Robust Deep Learning with Active Noise Cancellation for Spatial Computing

Li Chen,[1] David Yang,[2] Purvi Goel,[3] Ilknur Kabul[4]
Facebook
[1]lichen66@fb.com, [2]dzyang@fb.com, [3]purvigoel@fb.com, [4]ilknurkabul@fb.com

Abstract

This paper proposes CANC, a **C**o-teaching **A**ctive **N**oise **C**ancellation method, applied in spatial computing to address deep learning trained with extreme noisy labels. Deep learning algorithms have been successful in spatial computing for land or building footprint recognition. However a lot of noise exists in ground truth labels due to how labels are collected in spatial computing and satellite imagery. Existing methods to deal with extreme label noise conduct clean sample selection and do not utilize the remaining samples. Such techniques can be wasteful due to the cost of data retrieval. Our proposed CANC algorithm not only conserves high-cost training samples but also provides active label correction to better improve robust deep learning with extreme noisy labels. We demonstrate the effectiveness of CANC for building footprint recognition for spatial computing.

Introduction

Deep learning has shown success in a variety of computer vision applications. In spatial computing, deep learning has demonstrated superior performance in road or building footprint recognition as well as population density estimation. However one undesired property of deep learning can hinder its performance badly: in the presence of noisy labels, which can arise due to measurement errors, crowdsourcing, insufficient expertise and so on, deep learning tends to have poor generalization performance on test data. The neural networks fit to correctly-labeled training samples at earlier epochs, but eventually overfit to samples with noisy labels, leading to poor classification performance on the test set. This is the so-called memorization effect of deep learning in the presence of noisy labels. (Han et al. 2018; Arpit et al. 2017; Algan and Ulusoy 2020).

In spatial computing, noisy labels, largely inevitable, come from two major sources. The first source is crowdsourcing, a collaborative initiative to create a free map with participants contribute to labeling the map. One famous and widely-used database for spatial computing is Open Street Map (OSM) (OpenStreetMap contributors 2018), a free and editable map of the whole world that is being built by volunteers largely from scratch and released with an open-content license. The labels can be contributed by mappers either on foot or bicycle or in a car or boat. If a location is not widely explored by the mappers, it is likely the labels are noted wrongly. This type of noise can be extreme, meaning the majority of the labels are either sparse or wrong, as seen in Figure 1. The second major source of noisy labels in spatial computing is measurement error. Imagery providers usually do a good job at geo-referencing their imagery, but occasionally the images can be out of position by a few meters or more. Particularly in hilly or mountainous areas, it is difficult to stretch a flat image over an area of the Earth with many contours. Then even with professional annotators, the resulting ground truth can be offset due to the imagery misalignment.

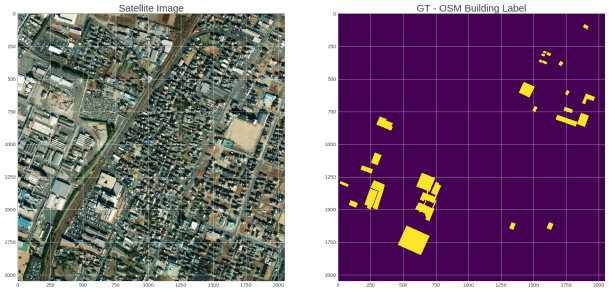

Figure 1: Left: satellite image shows a large number of buildings in the region. Right: Labels from Open Street Map (OSM). OSM only has labels for a few of them. A lot of building footprints are missing due to label crowdsourcing. This is considered as extreme noise with the majority of labels being sparse or missing.

Co-teaching, proposed in (Han et al. 2018), has demonstrated efficacy to improve deep learning robustness in the presence of extreme noisy labels. It involves training two neural networks simultaneously such that each network teaches the other for clean sample selection. As a clean data selection method, while co-teaching outperforms a variety of noisy label learning methods such as MentorNet (Jiang et al. 2018b), Decoupling (Malach and Shalev-Shwartz 2017), S-net (Goldberger and Ben-Reuven 2016), Bootstrap (Reed

et al. 2014), it can filter a large number of training samples via a parameter called remember rate. It is trained only on a smaller portion of the data samples as epochs grow to unsure robust learning. However, such procedure may not take full advantage of high-cost commercial satellite imageries in spatial computing. In addition, co-teaching can help to find the data with less noise and use them more often during training, but it could not actively correct the noise. In the case of spatial computing where there is not enough data with low noise , often estimated to be less than 1%, co-teaching might not achieve good performance due to the lack of data with low noise.

Hence, in order to address learning with extreme noisy labels and conserving valuable data samples in spatial computing, we propose CANC, **C**o-teaching **A**ctive **N**oise **C**ancellation, a training paradigm to not only perform clean sample filtering but also actively correct unreliable labels. Our method makes better use of training dataset compared to co-teaching. Especially under higher noise strengths, CANC outperforms co-teaching because CANC's active label correction capability enables utilizing more training data to train a resilient deep learning model. We also notice, however, when the noise strength is less, CANC does not outperform co-teaching. In this paper, we demonstrate the effectiveness of CANC in building footprint recognition under extreme noise and evaluate the performance using accuracy, precision, recall, F1 score and super-pixel intersection over union.

Related Work

Deep learning has the tendency to overfit data and even memorize completely random noise. Such undesired property of deep learning inspired several research directions in robust deep learning with noisy labels. Two main categories of methods exist, namely noise-model-based and noise-model-free methods. The task of noise-model-based methods is to identify the best estimator via minimizing a risk function to describe the effects of noisy samples. Existing methods include noise channel estimation, label noise cleansing, dataset pruning and sample importance weighting. The authors in (Northcutt, Wu, and Chuang 2017) estimate the noise rates according to the sizes of a confidently clean and a noisy subset. (Wu et al. 2018) uses transfer learning trained on a clean dataset and fine-tunes it on noisy dataset for relabeling and then the network is retrained on relabeled data to re-sample to dataset to construct a final clean dataset. The authors in (Jiang et al. 2018a) propose to randomly split dataset to labeled and unlabeled subgroups. Then, labels are propagated to unlabeled data using similarity index among instances. (Wang et al. 2018) uses the Siamese network to detect noisy labels by learning discriminating features to apart clean and noisy data. Noisy samples are detected and pulled from clean samples. Then, each iteration weighting factor is recalculated for noisy samples, and the base classifier is trained on whole dataset.

On the other hand, the task of noise-model free methods is to design robust algorithms rather than modeling the noise. In that sense, label noise is not decoupled from classification. Non-convex loss function is more noise tolerant than convex losses as studied in (Manwani and Sastry 2013), (Ghosh, Manwani, and Sastry 2015), (Charoenphakdee, Lee, and Sugiyama 2019). (Xu et al. 2019) proposed to use information-theoretic loss. (Van Rooyen, Menon, and Williamson 2015) proposed to use classification-calibrated loss function. Regularization methods such as dropout, adversarial training, label smoothing are useful for addressing noisy label problems. Boosting algorithms such as BrownBoost and LogitBoost are shown to be more robust to noisy labels (McDonald, Hand, and Eckley 2003).

Building footprints are critical for many applications, such as mapping, city and government planning, population density estimation (Tiecke et al. 2017), and natural disaster estimation. Manual building detection for the whole world takes a long time and requires a huge amount of human resources such that it might be infeasible. Recently, deep learning has been applied to extract map features from satellite images, such as roads and buildings (Kaiser et al. 2017) (Bonafilia et al. 2019) (Batra et al. 2019) (Chawda, Aghav, and Udar 2018) (Pan et al. 2019). Among these works, (Batra et al. 2019), (Chawda, Aghav, and Udar 2018), and (Pan et al. 2019) used professionally labeled training dataset with little noise, such as DeepGlobe dataset (Demir et al. 2018), SpaceNet dataset (Etten, Lindenbaum, and Bacastow 2019), and Massachusetts Building dataset (Mnih 2013). As these datasets are limited to a few selected areas, the models trained on them might not generalize well across the globe. In contrast, (Kaiser et al. 2017) and (Bonafilia et al. 2019) used the crowd-sourced map dataset OSM for training, but the OSM dataset has high noise-to-signal ratio.

Framework and Methodology

Our proposed framework, as seen in Figure 2, consists of the following four main steps: 1. Mask creation and label generation, 2. Noisy label simulation 3. CANC: co-teaching with active noise cancellation 4. Evaluation. We describe each component in detail.

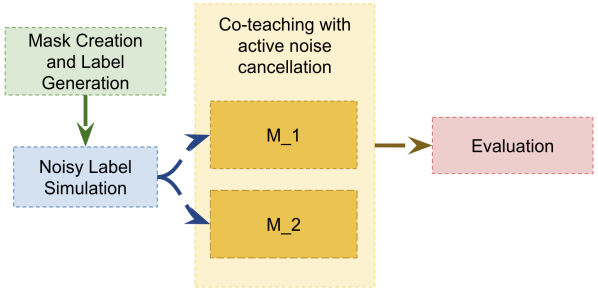

Figure 2: Overview of our proposed framework. We consider a sub-task called mask classification task, where the masks and associated labels are created from the original satellite imagery. Then we perturb the labels via a noise transition matrix. CANC is trained on the noisy data to produce a robust deep learning algorithm. Finally we evaluate the performance using accuracy, precision, recall, F1 and super-pixel intersection over union.

Mask Creation

Given a high-resolution satellite image, we generate masks of size $m \times m \times 3$ by cropping the original image of size $n \times n \times 3$, where m is divisible by n. Hence each satellite image is represented by $[\frac{n}{m}]^2$ masks.

Similarly on each ground truth satellite image, we create grey-scale masks with size $m \times m \times 1$ to represent the annotated area. The label for each mask is computed as follows:

$$l = \mathbb{1}_{\{\frac{S_1}{m^2} \geq \tau\}}, \tag{1}$$

where $\mathbb{1}$ is the identity function, S_1 is the sum of the pixels that belong to building class, and τ is a threshold set to represent the percentage of pixels needed in one mask for a professional annotator to recognize a building in it. $\tau = 0.5\%$ or 1% is usually a reasonable value.

Noise Simulation

Our framework then simulates noise to ground truth so we can use as a baseline to assess the effectiveness of the framework. TWe consider two types of noise simulation schemes: symmetric and anti-symmetric. Both of these schemes require manipulation of a square noise transition matrix used to choose which data samples to mislabel.

In the symmetric scheme, our transition matrix has a diagonal entries of (1 - ϵ), where ϵ is a predefined noise ratio indicating the percentage of data samples to mislabel. Everywhere else but the diagonal, the matrix has entries of $(1-\epsilon)/(n-1)$ where n is the number of classes. Intuitively, this transition matrix means that data samples can be mislabeled as any other class with equal probability.

In the anti-symmetric scheme, we use a transition matrix

$$\begin{bmatrix} 1-\epsilon & 0 \\ \epsilon & 1 \end{bmatrix} \tag{2}$$

Again, ϵ is the chosen noise ratio. Intuitively, this matrix suggests that all of class 1 will stay the same, and class 0 is flipped to class 1 with a probability equal to $(1 - \epsilon)$.

Co-teaching with Active Noise Cancellation

We present our proposed Co-teaching with Active Noise Cancellation (CANC) as seen in Algorithm 2 in this subsection. We leverage co-teaching (Han et al. 2018), which is a training paradigm to address noisy label learning robustness. In co-teaching training scheme, there are two neural networks where two networks are simultaneously trained and taught by each other given every mini batch. It essentially consists of three steps: i) each neural network feeds forward all data and performs clean sample selection based on losses, ii) two networks exchange what data shall be used for training, iii) each network trains on the data selected by the other teacher and updates itself. Co-teaching has shown to be robust especially under the extreme noisy label scenario.

Our goal is to not only train a robust deep learning algorithm in the presence of noisy labels, but also actively correct the noise during the training process without hurting classification performance. To make better use of the data, instead of forgetting the remaining samples in the batch as in co-teaching, we consider swapping the labels of the samples that are most likely to have noisy labels. Since our classification problem is binary, the swapping process becomes a binary decision. Our method inherits co-teaching such that the loss is sorted in each mini-batch by the first network which sends the top samples with least amount of loss to the other network. Additionally our method would select the bottom samples with the highest amount of loss, correct their labels and combine with the first subset to update the parameters of the second network. The same applies for updating the parameters of the first network. The algorithm is stated in Algorithm 2.

Algorithm 1 Co-teaching with active noise cancellation (CANC)

Goal: Robust training on extreme noisy labels with active noise correction
Input: w_{M_1}, w_{M_2}, learning rate η, epoch T_k, $T_{\max}$, iteration $N_{\max}$, remember rate $R(T)$, swap rate S;
for $T = 1, 2, ..., T_{\max}$ **do**
  for $N = 1, ..., N_{\max}$ **do**
    Step 1: Feed mini-batch $\mathcal{D}$ to both M_1 and M_2
    Step 2: Obtain
    $\hat{D}_{M_1}^{clean} = \arg\min_{D':|D'|\geq R(T)|D'|} l(M_1, D')$
    Step 2: Obtain
    $\hat{D}_{M_1}^{swap} = \arg\max_{D':|D'|\geq S|D'|} l(M_1, D')$
    Step 3: Define $\hat{D}_{M_1} := \hat{D}_{M_1}^{clean} \cup \hat{D}_{M_1}^{swap}$ and update $w_{M_2} \leftarrow w_{M_2} - \eta\nabla(M_2, D_{M_1})$
    Step 4: Obtain
    $\hat{D}_{M_2}^{clean} = \arg\min_{D':|D'|\geq R(T)|D'|} l(M_2, D')$
    Step 5: Obtain
    $\hat{D}_{M_2}^{swap} = \arg\max_{D':|D'|\geq S|D'|} l(M_2, D')$
    Step 6: Define $\hat{D}_{M_2} := \hat{D}_{M_2}^{clean} \cup \hat{D}_{M_2}^{swap}$ and update $w_{M_1} \leftarrow w_{M_1} - \eta\nabla(M_1, D_{M_2})$
  end for
  Step 7: Update $R(T) = 1 - \min\{\frac{T}{T_k}\tau, \tau\}$
end for
Output: w_{M_1} and w_{M_2}

Evaluation

We evaluate the accuracy, precision, recall, F_1 score and super-pixel intersection over union (SP-IOU) to assess the performance of our algorithm on spatial computing, where each super-pixel corresponds to the generated mask. We define a smoothed SP-IoU as the following:

$$\text{SP-IoU} = \frac{intersection + smooth}{union + smooth}, \tag{3}$$

where $intersection = \|\{y_i == \hat{y}_i\}\|_{\{1\leq i\leq N\}}$, $union = \|\{y_i\}\|_{\{1\leq i\leq N\}} + \|\{\hat{y}_i\}\|_{\{1\leq i\leq N\}} - intersection$, y_i and $\hat{y}_i$ are the label and predicted label of mask i respectively and $\|\cdot\|$ denotes the cardinality. We set $smooth$ parameter to be 1.

Data Description

The inputs are high resolution satellite images from Maxar. Each pixel in the images represents an area of 50cm by 50cm

Algorithm 2 Co-teaching with active noise cancellation (CANC) Revised version

Goal: Robust training on extreme noisy labels with active noise correction
Input: w_{M_1}, w_{M_2}, learning rate η, epoch T_k, $T_{\max}$, $T_{k'}$, iteration $N_{\max}$,τ,cal τ'
for $T = 1, 2, ..., T_{\max}$ **do**
  for $N = 1, ..., N_{\max}$ **do**
    Step 1: Feed mini-batch $\mathcal{D}$ to both M_1 and M_2
    Step 2: Obtain $\hat{D}_{M_1}^{clean} = \arg\min_{D':|D'|\geq R(T)|D'|} l(M_1, D')$
    Step 2: Obtain $\hat{D}_{M_1}^{swap} = \arg\max_{D':|D'|\geq S(T)|D'|} l(M_1, D')$
    Step 3: Define $\hat{D}_{M_1} := \hat{D}_{M_1}^{clean} \cup \hat{D}_{M_1}^{swap}$ and update $w_{M_2} \leftarrow w_{M_2} - \eta\nabla(M_2, \hat{D}_{M_1})$
    Step 4: Obtain $\hat{D}_{M_2}^{clean} = \arg\min_{D':|D'|\geq R(T)|D'|} l(M_2, D')$
    Step 5: Obtain $\hat{D}_{M_2}^{swap} = \arg\max_{D':|D'|\geq S(T)|D'|} l(M_2, D')$
    Step 6: Define $\hat{D}_{M_2} := \hat{D}_{M_2}^{clean} \cup \hat{D}_{M_2}^{swap}$ and update $w_{M_1} \leftarrow w_{M_1} - \eta\nabla(M_1, \hat{D}_{M_2})$
  end for
  Step 7: Update $R(T) = 1 - \min\{\frac{T}{T_k}\tau, \tau\}$, $S(T) = 1 - \min\{\frac{T}{T_{k'}}\tau', \tau'\}$
end for
Output: w_{M_1} and w_{M_2}

on the ground. The satellite images are tiled at zoom level 15 with the size of 2048 by 2048, so each satellite image represents about 1km by 1km on the ground. The building labels of the training dataset were collected from OpenStreetMap (OSM). OpenStreetMap is a free, editable map of the whole world that is being built by volunteers largely from scratch and released with an open-content license. The building labels, in the form of polygons, are rasterized to create the training label images.

Experiments

Data Preparation

We obtain 81 satellite imageries with professional annotation, which means we have minimum noisy labels. We put 50 satellite images as training, 15 as validation and 15 as testing. The original satellite imageries are of size $8192 \times 8192 \times 3$. We set the mask size to be 32×32. Then each satellite image is segmented into 256×256 masks. We further set the threshold $\tau = 1\%$ since with approximately 8 pixels, the annotators can recognize this part of a building. Hence after preprocessing, we obtain 3.34 million masks for training, 983k masks for validation and 15 images for another testing.

We report accuracy, precision, recall and F_1 score on the validation masks and report the super-pixel IOU on the 15 test images.

Label Noise Simulation

We use several methods to apply label noise to our clean training labels. Because of our data preparation, in which we cropped and labelled images with either a 0, for no buildings, or a 1, for containing buildings, we've reparameterized the DNN's decision into a binary classification problem. As follows, label noise can be added by perturbing these 0/1 labels rather than the position or presence of building polygons themselves.

Symmetric Noise Simulation. In the symmetric noise strategy, we flip one label to another based on a transition probability matrix T. The length and width is equal to the number of classes; in this binary classification case, the matrix dimensions are 2 x 2. The diagonal entries of the transition matix are (1 - N), where N is some specified noise ratio between 0 and 1, which indicates a rough percentage of labels to be perturbed. We distribute the remaining probability equally per-row as: (N) / c, where c is the number of classes. An example transition matrix, with noise ratio. An example matrix, calculated with a noise ratio of 0.35, is shown below:

$$\begin{bmatrix} 0.65 & 0.35 \\ 0.35 & 0.65 \end{bmatrix} \tag{4}$$

Antisymmetric Noise Simulation. In the antisymmetric noise strategy, we add noise to only single class. As before, we use a 2x2 noise transition matrix. This time, the matrix flips the class with probability equal to (1 - noise ratio), using a matrix such as the following for a noise ratio of 0.35.

$$\begin{bmatrix} 0.65 & 0 \\ 0.35 & 1 \end{bmatrix} \tag{5}$$

We perform our experiments with varying noise strengths: 0.15, 0.35, 0.45, and 0.55.

Results

We apply CANC on the noisy training masks and compare the performance with deep learning directly training with noisy labels and co-teaching without active noise cancellation. We consider both transfer learning via ResNet18 (He et al. 2016) and training from scratch utilizing a 9-layer CNN with LReLU architecture as in (Han et al. 2018; Laine and Aila 2016) on the dataset, and do not observe significant performance difference between training from scratch and transfer learning in our use case. Within the 3.34 million training masks, we use 80% to update the weights and 20% as validation for model selection at the best validation accuracy. On a second validation set consisting 983k masks, we evaluate the model performance and report accuracy, precision, recall and F_1 score. Finally we evaluate the super-pixel IoU per satellite image from the test set which contains another 15 out-of-sample satellite images.

The performance of CANC, co-teaching and vanilla deep learning is shown in Table 1. The clean baseline indicates the model's performance on validation set when the model was trained with training data with clean labels. The noisy baseline indicates the model's performance on validation set when the model was trained with training data with noisy labels generated by different noise types and at different noise

strength. For lower noise ratio at 0.15, CANC does not outperform co-teaching. As the noise ratio is increased higher to 0.45 and 0.55, CANC improves network performance compared with directly training on noisy labels and outperforms co0teaching in accuracy, precision, recall and F1 scores.

On the 15 images in test set, we plot the super-pixel IoUs and compare this metric as seen in Figure 3 and 4 under two noise types and strengths respectively.

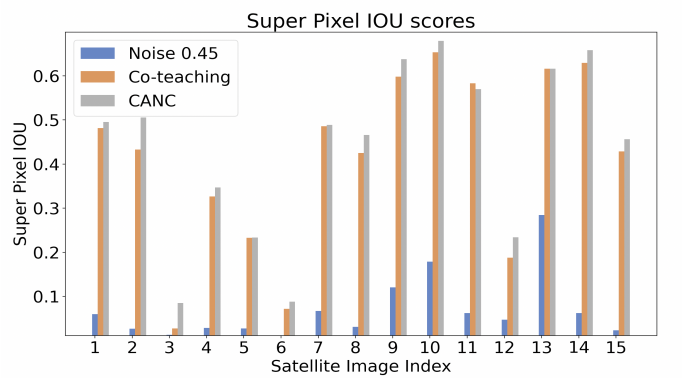

Figure 3: We calculate the super-pixel IoU under noise ratio of 0.45 for anti-symmetric noise for 15 test images. CANC with improves the SP-IoU compared with deep learning without addressing noisy labels. It also outperforms co-teaching on most of the satellite images.

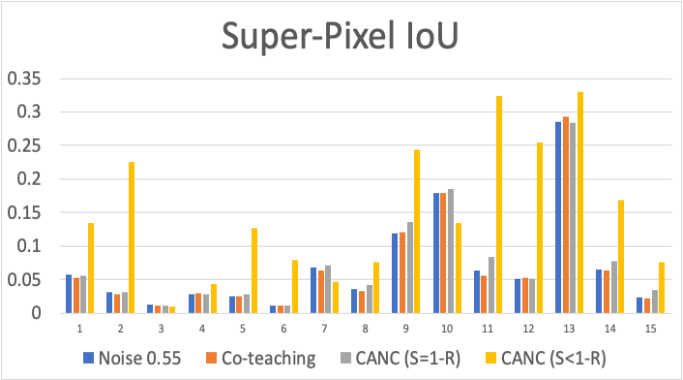

Figure 4: We calculate the super-pixel IoU under noise ratio of 0.55 for symmetric noise for 15 test images. In this experiment, we additionally considered the case when the swap rate is the same as the initial forget rate in co-teaching, which we denote as CANC with $S = 1 - R$. CANC with $S < 1 - R$ improves the SP-IoU the most. Both CANC methods can improve super-pixel IOU when training with noisy labels.

Inspection

We examine which images benefit the most from CANC in terms of SP-IoU when the model has been learned with symmetric noise at 0.55 noise strength. The improved imagery has very different geo-condition compared with the worst improved imagery. The best improved image has buildings in a much smaller location. We think the performance improvement is due to the model being fed by corrected information about buildings in such geo-condition. However we acknowledge that the false positive rate also increases. We plan to do more investigation on interpreting the results.

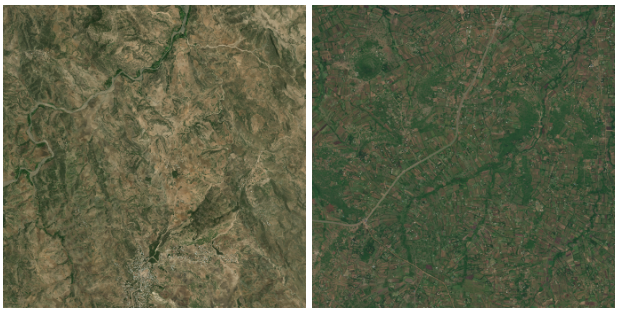

Figure 5: (Left) The satellite image with best SP-IoU improvement (7.35x better) with CANC under noise ratio 0.55 of symmetric noise. (Right) The satellite image with worst SP-IoU improvement (0.67x better) with CANC under noise ratio 0.55 of symmetric noise. We think the performance improvement is due to the model being fed by corrected information about buildings in such geo-condition. However we acknowledge that the false positive rate also increases.

Conclusion

Having high quality training data is an important part for the success of the training methods in machine learning. For classification problems, label quality plays an important role for the data quality. Obtaining high quality ground truth labels is a very challenging problem due to the scarcity of labelling resources, inconsistencies between the labellers, as well as the measurement errors that happens during the label collection process. In this paper, by using CANC we are able to mitigate some of these issues and improve the predictive accuracy of our models. We validated the proposed method on real data and obtained promising results. We showed that the proposed technique has advantages over co-teaching in that it corrects some of the labelling errors in the training data during the training process rather than removing them. We also acknowledge that under certain noise strength, co-teaching demonstrates more benefits over CANC.

Future Work

In our noise simulation scheme, we applied noise directly to the mask labels. We are working on a noise simulation process directly applied on the building polygons. In this simulation process, we artificially add buildings to images, where the buildings are selected from aggregated statistical analyses on all the real-world buildings. Within a single given satellite image, we collect all the ground-truth buildings shapes and sort them, based on their image-space polygonal area, into a histogram. We normalize the histogram to convert it into a distribution of per-image building areas and

Scenario	Noise strength	Noise type	Accuracy	Precision	Recall	F_1
Clean baseline	Clean	Clean	0.9350	0.9844	0.9324	0.9577
Noisy baseline	0.15	Anti-symmetric	0.8226	0.9915	0.8089	0.8910
CANC	0.15	Anti-symmetric	0.6899	0.8702	0.7684	0.8161
Coteaching	0.15	Anti-symmetric	0.9209	0.9686	0.9422	0.9552
Noisy baseline	0.35	Anti-symmetric	0.1164	0.9736	0.0142	0.0280
Noisy baseline	0.35	Symmetric	0.3684	nan	0	nan
Coteaching	0.35	Symmetric	0.9290	0.9514	0.9739	0.9625
Noisy baseline	0.45	Anti-symmetric	0.1040	nan	0	nan
CANC	0.45	Anti-symmetric	0.8870	0.9833	0.8889	0.9337
Coteaching	0.45	Anti-symmetric	0.8649	0.9827	0.8643	0.9197
Noisy baseline	0.55	Anti-symmetric	0.1040	nan	0	nan
CANC	0.55	Anti-symmetric	0.8944	0.8958	0.9982	0.9442
Noisy baseline	0.55	Symmetric	0.1040	0.1176	0	0
CANC	0.55	Symmetric	0.9190	0.9291	0.9847	0.9561
CANC (S = 1-R)	0.55	Symmetric	0.1637	0.9866	0.0675	0.1263
Coteaching	0.55	Symmetric	0.1040	nan	0	nan

Table 1: Performance on validation set when training with various noise strengths and noise types.

randomly sample a bucket from the distribution, then sample a building shape from that bucket. The building is placed at a randomly chosen coordinate within the image. We continue this process of sampling-and-adding until the ratio of pixels labeled as buildings to the total number of pixels is roughly equal to a predefined threshold. The challenge to make such simulation process realistic is due to building boundaries. An added polygon shall not have boundaries to overlap with existing polygons. While we can employ an acceptance-rejection method, it significantly increases the noise simulation run time. We plan to investigate an efficient method to perturb labels directly in the building polygon-level.

In this use case, we mainly focus on binary classification problem, in which we rely on the loss in the mini-batch to flip the training data labels. For future work, we would also like to extend it to multi-family classification. We plan to consider a weighted decision to flip the labels based on the predicted probabilities.

Our active noise cancellation process via a swapping procedure is not sophisticated. We would like to investigate in the future is incorporating uncertainty estimation to CANC to improve the robustness of the models to label noise. We would like to investigate both epistemic and aleatory uncertainty techniques in this research direction.

References

Algan, G., and Ulusoy, İ. 2020. Label noise types and their effects on deep learning. *arXiv preprint arXiv:2003.10471*.

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. *arXiv preprint arXiv:1706.05394*.

Batra, A.; Singh, S.; Pang, G.; Basu, S.; Jawahar, C.; and Paluri, M. 2019. Improved road connectivity by joint learning of orientation and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bonafilia, D.; Gill, J.; Basu, S.; and Yang, D. 2019. Building high resolution maps for humanitarian aid and development with weakly- and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Charoenphakdee, N.; Lee, J.; and Sugiyama, M. 2019. On symmetric losses for learning from corrupted labels. *arXiv preprint arXiv:1901.09314*.

Chawda, C.; Aghav, J.; and Udar, S. 2018. Extracting building footprints from satellite images using convolutional neural networks. 572–577.

Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; and Raskar, R. 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Etten, A. V.; Lindenbaum, D.; and Bacastow, T. M. 2019. Spacenet: A remote sensing dataset and challenge series.

Ghosh, A.; Manwani, N.; and Sastry, P. 2015. Making risk minimization tolerant to label noise. *Neurocomputing* 160:93–107.

Goldberger, J., and Ben-Reuven, E. 2016. Training deep neural-networks using a noise adaptation layer.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, 8527–8537.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jiang, J.; Ma, J.; Wang, Z.; Chen, C.; and Liu, X. 2018a. Hyperspectral image classification in the presence of noisy

labels. *IEEE Transactions on Geoscience and Remote Sensing* 57(2):851–865.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018b. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, 2304–2313.

Kaiser, P.; Wegner, J. D.; Lucchi, A.; Jaggi, M.; Hofmann, T.; and Schindler, K. 2017. Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing* 55(11):6054–6068.

Laine, S., and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Malach, E., and Shalev-Shwartz, S. 2017. Decoupling" when to update" from" how to update". In *Advances in Neural Information Processing Systems*, 960–970.

Manwani, N., and Sastry, P. 2013. Noise tolerance under risk minimization. *IEEE transactions on cybernetics* 43(3):1146–1151.

McDonald, R. A.; Hand, D. J.; and Eckley, I. A. 2003. An empirical comparison of three boosting algorithms on real data sets with artificial class noise. In *International Workshop on Multiple Classifier Systems*, 35–44. Springer.

Mnih, V. 2013. *Machine Learning for Aerial Image Labeling*. Ph.D. Dissertation, CAN.

Northcutt, C. G.; Wu, T.; and Chuang, I. L. 2017. Learning with confident examples: Rank pruning for robust classification with noisy labels. *arXiv preprint arXiv:1705.01936*.

OpenStreetMap contributors. 2018. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

Pan, X.; Yang, F.; Gao, L.; Chen, Z.; Fan, H.; and Ren, J. 2019. Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms. *Remote Sensing* 11:917.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Tiecke, T. G.; Liu, X.; Zhang, A.; Gros, A.; Li, N.; Yetman, G.; Kilic, T.; Murray, S.; Blankespoor, B.; Prydz, E. B.; and Dang, H. H. 2017. Mapping the world population one building at a time. *CoRR* abs/1712.05839.

Van Rooyen, B.; Menon, A.; and Williamson, R. C. 2015. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, 10–18.

Wang, Y.; Liu, W.; Ma, X.; Bailey, J.; Zha, H.; Song, L.; and Xia, S.-T. 2018. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8688–8696.

Wu, X.; He, R.; Sun, Z.; and Tan, T. 2018. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security* 13(11):2884–2896.

Xu, Y.; Cao, P.; Kong, Y.; and Wang, Y. 2019. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *Advances in Neural Information Processing Systems*, 6225–6236.