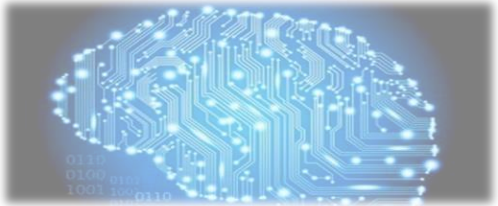
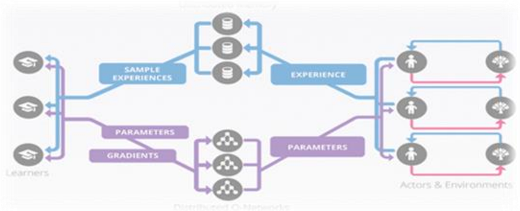



Deep Multimodal Neural Architecture Search

AutoML 연구실
석사과정 문아성

21.06.08

Deep Multimodal Neural Architecture Search

- Deep Multimodal Neural Architecture Search

Yu, Zhou, et al. "Deep Multimodal Neural Architecture Search."

Proceedings of the 28th ACM International Conference on Multimedia. 2020.

- Task

Visual Question Answering task: {images, questions, answers}

Image-Text matching task: {images, captions}

Visual Grounding task: {images, query}

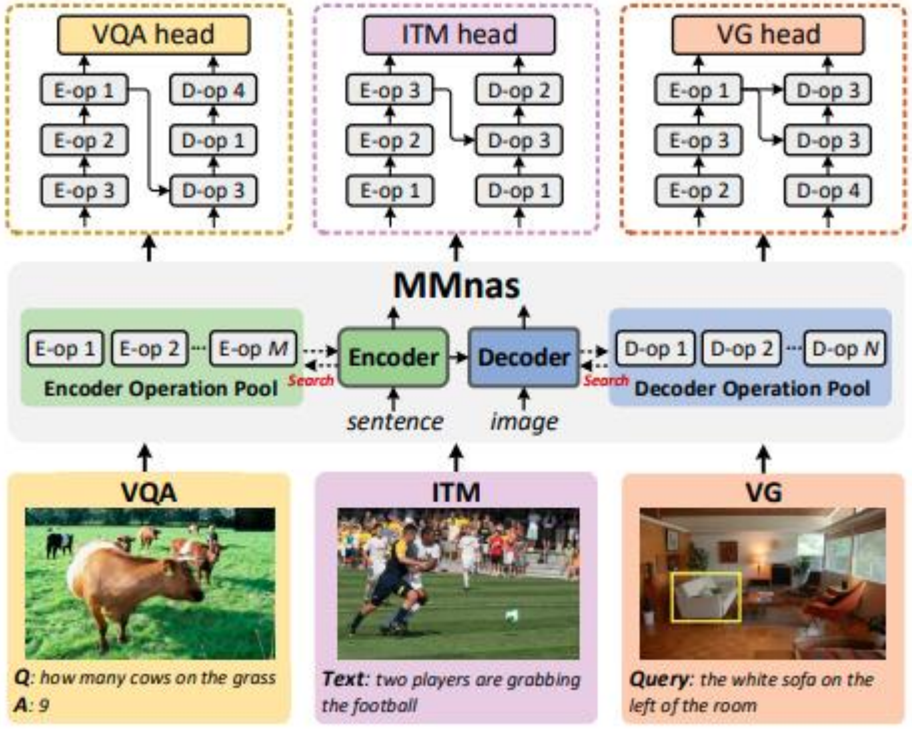
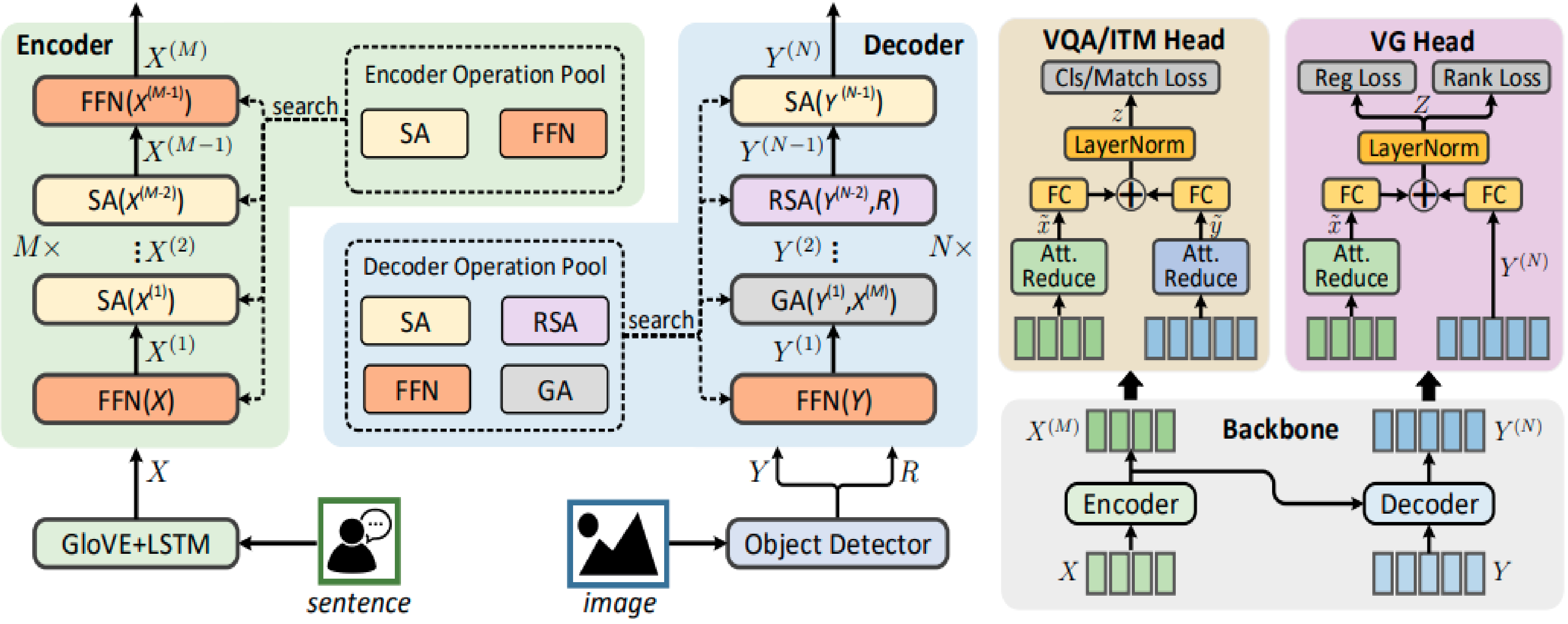


Figure 1: Schematic of the proposed generalized MMnas framework, which searches for the optimal architectures for the VQA, image-text matching, and visual grounding tasks.

Deep Multimodal Neural Architecture Search

- MMnas Framework



(a) Unified Encoder-Decoder Backbone

(b) Task-specific Heads

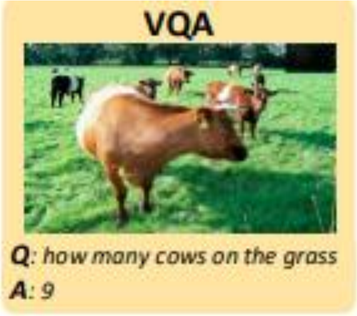
Figure 2: The flowchart of the MMnas framework, which consists of (a) unified encoder-decoder backbone and (b) task-specifi heads on top the backbone for visual question answer (VQA), image-text matching (ITM), and visual grounding (VG).

Deep Multimodal Neural Architecture Search

- Experiment Settings
  - Dataset

VQA-v2: Visual Question Answering {images, questions, answers}

200,000 images, 1,106,000 questions

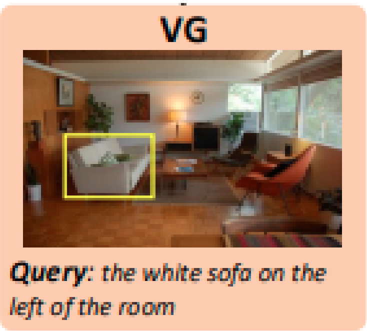


Flickr 30K*: image with caption from Flickr website {images, captions}*

31,000 images

RefCOCO, RefCOCO+ and RefCOCOg: *{images, query}*

66,600 images and queries



Deep Multimodal Neural Architecture Search

- Experiment Settings
  - MMnas search space module

MHA: Multi-head Attention

Self-attention operation

$$Z = \mathrm{SA}(X) = \mathrm{MHA}(X, X, X, \mathbf{0}) \quad (4)$$

Guided-attention operation

$$Z = \mathrm{GA}(X, Y) = \mathrm{MHA}(X, Y, Y, \mathbf{0}) \quad (5)$$

Feed-forward network operation

$$Z = \mathrm{FFN}(X) = \mathrm{FC}_d \circ \mathrm{Drop}_{0.1} \circ \mathrm{ReLU} \circ \mathrm{FC}_{4d}(X) \quad (6)$$

Relation self-attention operation

$$\begin{aligned} \mathrm{MLP}(R) &= \mathrm{ReLU} \circ \mathrm{FC}_1 \circ \mathrm{ReLU} \circ \mathrm{FC}_{d_h}(R) \\ Z = \mathrm{RSA}(X, R) &= \mathrm{MHA}(X, X, X, \log(\mathrm{MLP}(R) + \epsilon)) \end{aligned} \quad (7)$$

Deep Multimodal Neural Architecture Search

- MMnas Search Algorithm

$$(\theta^*, W^*) = \underset{\theta, W}{\operatorname{argmin}} \ \mathbb{E}_{a \sim \mathcal{A}(\theta)} [\mathcal{L}(\mathcal{N}(a, W))] \qquad (16)$$

θ : Architecture weights

W : Model weights

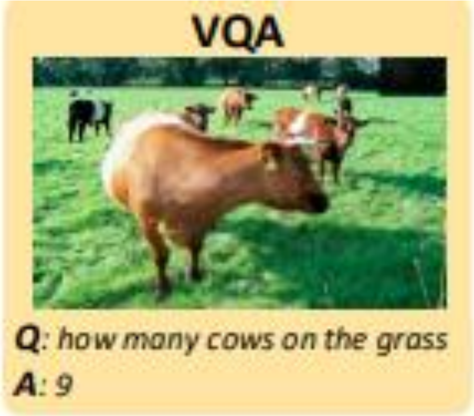
Algorithm 1: Search Algorithm for MMnasNet.

Input: A supernet parameterized by the architecture weights θ and the model weights W. Training set $\mathcal{D}_a$ and $\mathcal{D}_m$ are used to optimize θ and W, respectively. T_w and T_j denote the number of epochs for the warming-up and iterative optimization stages, respectively. u is a factor to balance the update frequencies of θ and W.
Output: The searched optimal architecture a^*
Random initialize θ and W;
 # The warming-up stage;
for $t = 1$ **to** T_w **do**
  Random sample an architecture $a \sim \mathcal{A}$;
  Random sample a mini-batch $d_m \subseteq \mathcal{D}_m$;
  Update W by descending $\nabla_W \mathcal{L}_{\text{train}}(\mathcal{N}(a, W))$ on d_m;
end
 # The iterative optimization stage;
for $t = 1$ **to** T_j **do**
  for $i = 1$ **to** u **do**
    Random sample a mini-batch $d_m \subseteq \mathcal{D}_m$;
    Sample an architecture $a \sim \mathcal{A}(\theta)$ with respect to θ;
    Update W by descending $\nabla_W \mathcal{L}_{\text{train}}(\mathcal{N}(a, W))$ on d_m;
  end
  Random sample a mini-batch $d_a \subseteq \mathcal{D}_a$;
  Sample an architecture $a \sim \mathcal{A}(\theta)$ with respect to θ;
  Update θ by descending $\nabla_\theta \mathcal{L}_{\text{train}}(\mathcal{N}(a, W))$ on d_a;
end
Return a^* by picking the operation with the largest value in θ^* for each block.

Deep Multimodal Neural Architecture Search

- Experiment Results : *Visual Question Answering Task*

Method	All	Yes/No	Num	Other
Bottom-Up	65.32	81.82	44.21	56.05
MFH+CoAtt	68.76	84.27	49.56	59.89
BAN-8	69.52	85.31	50.93	60.26
BAN-8 (+G+C)	70.04	85.42	54.04	60.52
DFAF-8	70.22	86.09	53.32	60.49
MCAN-6	70.63	86.82	53.26	60.72
MMnasNet (paper)	**71.24**	**87.27**	**55.68**	**61.05**
MMnasNet (me)	**67.53**	**84.93**	**51.74**	**58.47**



Number

Deep Multimodal Neural Architecture Search

- Experiment Results : *Visual Grounding Task*

Method	Object Detector			Validation Acc
	Dataset	Model	Backbone	
Spe.+Lis.+Rein.+MMI	COCO	SSD	VGG-16	69.50
MAttNet	COCO	MRCN	ResNet-101	76.7
DDPN	Genome	FRCN	ResNet-101	76.8
MMnasNet (paper)	**COCO**	**MRCN**	**ResNet-101**	**81.5**
MMnasNet (me)	**COCO**	**MRCN**	**ResNet-101**	**81.2**