Co-occurrent Features in Semantic Segmentation

Hang Zhang 1  Han Zhang 2  Chenguang Wang 1  Junyuan Xie 1

1Amazon Web Services  2Google Brain

{hzaws, chgwang, junyuanx}@amazon.com, zhanghan@google.com

Abstract

Recent work has achieved great success in utilizing global contextual information for semantic segmentation, including increasing the receptive field and aggregating pyramid feature representations. In this paper, we go beyond global context and explore the fine-grained representation using co-occurrent features by introducing Co-occurrent Feature Model, which predicts the distribution of co-occurrent features for a given target. To leverage the semantic context in the co-occurrent features, we build an Aggregated Co-occurrent Feature (ACF) Module by aggregating the probability of the co-occurrent feature within the co-occurrent context. ACF Module learns a fine-grained spatial invariant representation to capture co-occurrent context information across the scene. Our approach significantly improves the segmentation results using FCN and achieves superior performance 54.0% mIoU on Pascal Context, 87.2% mIoU on Pascal VOC 2012 and 44.89% mIoU on ADE20K datasets. The source code and complete system will be publicly available upon publication.

1. Introduction

Semantic segmentation provides per-pixel label of object categories for the given image, which is a challenging task requiring accurate prediction of the object category, location and shape. Successful approaches are usually based on Fully Convolutional Network (FCN) [31], with a Deep Convolutional Neural Network (CNN) [22, 23] as the base network. Recent work achieves great success in leveraging contextual information, including enlarging receptive field size with pyramid-based representations [6,29,53] and learning category specific scaling factors using context embedding [51].

Despite the success in incorporating global contextual information, in some challenging scenarios, a rough holistic global context might not be enough for the classification of ambiguous objects in the scene. In addition, we observe natural scenes usually have reasonable and coherent composition of objects. The presence of one object, even in a spatially disjoint region, can be compelling evidence of the existence of the other. The co-occurrence property among objects can improve the robustness of the recognition system and help resolve the ambiguity of object labels against noises such as occlusion and variations in pose and illumination. For example, as shown in Figure 1, sea, river and water are very similar in appearance and the global context as a city scene is not able to disambiguate these three as they can all exist near a city. But object co-occurrence asserts that sea is more likely to appear when boat is around.

Figure 1: Some object categories are difficult to distinguish based on local appearance and scene context. In this example, water, river and sea are visually similar and all fit this scene context. Human can utilize the presence of the boat to make the prediction, as it typically co-occurs with the sea. Motivated by this, we introduce Aggregated Co-occurrent Feature Module to relook at the relations with all the co-occurrent features before making the predictions. (More visual examples in Figure 2)

1Links can be found at http://hangzh.com/
Moreover, co-occurrence does not only exist between objects and it can also be generalized to different parts of an object. As shown in the 1st row of Figure 2, an armchair is composed of armrests, legs, back and seat. It is difficult to resolve the ambiguity between a chair with an armchair without noticing the co-occurring armrest parts. In general, co-occurrence features play an important role in recognizing the class labels of image pixels. Therefore, a powerful approach directly capturing the co-occurrence features and utilizing their dependencies is desirable for semantic segmentation.

Existing approaches are not capable to capture the dependencies between co-occurrence features due to their fixed spatial structure. The baseline FCN [31] has a relatively local receptive field and fails to utilize co-occurrence features in distant portions of the image. Recent work simply enlarges the receptive field by utilizing the multi-scale feature representations using pyramid pooling method [17, 53] or different atrous rates of convolutions [6]. So the same pooling or atrous convolution operation is applied everywhere in the feature map. However, the distribution of crucial features for the recognition of different image regions varies tremendously. Instead of having fixed spatial connection, the network should be able to capture co-occurrence features across different relative locations, in a spatial invariant manner.

As the first contribution of this work, the feature co-occurrence is modeled as a probability distribution over the feature space conditioned on a given target feature, which we refer to as Co-occurrence Feature Model (CFM). The CFM learns an inherent co-occurrence representation, where the similarities between features measure how likely the features would co-occur with the target in the same image. We therefore define a probability distribution conditioned on target feature using Softmax of the similarities between the target and co-occurrence features across the space, which inherits the spatial invariant nature. Moreover, we expect the co-occurrence features also capture the scene context. However, we find that the limitation in expressiveness of the Softmax distribution is a bottleneck for modeling the context information. For this, we propose a contextual prior as a conditional probability on the scene context. The CFM is then defined as a mixture of Softmaxes distribution with the contextual prior. With the proposed CFM, we build Aggregated Co-occurrence Feature Module to integrate the context-aware information within the co-occurrence features, which allows the network to recap the whole scene before making individual predictions (overview in Figure 3).

The second contribution of this paper is constructing Co-occurrence Feature Network (CFNet), the state-of-the-art semantic segmentation architecture. With the proposed ACF Module, we build CFNet with pre-trained ResNet [18] as the base network. The proposed CFNet with ResNet-101 base network achieves state-of-the-art results 54.0\% mIoU on Pascal Context [33], 87.2\% mIoU on Pascal VOC 2012 [12] and 44.89\% mIoU on ADE20K [56].

2. Co-occurrence Features

We refer to the features co-occurring with the target feature within the same input image/featuremap as co-occurrence features. In this section, we first introduce the Co-occurrence Features Model to capture the distribution of the co-occurrence features for a given target. We further introduce Aggregated Co-occurrence Feature Module to aggregate the contextual information of co-occurrence features across the scene as the output target feature representation.

2.1. Co-occurrence Feature Model

We tackle the feature co-occurrences as a probabilistic problem instead of predicting their presences, since the co-occurrence features for a given target are usually not deterministic. We build a Co-occurrence Feature Model, which learns an inherent representation via measuring the similarity between the co-occurrence feature and the target feature, indicating how likely they would co-occur. Then the probability distribution of the co-occurrence features conditioned on target feature can be defined using Softmax of the similarities across the space. Consider the input CNN feature map as \(N\) number of channel-dimensional features \(X = \{x_1, ..., x_N\}\), and \(x_i\) for \(i \in \{1, ..., N\}\) is the input feature at location \(i\). The probability of the co-occurrence feature \(x_c\) for a given target feature \(x_t\) is:

\[
p(x_c | x_t) = \frac{e^{s(x_c, x_t)}}{\sum_{i=1}^{N} e^{s(x_i, x_t)}},
\]

where \(s(x_c, x_t)\) is the similarity between the co-occurrence feature \(x_c\) and the target feature \(x_t\). A natural parameterization for the similarity function \(s\) is using dot product similarity \(s(x_c, x_t) = u_{x_c}^T v_{x_t}\), where \(v_{x_t}\) and \(u_{x_c}\) are the target and co-occurrence vector representations for feature \(x_t\) and \(x_c\). The vector representations are given by \(u_{x_c} = \Phi_c(x_c)\) and \(v_{x_t} = \Phi_t(x_t)\), where \(\Phi_c\) & \(\Phi_t\) are the learnable transformations using feed-forward networks. The proposed model in Eq. 1 is in the same spirit with the skip-gram model proposed in [32], which is used to capture the co-occurrence word representation.

**Contextual Prior.** We find it is difficult to model the co-occurrence features only using the target information without knowing the whole scene, because the distribution of the co-occurrence features for the same target varies in different context. For example, we may expect chairs or tables

\[\text{Inspired by the distributed hypothesis [16]: the target feature representations are modeled to predict well co-occurrence features in its context.}\]
Some challenging category labels are difficult to distinguish even using global semantic context, which requires understanding the fine-grained details in the co-occurrence features. In the 1st example, it is hard to know whether the chair is an armchair without noticing co-occurring arm. For the 2nd example, the baseline FCN fails to predict mirror parts that are far from sconce. Similarly, FCN fails to utilize the spatial layout to distinguish the cabinet with kitchen island. The proposed CFNet relooks at the relations with the co-occurrence features before classifying each pixel, which successfully segments the above mentioned objects and also distinguishes the road from sidewalk and dirt track, segments the mountain as a whole part in the last two examples. (Visual examples from ADE20K dataset [56].)

c-o-occurring with a human in the indoor scene, but expect vehicles and buildings instead in the outdoor scene. Recent study also shows that the Softmax-based models do not have enough capacity for high-rank problems [49]. We can hypothesize that predicting co-occurrence features is a high-rank problem in images, which we can show with empirical observations. If the co-occurrence features are low-rank, we could use finite number of bases to represent all possible co-occurrence features by a weighted combinations of these bases. However, this contradicts with our common sense about the varieties of the real-world images. Therefore, predicting co-occurrence features is a high-rank problem.

To tackle the above issues, we propose to model the scene context as contextual prior. Inspired by Yang et
al. [49], the contextual prior is defined as a Mixture of Softmaxes (MoS) to learn a prior distribution for the co-occurrent features conditioned on the contextual information. The MoS formulates the co-occurrent probability of \( x_t \) for target \( t \) as:

\[
p(x_t|x_t) = \sum_{k=1}^{K} \pi^k \sum_{i=1}^{N} e^{s^k(x_t,x_t)},
\]

where \( \pi^k \) is the prior or mixture weight of the \( k \)-th component, and \( s^k(x_t,x_t) \) is the similarity in \( k \)-th component for \( k \in \{1,...,K\} \). The vector representations \( u_{x_t} \) and \( v_{x_t} \) are chunked into \( K \) sub-components, and the similarity of each component is given by \( s^k(x_t,x_t) = u_{x_t}^\top k v_{x_t}^k \). The prior of each mixture is conditional on the contextual information, which can be parameterized as \( \pi^k = \frac{\exp(w_k^c v_{x_t})}{\sum_{k=1}^{K} \exp(w_k^c v_{x_t})} \), where \( v_{x_t} = \sum_{c=1}^{C} v_{c}^t w_c^t \) captures the contextual information and \( w_k \) is a learnable vector. The MoS allows the co-occurrent features under different semantic context can have different priors.

### 2.2. Aggregated Co-occurrent Feature Module

To utilize the co-occurrent features, we build Aggregated Co-occurrent Feature Module (ACF), which aggregates the co-occurrent contexts with their co-occurrent probabilities across the spatial locations in a self-attention [42] or non-local [3] manner:

\[
z_t = \sum_{c=1}^{C} p(x_c|x_t) \cdot \psi_c,
\]

where \( z_t \) is the aggregated feature output for target \( t \), \( p(x_c|x_t) \) is the co-occurrent probability given by Equation 2 and \( \psi_c \) is the co-occurrent context at location \( c \). The co-occurrent context is given by \( \psi_c = \Psi(x_c) \), and \( \Psi \) is a learnable transformation using feed-forward network. The ACF Module aggregates the co-occurrent feature distributions, and aggregates the contextual information with the co-occurrent probabilities.

#### Dropout and Multi-head Ensembles

Model combination almost always improves the performance for machine learning algorithms. Dropout [40] randomly drops units during the training, so that it learns “thinned” networks and averages the logits during the inference. Dropout can avoid the network adapting too much on the training data for overfitting. We apply dropout [40] on the co-occurrent features and expect the network to make correct predictions even if some of the concurrent features are missing, so that the network can generalize from limited patterns appeared in the training set. Another model combination we explore is using “multi-head” [42], which concatenates the features of module outputs using different weights to build a network ensemble. We adapt the multi-head strategy to further improve the model capacity.

#### Global Pooling Feature

Global pooling (GP) feature is commonly used in modern semantic segmentation approaches [6, 29, 53], which provides a global receptive field as a strong cue to distinguish category in confusing areas. The GP feature is captured by a global average pooling, fol-
followed by a $1 \times 1$ convolution, and then attached to each feature location. We extend the proposed Co-occurrence Feature Module with a global pooling feature branch to leverage the global context, as shown in Figure 3.

3. Co-occurrence Feature Network

With proposed Co-occurrence Feature Module, we build Co-occurrence Feature Network (CFNet) as shown in Figure 3. We use pre-trained ResNet [18] as the base network and apply dilated network strategy to Res-4 and Res-5 of ResNet, resulting stride-8 models. The proposed Aggregated Co-occurrence Feature Module and global pooling feature branch are added on top of the base network. ACF Module considers the input convolutional featuremap with feature branch are added on top of the base network. ACF Module with a global pooling feature branch to leverage the global context, as shown in Figure 3. We use pre-trained ResNet [18] as the base network (CFNet) as shown in Figure 3. We use pre-trained ResNet [18] as the base network.

Table 1: Imagenet [10] pretraining for the base networks. The top-1 and top-5 accuracy (%) on validation set use center crop on image size of $224 \times 224$ and $320 \times 320$.

<table>
<thead>
<tr>
<th>Network</th>
<th>$224^2$ center</th>
<th>$320^2$ center</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>top-1</td>
<td>top-5</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>78.55</td>
<td>94.17</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>80.24</td>
<td>95.12</td>
</tr>
</tbody>
</table>

Context Aggregation. Pioneering work demonstrates that combining global features with local patches can improve the segmentation results [37, 39, 41]. ParseNet [29] proposes to concatenate a global pooling feature with original featuremap to capture global context and increase the receptive field size. Pyramid Pooling Module (PPM) [17, 53] concatenates the global pooling features from a multi-scale pyramid. Atrous Spatial Pyramid Pooling (ASPP) [6] uses a set of different atrous rate convolutions to capture pyramid feature representations with different receptive field sizes.

These methods have the predefined spatial connections. Consider the convolution operation as a matrix multiplication $y = Wx$, where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ are the flatten input and output and $W \in \mathbb{R}^{mn}$ is a transform matrix depending on the convolution kernel [11]. Matrix $W$ has $k_h \cdot k_w$ non-zero elements in each row for the convolution kernel with the shape of $k_h \times k_w$. Combining different atrous rate of convolutions as in ASPP [6] is adding the non-zero entries to each row, but the overall spatial connections are still sparse and the representation is spatial sensitive. The proposed ACF Module can also be formulated as $y = W \cdot x$ and $W$ is the co-occurrence probability. Comparing to existing methods, the ACF Module captures the context across the whole scene with spatial invariant representation.

Featuremap Attention. Attention mechanism has achieved great success in natural language processing [35, 42], which captures the long-range information using a weighted sum of all the features in a sequence. Non-local neural network [45] brings the self-attention to the field of video classification and object detection in computer vision. A key difference between Co-occurrence Feature Model with

3 We refer to the network stages with the original strides of 16 and 32 as Res-4 and Res-5.
Table 2: Ablation study of CFNet on Pascal Context dataset. **ACF** indicates using Aggregated Co-occurrent Feature Module, **GP** means including global pooling feature branch, **Enc** represents Context Encoding Module [51]. Adding Co-occurrent Feature significantly improves the segmentation results, and including global pooling feature and Context Encoding can further boost the performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>BaseNet</th>
<th>ACF</th>
<th>GP</th>
<th>Enc</th>
<th>pixAcc</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>Res50</td>
<td></td>
<td></td>
<td></td>
<td>76.3</td>
<td>46.3</td>
</tr>
<tr>
<td>FCN</td>
<td>Res50</td>
<td>✓</td>
<td></td>
<td></td>
<td>79.0</td>
<td>49.8</td>
</tr>
<tr>
<td>CFNet</td>
<td>Res50</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>79.3</td>
<td>51.6</td>
</tr>
<tr>
<td>CFNet</td>
<td>Res50</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>79.8</td>
<td>52.4</td>
</tr>
<tr>
<td>CFNet</td>
<td>Res101</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>81.1</td>
<td>54.9</td>
</tr>
</tbody>
</table>

Table 3: Ablation Study of Contextual Prior and Multi-heads. We vary the number of mixtures K and number of multi-heads H and find H=4, K=2 gives the best performance.

<table>
<thead>
<tr>
<th>pixAcc/mIoU K=1</th>
<th>K=2</th>
<th>K=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>H=1</td>
<td>77.1/49.4</td>
<td>79.3/51.6</td>
</tr>
<tr>
<td>H=2</td>
<td>77.9/49.4</td>
<td>79.4/51.8</td>
</tr>
<tr>
<td>H=4</td>
<td>79.2/51.2</td>
<td>79.6/52.1</td>
</tr>
</tbody>
</table>

4.4. Implementation Details

For baseline FCN and proposed CFNet, we use ResNet [18] as the base network and apply dilation strategy for the pre-trained networks, resulting in stride-8 models. Following the prior work [51, 53], we use bilinear interpolation to upsample the network output logits for calculating the loss. We use standard SGD optimizer and set the momentum to 0.9 and weight decay to 0.0001. A "poly" like learning rate scheduling [5] is used $lr = base\_lr \times (1 - \frac{iter}{total\_iter})^{power}$. We set the base learning rate as 0.004 for ADE20K and Cityscapes datasets and the power is set to 0.9. We use base learning rate of 0.004 for COCO pre-training and reduce it to 0.001 when fine-tuning on Pascal VOC. We use the "sync-once" implementation of Cross-GPU Batch Normalization provided by Zhang et al. [51]. As ACF Module is compatible with existing FCN based approaches, we also study the performance when adding Context Encoding Module and Semantic Encoding Loss with default settings in EncNet [51]. Following the prior work [53], an auxiliary loss is added after Res-4 by adding an additional FCN head to Res-4, which is applied to all the experiments.

The networks are trained for 120 epochs for ADE20K dataset, 180 epochs on Cityscapes dataset, 30 epochs for COCO pretraining, 50 epochs on Pascal VOC and 80 epochs on Pascal Context dataset. The images and ground truth masks are randomly flipped and rescaled to the ratio of 0.5 to 2.0 and randomly cropped into the training sizes using zero padding if needed. We use the mini-batch size of 16 during the training. The samples are randomly shuffled, and the last batch is discarded if mini-batch size is less than 16.

**Evaluation and Metrics.** During the evaluation, we follow the best practice [51] to average the network predictions in multiple scales. We first resize the original image into different scales {0.5, 0.75, 1.0, 1.25, 1.5, 1.75}, then crop the scaled images into training image size and feed the images into the network with flipping. Finally, the predicted logits are averaged across different crops and scales. Since the multi-size evaluation improves the performance of all the methods, we adopt this strategy for all the experiments. We use the standard metrics of pixel accuracy (pixAcc) and mean intersection of union (mIoU) in this experiments. For
the scene parsing results on Pascal Context and ADE20K validation sets, we ignore the background pixels in calculating the evaluation metrics, following the standard benchmark [56]. For the semantic segmentation results on Pascal VOC and Cityscapes datasets, we use the public server for the evaluation.

**ImageNet Pretraining.** Similar as in the prior work [51, 53], we modify the standard ResNet [17] by replacing the first $7 \times 7$ convolution with 3 consequent $3 \times 3$ convolution. We follow the best practice of ImageNet training [19] to train our base networks. The Top-1 and Top-5 accuracy on ImageNet validation set using center crop with the crop size of $224 \times 224$ and $320 \times 320$ are shown in Table 1. The pre-trained models will be included in the public code system.

**4.2. Abalation Study on Pascal Context**

Pascal Context dataset [33] is a scene parsing dataset, containing the semantic labels for the entire image, with 4,998 training and 5,105 validation images. Following the practice in prior work [5, 26, 33, 51], we use the 59 most frequent categories for this benchmark and consider all the other classes as background.

**Ablation Study of CFNet.** We first break down the improvements of CFNet over FCN, by conducting a set of experiments by adding individual components step-by-step to the baseline FCN. We use 4 mixtures and 2 multiheads in ACF Module with atrous-rate of 4 for the transformation $\Phi$ in this study. The baseline FCN achieves 76.3% pixAcc and 46.3% mIoU. Adding the ACF Module improves the pixAcc and mIoU by 2.7% and 3.6%. Including global pooling feature yields 0.9% boost in mIoU. Further improvements are from adding Context Encoding Module [51] and using deeper base network (See results in Table 2).

**Ablation Study of ACF Module.** To explore the best performance of Aggregated Co-occurrent Feature Module, we conduct the experiments with different hyper-parameters and settings. We first study different instantiations of the transformation $\Phi$ using different feed-forward network architectures and empirically find using the atrous rate of 4 gives best performance (detailed study in the supplementary material). We also explore the influence of contextual prior and multi-heads in the ACF Module in Table 3. To keep the comparison fair, we reduce the feature dimension, when increasing the number of mixtures or the number of multi-heads, so that the total computation of ACF Module remains roughly the same. Varying the number of mixtures $K$ for contextual prior and the number of multi-heads $H$ in ACF Module, we can see that using contextual prior significantly improves the expressiveness of the Softmax model, and empirically find $K=2$ and $H=4$ gives the best performance.

**State-of-the-art Comparisons.** We consider the background as one of the categories in order to compare with prior work (60 classes in total). The results are shown in Table 4. CFNet with ResNet-50 already outperforms most of the previous work even using much shallower base network. CFNet (ResNet-101) achieves 54.0% mIoU on validation set, which surpasses other approaches by a large margin even without using deeper base network or COCO pre-training.

**4.3. Results on Pascal VOC 2012**

Pascal VOC 2012 segmentation dataset [12] is one of the gold-standard benchmarks for object segmentation. Following the work [5, 53], we utilize the augmented set [14] with 10,582, 1,449 and 1,456 images in training, validation and test set. The CFNet is first trained on the train + val sets on the augmented set and then finetuned on the original Pascal VOC 2012 images as in previous work [51]. For fair comparison with prior work, we use ResNet-101 as the base network. CFNet-101 achieves 84.2% mIoU on the test set, which outperforms all the previous work without COCO pre-training and achieves superior performance on most of the categories. State-of-the-art approaches typically pre-train the network using MS-COCO dataset [28]. We follow the prior work [6, 51] to generate semantic segmentation mask by merging the instance labels for the 20 categories shared with Pascal VOC 2012 dataset, and discard the labels for the other categories, which results in around 90K images with more than 1000 labeled pixels (from the training set of MS-COCO 2017). We first pre-train the CFNet on COCO dataset using learning rate of 0.004 and then finetune on the augmented and original training set. CFNet achieves the best result of 87.2% mIoU on the test set. The per-class comparison is shown in Table 5, and CFNet achieves superior performance on many categories. (The entries using larger base model such as Xeption, or extra than COCO & ImageNet [10] data for pre-training are not included in this benchmark[7, 25, 44].)

**4.4. Results on ADE20K**

ADE20K dataset [56] is a large scale scene parsing benchmark with 150 object and stuff categories, containing 20K training, 2K validation and 3K test images. We first train the baseline FCN and CFNet on the training set and evaluate the models on the validation set (results are shown in Table 6). Our baseline FCN using ResNet-50 achieves 39.28% mIoU using good pre-trained base network and multi-size evaluation. CFNet outperforms FCN by more
4.5. Results on Cityscapes Dataset

Cityscapes dataset [9] is a high-resolution city street scene parsing dataset, including 5K high-quality labeled frames (fine data) and 20K weakly annotated ones (coarse data). We only use the fine data in this experiment with 2,975, 500 and 1,525 number of images for training, validation, and testing. 19 object/stuff categories are used in the evaluation. We use ResNet-101 as the base network, and train our CFNet on the training set using 768 crop size, then evaluate it on the validation set. CFNet achieves 79.56% mIoU on the validation set. For performance on test set, we retrain CFNet-101 on train and validation set and submit the prediction on test set to the evaluation server. CFNet achieves 79.60% mIoU (IoU classes) on the test set only using fine-label data. We have not used online hard example mining (OHEM) strategy in this experiment, which can further improve the performance.

5. Conclusion

To capture and utilize the co-occurrence features, we introduce a Co-occurrence Feature Model, which predicts the probability distribution of co-occurrence features for the given target. To further utilize co-occurrence features in semantic segmentation, we introduce an Aggregated Co-occurrence Feature Module to aggregate the co-occurrence features. The proposed approach outperforms existing contextual modules achieving superior performance on gold-standard semantic segmentation benchmarks. We expect the co-occurrence feature representation and our state-of-the-art implementations will be beneficial to the segmentation work in the community.
References


