ICML 2025

Revisiting Neural Networks for Few-Shot Learning: A Zero-Cost NAS Perspective

Speaker: Haidong Kang

Authors: Haidong Kang

College of Software, Northeastern University

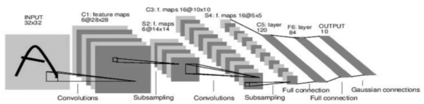
Outline

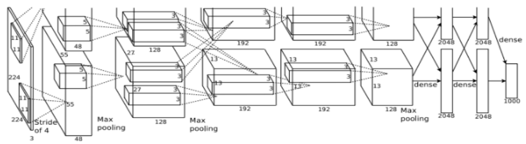
- Introduction
  - Few-shot Learning
  - Neural Architecture Search
- Goal and Motivation
- The IBFS Algorithm
- Experimental Results
- Conclusions

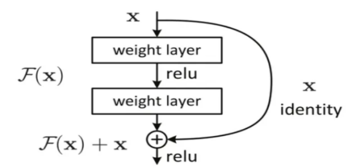
Outline

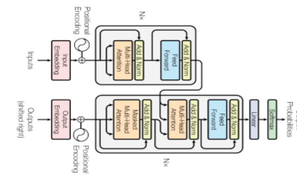
- Introduction
  - Neural Architecture Search (NAS)
  - Few-shot Learning (FSL)
- Goal and Motivation
- The IBFS Algorithm
- Experimental Results
- Conclusions

Introduction: NAS



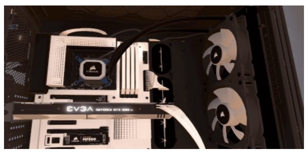


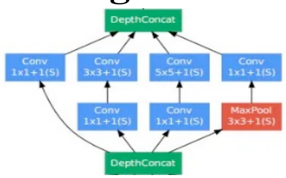


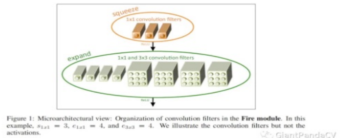


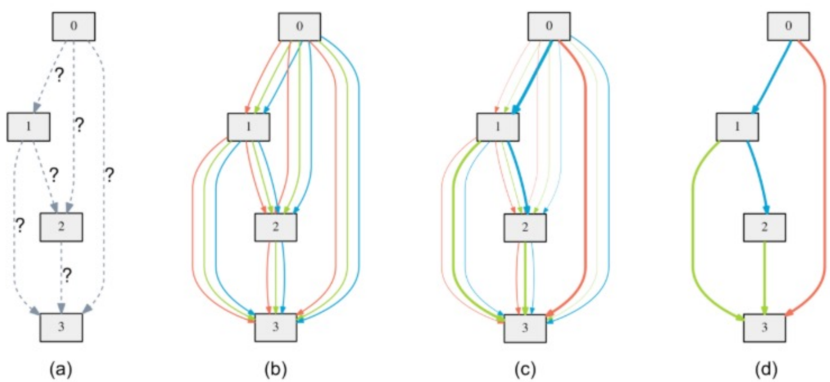
LeNet-5 | 2010 | AlexNet | 2013 | 2014 | ResNet | 2016 | Transformer

1988 | Dan Ciresan Net | 2012 | ✓ZF Net ✓Overfeat ✓NIN ✓R-CNN | GoogLeNet | 2015 | SqueezeNet | 2017









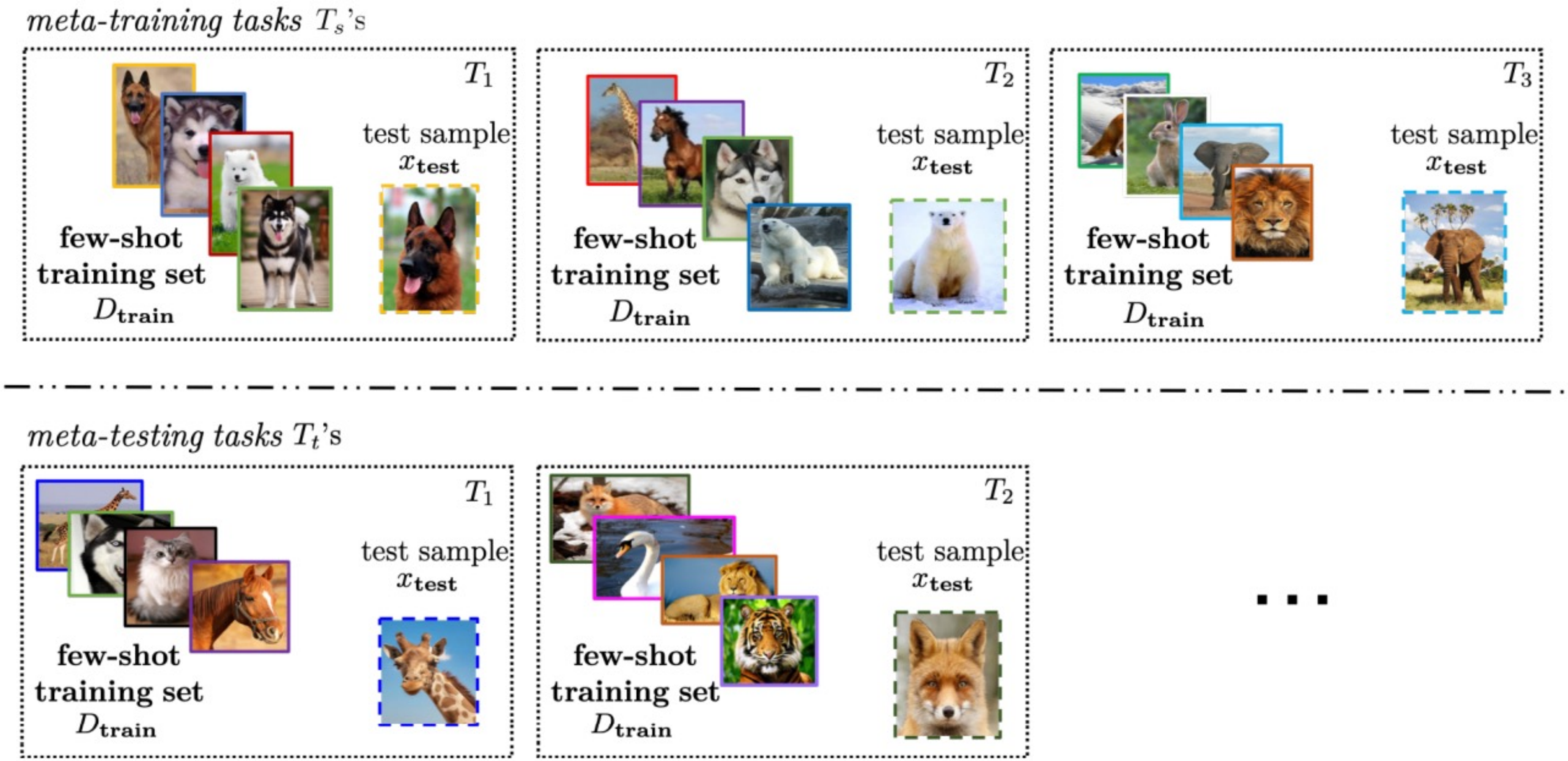
DARTS

$$\bar{o}^{(i,j)}(x) = \sum_{o \in \mathcal{O}} \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o' \in \mathcal{O}} \exp(\alpha_{o'}^{(i,j)})} o(x)$$

$$\min_{\alpha} \quad \mathcal{L}_{val}(w^*(\alpha), \alpha)$$
$$\text{s.t.} \quad w^*(\alpha) = \text{argmin}_w \; \mathcal{L}_{train}(w, \alpha)$$

★ NAS

Introduction: FSL



Solving the FSL problem by meta-learning

Outline

- Introduction
  - Neural Architecture Search (NAS)
  - Few-shot Learning (FSL)
- Goal and Motivation
- The IBFS Algorithm
- Experimental Results
- Conclusions

The Goal

Designing the best neural architectures **without involving any training** :

for $\left\{ \begin{array}{l} \text{Image classification} \\ \text{FSL} \end{array} \right.$

Answering two questions:

- Which properties of MAML impact the global convergence of FSL?
- Can we find a simple proxy for FSL?

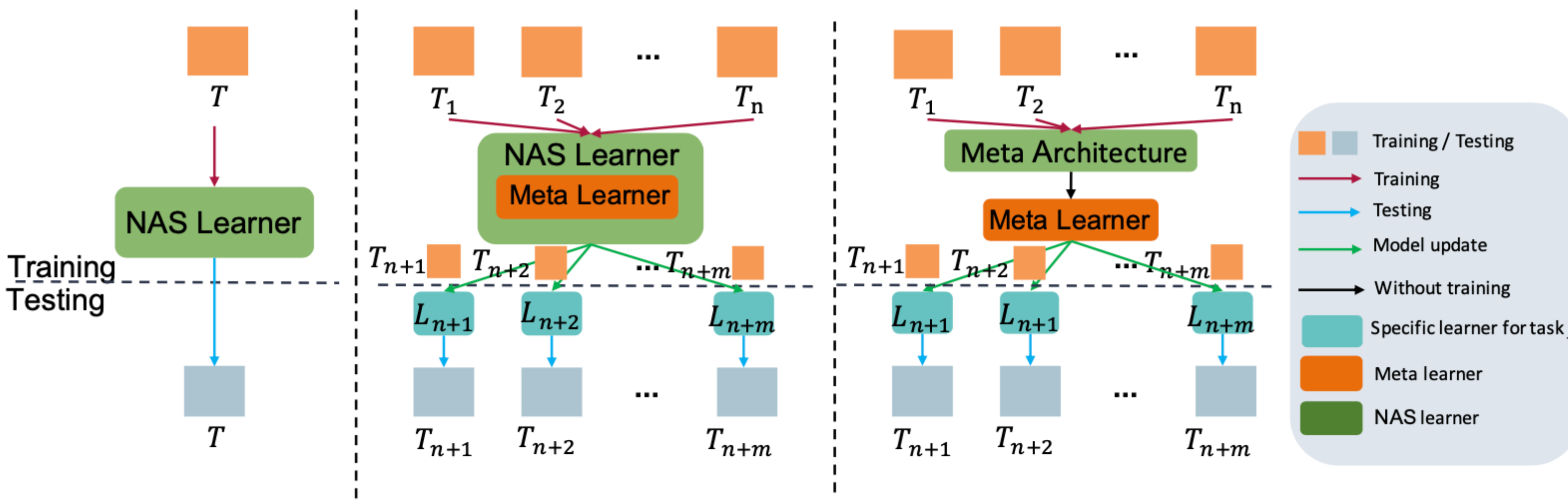
The Problems

Problems in NAS and FSL:

- Previous NAS works only target a pre-defined task.
- For a new task in few-shot learning (FSL) scenarios, the architecture is either searched from scratch, which is neither efficient nor flexible, or borrowed architecture from the ones obtained on other tasks, which may lead to sub-optimal.

Motivation

Illustration of our IBFS and related approaches. (a) Vanilla neural architecture search. (b) Adaption neural architecture search. (c) The proposed IBFS can find the best meta architecture without training for multiple unseen tasks.



(a) Neural Architecture Search (Vanilla)

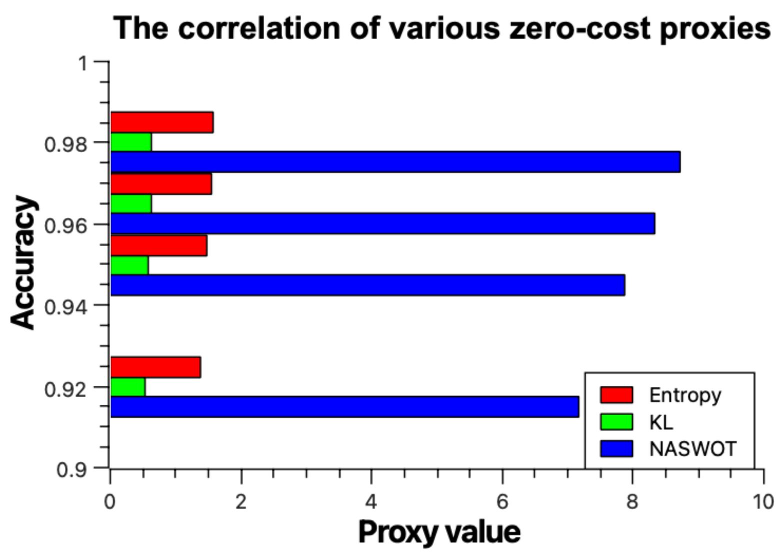
(b) Adaption Neural Architecture Search (meta-learning)

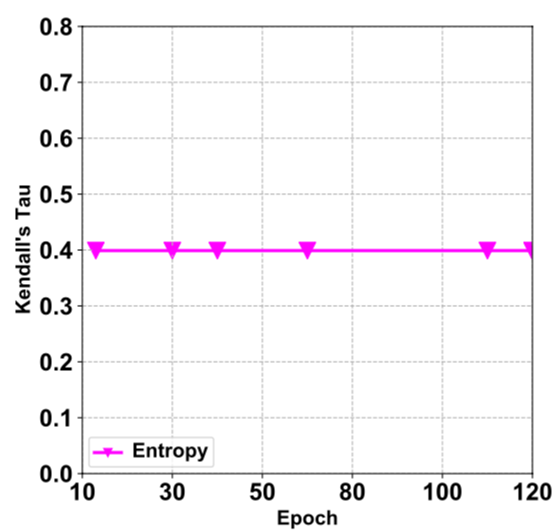
(c) Adaption Neural Architecture Search (Ours)

Motivation

Observations.

1. Existing proxies (i.e., NASWOT) suffer from larger score variance, which will degrade the accuracy.
2. Kendall’s Tau between accuracy and information entropy remains unchanged.





Outline

- Introduction
  - Neural Architecture Search (NAS)
  - Few-shot Learning (FSL)
- Goal and Motivation
- The IBFS Algorithm
- Experimental Results
- Conclusions

Design of FSL-Friendly Architectures

- **Global Convergence of MAML.**

To clearly illustrate the global convergence of MAML, we borrow one of the conclusions from MetaNTK-NAS, which provides a Neural Tangent Kernel (NTK) perspective to understand MAML. While MetaNTK-NAS provides valuable theoretical guarantees, its proof is highly intricate and relies heavily on NTK theory, which can be computationally expensive. In contrast, this paper presents a novel and more accessible proof for the global convergence of MAML.

$$\ell(W^t) \leq \left(1 - \tau \cdot \eta_0 \sigma_{min}(\Phi)\right)^{2t}, \quad \tau \in (0,1)$$

Design of FSL-Friendly Architectures

- **IBFS.**

$$I(R;Y) = \sum_y \sum_r p(y,r) \log \frac{p(y,r)}{p(y)p(r)} \leq I(X;Y),$$

$$\frac{\delta \mathcal{L}}{\delta p(r|x)} =$$

$$p(x) \left[\log \frac{p(r|x)}{p(r)} + \beta \sum_y p(y|x) \log \frac{p(y|x)}{p(y|r)} - \tilde{\lambda}(x) \right] = 0,$$

$$NN_{expressivity} = -\sum_{k=1}^{N} p \log p$$

$$\leq H * exp(-\mathbb{E}_{p(\boldsymbol{\varsigma_k})}[-\log p(\boldsymbol{\varsigma_k})]).$$

Outline

- Introduction
  - Neural Architecture Search (NAS)
  - Few-shot Learning (FSL)
- Goal and Motivation
- The IBFS Algorithm
- Experimental Results
- Conclusions

Experiments: FSL

- **Mini-ImageNet**
  - 60,000 RGB images of 84x84 pixels
  - 100 classes
  - Split 64:16:20
- **Tiered-ImageNet**
  - 779,165 RGB images of 84x84 pixels
  - 608 classes
  - Split 351:97:160

Experiments: NAS

- **NAS-Bench-201**
  - CIFAR-10
  - CIFAR-100
  - ImageNet-16-120
- **Larger Dataset**
  - ImageNet1K
- **Transformer Design**
  - ImageNet1K
  - TF-TAS-T

Results: NAS-Bench-201

Table 1. Comparison results on NAS-Bench-201. Red, blue, and orange indicate the best, second-best, and third-best results, respectively.

Method	Year	Search (s)	CIFAR-10		CIFAR-100		ImageNet-16-120		
			validation (%)	test (%)	validation (%)	test (%)	validation (%)	test (%)	Search Methods
ResNet(He et al., 2016)	CVPR2016		93.97			70.86		43.63	Manual
Non-weight sharing									
REA(Zoph et al., 2018)	CVPR2018	12000	91.19±0.31	93.92±0.30	71.81±1.12	71.84±0.99	45.15±0.89	45.54±1.03	EA
BOHB(Liu et al., 2018)	ECCV2018	12000	90.82±0.53	93.61±0.52	70.74±1.29	70.85±1.28	44.26±1.36	44.42±1.49	HPO
REINFORCE(Real et al., 2019b)	AAAI2019	12000	91.09±0.37	93.85±0.37	71.61±1.12	71.71±1.09	45.05±1.02	45.24±1.18	RL
Weight sharing									
SNAS (Xie et al., 2020)	ICLR2018	-	90.10±1.04	92.77±0.83	69.69±2.39	69.34±1.98	42.84±1.79	43.16±2.64	GD
ENAS(Pham et al., 2018)	ICML2018	13315	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00	RL
DARTS-V2(Liu et al., 2019b)	ICLR2019	29902	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00	GD
GDAS(Dong & Yang, 2019)	CVPR2019	28926	90.00±0.21	93.51±0.13	71.14±0.27	70.61±0.26	41.70±1.26	41.84±0.90	GD
DSNAS (Hu et al., 2020)	ICLR2019	-	89.66±0.29	93.08±0.13	30.87±16.40	31.01±16.38	40.61±0.09	41.07±0.09	GD
PC-DARTS (Xu et al., 2019)	ICLR2020	-	89.96±0.15	93.41±0.30	67.12±0.39	67.48±0.89	40.83±0.08	41.31±0.22	GD
RSPS(Li & Talwalkar, 2020)	UAI2020	7587	84.16±1.69	87.66±1.69	59.00±4.60	58.33±4.34	31.56±3.28	31.14±3.88	RS+WS
iDARTS (Zhang et al., 2021a)	ICML2021	-	89.86±0.60	93.58±0.32	70.57±0.24	70.83±0.48	40.38±0.59	40.89±0.68	GD
OLES (Jiang et al., 2023)	NeurIPS2023	-	90.88±0.10	93.70±0.15	70.56±0.28	70.40±0.22	44.17±0.49	43.97±0.38	GD
IS-DARTS (He et al., 2024)	AAAI2024	7200	91.55±0.00	94.36±0.00	73.49±0.00	73.31±0.00	46.37±0.00	46.34±0.00	GD
Training-free									
Random		-	83.20 ± 13.28	86.61 ± 13.46	60.70 ± 12.55	60.83 ± 12.58	33.34 ± 9.39	33.13 ± 9.66	Random
NASWOT (Mellor et al., 2021)	ICML2021	4.4	88.47 ± 1.33	91.53 ± 1.62	66.49 ± 3.08	66.63 ± 3.14	38.33 ± 4.98	38.33 ± 5.22	Zero-cost
GradSign (Zhang & Jia, 2022)	ICLR2022	30.38	-	93.52 ± 0.19	-	70.57 ± 0.31	-	41.89 ± 0.69	Zero-cost
ZiCo (Li et al., 2023)	ICLR2023	6.2	-	93.50 ± 0.18	-	70.62 ± 0.26	-	42.04 ± 0.82	zero-cost
AZ-NAS (Lee & Ham, 2024)	CVPR2024	0.71	-	93.53 ± 0.15 0.723	-	70.75 ± 0.48	-	45.43 ± 0.29	zero-cost
SWAP (Peng et al., 2024)	ICLR2024	4.7	87.31 ± 2.36	90.48 ± 0.94	65.92 ± 4.32	67.13 ± 1.83	33.85 ± 4.98	35.40 ± 3.96	Zero-cost
IBFS(ours)		3.82	91.55 ± 0.76	94.37 ± 0.34	73.31 ± 2.12	73.09 ± 2.08	45.59 ± 0.32	46.33 ± 1.2	Zero-cost
Optimal (NAS-Bench-201)		N/A		94.37		73.51		47.31	N/A

Results: Larger ImageNet1k

Table 2. Performance comparison on the DARTS search space with ImageNet1k dataset, where "Img" denotes network models directly searched in ImageNet1k, and "C10" and "C100" denotes network models searched in CIFAR-10 and CIFAR-100, respectively.

Method	Top-1(%)	Top-5(%)	# Params (M)	FLOPS (M)	Search Cost (GPU-days)	Search Method
ResNet50 (He et al., 2016)	75.3	92.2	25.6	4100	-	Manual
MobileNetV1 (Howard et al., 2017)	70.6	89.5	4.2	575	-	Manual
MobileNetV2 (Sandler et al., 2018)	74.7	91.0	3.4	300	-	Manual
ShuffleNetV2 (Ma et al., 2018)	72.6	-	3.5	299	-	Manual
AmoebaNet-A (Zoph et al., 2018)	74.5	92.4	6.4	555	3150	Evolution
ProxylessNAS-RL (Cai et al., 2018)	74.6	92.3	5.8	465	8.3	RL
EfficientNet-B0 (Tan & Le, 2019)	76.3	93.2	5.3	390	≈3000	RL
NASNet-A (Real et al., 2019b)	74.0	91.6	5.3	564	2000	RL
DARTS (Liu et al., 2019b)	73.3	91.3	4.7	574	4	Gradient
FBNet (Wu et al., 2019)	74.9	-	5.5	375	216	Gradient
PC-DARTS(Img) (Xu et al., 2019)	75.8	92.7	5.3	597	3.7	Gradient
P-DARTS(C100) (Chen et al., 2019)	75.3	92.5	5.1	577	0.3	Gradient
DARTS+ (Liang et al., 2019)	76.3	92.8	5.1	591	0.2	Gradient
DARTS-(Img) (Chu et al., 2020a)	76.2	93.0	4.9	467	4.5	Gradient
FairDARTS-B(Img) (Chu et al., 2020b)	75.1	92.5	4.8	541	-	Gradient
SNAS(C10) (Xie et al., 2020)	72.7	90.8	4.3	522	1.5	Gradient
DARTS+PT(C10) (Wang et al., 2020)	74.5	92.0	4.6	-	0.8	Gradient
DOTS(C10) (Gu et al., 2021)	75.7	92.6	5.2	581	0.3	Gradient
β-DARTS(C100) (Ye et al., 2022)	75.8	92.9	5.4	597	0.4	Gradient
Λ-DARTS (Movahedi et al., 2023)	75.7	-	5.2	-	-	Gradient
OLES (Jiang et al., 2023)	75.5	92.6	4.7	-	0.4	Gradient
FP-DARTS(C10) (Wang et al., 2023)	75.7	92.7	5.4	-	0.08	Gradient
PDARTS-AER^{ad} (Jing et al., 2023)	76.0	92.8	5.1	578	2.0	Gradient
IS-DARTS (He et al., 2024)	75.9	92.9	6.4	-	0.42	Gradient
NAO (Luo et al., 2018)	74.3	91.8	11.4	584	200	Proxy
SemiNAS (Luo et al., 2020)	76.5	93.2	6.3	599	4	Proxy
WeakNAS (Wu et al., 2021)	76.5	93.2	5.5	591	2.5	Proxy
TENAS (Chen et al., 2021c)	75.5	92.5	5.4	-	0.17	Training-free
NASI-ADA(C10) (Shu et al., 2022)	75.0	92.2	4.9	559	0.01	Training-free
SWAP (Shu et al., 2022) (Img)	75.0	92.4	5.8	-	0.006	Training-free
IBFS(C10)	76.7 ↑(0.2)	93.5 ↑(0.3)	5.2	587	0.0042 ↓(0.0022)	training-free

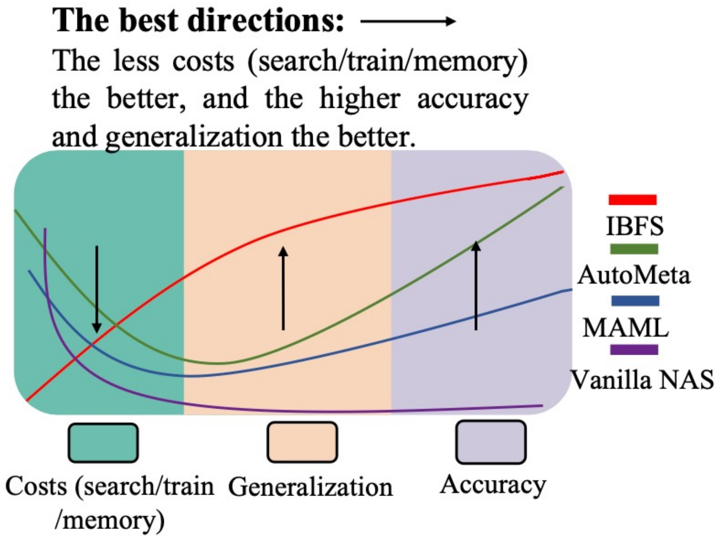
Results: FSL

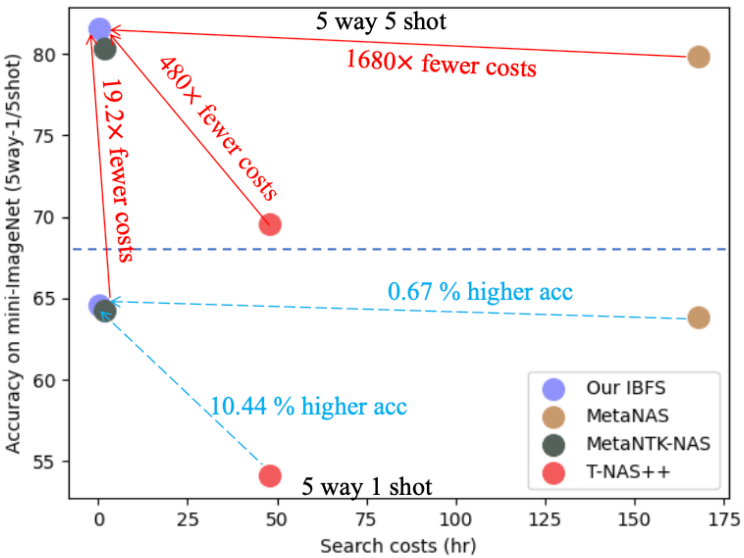
- SOTA accuracy
- Lightweight architecture
- Lowest search costs

Table 3. Comparison results on FSL benchmarks. Red, blue, and orange indicate the best, second-best, and third-best results, respectively.

					mini-ImageNet 5-way				tiered-ImageNet 5-way			
Method	**Year**	**Arch.**	**#Cells**	**Train**	**#Param.**	**Search Cost**	**1-shot (%)**	**5-shot (%)**	**#Param.**	**Search Cost**	**1-shot (%)**	**5-shot (%)**
MAML(Finn et al., 2017)	ICML17	Conv4	-	MAML	30k	-	48.70±1.84	63.11±0.92	30k	-	51.67±1.81	70.30±1.75
ANIL (Raghu et al., 2020)	ICLR20	Conv4	-	ANIL	30k	-	48.0±0.7	62.2±0.5	-	-	-	-
COMLN(Deleu et al., 2022)	LCLR22	Conv4	-	-	-	-	53.01±0.62	70.54±0.54	-	-	54.30±0.69	71.35±0.57
Meta-AdaM(Sun & Gao, 2024)	NeurIPS23	Conv4	-	Meta Learning	-	-	52.00±0.49	70.70±0.49	-	-	53.93±0.49	72.66±0.49
GAP(Kang et al., 2023)	CVPR23	Conv4	-	-	-	-	54.86±0.85	71.55 ±0.61	-	-	57.60±0.93	74.90±0.68
MetaDiff(Zhang et al., 2024)	AAAI24	Conv4	-	Meta Learning	-	-	55.06±0.81	73.18 ±0.64	-	-	57.77±0.90	75.46±0.69
MetaOptNet(Lee et al., 2019)	CVPR19	ResNet-12	-	MetaOptNet	12.5M	-	62.64±0.61	78.63±0.46	12.7M	-	65.99±0.72	81.56±0.53
CTM(Li et al., 2019)	CVPR19	ResNet-18	-	-	-	-	64.12±0.82	80.51±0.13	-	-	68.41±0.39	84.28±1.73
RFS(Tian et al., 2020)	ECCV20	ResNet-12	-	RFS	12.5M	-	62.02±0.63	79.64±0.44	12.7M	-	69.74±0.72	84.41±0.55
MAML+ALFA(Baik et al., 2020)	NeurIPS20	ResNet-12	-	MAML	-	-	59.74±0.49	77.96±0.47	-	-	64.62±0.49	82.48±0.39
Sparse-MAML(Von Oswald et al., 2021)	NeurIPS21	ResNet-12	-	MAML	-	-	56.39±0.38	73.01±0.24	-	-	53.47±0.53	68.83±0.65
MeTAL(Baik et al., 2021)	ICCV21	ResNet-12	-	Meta Learning	-	-	59.64±0.38	76.20±0.19	-	-	63.89±0.48	80.14±0.40
ClassifierBaseline(Chen et al., 2021d)	ICCV21	ResNet-12	-	Meta Learning	-	-	61.22±0.84	78.72±0.60	-	-	69.71±0.88	83.87 ±0.64
MetaQDA(Zhang et al., 2021b)	ICCV21	ResNet-12	-	Bayesian	-	-	65.12 ±0.66	80.98±0.75	-	-	69.97 ±0.52	85.51 ±0.58
MAML+SiMT(Tack et al., 2022)	NeurIPS22	ResNet-12	-	MAML	-	-	51.49±0.18	68.74±0.12	-	-	52.51±0.21	69.58±0.11
COMLN(Deleu et al., 2022)	LCLR22	ResNet-12	-	-	-	-	59.26±0.65	77.26±0.49	-	-	62.93±0.71	81.13 ±0.53
Meta-AdaM(Sun & Gao, 2024)	NeurIPS23	ResNet-12	-	Meta Learning	-	-	59.89±0.49	77.92±0.43	-	-	65.31±0.48	85.24±0.35
MetaDiff(Zhang et al., 2024)	AAAI24	ResNet-12	-	Meta Learning	-	-	64.99±0.77	81.21 ±0.56	-	-	72.33±0.92	86.31±0.62
AutoMeta(Kim et al., 2018)	NeurIPS18	Cells	-	Reptile	100k	2688 hr	57.6±0.2	74.7±0.2	-	-	-	-
T-NAS++(Lian et al., 2020)	ICLR20	Cells	2	FOMAML	27k	48 hr	54.11±1.35	69.59±0.85	-	-	-	-
MetaNAS(Elsken et al., 2020)	CVPR20	Cells	5	Reptile	1.1M	168 hr	63.1±0.3	79.5±0.2	-	-	-	-
MetaNAS(Elsken et al., 2020)(retrained)†	CVPR20	Cells	8	RFS	3.53M	168 hr	63.88±0.23	79.88±0.14	3.70M	168 hr	72.32±0.02	86.48±0.06
H-Meta-NAS(Zhao et al., 2022)	NeurIPS22	-	-	-	70.28K	-	57.36±1.11	77.53±0.77	-	-	-	-
MetaNTK-NAS(Wang et al., 2022a)	CVPR22	Cells	8	RFS	3.21M	1.92 hr	64.26±0.14	80.35±0.12	4.78M	2.73 hr	72.37±0.79	86.43±0.52
IBFS	-	Cells	8	RFS	3.50M	0.1 hr	64.55±0.02	81.52±0.08	4.50M	0.10 hr	72.56±0.02	86.73±0.08

Results: FSL





IBFS *vs.* Peer competitors in terms of costs, accuracy, and generalization.

Comparison to SOTAs in terms of accuracy of FSL and Search costs.

Results: Transformer Design

To further scrutinize the effectiveness of our method for transformer designing, conducting additional experiments on AutoFormer in a larger ImageNet dataset. Those empirical results show the strong generalizability of our method for transformer design.

Table 4. Results on AutoFormer benchmark in ImageNet.

Method	Year	Top-1 (%)	Search Cost	Model Type	Search Method
ViT-Ti(Dosovitskiy et al., 2020)	ICLR2020	74.5	-	Transformer	Manual
AutoFormer-T (Chen et al., 2021a)	CVPR2021	74.9	24	Transformer	Evolution
TF-TAS-T (Zhou et al., 2022)	CVPR2022	75.3	0.5	Transformer	Training-free
ViTAS-C (Su et al., 2022)	ECCV2022	74.7	32	Transformer	Evolution
Auto-Prox (Wei et al., 2024)	AAAI2024	75.6	0.1	Transformer	Training-free
IBFS	-	76.5	0.03	CNNs	Training-free

Outline

- Introduction
  - Neural Architecture Search (NAS)
  - Few-shot Learning (FSL)
- Goal and Motivation
- The IBFS Algorithm
- Experimental Results
- Conclusions

Why IBFS Works Well?

- Measuring first-order loss landscape of MAML.
  - Theory: On exact computation with an infinitely wide neural net [Arora, NeurlPS'19].
- How to design training-free method for FSL?
  - Good Theory (Information Bottleneck).
  - Good findings (larger score variance).
- Make it fast: training-free and acceleration.
  - Without any training.

Thank you!

Questions please?