WEAKLY-SUPERVISED CONTINUAL LEARNING FOR CLASS-INCREMENTAL SEGMENTATION
Gaston Lenczner, Adrien Chan-Hon-Tong, Nicola Luminari, Bertrand Le Saux

To cite this version:
Gaston Lenczner, Adrien Chan-Hon-Tong, Nicola Luminari, Bertrand Le Saux. WEAKLY-SUPERVISED CONTINUAL LEARNING FOR CLASS-INCREMENTAL SEGMENTATION. IEEE IGARSS 2022, Jul 2022, Kuala Lumpur, Malaysia. 10.1109/IGARSS46834.2022.9884547. hal-03856281

HAL Id: hal-03856281
https://hal.science/hal-03856281
Submitted on 16 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Transfer learning is a powerful way to adapt existing deep learning models to new emerging use-cases in remote sensing. Starting from a neural network already trained for semantic segmentation, we propose to modify its label space to swiftly adapt it to new classes under weak supervision. To alleviate the background shift and the catastrophic forgetting problems inherent to this form of continual learning, we compare different regularization terms and leverage a pseudo-label strategy. We experimentally show the relevance of our approach on three public remote sensing datasets.

Fig. 1. Add building in the \{background, road\} label space with clicked annotations. Initial prediction (a) leads to new-class annotations (b) to collect the new prediction (c).

Corresponding author: gaston.lenczner@alteia.com
to fully benefit from the sparse annotations. We also explore a pseudo-label strategy to alleviate catastrophic forgetting. To summarize, the contributions of this paper are:

- A framework to increase the label space of a DNN for semantic segmentation with sparse annotations.
- A pseudo-label strategy and a comparison of different regularizations for this weakly supervised fine-tuning.
- An experimental validation of the proposed approach on three public remote sensing datasets.

2. RELATED WORK

2.1. Semi and weak supervision in semantic segmentation

When labels are scarce or flawed, the common training paradigm is to learn to perform the segmentation task with the available labels while leveraging unlabeled data to learn a better inner representation as support. In semantic segmentation, weak labels can take various forms like incomplete ground-truth maps or bounding boxes. Recently, different works in remote sensing explore these different aspects. In [1], the authors learn from a mix of fully annotated images and unannotated ones with auxiliary tasks while [3] designs a custom regularization weighted by the proportion of labeled data to address semi-supervised domain adaptation. In weakly supervised learning, semantic segmentation can be achieved from bounding-boxes [4], rough polygons [6] or image-level labels [7]. Closely related to our kind of supervision, DISCA [5] and FESTA [2] leverage point annotations either as user input clicks or partial labels. We here use similar point annotations to add a new segmentation class to a pretrained DNN.

2.2. Class-incremental semantic segmentation

The goal of class incremental learning is to modify the output space of DNNs to add label classes. This problem was already considered two decades ago [8] and two main pitfalls are identified. First, like in other class-incremental tasks, it is important to prevent the catastrophic forgetting of the previously learned knowledge. Second, specific to semantic segmentation, the background-shift first tackled in [9] has to be dealt with. Indeed, the new class comes from the background, which causes a discrepancy with the previously learned background class. Many methods addressed these two problems by simply storing previous examples [10] but more recent works consider this constraint to be too restrictive due to limited storage or security reasons. Hence, we draw inspiration from [11, 12] which design customized regularization to address the two identified pitfalls. [11] proposes distillation losses which enforce statistical matches between the networks. Inspired by few-shot and contrastive learning, [12] relies on prototypes representing the different semantic classes to distinguish them.

3. METHODOLOGY

3.1. Scenario, baseline and constraints

A DNN pre-trained for semantic segmentation to detect $N-1$ classes is applied on a new image $I$. The goal is to add a new possible segmentation class and to provide an accurate segmentation map with respect to these $N$ classes. To this aim, $M$ annotations are provided to make the network interactively learn the $N^{th}$ segmentation class. In this work, we follow findings from [2] and use point annotations as they appear to be particularly handy for a potential user. In practice, the last layer of the DNN is modified to increment its output space. As we observed that freezing parts of the network lowered its accuracy, the entire network is then retrained with a cross-entropy (CE) loss on the annotated pixels. However, different problems arise with this simple baseline:

- Catastrophic forgetting since the new provided labels only belong to the new class.
- Background shift since the new class was previously classified as background. So, the network has to both learn to detect a new class and to modify the learned representation of the background class.
- Poor class representation since the network has to learn the new class representation from sparse annotations (i.e. highly incomplete ground-truth).

To address these problems, we notably rely on a pseudo-label strategy and compare different regularizations added to the cross-entropy loss.

Following notation from [10], we refer through the rest of this paper to the previous network with the $N-1$ class prediction as the memory network and to the new one as the updated network. We also refer to the set of semantic classes except the background as the classes of interest.

3.2. Learning from pseudo-labelling

The prediction of the memory network is used to sample old classes labels. This ensures that the user does not have to provide such old-class annotations while preventing the network from forgetting them. In addition, the interactive user inputs provide both the new class and the background labels. These background labels avoid the aforementioned background-shift issue since background predictions from the memory network could actually belong to the new class. To emphasize the $N$ annotations of the new class, only a maximum of $N$ of the most confident old-classes pixels are considered as annotations.

3.3. Regularization

In order to make the most of the sparse annotations, we consider the following regularization schemes.
DISCA. Since the new class was previously part of the background, the predictions over the other classes of interest should remain similar. To enforce this property, we follow [5] and consider adding a cross entropy regularization term over the pixels which are predicted as belonging to classes of interest by the memory network. This can be seen as an output-level knowledge distillation loss.

PodNet. Following PodNet [11], we enforce a statistic match at the encoder level over the channels between the memory network and the updated one with a $L^2$ intermediate-level knowledge distillation loss.

SDR. Inspired by few-shots and contrastive learning, we follow [12] to regularize the DNN latent space. This aims to reduce forgetting whilst improving the recognition of the new class. Concretely, we add a prototype-based regularization at the encoder level of the DNN. Prototypes are vectors representative of each segmentation class and are computed on the fly at each new learning step. This regularization is composed of three terms: a matching term enforces that the updated prototypes stay close to the previous ones, a repulsive term ensures that the prototypes are far from each other and an attracting term pushes the pixels to remain close to their associated prototype.

FESTA. To deal with sparse annotations for segmentation in remote sensing, [2] introduces an unsupervised loss that accounts for neighbourhood structures both in spatial and feature domains as it assumes that nearby pixels share labels.

4. EXPERIMENTS

4.1. Experimental set-up

We evaluate our approach on the ISPRS Potsdam & Vaihingen datasets [13] respectively composed of 38 ($6000 \times 6000$ pxls, 5cm res.) and 33 (various sizes, 9cm res.) images and on SemCity Toulouse [14] composed of 4 images ($3504 \times 3452$ pxls, 50 cm res.). We pretrain the network for old class (road\(^3\)) segmentation using dense ground-truth maps on 30 (Potsdam), 26 (Vaihingen) and 2 (Toulouse) images. We fine-tune for new class (building\(^2\)) using 300 clicked new-class and background annotations simulated from the ground-truth on the remaining data. To compare to a potential upper born, we consider a control network directly pretrained on dense \{road, building\} ground-truth maps (and not fine-tuned).

The considered metric is the Intersection over Union (IoU) averaged image-wise. Except when we analyze the impact of the number of annotations, we simulate 300 new-class and background annotations.

For all experiments, we use a LinkNet [15] architecture trained using Adam optimizer with a learning rate of $10^{-4}$ during 10 pseudo-epochs. Each pseudo-epoch consists in 10000 $256 \times 256$ labeled samples randomly chosen from training data. We infer using a $256 \times 256$ sliding window with an overlap of 50%. Due to the stochastic nature of the optimization process and the simulation of the annotations, all experiments are averaged on 3 runs to obtain statistically significant results.

During fine-tuning, we use an Adam optimizer with a learning rate of $2 \cdot 10^{-5}$. We fine-tune for 30 steps and select the best performances obtained in the last 15 steps. Each step consists of 10 back-propagation iterations.

4.2. Approach assessment

<table>
<thead>
<tr>
<th></th>
<th>Contr.</th>
<th>Bas.</th>
<th>FESTA</th>
<th>DISCA</th>
<th>PodNet</th>
<th>SDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pot.</td>
<td>76.4</td>
<td>68.7+</td>
<td>66.5+</td>
<td>68.2+</td>
<td>71.7+</td>
<td>72.0+</td>
</tr>
<tr>
<td>Vai.</td>
<td>84</td>
<td>76.2+</td>
<td>75.1+</td>
<td>74.8+</td>
<td>74.1+</td>
<td>79.7+</td>
</tr>
<tr>
<td>Toul.</td>
<td>74</td>
<td>63.2+</td>
<td>62.4+</td>
<td>54.0+</td>
<td>65.4+</td>
<td>67.7+</td>
</tr>
</tbody>
</table>

As indicates Table 1, the DNN is able to learn a new class with clicked annotations. Even without additional regularizations, it reaches an IoU over the three classes of 68.7% on Potsdam, 76.2% on Vaihingen and 63.2% on Toulouse.

\(^3\)Except when specified otherwise.
Appropriate regularization further improves the performances up to 4%. Indeed, SDR consistently improves the results on the three datasets: 3.3% on Potsdam, 3.5% on Vaihingen and 4.5% on Toulouse. Moreover, SDR tends to stabilize the results compared to the baseline, as testifies the standard deviation measures (e.g. the std on Toulouse is of 0.6 for SDR and of 3.2 for the baseline). As shows Figure 2, it visually translates into sharper contours, even though the baseline already produces visually accurate results. However, the other regularizations under-perform in this setting in regards to SDR and can even lead to worse results than the baseline on Vaihingen and Toulouse. Hence, the distillation losses and FESTA seem to be less relevant for this problem than the latent space regularization proposed by SDR.

4.3. Influence of the number of annotations

Fig. 3. Metric evolution on the three datasets w.r.t. the number of annotations. Control network performances are in dots.

To better apprehend our approach, we analyze the influence of the number of annotations over the performances. According to the results of the previous section, we fine-tune the network using SDR regularization. As we can observe on Figure 3, less than one hundred annotations leads to sub-optimal results and the performances then greatly improve with respect to the number of annotations. After 100 annotations for the two ISPRS datasets and 300 annotations for the SemCity dataset, the DNN almost catches up with the control network performances and then stabilizes on a plateau.

5. CONCLUSION

To conclude, we have shown that it is possible to make a DNN learn on the fly new segmentation classes after the initial training with point supervision using pseudo-labelling and relevant regularizations. We have assessed the method efficiency with experiments on three public remote sensing datasets. Moreover, with a sufficient amount of annotations, the performances even almost catch up with a DNN trained on the classes of interest in a fully supervised way.

6. REFERENCES