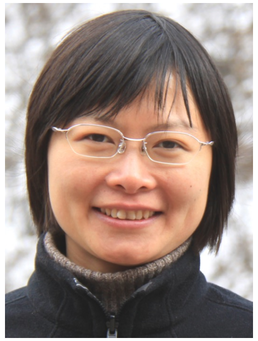
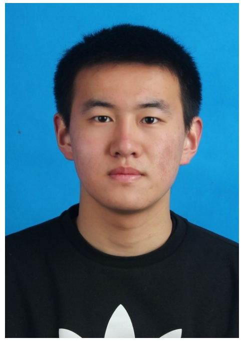
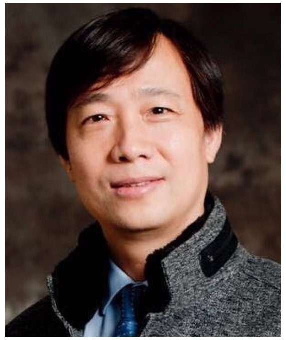
AGNAS: Attention-Guided Micro- and Macro-Architecture Search

Zihao Sun, Yu Hu, Shun Lu, Longxing Yang, Jillin Mei, Yinhe Han, Xiaowei Li

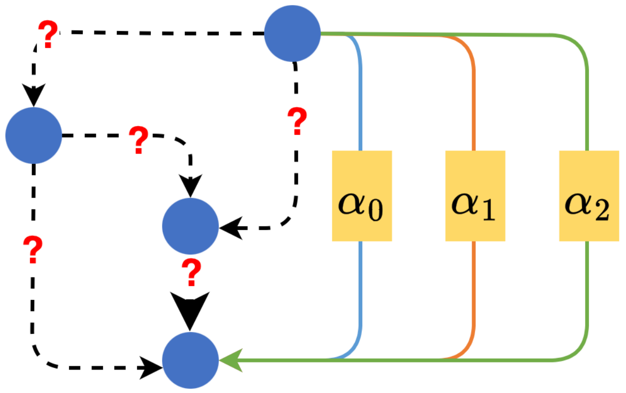








Motivation:

- Micro Search: Architecture parameters $\neq$ Operation strength [1]



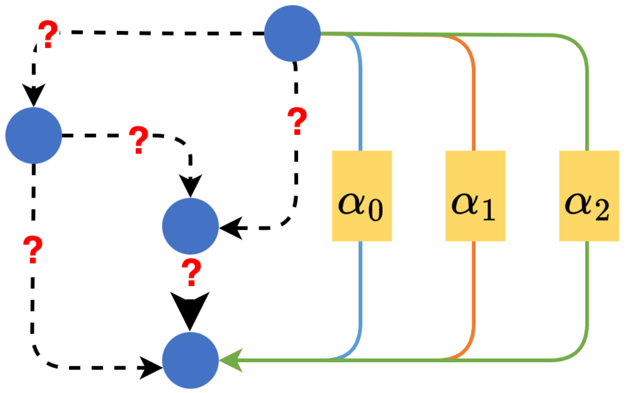
Micro Search

$$\min_{\alpha} \quad \mathcal{L}_{val}(w^*(\alpha), \alpha)$$
$$s.t. \quad w^*(\alpha) = \arg\min_w \mathcal{L}_{train}(w, \alpha)$$

[1] Wang R, Cheng M, Chen X, et al. Rethinking architecture selection in differentiable NAS. ICLR, 2021

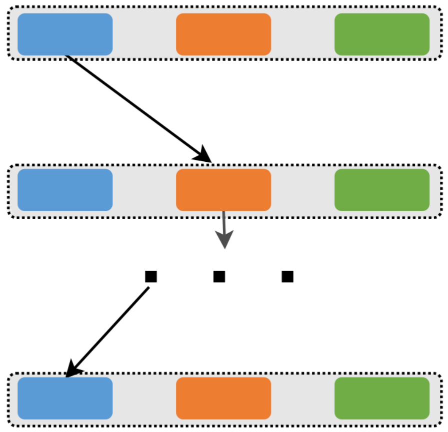
Motivation:

- Micro Search: Architecture parameters ≠ Operation strength [1]
- Macro Search: Path sampling evaluation ⟷ Time-consuming [2]



Micro Search

$$\min_{\alpha} \quad \mathcal{L}_{val}(w^*(\alpha), \alpha)$$
$$s.t. \quad w^*(\alpha) = \arg\min_w \mathcal{L}_{train}(w, \alpha)$$



Macro Search

$$w^*(a) = \arg\min_w \mathbb{E}_{a\sim\mathcal{A}} \mathcal{L}_{train}(w, a)$$
$$a^* = \arg\max_{a\sim\mathcal{A}} ACC_{val}(a, w^*(a))$$

[1] Wang R, Cheng M, Chen X, et al. Rethinking architecture selection in differentiable NAS. ICLR, 2021
[2] Guo Z, Zhang X, Mu H, et al. Single path one-shot neural architecture search with uniform sampling. ECCV, 2020

Motivation:

We ask for :

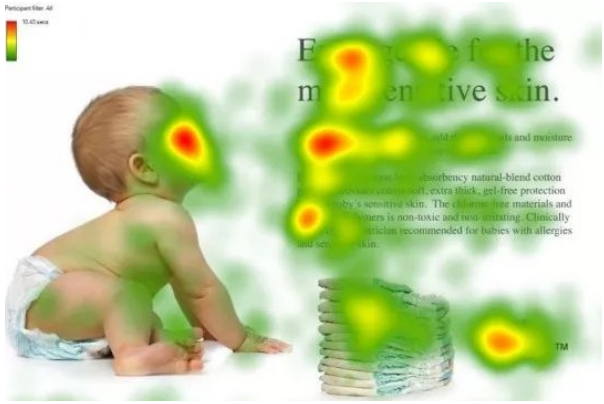
- Search paradigm of **accurate** and **end-to-end** ?

Motivation:

We ask for :

- Search paradigm of **accurate** and **end-to-end** ?

Inspired by :



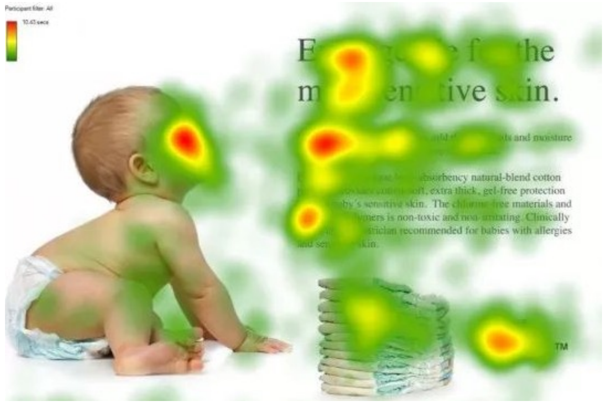
Attention

Motivation:

We ask for :

- Search paradigm of **accurate** and **end-to-end** ?

Inspired by :



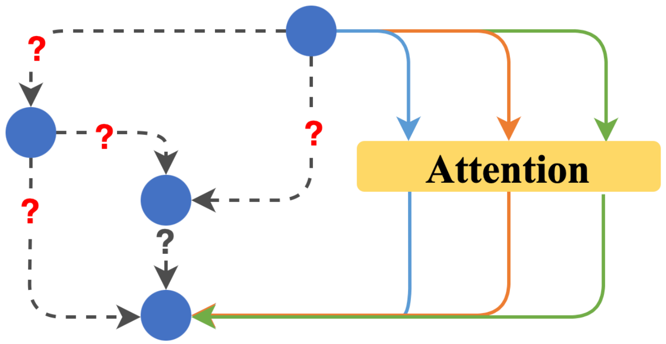
Attention + NAS = AGNAS

ICML
International Conference
On Machine Learning

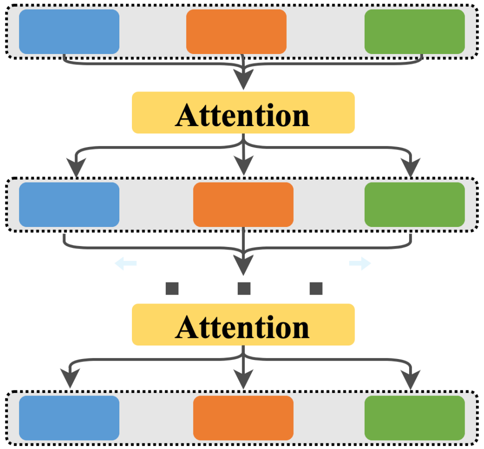
Motivation:

We ask for :

- Search paradigm of **accurate** and **end-to-end** ?



Micro Search

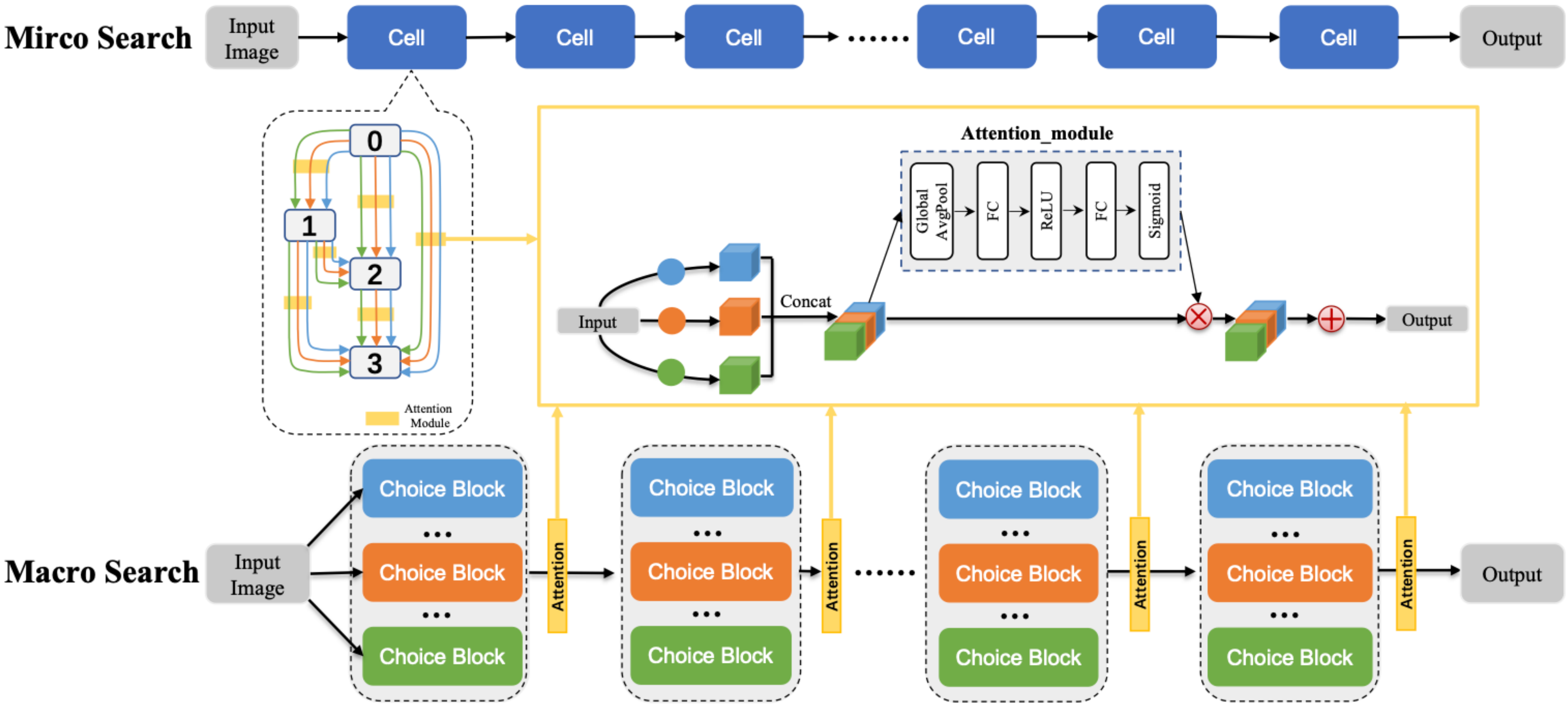


Macro Search

ANS@ICT

Method:

- Overview of AGNAS

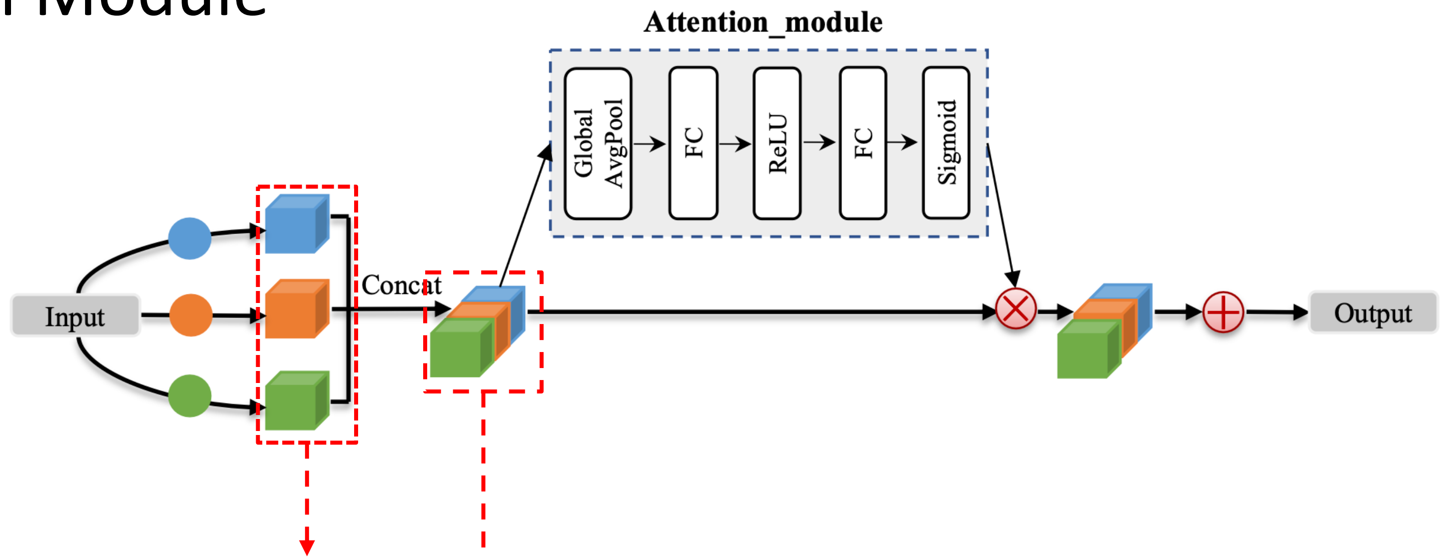


ANS@ICT

ICML International Conference On Machine Learning

Method:

- Attention Module



$F_k^o \in \mathbb{R}^{b\times c\times h\times w}$

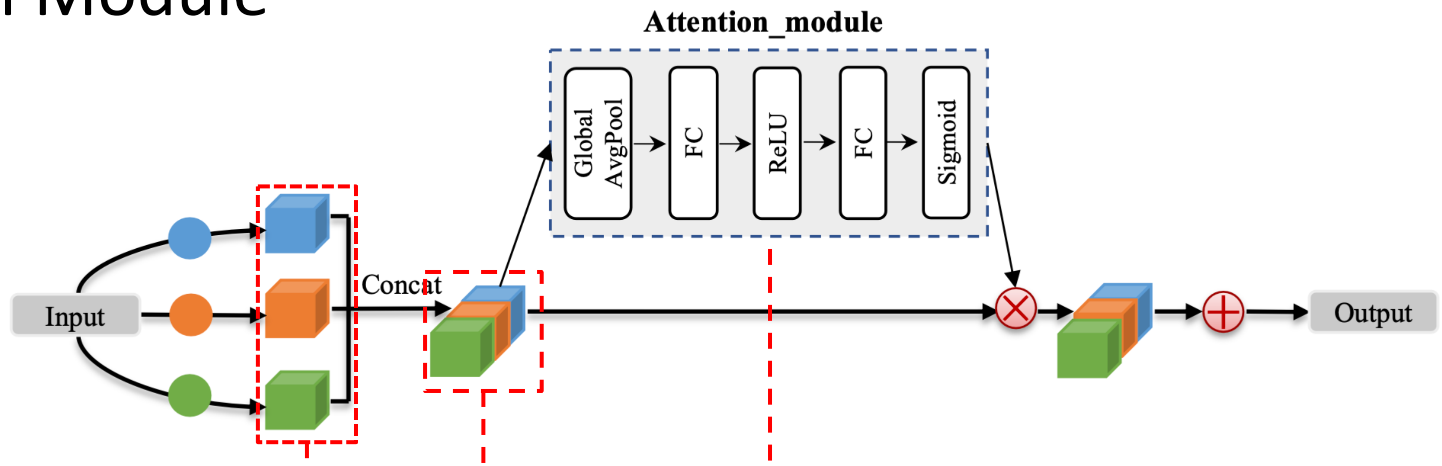
$F^{con} \in \mathbb{R}^{b\times (m*c)\times h\times w}$

Pre-Processing | Attention-Computing | Architecture-Evaluation

ANS@ICT

Method:

- Attention Module



$$F_k^o \in \mathbb{R}^{b\times c\times h\times w}$$

$$F^{con} \in \mathbb{R}^{b\times (m*c)\times h\times w}$$

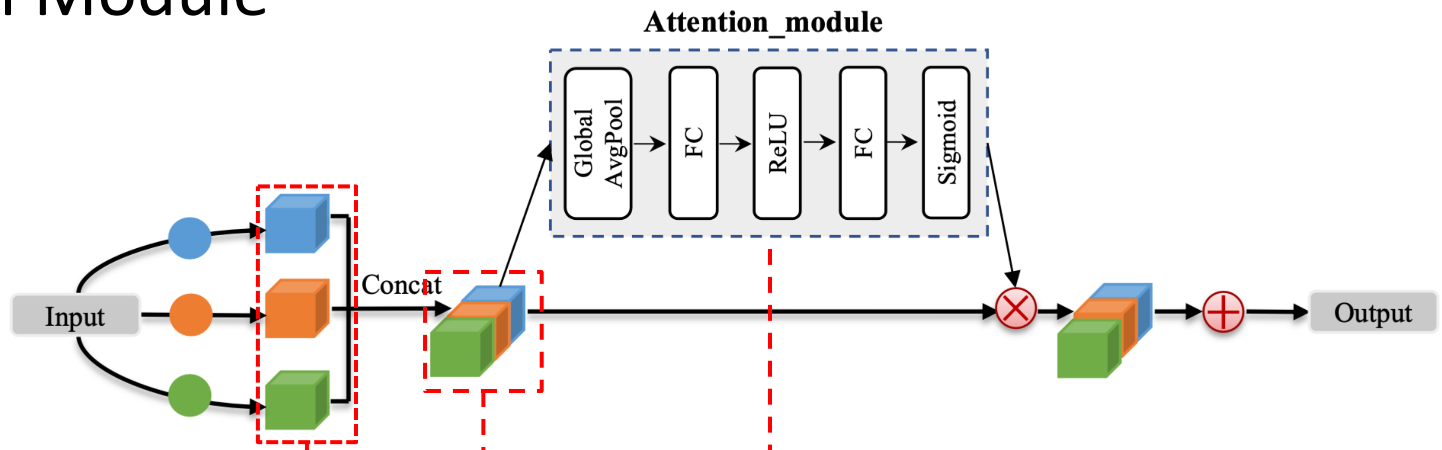
$$F_{gap}^{con} = \frac{1}{h\times w}\sum_{p=1}^{h}\sum_{q=1}^{w} F^{con}(p,q)$$

$$Atten = \sigma(W_2 Relu(W_1 F_{gap}^{con}))$$

Pre-Processing	Attention-Computing	Architecture-Evaluation

Method:

- Attention Module



$$F_k^o \in \mathbb{R}^{b\times c\times h\times w}$$

$$F^{con} \in \mathbb{R}^{b\times(m*c)\times h\times w}$$

$$F_{gap}^{con} = \frac{1}{h\times w}\sum_{p=1}^{h}\sum_{q=1}^{w}F^{con}(p,q)$$

$$Atten = \sigma(W_2 Relu(W_1 F_{gap}^{con}))$$

$$\begin{cases} A_1 &= \sum_{i=1}^{c} Atten(i) \\ A_2 &= \sum_{i=c}^{2c} Atten(i) \\ & \vdots \\ A_m &= \sum_{i=(m-1)c}^{mc} Atten(i) \end{cases}$$

$$A^* = argmax(A_1, A_2, \cdots, A_m)$$

Pre-Processing	Attention-Computing	Architecture-Evaluation

Theoretical Analysis:

F-Principle

- The Global Average Pooling (GAP) is the special case of Two-Dimensional Discrete Cosine Transform (2D-DCT)

$$f_{0,0}^{2d} = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} cos(\frac{0}{H}(i+\frac{1}{2}))cos(\frac{0}{W}(j+\frac{1}{2}))$$
$$= \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} = GAP(x^{2d})HW$$

- low-frequency components help to improve the generalization [1]

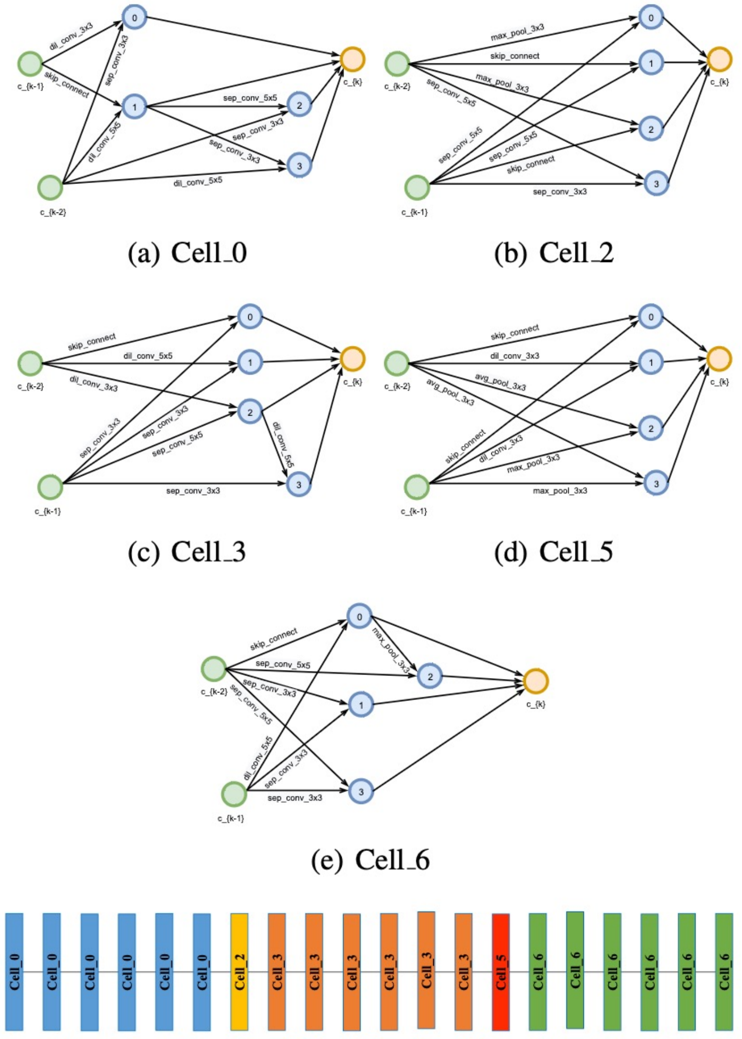
[1] Xu Z Q J, Zhang Y, Luo T, et al. Frequency principle: Fourier analysis sheds light on deep neural networks. ICML, 2019.

ICML
International Conference
On Machine Learning

Experimental Results:

- DARTS search space

Methods	Test Err.(%)	Params(M)	Search Cost (GPU-days)	Search Algorithm
NASNet-A (Zoph et al., 2018)	2.65	3.3	1800	RL
AmoebaNet-A (Real et al., 2019)	3.34±0.06	3.2	3150	EA
AmoebaNet-B (Real et al., 2019)	2.55±0.05	2.8	3150	EA
PNAS (Liu et al., 2018a)	3.41±0.09	3.2	225	SMBO
ENAS (Pham et al., 2018)	2.89	4.6	0.5	RL
DARTS (1st order) (Liu et al., 2018c)	3.00±0.14	3.3	1.5	Gradient
DARTS (2nd order) (Liu et al., 2018c)	2.76±0.09	3.3	4	Gradient
SNAS (Xie et al., 2018)	2.85±0.02	2.8	1.5	Gradient
GDAS (Dong & Yang, 2019)	2.93	3.4	0.21	Gradient
BayesNAS (Zhou et al., 2019)	2.81±0.04	3.4	0.2	Gradient
Robust-DARTS (Zela et al., 2020)	2.95±0.21	N/A	1.6	Gradient
PC-DARTS (Xu et al., 2019a)	2.57±0.07	3.6	0.1	Gradient
DATA (Chang et al., 2019)	2.59	3.4	1	Gradient
SGAS(Cri.1 avg.) (Li et al., 2020)	2.66±0.24	3.7	0.25	Gradient
SDARTS-ADV (Chen & Hsieh, 2020)	2.61±0.02	3.3	1.3	Gradient
DARTS+PT (Wang et al., 2021)	2.61±0.08	3.0	0.8	Gradient
AGNAS (avg.)	2.53±0.003	3.6	0.4	Gradient
AGNAS (best)	**2.46**	3.6	0.4	Gradient



(a) Cell_0 (b) Cell_2

(c) Cell_3 (d) Cell_5

(e) Cell_6

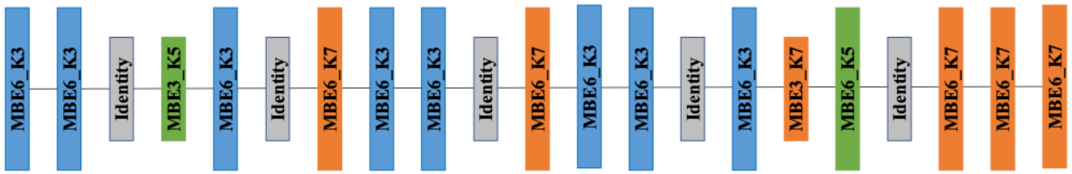
(f) Target network built on CIFAR-10

ANS@ICT

Experimental Results:

- ProxylessNAS search space

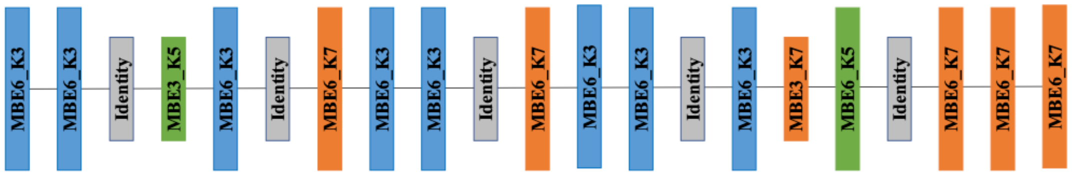
Methods	Test Err. (%)		Params	Search Cost	Search
	Top-1	Top-5	(M)	(GPU-days)	Algorithm
MnasNet (Tan et al., 2019)	26	8.2	4.2	2000	RL
NASNet (Zoph et al., 2018)	26.0	8.4	5.3	1800	RL
AmoebaNet (Real et al., 2019).	24.3	7.6	6.4	3150	EA
PNAS (Liu et al., 2018a)	25.8	8.1	5.1	225	SMBO
FBNet-C (Wu et al., 2019)	25.1	7.9	5.5	9	Gradient
ProxylessNAS(GPU) (Cai et al., 2018)	24.9	7.5	7.1	8.3	Gradient
SPOS (Guo et al., 2020)	26.0	8.4	5.3	11‡	Evolution
FairNAS-A (Chu et al., 2021)	24.66	7.8	4.6	16‡	Evolution
GreedyNAS-C (You et al., 2020)	23.8	7.5	4.7	8‡	Evolution
RLNAS (Zhang et al., 2021)	24.4	7.4	5.3	N/A	Evolution
AGNAS	**23.4**	**6.8**	6.7	3.3	Gradient



Experimental Results:

- ProxylessNAS search space

Methods	Test Err. (%) Top-1	Test Err. (%) Top-5	Params (M)	Search Cost (GPU-days)	Search Algorithm
MnasNet (Tan et al., 2019)	26	8.2	4.2	2000	RL
NASNet (Zoph et al., 2018)	26.0	8.4	5.3	1800	RL
AmoebaNet (Real et al., 2019).	24.3	7.6	6.4	3150	EA
PNAS (Liu et al., 2018a)	25.8	8.1	5.1	225	SMBO
FBNet-C (Wu et al., 2019)	25.1	7.9	5.5	9	Gradient
ProxylessNAS(GPU) (Cai et al., 2018)	24.9	7.5	7.1	8.3	Gradient
SPOS (Guo et al., 2020)	26.0	8.4	5.3	$11^{\ddagger}$	Evolution
FairNAS-A (Chu et al., 2021)	24.66	7.8	4.6	$16^{\ddagger}$	Evolution
GreedyNAS-C (You et al., 2020)	23.8	7.5	4.7	$8^{\ddagger}$	Evolution
RLNAS (Zhang et al., 2021)	24.4	7.4	5.3	N/A	Evolution
AGNAS	**23.4**	**6.8**	6.7	3.3	Gradient



- NAS-Bench-201 search space

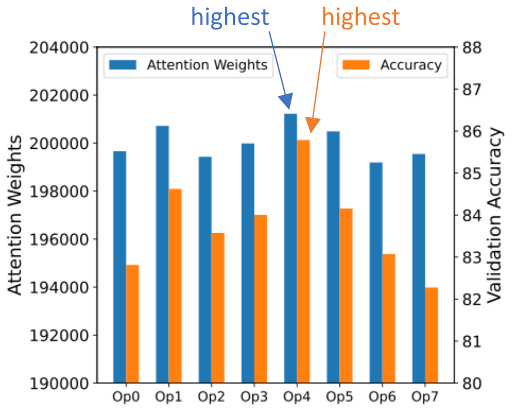
Methods	CIFAR-10 validation	CIFAR-10 test	CIFAR-100 validation	CIFAR-100 test	ImageNet-16-120 validation	ImageNet-16-120 test
Optimal	91.61	94.37	73.49	73.51	46.77	47.31
RSPS	80.42±3.58	84.07±3.61	52.12±5.55	52.31±5.77	27.22±3.24	26.28±3.09
DARTS	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00
GDAS	89.89±0.08	93.61±0.09	71.34±0.04	70.70±0.30	41.59±1.33	41.71±0.98
SETN	84.04±0.28	87.64±0.00	58.86±0.06	59.05±0.24	33.06±0.02	32.52±0.21
ENAS	37.51±3.19	53.89±0.58	13.37±2.35	13.96±2.33	15.06±1.95	14.84±2.10
AGNAS	**91.25±0.019**	**94.05±0.059**	**72.4±0.382**	**72.41±0.061**	**45.5±0.003**	**45.98±0.457**

ICML International Conference On Machine Learning

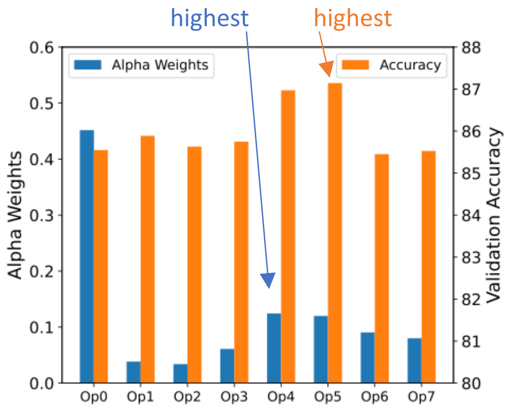
Discussion:

- Attention Weight vs. Architecture Parameter

Take the second edge in the first cell as example,



(a) AGNAS



(b) DARTS

Discretization Accuracy	Attention Weights	Operation	Alpha Weights	Discretization Accuracy
85.79	199663	none	0.452	87.14
84.62	200717	max_pool_3x3	0.038	86.97
84.16	199431	avg_pool_3x3	0.034	85.89
84.00	199984	skip_connect	0.061	85.75
83.58	201223	sep_conv_3x3	0.124	85.63
83.07	200489	sep_conv_5x5	0.120	85.55
82.81	199193	dil_conv_3x3	0.091	85.53
82.28	199547	dil_conv_5x5	0.080	85.45
AGNAS Kendall = 0.71			DARTS Kendall = 0	

(c) The ranking of attention weights or alpha weights against discretization accuracy.

- Attention weights **>** Architectural parameters
- Kendall τ : AGNAS **>** DARTS

ANS@ICT

Conclusion:

- We propose a novel paradigm that leverages the attention mechanism to guide micro- and macro-architecture search.
- AGNAS can truly reflect the operational importance and conduct end-to-end search.
- AGNAS significantly outperforms state-of-the-art approaches on various search space.

Thanks for your attention

Q&A

AGNAS: Attention-Guided Micro- and Macro-Architecture Search