SEMANTIC-EMBEDDED KNOWLEDGE ACQUISITION AND REASONING FOR IMAGE SEGMENTATION

Wei Liu,Huigang Zhang,Xiaojie Xia,Liuan Wang and Jun Sun

Fujitsu R&D Center, Co., LTD

ABSTRACT

Image segmentation is a difficult and challenging task because of the complex object appearance and diverse object categories. Traditional methods directly use visual features for segmentation but ignore the correlation between objects. We introduce a knowledge reasoning module (KRM) for external knowledge aggregation and leverage a graphic neural network to aggregate the knowledge feature, which is concatenated with a visual feature for semantic segmentation. To this end, we use word embedding of category names as semantic feature and establish the relationship between categories. Through iteration, the aggregated features can be enriched. In experiments, three well known semantic segmentation methods are used as baseline. Our experiment results outperform the baseline methods on the food dataset FoodSeg103 and Cityscapes, and demonstrate the effectiveness of our proposed method.

Index Terms— Knowledge reasoning, semantic segmentation

1. INTRODUCTION

Semantic segmentation is a main computer vision task, most of the existing segmentation algorithms are based on the visual features of encoder-decoder structures. These algorithms are restricted by the limited receptive field of convolution kernel. At present, the popular graph convolution network can expand the relationship between various regions in the image. The combination of external knowledge and graph convolution network can naturally solve the problem of insufficient receptive fields in semantic segmentation, thus enriching the features used in semantic segmentation.

Most semantic segmentation methods [1, 2, 3, 4, 5, 6] follow the encoder-decoder architecture based on fully convolution network [7]. Previous researchers have achieved good results on general semantic segmentation datasets such as PASCAL VOC2017, Cityscape and ADE20K, many of the efforts are addressed on enlarging the receptive field.In the present research, we aim to study semantic segmentation from a different perspective.

Recently GCN [8] has demonstrated its remarkable ability on computer vision tasks, which can alleviate the limited receptive field problem. The GCN-based algorithm exhibits good performance in handling non-local regions. GAT [9] introduces masked self-attentional layers and assigns the corresponding weights to different adjacent nodes to improve the shortcomings of the GCN. Sun et al.

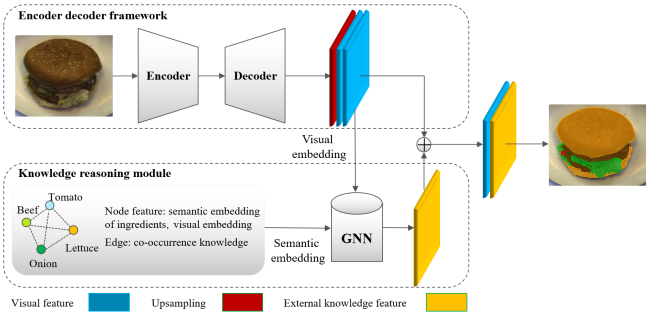


Fig. 1. Overview of the proposed knowledge reasoning network for semantic segmentation task. Our method is designed on the basic encoder decoder framework and can be flexibly applied to all networks based thereon. Firstly, the semantic knowledge of the category names is extracted by word embedding and visual knowledge from convolution weights as the node feature of the graphic convolution network, then the co-occurrence knowledge of categories in the dataset is used as the value of the edge that connects two nodes. Finally, this knowledge is fed into GNN and the external knowledge is extracted, and the feature is concatenated with visual feature for the final semantic segmentation.

Work has been undertaken on the inductive knowledge reasoning for CV tasks. Zhang et al. [10] introduces knowledge-based reasoning network for object detection, and leverages graphic neural network to build the knowledge relationship between objects. Chen et al. [11] introduces KR-Net to establish the prior semantic relationship between the objects in segmentation task, but ignores the visual feature relationship.

To address the difficulties of segmentation task, we propose a knowledge reasoning network for segmentation task. Firstly, the semantic features and visual features of objects are extracted as the nodes of the graph convolution network, we then establish the relationship between these objects. Fi-

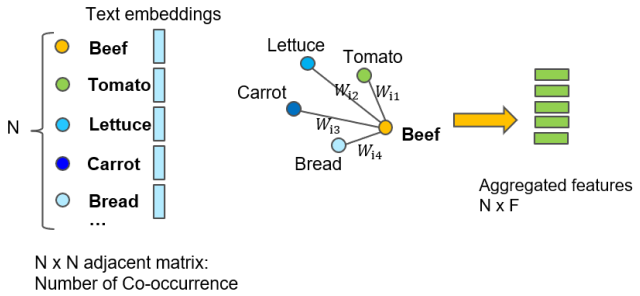


Fig. 2. Knowledge reasoning module process.

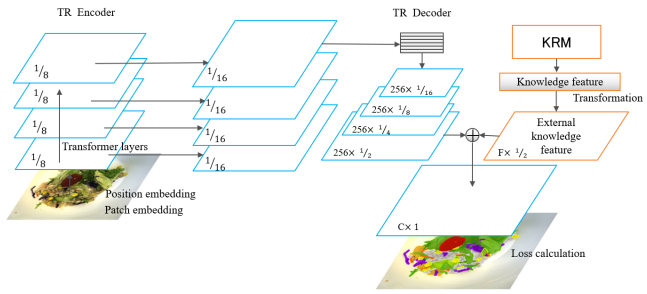


Fig. 3. SETR plug-in: proposed knowledge reasoning module.

nally, a graph convolution network is used to aggregate the information from the node features of these objects and the edge to produce the external knowledge features. The external features are optimised iteratively, and finally aggregated with visual features for food image semantic segmentation.

The main contribution of this paper are as follows:

- A novel knowledge reasoning module is proposed for semantic segmentation task by leveraging external intuitive knowledge. GCN [8] is introduced to aggregate the external knowledge to enhance the features for segmentation task.
- The proposed module can be flexibly integrated with other semantic segmentation framework.
- The effectiveness of the proposed method is proved by the evaluation on the public dataset FoodSeg103 and Cityscapes.

2. METHODOLOGY

We propose a simple yet effective module to aggregate the external knowledge and obtain the knowledge features as an auxiliary of visual features for semantic segmentation. Our main framework is illustrated in Fig. 1. The supplementary knowledge features are appended in specific layer of decoder module, which enrich the features for final segmentation. The representative framework SETR [6] are illustrated in Fig. 3.

2.1. External knowledge reasoning module

As described above, we target image semantic segmentation with external knowledge reasoning. The task can be formulated thus: $D_t = (X_{it}, L_{it}, S_{it}, E_{it})$, where X_{it} is the i_{th} input sample, L_{it} is the segmentation labels, S_{it} is the visual semantic representation, and E_{it} is the external knowledge representation. $C_t = \{C_{1t}, C_{2t} \dots C_{Nt}\}$ is the set of objects categories. In this research, we propose to use GNN to aggregate the external knowledge. The inputs of GNN consist of the nodes and edges, which are described as below:

Node representation. The node representation comes from word embedding of the object name or visual embedding of convolution transformation weights. The word embedding can be obtained from pretrained text encoder, such as GLoVe [12] or CLIP [13]. For each object category, the text representation can be represented as $F_i \in R^{d^T}$, d^T is the vector length. Since we have N object categories, the final text representation $F^T \in R^{N \times d^T}$ is determined. The visual embedding is realized from convolution transformation weights using a 1×1 convolution when converting the channels, it can be denoted by $F^{T^{'}} \in R^{N \times C}$, where C is the number of channels in the penultimate convolution of the decoder module.

Edge representation. The edge between two nodes represent the relationship of two objects in the image (e.g., beef and bread often appear together). A simple method is used to measure the relationship by counting the number of co-occurrences between different categories in the training datasets. We thereby obtain the symmetric relationship matrix $M_{N \times N}$. After normalization, the adjacent matrix A is obtained and used as the edge of GCN.

The GCN is initialized with the text or visual representation H and the adjacent matrix A. Our target is to attain an enhanced feature as a supplement of visual semantic feature. As shown in Fig. 2, GCN provides a strategy that the node features can be iteratively optimised through the weight calculation with related categories. The forward propagation process can be expressed using the following formula:

$$h_i^{'} = \sigma(\sum_{(j \in N)} a_{ij} h_i \omega_{ij}) \tag{1}$$

where a_{ij} is the correlation between h_i and h_j, h_i is the current head, ω_{ij} represents the weight of GCN and σ is the nonlinear activation function.

2.2. Multi-modality knowledge fusion

To enrich the features of external knowledge, three different methods are used to extract the knowledge features, NLP

semantic feature and CLIP [13] based multi-modality feature and visual knowledge feature. NLP semantic feature use single word of the category name to transfer the word to vector, each text embedding is represented as $f \in R^{d^T}$, where d^T is the length of the vector. CLIP-based feature [13] first make a sentence using the category name(e.g., This is a picture of tomato), then the text encoder of CLIP is employed to extract the text feature, the pretrained CLIP has already made a good alignment between the real-world images and the description sentences, which is auxiliary information for our knowledge reasoning module. The visual knowledge feature use convolutional weights from visual features transformation in the decoder module. All the three modules use the same $N \times N$ adjacent matrix as the the edges of GCN. The features extracted by the three modules are supplement to the visual features and effectively reduce the data deviation. We will show the influences of different knowledge embedding modules in the section describing the experimental work.

2.3. Feature enhancement for the semantic segmentation

As shown in Fig.1, after obtaining the feature from the knowledge reasoning module, we add these knowledge features to the visual features as in [10]. Because the knowledge features represent the relationship between the objects which belong to high-level features, the knowledge features are added to the penultimate decoder layer.

The knowledge feature can be denoted by $F_k \in R^{N \times D}$, where N is the number of the categories including background, D is the size of the feature vector in each category. While the last layers feature of the decoder is $F^l \in R^{B \times N \times W \times H}$, where B denotes the batch size, N is the number of categories including background, and the penultimate layers feature is $F^{l-1} \in R^{B \times C \times W \times H}$, where C represents the number of channels in the penultimate decoder layer. We make the transformation as follows, each of image features in the batch is iterated, where F^{li} is i_{th} sample feature F^l. The enhanced feature can be denoted thus:

$$\mathcal{F}^{cmb} = \mu(\varphi(F^{li}, \gamma(F^k)), F^{(l-1)i}) \tag{2}$$

where $F^{cmb} \in R^{(D+C) \times W \times H}$, γ is the repeat operation (H times), φ is the multiplication operation and μ is concatenate operation.

After all features are iterated with the above operations in the batch, finally we obtain the $F^{batch} \in R^{B \times (D+C) \times W \times H}$ featuresand make a 1×1 convolution to change the channels to $F^{new} \in R^{B \times N \times W \times H}$, the generated new features will participate in the loss calculation.

3. EXPERIMENTS

Our experiments are conducted with three popular baseline methods sem-FPN [3], CCNet [4] and SeTR [6] on food segmentation dataset FoodSeg103 and Cityscapes, and the knowledge reasoning module is implemented to demonstrate the effectiveness of the proposed method.

Method	mIoU	Model size
FPN [3](ResNet50)	27.3	218M
FPN-KRM(CLIP+GCN)	**28.3**	227M
CCNet [4](ResNet50)	35.1	381M
CCNet-KRM(CLIP+GCN)	**36.4**	399M
SETR [6](Vit-16/B)	44.6	776M
SETR-KRM(CLIP+GCN)	**45.7**	805M

Table 1. Segmentation evaluation results of MIou on FoodSeg103 dataset

Method	mIoU	Model size
FPN [3](ResNet50)	74.5	218M
FPN-KRM(CLIP+GCN)	**76.4**	227M
CCNet [4](ResNet50)	79.3	381M
CCNet-KRM(CLIP+GCN)	**80.5**	399M
SETR [6](Vit-16/B)	78.1	776M
SETR-KRM(CLIP+GCN)	**79.7**	805M

Table 2. Segmentation evaluation results of MIou on Cityscapes(test), training schedule with 80k

3.1. Implementation details

FoodSeg103 contains 7118 food RGB images, including 4983 training images and 2135 testing images, the dataset has 103 food ingredient categories, all the images come from dataset Recipe1M. 4983 images are used as training samples and 2135 images are used as validation samples. **Cityscapes** contains 19 object categories of urban scenes. It includes 5000 finely annotated images, with 2975, 500 and 1525 for training, validation and testing respectively.

Encoder decoder settings: the baseline methods on Food-Seg103 are conducted based on the MMSegmentation platform [14]. We all use the pretrained model to improve the performance. FPN and CCNet use Resnet-50 as backbone, while SeTR use ViT-16/B. In the decoder part of SeTR, the five-layer up-sampling structure is selected as the decoder to facilitate the experiment of adding knowledge features.

Training strategy settings: to allow fair comparison, the same batch size and pretrained model are used for both proposed method and baseline method with a batch size 4, four GPUs. We use 0.01 as the initial learning rate for each baseline method. For our proposed knowledge reasoning methods, the initial learning rate is set as 0.0004 to ensure network converge.

Knowledge reasoning module setting: we use two different graph network GCN and GAT to aggregate the external

Method	mIoU
FPN+CLIP+GCN	**28.3**
FPN+CLIP+GAT	27.9
CCNet+CLIP+GCN	**36.4**
CCNet+CLIP+GAT	36.1
SETR+CLIP+GCN	**45.7**
SETR+CLIP+GAT	44.9

Table 3. Influcences of Graphic networks on FoodSeg103

knowledge features to compare the performance.

3.2. Results and evaluation metrics

For all the categories, mIoU (mean of all categories IoU) and mAcc (mean of all categories Acc) are used to evaluate the performance of each segmenter. The experiments are conducted on three popular baseline methods (FPN, CCNet and SeTR). In our knowledge module, CLIP based knowledge extraction module performs best and GCN performs better than GAT for knowledge feature generation.

Results on FoodSeg103 Table 1 Compares the segmentation results on FoodSeg103, our proposed knowledge reasoning module achieves a better performance (mIoU: +1.0%, +1.3%, +1.1%).

Results on Cityscapes Table 2 Compares the segmentation results on Cityscapes, SETR-Naive structure is adopted and our proposed method significantly outperforms the baseline with a large margin (mIoU: +1.9%, +1.2%, +1.6%), proving the effectiveness of the proposed method.

3.3. Influence of different GNN models

In this section, different graphic networks such as GCN and GAT are used to examine the influences of GNN models.

3.4. Ablation study

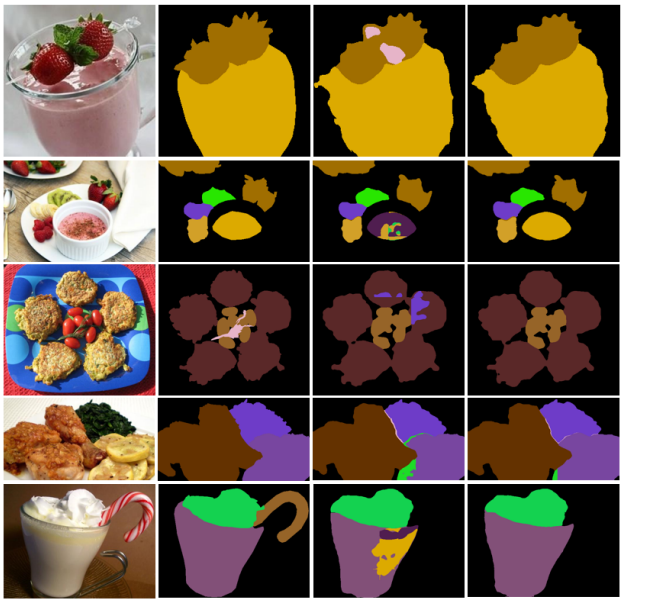
GCN aggregation module can achieve better performance and the CLIP-based multi-modality is better than other knowledge embedding methods. In this section, we select SETR and GCN model as baseline and use different combinations of knowledge embedding to evaluate the performance. Experimental results (Table 4) shows that a combination of CLIP-based module and visual knowledge module can achieve the best mIoU performance.

3.5. Visualization

In Fig. 4, some visualization examples are demonstrated. SETR is used as the baseline method in the third column, the fourth column is proposed method. From the visualization observations, the SETR-knowledge reasoning module (KRM) achieves better performance and more detailed results. The

Method	mIoU
SETR[6](Vit-16/B)	44.6
SETR-KRM(GCN+CLIP)	45.7
SETR-KRM(GCN+CLIP+GLOVE)	45.3
SETR-KRM(GCN+CLIP+VISUAL)	**45.8**
SETR-KRM(GCN+CLIP+GLOVE+VISUAL)	45.2

Table 4. Results of different modules on FoodSeg103



(a).source image (b).ground truth (c). SETR baseline (d).SETR with knowledge

Fig. 4. Visualization of testing samples in FoodSeg103: SETR with knowledge achieves better performance.

SETR-KRM method also focuses on macroscopic relationships between objects in one image, which help improve the segmentation performance of objects that visually have large intra-class variance.For example, in the first row the green leaves belong to the strawberry, our SETR-KRM model can recognize the green leaves as part of a strawberry.

4. CONCLUSION

In this paper, a semantic segmentation framework that incorporates the KRM for image segmentation task is introduced. The category name text semantic feature and visual feature are used as auxiliary information, and the relationship between the objects is established, then GCN is used to aggregate the external knowledge. Three popular baseline methods are selected for comparisons with our proposed method with plug-in KRM module. Experiments show that our proposed method outperforms the baseline methods on the FoodSeg103 and Cityscapes datasets. We hope our proposed method can contribute to the community in semantic segmentation tasks.

5. REFERENCES

[1] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[2] 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," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[3] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6399–6408.

[4] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612.

[5] Yuhui Yuan, Xilin Chen, and Jingdong Wang, "Object-contextual representations for semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 173–190.

[6] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

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

[8] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[9] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[10] Huigang Zhang, Liuan Wang, and Jun Sun, "Knowledge-based reasoning network for object detection," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 1579–1583.

[11] Shengjia Chen, Zhixin Li, and Xiwei Yang, "Knowledge reasoning for semantic segmentation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2340–2344.

[12] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[14] MMSegmentation Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.