Deep Multimodal Fusion by Channel Exchanging

Yikai Wang[1], Wenbing Huang[1], Fuchun Sun[1],
Tingyang Xu[2], Yu Rong[2], Junzhou Huang[2]

[1] Tsinghua University [2] Tencent AI Lab

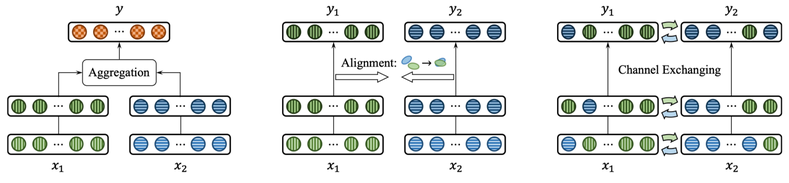
Code available at: https://github.com/yikaiw/CEN

NEURAL INFORMATION PROCESSING SYSTEMS

Summary of our work

This work proposes Channel-Exchanging-Network (CEN) for multimodal fusion, which has two basic advantages,

- Is self-guided with a global criterion and no more trails for fusion positions like existing methods;
- Balances the trade-off between inter-modal fusion and intra-modal processing.



(a) Aggregation-based fusion **(b)** Alignment-based fusion **(c)** Ours

A sketched comparison between existing fusion methods and ours

Background

The goal of deep multimodal fusion is to determine a multi-layer network (particularly CNN in this paper) whose output is expected to fit the target as much as possible. This can be implemented by minimizing the empirical loss as:

$$\min_{f} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\hat{\boldsymbol{y}}^{(i)} = f(\boldsymbol{x}^{(i)}), \boldsymbol{y}^{(i)}\right).$$

Two typical kinds of instantiations of this equation could be the aggregation-based fusion and the alignment-based fusion:

1) **Aggregation-based fusion** (processes each modality with a separate sub-network and then combine all their outputs via an aggregation operation followed by a global mapping.)

$$\hat{\boldsymbol{y}}^{(i)} = f(\boldsymbol{x}^{(i)}) = h(\mathrm{Agg}(f_1(\boldsymbol{x}_1^{(i)}), \cdots, f_M(\boldsymbol{x}_M^{(i)}))),$$

2) **Alignment-based fusion** (leverages an alignment loss for capturing the inter-modal concordance while keeping the outputs of all sub-networks.)

$$\min_{f_{1:M}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}_m^{(i)}), \boldsymbol{y}^{(i)}\right) + \mathrm{Alig}_{f_{1:M}}(\boldsymbol{x}^{(i)}), \quad s.t. \sum_{m=1}^{M} \alpha_m = 1,$$

Our method

The whole optimization objective of our method is,

$$\min_{f_{1:M}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}^{(i)}), \boldsymbol{y}^{(i)}\right) + \lambda \sum_{m=1}^{M} \sum_{l=1}^{L} |\hat{\gamma}_{m,l}|, \quad s.t. \sum_{m=1}^{M} \alpha_m = 1,$$

where,

- Each sub-network is equipped with BN layers containing the scaling factors, and we will penalize the L1 norm of their certain portion of the scaling factors for sparsity,
- The sub-network shares the same multimodal parameters except BN layers to facilitate the channel exchanging as well as to compact the architecture further.

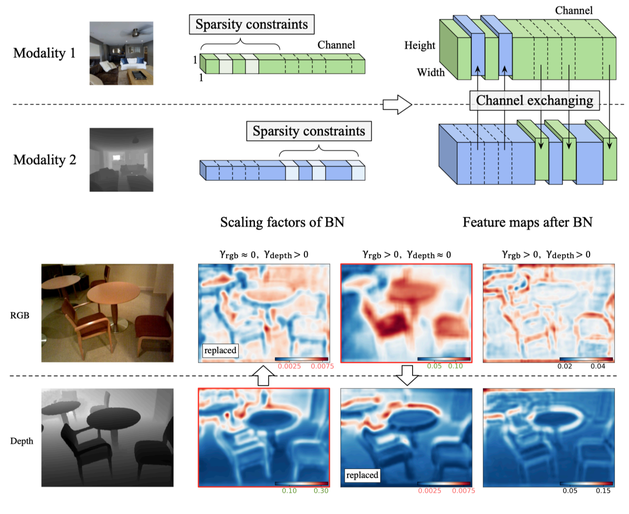
We replace the channels of small scaling factors with the ones of other sub-networks, since those channels potentially are redundant,

$$\boldsymbol{x}'_{m,l,c} = \begin{cases} \gamma_{m,l,c} \frac{\boldsymbol{x}_{m,l,c} - \mu_{m,l,c}}{\sqrt{\sigma^2_{m,l,c} + \epsilon}} + \beta_{m,l,c}, & \text{if} \quad \gamma_{m,l,c} > \theta; \\ \frac{1}{M-1} \sum_{m' \neq m}^{M} \gamma_{m',l,c} \frac{\boldsymbol{x}_{m',l,c} - \mu_{m',l,c}}{\sqrt{\sigma^2_{m',l,c} + \epsilon}} + \beta_{m',l,c}, & \text{else;} \end{cases}$$

To summary, our method has two steps,

- Create sparse activations by using a L1 norm over the BN scaling factors;
- Exchange an activation if its BN scaling factor is lower than a threshold.

Following figure illustrates our channel exchanging process.



Illustrations of our multimodal fusion strategy

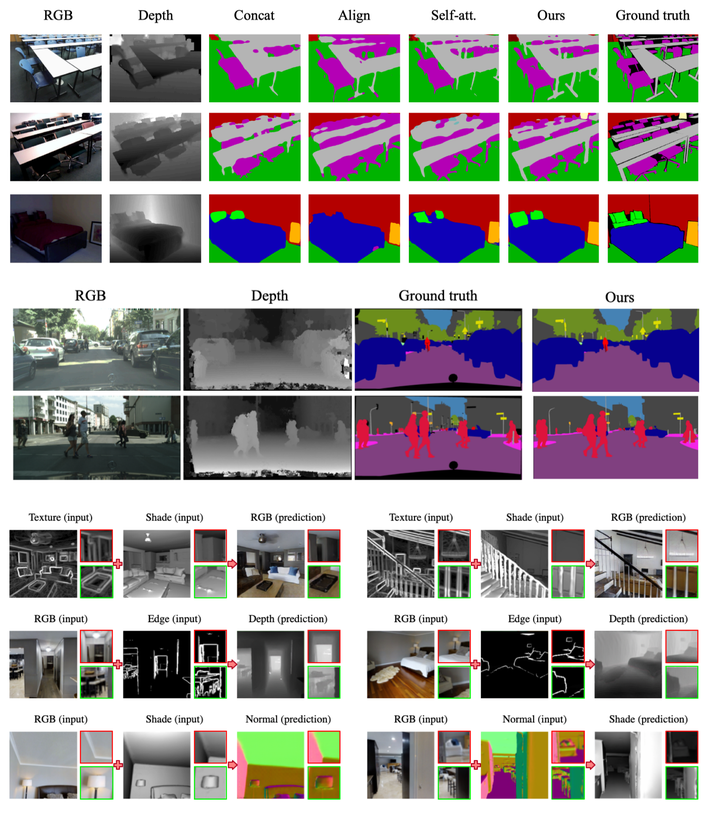
Analysis

Theorem 1. Suppose $\{\gamma_{m,l,c}\}_{m,l,c}$ are the BN scaling factors of a multimodal fusion network (without channel exchanging) optimized by our loss function. Then the probability of $\gamma_{m,l,c}$ being attracted to $\gamma_{m,l,c} = 0$ during training (a.k.a. $\gamma_{m,l,c} = 0$ is the local minimum) is equal to $2\Phi(\lambda|\frac{\partial L}{\partial x'_{m,l,c}}| - 1)$.

Corollary 1. If the minimal of our loss function implies $\gamma_{m,l,c} = 0$, then the channel exchanging (assumed no crossmodal parameter sharing) will only decrease the training loss, given the sufficiently expressive ability.

Experiments

We contrast the performance of CEN against existing multimodal fusion methods on two different tasks: **semantic segmentation** (surpass SOTAs on NYUDv2 and SUN-RGBD) and **image-to-image translation**.



Results visualization