HIERARCHICAL COMPLEMENT OBJECTIVE TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Hierarchical label structures widely exist in many machine learning tasks, ranging from those with explicit label hierarchies such as image classification to the ones that have latent label hierarchies such as semantic segmentation. Unfortunately, state-of-the-art methods often utilize cross-entropy loss which in-explicitly assumes the independence among class labels. Motivated by the fact that class members from the same hierarchy need to be similar to each other, we design a new training diagram called Hierarchical Complement Objective Training (HCOT). In HCOT, in addition to maximizing the probability of the ground truth class, we also neutralize the probabilities of rest of the classes in a hierarchical fashion, making the model take advantage of the label hierarchy explicitly. We conduct our method on both image classification and semantic segmentation. Results show that HCOT outperforms state-of-the-art models in CIFAR100, Imagenet, and PASCAL-context. Our experiments also demonstrate that HCOT can be applied on tasks with latent label hierarchies, which is a common characteristic in many machine learning tasks.

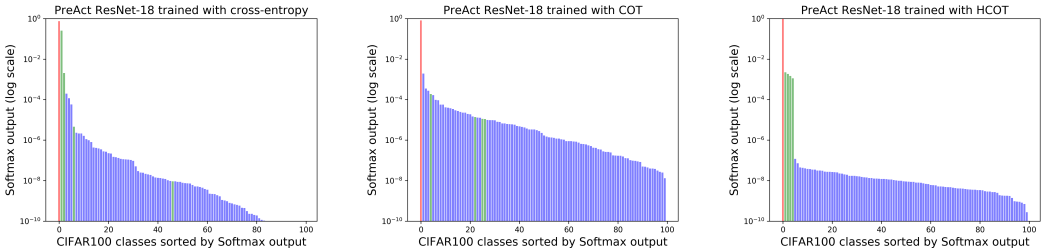
1 INTRODUCTION

Many machine learning tasks involve making predictions on classes that have an inherent hierarchical structure. One example would be image classification with hierarchical categories, where a category shares the same *parental category* with other ones. For example, the categories with label "dog" and "cat" might share a common parental category "pet", which forms a *explicit label hierarchy*. Another example would be in the task of semantic segmentation, where "beach", and "sea" are under the same theme "scenery" which forms a *latent label hierarchy*, while "people", and "pets" forms another one of "portrait." In this work, we call a parental category a *coarse(-level) category*, while a category under a coarse category is called a *fine(-level) category*.

Many successful deep learning models are built and trained with cross-entropy loss that assumes prediction classes to be mutually independent. This assumption works well for many tasks such as traditional image classifications where no hierarchical information is present. In the explicitly hierarchical setting, however, one problem is that learning with objectives that pose such a strong assumption makes the model difficult to utilize the hierarchical structure in the label space. Another challenge in modeling hierarchical labels is that many tasks sometime exhibit latent label hierarchy. Take semantic segmentation for example, an inherent hierarchical structure has been explored by (Zhang et al., 2018a) as "'global context". However, the dataset itself does not contain hierarchical information.

In this paper, we develop techniques that are capable of leveraging the information in a label hierarchy, through proposing new training objectives. Our proposed technique is different from previous methods (Yan et al., 2015; Murdock et al., 2016; Guo et al., 2018; Zhang et al., 2018a) which exploit the label hierarchy by changing model architectures but not the objectives. The general idea we propose is to penalize incorrect classes at different granularity levels: the classes that are "obviously wrong"—different from not only the ground truth but also the parental category of ground truth—should receive larger penalty than the ones that share the same parental categories of ground truth. Such a mechanism allows us to take advantage of the information in the label hierarchy during training.

To achieve this goal of training with hierarchy information, we introduce the concept of Complement Objective Training (COT) (Chen et al., 2019b;a) into label hierarchy. In COT, the probability of



(a) Baseline (cross-entropy) (b) COT (c) HCOT

Figure 1: Sorted predicted probabilities (denoted as $\hat{y}$) from three different training paradigms evaluated on CIFAR-100 dataset using PreAct ResNet-18. The red bar indicates the probability of the ground-truth (denoted as $\hat{y}_g$), the green bars are the probabilities of classes in the same parental category as the ground-truth (denoted as $\hat{y}_{G\setminus\{g\}}$), and blue bars are the probabilities of the rest classes (denoted as $\hat{y}_{K\setminus G}$, see Sec. 3 for detailed notation definition). Notice the "staircase shape" in (c) showing the significant difference between $\hat{y}_g$ and $\hat{y}_{G\setminus\{g\}}$, and then between $\hat{y}_{G\setminus\{g\}}$ and $\hat{y}_{K\setminus G}$, which confirms HCOT well captures the label hierarchy.

the correct class is maximized by a primary objective (i.e., cross-entropy), while the probability of incorrect classes are neutralized by a complement objective (Chen et al., 2019b). This training paradigm aims at widening the gaps between the predicted probability value of the ground truth and those of the incorrect classes. In this paper, we propose **H**ierarchical **C**omplement **O**bjective **T**raining (HCOT) with a novel complement objective called "Hierarchical Complement Entropy" (defined in Sec. 3), by applying the idea of the complement objective on both the fine-level class and its corresponding coarse-level classes.

HCOT learns the class probabilities by three folds: (a) maximizing the predicted probability of ground truth, (b) neutralizing the predicted probabilities of incorrect classes sharing the same coarse-level category as the ground truth, and (c) further penalizing others that are on different branches (in the label hierarchy) to the ground-truth class. Figure 1 illustrates the general idea of HCOT compared to cross-entropy and COT, which shows HCOT leads to both confident prediction for the ground-truth class and the predicted distribution that better reflects the label hierarchy (and therefore closer to the true data distribution). Particularly, the probability mass of the classes belonging to the parental category of the ground truth (in green) to be significantly higher than the rest of the classes (in blue). In other words, the model is trained to strongly penalize the obviously wrong classes that are completely irrelevant to both the ground-truth class and other classes belonging to the same parental category.

We conduct HCOT on two important problems: image classification and semantic segmentation. Experimental results show that models trained with the Hierarchical complement entropy achieve significantly better performance over both cross-entropy and COT, across a wide range of state-of-the-art methods. We also show that HCOT improves model performance when predicting the coarse-level classes. And finally, we show that HCOT can deal with not only tasks with *explicit label hierarchy* but also those with *latent label hierarchy*. To the best of our knowledge, HCOT is the first paradigm that trains deep neural models using an objective to leverage information from a label hierarchy, and leads to significant performance improvement.

2 Background

Explicit Label Hierarchy. Many tasks exhibit explicit label hierarchy that are presented as part of the dataset. Explicit hierarchical structures exist among the class labels for a wide range of problems. Taking visual recognition as an example, there have been many prior arts on non-neural models focused on exploiting the hierarchical structure in categories (Tousch et al., 2012). For neural models, HD-CNN (Yan et al., 2015) is an early work using the category hierarchy to improve performance over the flat N-way deep-network classifiers. The network architecture of HD-CNN contains a coarse component and several fine-grained components for learning from labels of different levels.

Unlike HD-CNN which uses one fixed model, Blockout (Murdock et al., 2016) uses a regularization framework that learns both the model parameters and the sub-networks within a deep neural network, to capture the information in a label hierarchy. Another prior art (Guo et al., 2018) combines the CNN-based classifier with a Recurrent Neural Network to exploit the hierarchical relationship, sequentially from the coarse categories to the fine ones. All of the above-mentioned approaches rely on modifying model architectures to capture the hierarchical structures among the class labels. This raises an intriguing question: Is it possible to design a training objective, rather than proposing a new model architecture, for a deep neural network to effectively capture the information contained in a label hierarchy?

Latent Label Hierarchy. Another group of tasks are rather exclusive on the hierarchical information but has an underlying assumption on an inherent label structure. Semantic segmentation is one of such tasks where co-occurrence of the class labels forms a latent label hierarchy. This hierarchy is not directly observed in the data but can be inferred from the data. In semantic segmentation, the goal is to assign a semantic label to each pixel of an image. Typically, when training a deep network model for semantic segmentation, the information of individual pixels are usually taken in isolation. That is, the per-pixel cross-entropy loss is calculated for an image, with respect to the ground truth labels. To consider the global information, EncNet (Zhang et al., 2018a) first utilizes the semantic context of scenes by exploiting model structures and provides a strong baseline in semantic segmentation. However, we argue that the potential of leveraging global information on the labeling space is still not discovered.

3 Hierarchical Complement Objective Training

In this section, we introduce the proposed Hierarchical Complement Objective Training (HCOT), which is a new training paradigm for leveraging information in a label hierarchy. Specifically, a novel training objective, Hierarchical Complement Entropy (HCE), is defined as the complement objective for HCOT. In the following, we first review the concept of the complement objective, and then provide the mathematical formulation of HCE.

Complement Objective. In Complement Objective Training (COT) (Chen et al., 2019b), a neural model is trained with both a primary objective and a complement objective: the primary objective (e.g., cross-entropy) is for maximizing the predicted probability of the ground-truth class, whereas the complement objective (e.g., complement entropy (Chen et al., 2019b)) is designed to neutralize the predicted probabilities of incorrect classes, which intuitively makes a model more confident about the ground-truth class. Eq(1) gives the definition of the complement entropy:

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{H}(P_{K\setminus\{g\}}(z_i)) \tag{1}$$

where N is the total number of samples, K is the set of labels. For the i^{th} sample, z_i is the vector of logits for the sample. Let g be the corresponding ground-truth class for the i^{th} sample, so $K \setminus \{g\}$ represents the set of incorrect classes. We use $\mathcal{H}$ to annotate the Shannon entropy function (Shannon, 1948) over the probability $P_{K\setminus\{g\}}(z_i)$, defined below.

$$\mathcal{H}(P_{K\setminus\{g\}}(z_i)) = -\sum_{j}\left(P_{K\setminus\{g\}}(z_i)_j\right)\log\left(P_{K\setminus\{g\}}(z_i)_j\right) \tag{2}$$

where $j \in K \setminus \{g\}$, and the probability function $P_{K\setminus\{g\}}(z_i)_j$ is defined as the output of the softmax function:

$$P_{K\setminus\{g\}}(z_i)_j = \frac{e^{z_{i,j}}}{\sum_{k\in K\setminus\{g\}} e^{z_{i,k}}}. \tag{3}$$

Intuitively, $P_{K\setminus\{g\}}(z_i)_j$ is the j^{th} dimension of a multinomial distribution normalized among the incorrect classes over logits z_i (that is, excluding the probability mass of the ground-truth class). Please note that the alternative definition of complement entropy is mathematically equivalent to the one presented in (Chen et al., 2019b).

Despite the good performance by maximizing Complement entropy to make complement events equally like to occur, this approach do not consider the generalization gap between predicted distributions and true data distributions. For example, if the ground-truth class is "dog", to flatten the predicted probabilities on irrelevant classes such as "cat" and "truck" is counter-intuitive.

Hierarchical Complement Entropy. The proposed Hierarchical Complement Entropy (HCE) regulates the probability masses, similar to what the complement entropy does, but in a hierarchical fashion. Let a subgroup G be a set that contains the sibling classes that belong to the same parental class of the ground-truth class, that is, $g \in G$ and $G \subseteq K$. HCE will first regulate complement entropy between the subgroup G and the ground truth g followed by the complement entropy between label space K and subgroup G. Detailed definition can be found in Eq(3). The proposed HCE is defined as the following with θ being the model parameters:

$$HCE(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathcal{H}(P_{G\setminus\{g\}}(z_i)) + \mathcal{H}(P_{K\setminus G}(z_i))\right] \tag{4}$$

It is not hard to see that Eq(4) is a direct implementation of the predicted probabilities trained with HCOT procedure in Figure 1, which impose probability regulation based on the hierarchical structure of the labels. $\mathcal{H}(P_{G\setminus\{g\}}(z_i))$ regulates inner hierarchy, which corresponds to the relationship between the probability masses marked as red and green. The second term, $\mathcal{H}(P_{K\setminus G}(z_i))$, regulates the outer hierarchy, which corresponds to the relationship between the green and blue class labels. Hierarchical complement entropy ensured that the gaps between each of the hierarchies are as wide as possible to enforce the hierarchical structure during training. In the extreme case when $K = G$, the second term in Eq(4) disappears and the Hierarchical complement entropy degenerates to Complement entropy.

Optimization. Our loss function consists of two terms: the normal cross entropy term (i.e., $XE(\theta)$), and the complement objective term $HCE(\theta)$.

$$\mathcal{L}(\theta) = XE(\theta) - HCE(\theta) \tag{5}$$

In *Direct optimization*, we optimize Eq(5) directly using SGD. An alternative approach is *Alternative optimization*, which optimizes the cross-entropy term and the complement objective term interleaved. This is done by maximizing HCE followed by minimizing XE for a single training iteration. In our paper, we choose between these two methods to achieve the best performance for our models.

4 IMAGE CLASSIFICATION

Classification usually comes with explicit hierarchy. We evaluate HCOT on image classification. Experiments are conducted with two widely-used datasets that contain label hierarchy: CIFAR-100 (Krizhevsky, 2009) and ImageNet-2012 (Krizhevsky et al., 2012).

4.1 CIFAR-100

CIFAR-100 is a dataset consisting of 60k colored natural images of 32x32 pixels equally divided into 100 classes. There are 50k images for training and 10k images for testing. The official guide CIFAR-100 (Krizhevsky, 2009) further group the 100 classes into 20 coarse classes where each coarse class contains five fine classes, forming the label hierarchy. Therefore, each image sample has one fine label and one coarse label. Here we follow the standard data augmentation techniques (He et al., 2016b) to pre-process the dataset. During training, zero-padding, random cropping, and horizontal mirroring are applied to the images. For the testing images, we use the original images of 32×32 pixels.

Experimental Setup. For CIFAR-100, we follow the same settings as the original ResNet paper (He et al., 2016b). Specifically, the models are trained using SGD optimizer with momentum of 0.9; weight decay is set to be 0.0001 and learning rate starts at 0.1, then being divided by 10 at the

100^{th} and 150^{th} epoch. The models are trained for 200 epochs, with a mini-batch size of 128. For training WideResNet, we follow the settings described in (Zagoruyko & Komodakis, 2016), and the learning rate is divided by 10 at the 60^{th}, 120^{th} and 180^{th} epoch. In addition, no dropout (Srivastava et al., 2014) is applied to any baseline according to the best practices in (Ioffe & Szegedy, 2015). We follow alternating training (Chen et al., 2019b), where models are trained by alternating between the primary objective (i.e., cross-entropy) and the complement objective (i.e., Hierarchical Complement Entropy).

Results. Our method demonstrates improvements over all of the state-of-the-art models compared to baseline and COT, improving error rates by a significant margin. These models range from the widely used ResNet to the SE-ResNet (Hu et al., 2018), which is the winner of the ILSVRC 2017 classification competition. SE-ResNet considers novel architecture units named Squeeze-and-Excitation block (SE block) in ResNet framework for explicitly capturing the inter-dependencies between channels of convolutional layers. Results are shown in Table 1.

Table 1: Error rates (%) on CIFAR-100 using ResNet, SE-ResNet, and variants of ResNet.

Model	Baseline	COT	**HCOT**
ResNet-56 (He et al., 2016b)	29.41	27.76	**27.3**
ResNet-110 (He et al., 2016b)	27.93	27.24	**26.46**
SE-ResNet-56 (Hu et al., 2018)	28.11	27.04	**26.54**
SE-ResNet-110 (Hu et al., 2018)	26.49	26.09	**25.49**
PreAct ResNet-18 (He et al., 2016a)	25.44	24.73	**23.8**
ResNeXt-29 (2×64d) (Xie et al., 2017)	23.45	21.9	**21.64**
WideResNet-28-10 (Zagoruyko & Komodakis, 2016)	21.91	20.99	**20.32**

Results with Mixup and Cutout. We also show that HCOT can be applied in synergy with other commonly-used techniques to further improve model performance. We conduct experiments on ResNet-110 with "Cutout" (Devries & Taylor, 2017) for input masking and "Mixup " (Zhang et al., 2018b) for data augmentation. Table 2 shows the accuracy of models trained with HCOT consistently outperform the baseline and the models trained with COT.

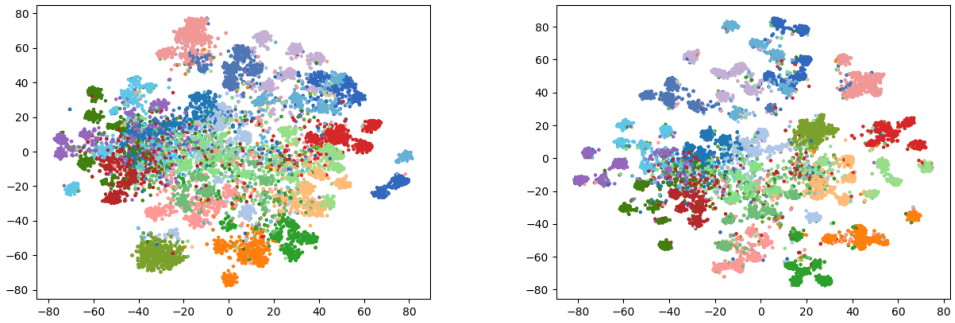
Table 2: Error rates (%) on CIFAR-100 using ResNet with Cutout and Mixup techniques.

Model	Baseline	COT	**HCOT**
ResNet-110 + Cutout	24.61	23.93	**23.85**
ResNet-110 + Mixup	24.46	23.82	**23.33**

Analysis on Coarse-level Labels. To understand the places where performance improvements of HCOT coming from, we show the results by splitting them into coarse and fine labels in Table 3. Here we see that HCOT improves the performance significantly on the coarse-level labels, where COT hardly improves. Such a performance improvement is a direct result of modeling label hierarchies, which is not taken into account in either baseline or COT. Surprisingly, HCOT also improves fine-level labels significantly, over the already improved the results of COT. This suggests that modeling of the fine-level labels can benefit from modeling label hierarchies.

Table 3: Error rates (%) on both coarse and fine classes on CIFAR-100 using SE-PreAct ResNet-18.

Label	Baseline	COT	**HCOT**
Coarse	15.08	15.05	**14.02**
Fine	24.21	23.33	**22.64**



(a) Embeddings trained with cross-entropy

(b) Embeddings trained with HCOT

Figure 2: Embeddings from 20 coarse classes of CIFAR-100 test images. The embedding of each sample is from the penultimate layer and projected to two dimensions (by t-SNE) for visualization. Notice in (b) the clusters are more distinct, with cleaner and well-separated boundaries, by which we conjecture that the model generalizes better.

Embedding Space Visualization. A visualization of logits of the coarse-level labels are shown in Figure 2. Here we compare it against the visualisation from the baseline SE-PreAct ResNet-18 (Hu et al., 2018) trained using cross-entropy. Compared to the baseline, the HCOT seems to form more distinct clusters in the embedding space that have clear separable boundaries, by which we conjecture that the model generalizes better and therefore achieves better performance.

4.2 IMAGENET-2012

ImageNet-2012 (Krizhevsky et al., 2012) is a large-scale dataset for image classification with 1k categories. This dataset consists of roughly 1.3 million training images and 50k validating images of 256×256 pixels. To study the hierarchical structure on labeling space, we follow (Redmon, 2017) to take proper synsets as our coarse label candidates from WordNet (Miller, 1995). With the hierarchy comprising coarse categories and fine categories, we are able to control the granularity of the coarse-level categories for further analysis.

Experimental Setup. To prepare for experiments, we apply random crops and horizontal flips during training, while images in the testing set use 224×224 center crops (1-crop testing) for data augmentation (He et al., 2016b). We follow (Goyal et al., 2017) as our experimental setup: 256 minibatch size, 90 total training epochs, and 0.1 as the initial learning rate starting that is decayed by dividing 10 at the 30th, 60th and 80th epoch. We use the same alternating training as we did in the CIFAR-100 dataset (Chen et al., 2019b).

Results. As the main result, we conduct HCOT with 52 coarse categories. Results in Table 4 shows significant improvements on both top-1 and top-5 error rates compared to COT and the baseline (ResNet-50 using cross-entropy). We note that top-5 error in-explicitly tested the model's abilities for hierarchical labels.

Table 4: Validation error rates (%) on ImageNet-2012 using ResNet-50.

	Baseline	COT	**HCOT**
Top-1 Error	24.7	24.4	**24.0**
Top-5 Error	7.6	7.4	**7.1**

Ablation study. To explore the effect to HCOT over different granularity of the coarse classes, we conduct a study on performance of our model over a range of coarse classes $N_c = \{1, 20, 52, 145,$

1000} on ImageNet-2012. We observe that HCOT perform best over the sufficient information of category hierarchies. In this case, the performance on Top-1 error peaked at $N_c = 52$. When the N_c goes to the extremes (i.e., 1 or 1000), HCOT degrades to COT. This can be illustrated in Eq(4). When $N_c = 1$, the left term in the equation will disappear, making it the same as COT. Similarly, when $N_c = 1000$, the right term will disappear as there are only 1000 classes.

Table 5: Validation error (%) of different numbers of coarse classes on ImageNet-2012 using ResNet-50.

N_c	1	20	52	145	1000
Top-1 Error	24.4	24.3	**24.0**	24.2	24.4

5 SEMANTIC SEGMENTATION

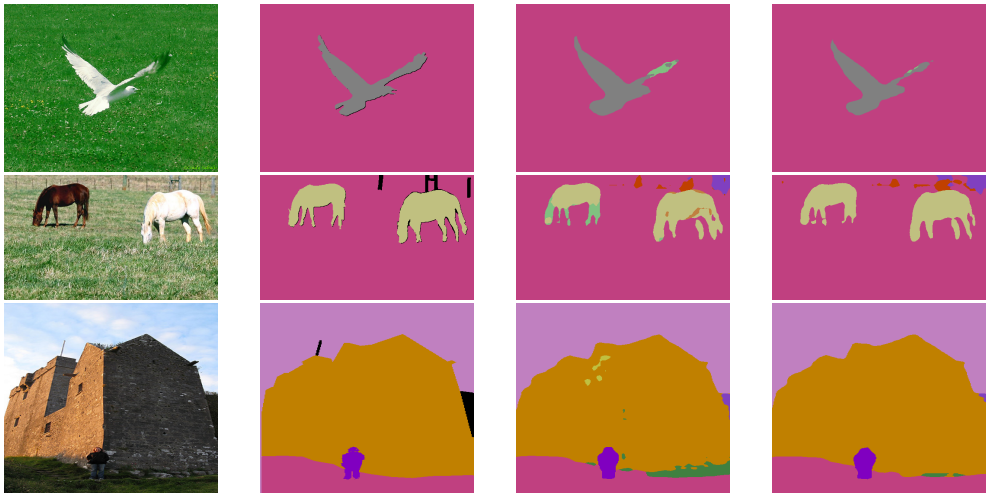
Semantic segmentation is a form of task that contains latent information about label hierarchy (Zhang et al., 2018a). Hierarchies do not exist explicitly in the labels but are rather inferred from the dataset. Applying HCOT can make effective use of this inferred information in the labeling space. In particular, the proposed HCOT procedure can achieve both high confidence of ground-truth and attention of global scene information for each label, which maintains the hierarchy between each semantic and the corresponding theme in a same image sample and helps to provide more accurate semantic segmentation.

Dataset. We experiment HCOT on the widely-used "Pascal-Context" dataset (Mottaghi et al., 2014). The PASCAL-Context dataset provides dense semantic labels for the entire scene of each given image. The dataset contains 4,998 images for training and 5,105 for testing. We follow the prior arts (Mottaghi et al., 2014; Lin et al., 2016; Chen et al., 2016) to create a set of 60 semantic labels for segmentation; these 60 semantic labels represent the most frequent 59 object categories, plus the "background" category.

Experimental Setup. We first take EncNet (Context Encoding Module) (Zhang et al., 2018a) to be the baseline. Here we follow the previous work (Chen et al., 2017; Yu et al., 2017; Zhao et al., 2017) for using the dilated network strategy on the pretrained network. In addition, we perform the Joint Pyramid Upsampling (JPU) (Wu et al., 2019) over EncNet denoted as "EncNet+JPU" to reproduce the state-of-the-art results in semantic segmentation. The training details are the same as described in (Zhang et al., 2018a; Wu et al., 2019). We train the model for 80^{th} epochs with SGD and set the initial learning rate as 0.001. The images are then cropped to 480×480 and grouped with batch size 16. We also use the polynomial learning rate scheduling as mentioned in (Zhao et al., 2017). Different from the training procedure on classification, here we adopt *direct optimization* which training our model by combing the complement loss and primary loss together, which achieve a better empirical performance.

Evaluation. We use the pixel accuracy (PixAcc) and mean Intersection of Union (mIoU) as the evaluation metrics with single scale evaluation. Specifically, for the PASCAL-Context dataset, we follow the procedure in the standard competition benchmark (Zhou et al., 2017) and calculate mIoU by ignoring the pixels that are labeled as "background".

Results. We evaluate the quality of the segmentation from the models trained with pixel-wise cross-entropy (as baseline) and trained with HCOT, by quantitatively calculating the PixAcc and mIoU scores and visually inspecting the output image segments. Experimental results show that HCOT achieves better performance than baseline (cross-entropy) as shown in Table 6a. We also form "EncNet+JPU" as another baseline, and the HCOT again significantly outperforms cross-entropy (as shown in Table 6b). As segmentation does not have inherent label hierarchies, hierarchical structures among labels will have to be inferred from the data. Images occur frequently together as a theme will in-explicitly form a label hierarchy that will be learned to improve the performance of the model.



(a) Real image (b) Ground truth (c) EncNet+JPU (d) EncNet+JPU+HCOT

Figure 3: Examples of segmentation results in PASCAL-Context. Notice the main objects (bird, horses, building) in each image are well-segmented in (d) and visually much closer to the ground truth in (b). We believe that HCOT retains the latent label hierarchies so the segmentation is clearer and without much irrelevant semantics.

Visualizations. In Figure 3, we show segmentation results from two test images on PASCAL-Context dataset. In addition to the input images (Figure 3a), we show the ground-truth segmentation (Figure 3b) and the results from EncNet+JPU model trained with cross-entropy (Figure 3c) and trained with the proposed HCOT (Figure 3d). The segments generated by the proposed HCOT are less fragmented and have less noises.

Table 6: Segmentation results of models trained with cross-entropy (denoted as XE) versus HCOT on PASCAL-Context dataset.

(a) EncNet

Method	PixAcc	mIoU%
XE	0.7835	49.70
HCOT	**0.7862**	**49.86**

(b) EncNet+JPU

Method	PixAcc	MIoU%
XE	0.7880	51.05
HCOT	**0.7918**	**51.35**

6 Conclusion

In this paper, we propose Hierarchical Complement Objective Training (HCOT) to answer the motivational question. HCOT is a new training paradigm that deploys Hierarchical Complement Entropy as the training objective to leverage information from label hierarchy. HCOT neutralizes the probabilities of incorrect classes at different granularity: under the same parental category as the ground-truth class or not belong to the same branch. HCOT has been extensively evaluated on image classification and semantic segmentation tasks, and experimental results confirm that models trained with HCOT significantly outperform the state-of-the-arts. A straight-line future work is to extend HCOT into Natural Language Processing tasks which involve rich hierarchical information.

REFERENCES

Hao-Yun Chen, Jhao-Hong Liang, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. Improving adversarial robustness via guided complement entropy. In *ICCV'19*, 2019a.

Hao-Yun Chen, Pei-Hsin Wang, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. Complement objective training. In *ICLR'19*, 2019b.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Yanming Guo, Yu Liu, Erwin M. Bakker, Yuanhao Guo, and Michael S. Lew. Cnn-rnn: a large-scale hierarchical image classification framework. *Multimedia Tools and Applications*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV'16*, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR'16*, 2016b.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR'18*, June 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML'15*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS'12*, 2012.

Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *arXiv preprint arXiv:1611.06612*, 2016.

George A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 1995.

R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR'14*, 2014.

Calvin Murdock, Zhen Li, Howard Zhou, and Tom Duerig. Blockout: Dynamic model selection for hierarchical deep networks. In *CVPR'16*, 2016.

Farhadi Redmon. Yolo9000: Better, faster, stronger. In *CVPR'17*, 2017.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 1948.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014.

Anne-Marie Tousch, Stphane Herbin, and Jean-Yves Audibert. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 2012.

Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yizhou Yu. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv preprint arXiv: 1903.11816*, 2019.

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR'17*, 2017.

Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition. In *ICCV'15*, 2015.

Fisher Yu, Vladlen Koltun, and Thomas A. Funkhouser. Dilated residual networks. *CVPR'17*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC'16*, 2016.

Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR'18*, 2018a.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *ICLR'18*, 2018b.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR'17*, July 2017.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR'17*, 2017.