社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

A Study on Fixed Orthogonal Prototype Classifier for Semantic Segmentation

Jialei CHEN†, Daisuke DEGUCHI†, Chenkai ZHANG†, Xu ZHENG††, and Hiroshi MURASE†

† Faculty of Information, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
†† No.1 Du Xue Rd, Nansha District, Guangzhou

Abstract Semantic segmentation models consist of an image encoder and a learnable classifier to do this task. Though various image encoders are proposed, few works concentrate on designing classifiers. In this paper, we observe that frozen orthogonal prototypes can work even better than learnable ones with the help of constraints for shaping feature spaces. Therefore, we also propose a loss function to better shape the feature space. Experiments on the ADE20k dataset show impressive results.

Key words orthogonal prototype, semantic segmentation, representation learning

1. Introduction

As a significant task in computer vision, semantic segmentation has attracted the attention of many research endeavors in recent years. Thanks to the rapid development of deep learning technology, semantic segmentation also entered a new era. Since FCN [1], the structure of end-to-end semantic segmentation has been fixed, *i.e.*, an image encoder and a 1×1 convolutional layers. Then the community paid much attention to designing a more powerful image encoder, *e.g.*, DeepLab V3 [2] applies a novel image encoder to enlarge the receptive fields. However, few researchers focus on designing a new type of classifier.

To design a new type of classifier, we propose Orthogonal Prototypes (OP). Specifically, we orthogonalize a group of randomly initialized vectors and fix them during the progress of training. Moreover, we further propose a loss to enhance the segmentation ability and better shape the feature space based on InforNCE [3]. We conduct experiments on the ADE20k dataset [4], and the results show that the proposed methods achieve the expected improvements.

2. Related Works

2.1 Semantic Segmentaiton

Aiming to assign each pixel to its corresponding class, semantic segmentation plays an important role in computer vision. Since the 1×1 convolutional layer is proposed as the classifier in FCN [1], the community focuses on the research of image encoder and achieved many impressive results, *e.g.*, DeepLab V3 [2] and segformer [5]. Different from these works, we propose a new design for the classifier and note that the classifier is always fixed.

2.2 Contrastive Learning

As a novel method of learning more representative features, contrastive learning aims to pull the positive pairs together, *e.g.*, features belonging to the same object, and pushes negative pairs apart. Since InfoNCE [3] is proposed, many works, *e.g.*, moco [6], simclr [7], are proposed and obtain unforgettable results. The proposed regularization term is also based on contrastive learning. However, different from the works above, the contrastive of our method is applied between the classifier and the image features.

3. Methods

In this section, we first introduce the orthogonal prototypes and then describe the regularization term. Finally, introduce the training objectives of our method.

3.1 Orthogonal Prototypes

First, we revisit the normal procedures to do semantic segmentation. Given a dataset $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{M}$ where x and y indicate the image and its pixel-level annotations and M denotes the number of images in a dataset. Then we forward the $\mathbf{X}$ into the image encoder E to obtain the representation $R^{B*C*H*W} \in \mathcal{R}$ where B is the input batch size. C represents the channel number of the representation and H, W mean the height and the width. Finally, $\mathbf{R}$ is fed into the classifier G to get the pixel-level classification results $\mathbf{P}$. Then $\mathbf{P}$ is supervised by the one-hot label y,

$$d = dis(\mathbf{P}, \mathbf{Y}) \tag{1}$$

where d indicates the distances between the label and the prediction, dis indicates the functions, *e.g.*, cross-entropy, to estimate the distances between the prediction and annotations.

One of the properties that the one-hot vectors maintain is orthogonality, *i.e.*, the inner product between any label belonging to different classes is 0 and the same class 1. However, the learned classifier can not guarantee this property leading to a gap between the learning and supervision.

To bridge this gap, we propose Orthogonal Prototypes (OP). Specifically, we first randomly initialize a group of vectors $\mathbf{S}^{N*C}$ where N indicates the number of classes in a dataset. Then based on the Gram-Schmidt algorithm [8], we orthogonalize and normalize $\mathbf{S}$, and $\mathbf{S}$ is always fixed and not updated.

Table 1: Ablation on the effect of OP and Regularization Term (RT).

OP	RT	mIoU	mAcc
-	-	33.4	44.4
✓	-	33.5	44.4
-	✓	34.3	45.6
✓	✓	**35.1**	**48.0**

3.2 Regularization Term

Though bridging the gap, as the decrease of network depth, the capability of the network may also decrease. To address this issue, we propose a new regularization term to better shape the feature space, *i.e.*, make the features close to their corresponding prototype.

First, we calculate the mean feature of a specified category in $\mathbf{R}$,

$$v_l = \frac{\sum_{B,H,W} \mathbf{R} * \mathbb{1}(\mathbf{Y} = l)}{\sum_{B,H,W} \mathbb{1}(\mathbf{Y} = l)}, \tag{2}$$

where $\mathbb{1}(Y = l)$ indicates where the $\mathbf{Y}$ belongs to class l and convert l to 1 or 0. Then the regularization term is,

$$\mathcal{L}_c = \Sigma_i^N \frac{\exp(\mathbf{v}_i \cdot \mathbf{s}_i/\tau)}{\Sigma_{j\neq i}^N \exp(\mathbf{v}_j \cdot \mathbf{s}_j/\tau) + \exp(\mathbf{v}_i \cdot \mathbf{s}_i/\tau)}, \tag{3}$$

where $s_i \in \mathbf{S}$ depicts the ith prototype and τ is a hyperparameter.

3.3 Training Objectives

The total loss function is,

$$\mathcal{L} = CE(E(\mathbf{X}) \cdot \mathbf{S}, \mathbf{Y}) + \lambda * \mathcal{L}_c, \tag{4}$$

where λ is a hyperparameter to control the scale of regularization, and CE is cross entropy.

4. Experiments

To estimate the effectiveness of the proposed methods, we conduct experiments on the widely known ADE20k dataset [4]. In this dataset, there are 20K images for training 2K images for valuation, and 3K images for testing. Besides there are 150 classes.

4.1 Implementation Details

Our codes are based on the MMsegmentation [9]. The baseline model we use is Segformer-B0. In addition, the batch size for both datasets is set to 16. The crop size is set to 512 pixels × 512 pixels. The models are trained 40K iterations on ADE20K. τ is set to 0.07 and the length of OP is set as 2 by default.

4.2 Ablation Studies

To evaluate the effectiveness of the proposed method, we conduct experiments compared with baseline methods shown in Table 1. Compared with baseline methods, *i.e.*, no OP and RT, with both OP and RT, the performance is 1.7% higher in mIoU and 3.6% in mAcc, which is a large margin. When removing RT, the performance drops drastically, *i.e.*mIoU to 33.5% and mAcc to 44.4% which is very close to the baseline methods. If we only consider the impact of OP the mIoU drops to 34.3% and mAcc to 45.6%.

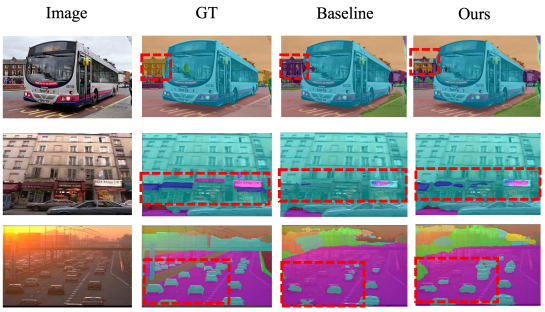

Figure 1: Visualization of on ADE20K dataset.

4.3 Visualization

We further visualize the prediction of our methods as shown in Fig. 1 where our methods perform better than the baseline.

5. Conclusion

In this paper, we observe that with regularization terms, the orthogonal prototypes perform better than learnable classifiers. This work may provoke the design of classifiers in semantic segmentation.

6. Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 23H03474, and JST CREST Grant Number JPMJCR22D1. The computations are carried out on the supercomputer "Flow" at the Information Technology Center, Nagoya University. The author Jielei Chen is sponsored by the China Scholarship Council.

Reference

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[3] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[4] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

[5] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[8] Å. Björck, "Numerics of gram-schmidt orthogonalization," *Linear Algebra and Its Applications*, vol. 197, pp. 297–316, 1994.

[9] M. Contributors, "Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark," 2020.