BM-NAS: Bilevel Multimodal Neural Architecture Search

Nanyang Technological University[1], Harvard University[2], The Pennsylvania State University[3]

Yihang Yin[1], Siyu Huang[2], Xiang Zhang[3]

AAAI-22

Source Code: https://github.com/Somedaywilldo/BM-NAS

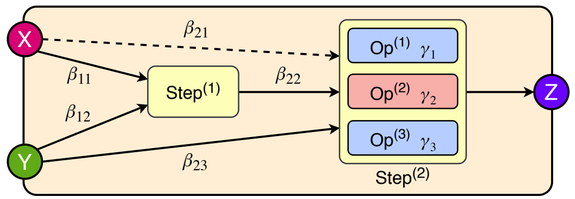
Our Contribution

- Towards a more generalized and flexible design of DNNs for multimodal learning, we propose a new paradigm that employs NAS to search both the unimodal feature selection strategy and the multimodal fusion strategy.
- We present a novel BM-NAS framework to address the proposed paradigm. BM-NAS makes the architecture of multimodal fusion models fully searchable via a bilevel searching scheme. And it is end-to-end differentiable.
- We conduct extensive experiments on three multimodal learning tasks to evaluate the proposed BM-NAS framework. Empirical evidences indicate that both the unimodal feature selection strategy and the multimodal fusion method are significant to the performance of multimodal DNNs.

Motivation

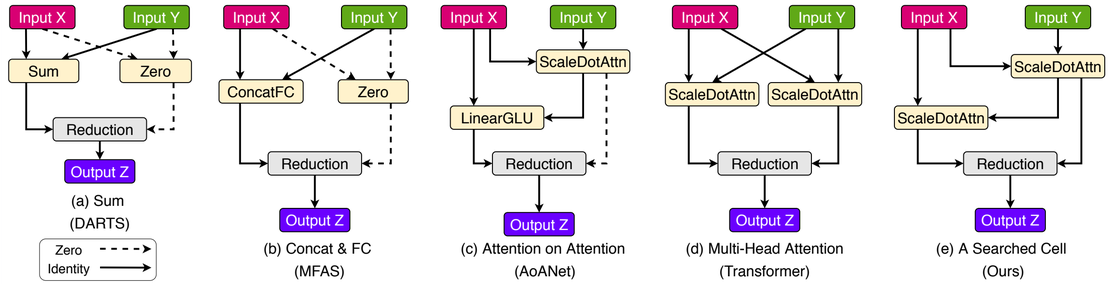
- DARTS only supports the search of unary operations (on the edge). The only operation on the step nodes to combine previous input is summation.
- To support attention mechanism, we need to support the search of binary/ternary operations on the step nodes. We designed a bilevel structure to achieve this.
- The lower level (steps) search fusion strategy, and the upper level (cells) search for feature selection strategy.

Lower Level: Fusion Strategy



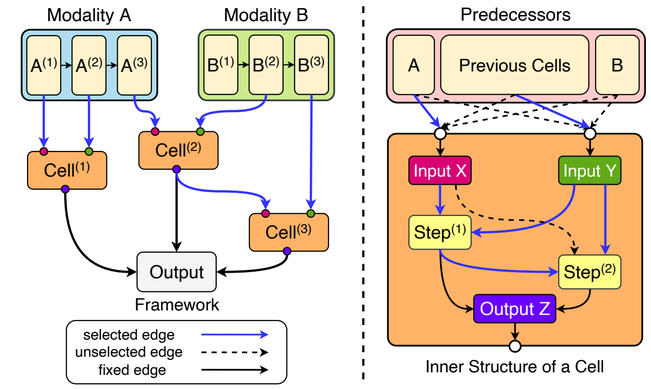
- Unary Operations (on the edges):
  - $\text{Identity}(x) = x$
  - $\text{Zero}(x) = 0$
- Binary Operations (on the step nodes):
  - $\text{Zero}(x, y) = 0$
  - $\text{Sum}(x, y) = x + y$
  - $\text{Attention}(x, y) = \text{Softmax}(\frac{xy^T}{\sqrt{C}}y)$
  - $\text{LinearGLU}(x, y) = \text{GLU}(xW_1, yW_2) = xW_1 \odot \text{Sigmoid}(yW_2)$
  - $\text{ConcatFC}(x, y) = \text{ReLU}(\text{Concat}(x, y)W + b)$

Lower Level Search Space



- The primitive operations are elaborately designed so they can be flexibly combined to accommodate various effective feature fusion modules such as multi-head attention (Transformer) and Attention on Attention (AoA), yet with a large space left for BM-NAS to discover better fusion operations.

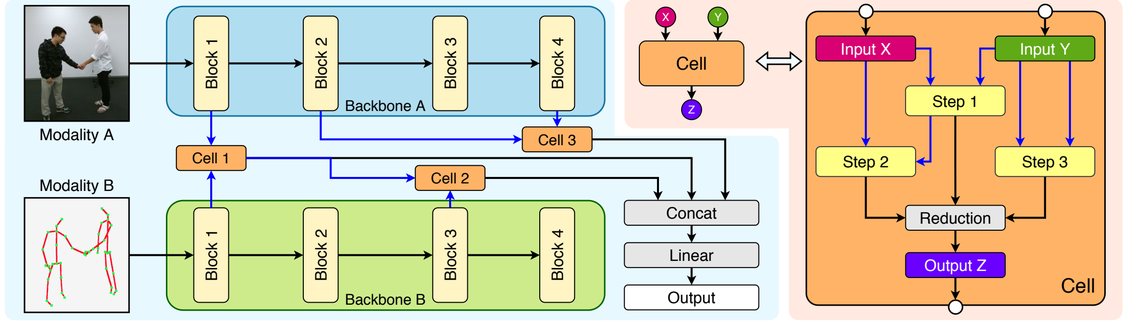
Upper Level: Feature Selection Strategy



Algorithm 1: Bilevel Multimodal NAS (BM-NAS)

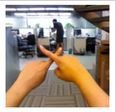
Result: The genotype of fusion networks.
Initialize architecture parameters α, β, γ and model parameters w;
Initialize *genotype* based on α, β, γ, set *genotype_best* = *genotype*;
Construct *hypernet* based on *genotype_best*;
while $\mathcal{L}$ *not converged* **do**
  Update ω on training set;
  Update (α, β, γ) on validation set;
  Derive upper level *genotype* based on α, derive lower level *genotype* based on β, γ;
  Update *hypernet* based on *genotype*;
  if *higher validation accuracy is reached* **then**
    Update *genotype_best* using *genotype*;
  end
end
Return *genotype_best*;

Framework



Experiments

- We performed multimodal classification tasks on MM-IMDB, NTU RGB-D and EgoGesture dataset.
- Involved modalities are image, text, RGB video, skeleton action sequence, depth video. Our framework can be generalized to other modalities as well.

Dataset	MM-IMDB	NTU RGB+D	EgoGesture
Modality A			
Modality B	Freedom fighters Neo, Trinity and Morpheus continue to lead the revolt against the Machine Army, unleashing their arsenal of extraordinary skills and weaponry against the systematic forces of repression and exploitation.		
Label	Action, Sci-Fi	Shaking Hands	Cross Index Fingers

MM-IMDB Dataset

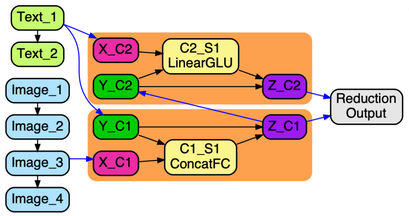


Table 1: Multi-label genre classification results on MM-IMDB dataset. Weighted F1 (F1-W) is reported.

Method	Modality	F1-W(%)
Unimodal Methods		
Maxout MLP (ICML13)	Text	57.54
VGG Transfer (ICLR15)	Image	49.21
Multimodal Methods		
Two-stream (NIPS14)	Image + Text	60.81
GMU (ICLR17)	Image + Text	61.70
CentralNet (ECCV18)	Image + Text	62.23
MFAS (CVPR19)	Image + Text	62.50
BM-NAS (ours)	Image + Text	**62.92** $\pm$ **0.03**

NTU RGB-D Dataset

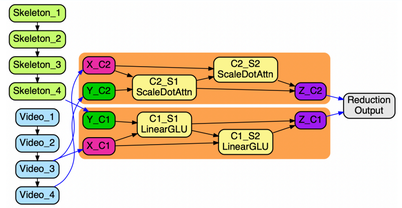


Table 2: Action recognition results on NTU RGB-D dataset.

Method	Modality	Acc(%)
Unimodal Methods		
Inflated ResNet-50 (CVPR18)	Video	83.91
Co-occurence (IJCAI18)	Pose	85.24
Multimodal Methods		
Two-stream (NIPS14)	Video + Pose	88.60
GMU (ICLR17)	Video + Pose	85.80
MMTM (CVPR20)	Video + Pose	88.92
CentralNet (ECCV18)	Video + Pose	89.36
MFAS (CVPR19)	Video + Pose	89.50 $\pm$ 0.60
BM-NAS (ours)	Video + Pose	**90.48** $\pm$ **0.24**

EgoGesture Dataset

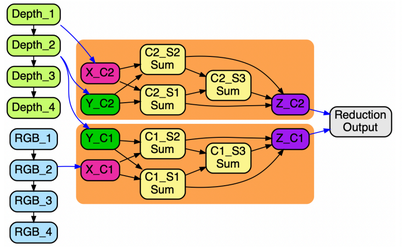


Table 3: Gesture recognition results on EgoGesture dataset. We use ResNext-101 as backbones for both RGB and depth modality for our BM-NAS method.

Method	Modality	Acc(%)
Unimodal Methods		
VGG-16 + LSTM (NIPS14)	RGB	74.70
C3D + LSTM + RSTTM (ICCV15)	RGB	89.30
I3D (CVPR17)	RGB	90.33
ResNext-101 (FG19)	RGB	93.75
VGG-16 + LSTM (CVPR14)	Depth	77.70
C3D + LSTM + RSTTM (CVPR16)	Depth	90.60
I3D (CVPR17)	Depth	89.47
ResNeXt-101 (FG19)	Depth	94.03
Multimodal Methods		
VGG-16 + LSTM (CVPR17)	RGB + Depth	81.40
C3D + LSTM + RSTTM (CVPR19)	RGB + Depth	92.20
I3D (CVPR17)	RGB + Depth	92.78
MMTM (CVPR20)	RGB + Depth	93.51
MTUT (3DV19)	RGB + Depth	93.87
3D-CDC-NAS2 (TIP21)	RGB + Depth	94.38
BM-NAS (ours)	RGB + Depth	**94.96** $\pm$ **0.07**