Learning Unbiased Zero-Shot Semantic Segmentation Networks Via Transductive Transfer

Fengmao Lv, Haiyang Liu, Yichen Wang, Jiayi Zhao, and Guowu Yang

Abstract—Semantic segmentation aims to obtain a detailed understanding of images. Deep learning has achieved great advances in semantic segmentation over the past years. In practice, however, the classes do not always correspond to the ones in the training stage. Since it is impractical to collect sufficient labeled data for all classes, zero-shot semantic segmentation has received increasing attentions recently. Although semantic segmentation neural networks can transfer knowledge from seen classes to unseen classes by incorporating the class-level semantic information, it shows a strong bias towards seen classes. In this letter, we propose an easy-to-implement transductive approach to alleviate the prediction bias in zero-shot semantic segmentation. We assume that both source images with full pixel-level labels and unlabeled target images are available for training. The source images are used to build the relationship between visual images and class-level semantic embeddings. On the other hand, the target images are used to alleviate the bias towards seen classes. Comprehensive experiments over the PASCAL dataset clearly demonstrate the effectiveness of our approach.

Index Terms—Semantic segmentation, transductive learning, zero-shot learning.

I. INTRODUCTION

RECENTLY, deep learning has achieved great advances in semantic segmentation. To train a good semantic segmentation neural network, we need to collect sufficient images with full pixel-level labels for training. In practice, however, the categories do not always correspond to the ones in the training stage. In this case, the semantic segmentation neural networks fails to make correct predictions for the unseen classes. Hence, it is a crucial problem to train semantic segmentation neural networks in the zero-shot setting.

Zero-Shot Learning (ZSL) aims to recognize unseen (target) classes by transferring knowledge from seen (source) classes [1], [2]. ZSL was mainly studied in the context of image recognition [1], [3]. The recent works propose to solve zero-shot semantic segmentation by incorporating class-level semantic information into semantic segmentation neural networks [4].

The previous approaches build the relationship between visual images and class-level semantic embeddings by mapping the image pixels into intermediate feature maps that lie in a semantic word embedding space shared by all the classes [4]. However, as the relationship is entirely built from the source data, semantic segmentation neural networks tends to map the pixels of unseen images to the points close to the semantic word embeddings specified by seen classes [5]. As a result, the semantic segmentation neural networks shows a strong bias towards seen classes in the Generalized ZSL (GZSL) setting. This work proposes an easy-to-implement transductive approach to alleviate the prediction bias in zero-shot semantic segmentation [6]. We assume that both source images with full pixel-level labels and unlabeled target images are available for training. We alleviate the prediction bias by forcing the pixels of target images to be mapped into the region specified by any target class. Our method can also be useful for the conventional ZSL setting since it regularizes the semantic segmentation neural networks to make more reasonable predictions for target images.

In summary, the contribution of this work is mainly two-fold. One is that we introduce unlabeled target images in zero-shot semantic segmentation to alleviate the prediction bias towards seen classes. To the best of our knowledge, this is the first work that focuses on transductive zero-shot semantic segmentation. The other is that we propose an effective approach to rectify the prediction bias of zero-shot semantic segmentation neural networks. Comprehensive experiments over the PASCAL dataset clearly demonstrate the effectiveness of our approach.

II. RELATED WORK

Semantic segmentation is an essential task in computer vision. The existing works are mainly based on Fully Convolutional Networks (FCNs) [6]. FCNs is trained in a pixel-to-pixel manner. It uses convolutional layers to process the image pixels into intermediate feature maps for final prediction. On the basis of FCNs, different components, such as atrous convolution [5], multi-scale aggregation [7], context relation [8], etc., are proposed to improve the performance.

Since the classes do not always correspond to the ones in the training stage, zero-shot semantic segmentation has received increasing attentions over the recent years. The key insight in zero-shot semantic segmentation is to build the relationship
between visual images and class-level semantic embeddings. Xian et al. propose to map the image pixels to the semantic word embedding space [4]. On the other hand, Kato et al. propose to map the class-level semantic embeddings to the visual space. Bucher et al. propose to generate visual representations of image pixels from the semantic word embeddings based on generative adversarial networks [9].

III. PROBLEM STATEMENT & MOTIVATION

Denote by \( x_s \in \mathbb{R}^{H \times W \times 3} \) the source images and \( x_t \in \mathbb{R}^{H \times W \times 3} \) the target images, where \( H \) and \( W \) are the height and width, respectively. \( y_s \in \mathcal{Y}_s^{H \times W} \) and \( y_t \in \mathcal{Y}_t^{H \times W} \) are the corresponding pixel-level labels. In zero-shot semantic segmentation, the source and target images are from different classes: \( \mathcal{Y}_s \cap \mathcal{Y}_t = \emptyset \). Each class \( y \in \mathcal{Y}_s \cup \mathcal{Y}_t \) has a semantic word embedding denoted by \( \phi(y) \in \mathbb{R}^d \). In this work, we assume that both labeled source images \( \mathcal{D}_s = \{(x_s, y_s)\}_{s \in S} \) and unlabeled target images \( \mathcal{D}_t = \{(x_t)\}_{t \in T} \) are available for training. Our goal is to obtain good pixel-level predictions for target images in the conventional ZSL setting and for both source and target images in the generalized ZSL setting.

A crucial point lies in how to effectively leverage the unlabeled target images to alleviate the prediction bias towards seen classes. To this end, a natural idea is to conduct self-training with the unlabeled target images. However, it cannot be directly suitable for zero-shot learning. The performance of self-training heavily depends on the initial network weights. As indicated above, the semantic segmentation neural networks has a strong prediction bias towards seen classes. Hence, the produced pseudo labels for target images can be relatively poor and cause the problem of negative transfer. To tackle this issue, our approach forces the semantic segmentation neural networks to project the visual pixels of target images into the region specified by any target class in the semantic word embedding space, instead of the fixed semantic word embeddings specified by the pseudo labels. Compared with the self-training approaches, our approach does not suffer from the problem of negative transfer.

IV. METHODOLOGY

The overall architecture of our approach is shown in Figure 1. The source images are used to construct the relationship between visual images and class-level semantic embeddings. On the other hand, the target images are used to alleviate the prediction bias towards seen classes.

A. Build Visual-Semantic Relationship

The relationship between visual images and class-level semantic embeddings reveals which class the visual pixels of an image belong to. To build such the relationship, our approach forces FCNs to project the pixels of input images into the semantic word embedding space shared by both source and target classes. To this end, we set the dimension of intermediate feature maps to that of the semantic word embeddings. Denote by \( F(x) \in \mathbb{R}^d \) the representations of image pixels over the semantic word embedding space. The inner product between the intermediate representations of image pixels and the semantic embeddings of each class can be used to calculate the class probabilities:

\[
p(y_i \mid x) = \frac{F(x)^T \phi(y)}{\sum_{y \in \mathcal{Y}_s \cup \mathcal{Y}_t} F(x)^T \phi(y)},
\]

where \( \hat{y}_{ij} \) denotes the prediction for pixel \( x_{ij} \). Immediately, we can train the segmentation model with the following loss:

\[
\mathcal{L}_r = \sum_{s \in S} \mathcal{L}_{seg}(y_s, p(\hat{y}_s)),
\]

where \( p(\hat{y}_s) \in \mathbb{R}^{H \times W \times |\mathcal{Y}_s \cup \mathcal{Y}_t|} \) and \( \mathcal{L}_{seg} \) denote the pixel-wise class probabilities and the pixel-wise cross-entropy loss, respectively. The \( \mathcal{L}_r \) loss forces the networks to map the image pixels into the semantic word embeddings specified by the corresponding classes. As the target images are unlabeled, \( \mathcal{L}_r \) does not involve the target images.

B. Alleviate Prediction Bias

To alleviate the prediction bias towards seen classes, our approach forces the semantic segmentation neural networks to project the pixels of target images into the region specified by any target class in the semantic word embedding space. To this end, we set the dimension of intermediate feature maps to that of the semantic word embeddings. The inner product between the intermediate representations of image pixels and the semantic embeddings of each class can be used to calculate the class probabilities:

\[
p(y_i \mid x) = \frac{F(x)^T \phi(y)}{\sum_{y \in \mathcal{Y}_s \cup \mathcal{Y}_t} F(x)^T \phi(y)},
\]

where \( p(y_i) \) denotes the prediction for pixel \( x_i \). Immediately, we can train the segmentation model with the following loss:

\[
\mathcal{L}_r = \sum_{s \in S} \mathcal{L}_{seg}(y_s, p(\hat{y}_s)),
\]

where \( p(\hat{y}_s) \in \mathbb{R}^{H \times W \times |\mathcal{Y}_s \cup \mathcal{Y}_t|} \) and \( \mathcal{L}_{seg} \) denote the pixel-wise class probabilities and the pixel-wise cross-entropy loss, respectively. The \( \mathcal{L}_r \) loss forces the networks to map the image pixels into the semantic word embeddings specified by the corresponding classes. As the target images are unlabeled, \( \mathcal{L}_r \) does not involve the target images.
end, we encourage the pixels of target images to have large probabilities of being any target class:

\[ L_b = -\sum_{t \in T} \sum_{i=1}^{H} \sum_{j=1}^{W} \ln \sum_{k \in \mathcal{V}_t} p(y_{ij} = k | x_t). \]

Compared with self-training, our approach does not use pseudo labels to rectify the prediction bias. Hence, it does not suffer from the negative transfer problem caused by incorrect pseudo labels. Although designed primarily for the generalized ZSL setting, our approach can also be useful for the conventional ZSL setting since it regularizes the segmentation model to make more reasonable predictions for target images.

C. Model Overview

To sum up, we have the following objective function:

\[ \min \ L_r + \lambda L_b, \]

where \( \lambda \) is the trade-off hyper-parameter that balances the importance of \( L_b \).

\( L_b \) encourages the semantic segmentation neural networks to have better prediction performance for the target images. As a result, the self-training strategies can produce more high-quality pseudo labels for the target images. Hence, we can progressively improve the performance with self-training after we rectify the prediction bias with \( L_b \).

V. Experiments

A. Dataset

In this work, the PASCAL-VOC dataset is used as the benchmark [13]. PASCAL-VOC consists of 12,031 images with pixel-level labels from 20 classes. We conduct experiments over different data splits of PASCAL-VOC as shown in Table I. Each data split contains 15 seen classes and 5 unseen classes.

Following [14], we evaluate the performance of our approach on the validation set of PASCAL-VOC.

B. Baseline

We compare our approach with the following baselines:

- **One-shot baselines** include One-shot Learning for Semantic Segmentation (OSLSM) from [14], and conditional Fully Convolutional Networks (co-FCN) from [17].
- **Few-shot baselines** include Foreground-Background (FG-BG) from [18], and Multi-scale Discriminative Location-aware network (MDL) from [19].
- **Inductive zero-shot baselines** include Semantic Projecting Network (SPNet) from [4] and Variational Mapping (VM) from [20].
- **Transductive zero-shot baselines** include Self-Training (ST) from [15] and Class-Balanced Self-Training (CBST) from [16].

All the above baselines use VGG-16 as the backbone network.

C. Performance Comparison

**GeneralizedZSL**: Table II shows the comparison of different approaches in the generalized ZSL setting. Following [4], we use the mean Intersection-Over-Union (mIOU) on seen classes, the mIOU on unseen classes and the harmonic mean of them as the evaluation metric.

We draw the following observations. First, our approach obtains better performance than the inductive ZSL baseline, i.e. SPNet. Such results clearly demonstrate the benefit of introducing unlabeled target images to alleviate the prediction bias towards source classes. However, the performance of the compared transductive ZSL baselines, i.e. ST and CBST, are far from satisfactory. Their performance, especially for unseen classes, can be even worse than that of SPNet in some data splits. We ascribe it to the poor pseudo labels caused by the model’s prediction bias towards seen classes. From the last row, we can see that self-training can progressively improve the performance after we rectify the prediction bias. However, performance drop can be observed on the data split P → 5\(^2\). It can be attributed to the fact that the “diningtable” and “person” classes have a weak relation with the other classes. Hence, the semantic segmentation neural networks cannot produce good pseudo labels for these classes.

**ConventionalZSL**: Table III shows the comparison of different approaches in the conventional ZSL setting. We can see that our approach can also be effective for the conventional ZSL setting. Similar observations can be drawn as in the generalized ZSL setting. The performance of our approach is slightly worse than that of the few-shot baselines, i.e. co-FCN and MDL, on the data split P → 5\(^2\). Such results can also be attributed to the weak relation between the source and target classes. The few-shot baselines overcome the problem through explicitly learning from a few labeled target images.

D. Discussion

In Fig. 2, we display the qualitative results on the data split P → 5\(^2\). It is clear that our approach obtains better segmentation...
TABLE II
Comparison of Different Approaches in the Generalized ZSL Setting. The Notations “Seen”, “Unseen” and “H” denote mIoU on Seen Classes, mIoU on Unseen Classes and Harmonic Mean of Them, Respectively

| Method          | Seen | Unseen | H     | Seen | Unseen | H     | Seen | Unseen | H     | Seen | Unseen | H     |
|-----------------|------|--------|-------|------|--------|-------|------|--------|-------|------|--------|-------|
|                 | P - 5^1 |        |       | P - 5^2 |        |       | P - 5^3 |        |       | mean  |        |       |
| SPNet [4]       | 60.1  | 7.0    | 12.5  | 55.2 | 23.0   | 34.4  | 55.7 | 14.3   | 22.8  | 61.6 | 14.2   | 23.1  |
| ST [5]          | 63.8  | 1.7    | 3.3   | 71.0 | 24.8   | 36.7  | 66.5 | 15.3   | 24.9  | 71.3 | 35.3   | 47.2  |
| CBST [16]       | 66.4  | 12.7   | 21.4  | 65.3 | 13.4   | 22.3  | 65.3 | 10.6   | 18.3  | 69.1 | 21.4   | 32.7  |
| Ours            | 68.4  | 40.0   | 50.5  | 70.4 | 53.5   | 60.8  | 63.7 | 24.3   | 35.1  | 69.4 | 43.9   | 53.8  |
| Ours + ST       | 69.7  | 53.5   | 60.6  | 71.6 | 58.7   | 64.6  | 65.9 | 19.1   | 29.6  | 71.4 | 53.8   | 61.3  |

Fig. 2. The qualitative results on P - 5^3 in the generalized ZSL setting. (a) the input images. (b, c, d) the results of the compared baselines. (e) the results of our approach. (f) the pixel-level ground truth.

TABLE III
Comparison of Different Approaches in the Conventional ZSL Setting in Terms of mIoU. The Notations †, ‡ and § Denote the One/Few-Shot Baselines, Inductive ZSL Baselines and Transductive ZSL Baselines, Respectively

| Method          | P - 5^1 | P - 5^2 | P - 5^3 | mIoU  |
|-----------------|---------|---------|---------|-------|
| OLSLM [14]      | 33.6    | 53.3    | 40.9    | 33.5  | 30.8  |
| co-FCN [17]     | 36.7    | 50.6    | 44.9    | 32.4  | 41.2  |
| FG-BG [18]      | 27.4    | 51.7    | 34.0    | 26.4  | 34.9  |
| MDL [19]        | 40.6    | 58.5    | 47.7    | 36.6  | 45.9  |
| SPNet [4]       | 27.5    | 71.8    | 32.9    | 47.8  | 44.9  |
| VM [20]         | 39.6    | 52.6    | 41.0    | 35.6  | 48.2  |
| ST [15]         | 2.0     | 40.7    | 18.5    | 49.4  | 27.7  |
| CBST [16]       | 19.6    | 25.8    | 15.4    | 43.8  | 26.2  |
| Ours            | 58.4    | 74.1    | 41.2    | 65.9  | 59.9  |
| Ours + ST       | 65.7    | 67.8    | 26.6    | 70.8  | 62.7  |

Fig. 3. The performance obtained by our approach by varying the value of λ.

In this work, we propose an easy-to-implement transductive zero-shot approach for semantic segmentation. We assume that both source images with full pixel-level labels and unlabeled target images are available for training. Our approach uses the unlabeled target images to alleviate the model’s prediction bias towards seen classes by forcing the semantic segmentation neural networks to project the pixels of target images into the region specified by any target class in the semantic word embedding space. Extensive experiments over the PASCAL dataset clearly demonstrate the effectiveness of our approach.

VI. Conclusion

The Λ_b loss will have a small effect on the training process. Hence, the semantic segmentation neural networks cannot build a strong relationship between visual images and semantic word embeddings.

In this work, we propose an easy-to-implement transductive zero-shot approach for semantic segmentation. We assume that both source images with full pixel-level labels and unlabeled target images are available for training. Our approach uses the unlabeled target images to alleviate the model’s prediction bias towards seen classes by forcing the semantic segmentation neural networks to project the pixels of target images into the region specified by any target class in the semantic word embedding space. Extensive experiments over the PASCAL dataset clearly demonstrate the effectiveness of our approach.
REFERENCES

[1] E. Kodirov, T. Xiang, and S. Gong, “Semantic autoencoder for zero-shot learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4447–4456.

[2] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, “Feature generating networks for zero-shot learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5542–5551.

[3] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, “Leveraging the invariant side of generative zero-shot learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Long Beach, CA, USA, Jun. 16-20, 2019, pp. 7402–7411.

[4] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, “Semantic projection network for zero- and few-label semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 8256–8265.

[5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.

[6] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3431–3440.

[7] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in Proc. 4th Int. Conf. Learn. Representations, San Juan, Puerto Rico, 2016.

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6230–6239.

[9] M. Bucher, T. Vu, M. Cord, and P. Pérèz, “Zero-shot semantic segmentation,” in Proc. Adv. Neural Inform. Process. Syst. 32: Annu. Conf. Neural Inform. Process. Syst. 2019, NeurIPS, Vancouver, BC, Canada, Dec. 2019, pp. 466–477.

[10] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. jégou, and T. Mikolov, “Fasttext.zip: Compressing text classification models,” 2016, ArXiv: abs/1612.03651.

[11] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.

[12] F. Jurie, M. Bucher, and S. Herbin, “Generating visual representations for zero-shot classification,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops, 2017, pp. 2666–2673.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, Jun. 2010.

[14] A. Shaban, B. Bansal, Z. Liu, I. Essa, and B. Boots, “One-shot learning for semantic segmentation,” in Proc. Brit. Mach. Vis. Conf., 2017, pp. 167.1–167.13.

[15] X. Zhu and J. D. Lafferty, “Harmonic mixtures: Combining mixture models and graph-based methods for inductive and scalable semi-supervised learning,” in Proc. ICML, 2005, pp. 1052–1059.

[16] Y. Zou, Z. Yu, B. V. K. V. Kumar, and J. Wang, “Unsupervised domain adaptation for semantic segmentation via class-balanced self-training,” in Proc. ECCV, 2018, pp. 297–313.

[17] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, “Conditional networks for few-shot semantic segmentation,” in Proc. Int. Conf. Learn. Representations, 2018.

[18] S. Caelles, K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. V. Gool, “One-shot video object segmentation,” in Proc. Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5320–5329.

[19] Z. Dong, R. Zhang, X. Shao, and H. Zhou, “Multi-scale discriminative location-aware network for few-shot semantic segmentation,” in Proc. 43rd IEEE Annu. Comput. Softw. Appl. Conf., Milwaukee, WI, USA, 2019, pp. 42–47.

[20] N. Kato, T. Yamasaki, and K. Aizawa, “Zero-shot semantic segmentation via variational mapping,” in Proc. Int. Conf. Comput. Vis. Workshop, Oct. 2019, pp. 1363–1370.