self-labelling. We also demonstrate how fast-learning high inductive bias CNN teacher models can quickly learn specific sub-tasks and distill their knowledge to a higher capacity but slow learning transformer model, which is eventually deployed in production. Our method allows the final model to master all the latent moderation sub-tasks without needing any additional manual labelling for noise removal or subclass discovery.

We benchmark the efficacy of our methods on an image moderation task that involves intercepting sexually explicit advert images. Additionally, we measure the direct improvements to the retail shopper experience enabled by our methods, by detailing the relative reduction in the prevalence of such images in the weeks following model deployment.

Acknowledgments

We would like to thank *Manisha Verma* for proofreading the work and providing constructive feedback.

[9] https://aws.amazon.com/cdk

References

[1] David Arthur and Sergei Vassilvitskii. 2006. *k-means++: The advantages of careful seeding*. Technical Report. Stanford.

[2] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. 2019. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371* (2019).

[3] Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).

[4] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. 2020. ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HklkeR4KPB

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems* 32 (2019).

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 132–149.

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882* (2020).

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294* (2021).

[9] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. 2020. On the Relationship between Self-Attention and Convolutional Layers. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HJlnC1rKPB

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 702–703.

[11] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* 26 (2013), 2292–2300.

[12] Zihang Dai, Hanxiao Liu, Quoc Le, and Mingxing Tan. 2021. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems* 34 (2021).

[13] Thomas G Dietterich. 2000. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of artificial intelligence research* 13 (2000), 227–303.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

[15] Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang. 2020. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *advances in neural information processing systems* 33 (2020), 12345–12355.

[16] AC Earle, A Saxe, and B Rosman. 2018. Hierarchical subtask discovery with non-negative matrix factorization. In *Proceedings of the ICLR Conference 2018*. OpenReview.

[17] Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. Efficiently Identifying Task Groupings for Multi-Task Learning. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=hqDb8d65Vfh

[18] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017. Efficient Knowledge Distillation from an Ensemble of Teachers.. In *Interspeech*. 3697–3701.

[19] Eshwar Shamanna Girishekar, Shiv Surya, Nishant Nikhil, Dyut Kumar Sil, Sumit Negi, and Aruna Rajan. 2021. Training Language Models under Resource Constraints for Adversarial Advertisement Detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*. 280–287.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[21] Jean-Michel Hupé. 2015. Statistical inferences under the Null hypothesis: common mistakes and pitfalls in neuroimaging studies. *Frontiers in neuroscience* 9 (2015), 18.

[22] Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. 2019. Knowledge distillation via route constrained optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1345–1354.

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems* 33 (2020), 18661–18673.

[24] Thomas Kipf, Yujia Li, Hanjun Dai, Vinicius Zambaldi, Alvaro Sanchez-Gonzalez, Edward Grefenstette, Pushmeet Kohli, and Peter Battaglia. 2019. Compile: Compositional imitation learning and execution. In *International Conference on Machine Learning*. PMLR, 3418–3428.

[25] Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. 2009. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience* 12, 5 (2009), 535–540.

[26] Kisoo Kwon, Hwidong Na, Hoshik Lee, and Nam Soo Kim. 2020. Adaptive Knowledge Distillation Based on Entropy. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7409–7413. https://doi.org/10.1109/ICASSP40776.2020.9054698

[27] Daniel Lee and H Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems* 13 (2000).

[28] Sang-Hyun Lee and Seung-Woo Seo. 2020. Learning compound tasks without task-specific knowledge via imitation and self-supervised learning. In *International Conference on Machine Learning*. PMLR, 5747–5756.

[29] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. 2021. Prototypical Contrastive Learning of Unsupervised Representations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=KmykpuSrjcq

[30] Wei-Hong Li and Hakan Bilen. 2020. Knowledge distillation for multi-task learning. In *European Conference on Computer Vision*. Springer, 163–176.

[31] Frank Lin and William W Cohen. 2010. Power iteration clustering. In *ICML*.

[32] Yuang Liu, Wei Zhang, and Jun Wang. 2020. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing* 415 (2020), 106–113.

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[34] Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam. (2018).

[35] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2020. Subclass distillation. *arXiv preprint arXiv:2002.03936* (2020).

[36] Behnam Neyshabur. 2020. Towards Learning Convolutions from Scratch. *Advances in Neural Information Processing Systems* 33 (2020).

[37] James B Orlin. 1997. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming* 78, 2 (1997), 109–129.

[38] Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. 2015. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5492–5500.

[39] Jason Tyler Rolfe. 2017. Discrete Variational Autoencoders. *ArXiv* abs/1609.02200 (2017).

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[41] Andrew M. Saxe, Adam C. Earle, and Benjamin Rosman. 2017. Hierarchy Through Composition with Multitask LMDPs. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3017–3026. https://proceedings.mlr.press/v70/saxe17a.html

[42] D Sculley, Matthew Eric Otey, Michael Pohl, Bridget Spitznagel, John Hainsworth, and Yunkai Zhou. 2011. Detecting adversarial advertisements in the wild. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 274–282.

[43] Richard Sinkhorn. 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics* 35, 2 (1964), 876–879.

[44] Richard Sinkhorn and Paul Knopp. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.* 21, 2 (1967), 343–348.

[45] Sungryull Sohn, Junhyuk Oh, and Honglak Lee. 2018. Hierarchical reinforcement learning for zero-shot generalization with subtask dependencies. *Advances in Neural Information Processing Systems* 31 (2018).

[46] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. 2021. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270* (2021).

[47] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2020. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877* (2020).

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 10347–10357.

[49] Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. 2021. {RODE}: Learning Roles to Decompose Multi-Agent Tasks. In *International Conference on Learning Representations*. https://openreview.net/

forum?id=TTUVg6vkNjK

[50] Li Weng and Bart Preneel. 2011. A secure perceptual hash algorithm for image content authentication. In *IFIP International Conference on Communications and Multimedia Security*. Springer, 108–121.

[51] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10687–10698.

[52] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1492–1500.

[53] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546* (2019).

[54] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1285–1294.

[55] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3712–3722.

[56] Yi Zhang, Sujay Kumar Jauhar, Julia Kiseleva, Ryen White, and Dan Roth. 2021. Learning to decompose and organize complex tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2726–2735.

[57] Qingqing Zhu, Xiuying Chen, Pengfei Wu, JunFei Liu, and Dongyan Zhao. 2021. Combining Curriculum Learning and Knowledge Distillation for Dialogue Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 1284–1295. https://doi.org/10.18653/v1/2021.findings-emnlp.111