SparseMask: Differentiable Connectivity Learning for Dense Image Prediction
Supplementary Material

Huikai Wu* Junge Zhang Kaiqi Huang†
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
{huikai.wu, jgzhang, kaiqi.huang}@nlpr.ia.ac.cn

1. Theorems

We present two theorems in Section 3.1.2 (main text), of which the proofs are given in this section.

Theorem 1. Concatenating the features and then applying convolution is equal to applying convolution to each feature and then take a summation.

Proof. Given \( M \) input features \( F_m \) with shape \( N \times C_{m_{in}} \times H \times W \), the concatenated feature is noted as \( F_{in} \) with shape \( N \times C_{in} \times H \times W \), where \( C_{in} = \sum_{m=0}^{M-1} C_{m_{in}} \). The corresponding convolution kernel is noted as \( W \) with shape \( C_{out} \times C_{in} \times KH \times KW \), which can be split into \( M \) weights \( W_m \) with shape \( C_{out} \times C_{in} \times KH \times KW \). The output feature \( F_{out} \) is represented as following:

\[
F_{out}[n, c_{out}, h, w] = \text{conv}(F_{in}, W)[n, c_{out}, h, w]
= \sum_{kh, kw} \sum_{c_{in}=0}^{C_{in}-1} W[c_{out}, c_{in}, kh, kw]F_{in}[n, c_{in}, h + kh, w + kw]
= \sum_{kh, kw} \sum_{m=0}^{M-1} \sum_{c_{in}=0}^{C_{m_{in}}-1} W^m[c_{out}, c_{in}, kh, kw]F^m_{in}[n, c_{in}, h + kh, w + kw]
= \sum_{m=0}^{M-1} \sum_{kh, kw} \sum_{c_{in}=0}^{C_{m_{in}}-1} W^m[c_{out}, c_{in}, kh, kw]F^m_{in}[n, c_{in}, h + kh, w + kw]
= \sum_{m=0}^{M-1} \text{conv}(F^m_{in}, W^m)[n, c_{out}, h, w].
\] (1)

Theorem 2. The order of bilinear upsampling and point-wise convolution is changeable.

Proof. The input feature is \( F_{in} \) with shape \( N \times C_{in} \times H_{in} \times W_{in} \), while the corresponding convolution kernel is \( W \) with

---

*This paper is based on the results obtained from the author’s internship at Preferred Networks, Inc.
†Also with CAS Center for Excellence in Brain Science and Intelligence Technology.
Figure 1: **Fully Dense Network based on MobileNet-V2**[2]. The inputs to the red circle (U) are multiple feature sets, while the output is the union of all the sets. F is the decoder stage, which takes a feature set as the input. Best viewed in color.

\[
F_{out}[n, c_{out}, h_{out}, w_{out}] = conv(f_{\uparrow}(F_{in}), W)[n, c_{out}, h_{out}, w_{out}]
\]

\[
= \sum_{c_{in}} W[c_{out}, c_{in}, 0, 0] f_{\uparrow}(F_{in})[n, c_{in}, h_{out}, w_{out}]
\]

\[
= \sum_{c_{in}} W[c_{out}, c_{in}, 0, 0] \sum_{i=0}^{3} [h_{in} - h_{in}^i][w_{in} - w_{in}^i] F_{in}[n, c_{in}, h_{in}^i, w_{in}^i]
\]

\[
= \sum_{i=0}^{3} \sum_{c_{in}} W[c_{out}, c_{in}, 0, 0] [h_{in} - h_{in}^i][w_{in} - w_{in}^i] F_{in}[n, c_{in}, h_{in}^i, w_{in}^i]
\]

\[
= \sum_{i=0}^{3} [h_{in} - h_{in}^i][w_{in} - w_{in}^i] \sum_{c_{in}} W[c_{out}, c_{in}, 0, 0] F_{in}[n, c_{in}, h_{in}^i, w_{in}^i]
\]

\[
= \sum_{i=0}^{3} [h_{in} - h_{in}^i][w_{in} - w_{in}^i] conv(F_{in}, W)[n, c_{out}, h_{in}^i, w_{in}^i]
\]

\[
= f_{\uparrow}(conv(F_{in}, W))[n, c_{out}, h_{out}, w_{out}],
\]

where \(f_{\uparrow}(\cdot)\) is bilinear upsampling, \(h_{in} = h_{out} / H_{out} \times H_{in}\) and \(w_{in} = w_{out} / W_{out} \times W_{in}\). \(h_{in}^i\) and \(w_{in}^i\) is calculated as follows:

\[
h_{in}^0 = [h_{in}], w_{in}^0 = [w_{in}]; h_{in}^1 = [h_{in}], w_{in}^1 = [w_{in}]
\]

\[
h_{in}^2 = [h_{in}], w_{in}^2 = [w_{in}]; h_{in}^3 = [h_{in}], w_{in}^3 = [w_{in}].
\]

2. **Fully Dense Network based on MobileNet-V2**

Figure 1 presents the Fully Dense Network based on MobileNet-V2. The inputs to the red circle (U) are multiple feature sets, while the output is the union of all the sets. F is the decoder stage, which takes a feature set as the input.

3. **Visual Results**

The visual results for experiments in Section 5 (main text) are shown in Figure 2.
Figure 2: Qualitative Results. Our method is not only quantitively but also qualitatively comparable to the baseline method.

References


