Analyzing and Mitigating Interference in Neural Architecture Search

Jin Xu[1], Xu Tan[2], Kaitao Song[2], Renqian Luo[2], Yichong Leng[3], Tao Qin[2], Tie-Yan Liu[2], Jian Li[1]

[1]Tsinghua University
[2]Microsoft Asia, [3]University of Science and Technology of China

Introduction

- Neural Architecture Search (NAS) has achieved state-of-the-art results on many domains
- **Weight sharing**
  - Re-use the weights of shared operators from previously trained child models
  - Reduce the cost of neural architecture search
- However, rank correlation is low due to the **interference** among different child models
  - The shared operators receive different gradient directions from child models with different architecture

Interference In Weight Sharing and Related Work

- Interference: Gradient interference on shared operators
- **Analyzing** Interference
  - Notice the interference issue *(Berder et al., 2018; Guo et al., 2020; Lanbe & Zell 2021; Xie et al., 2020)*
  - Sampling child models cause high variance of the rank *(Zhang et al. 2020a)*
- **Mitigating** Interference
  - Shrink search space *(Zhang et al., 2020bl Hu et al., 2020; Xu et al., 2021)*
  - Remove affine in batchnorm *(Ning et al., 2021)*

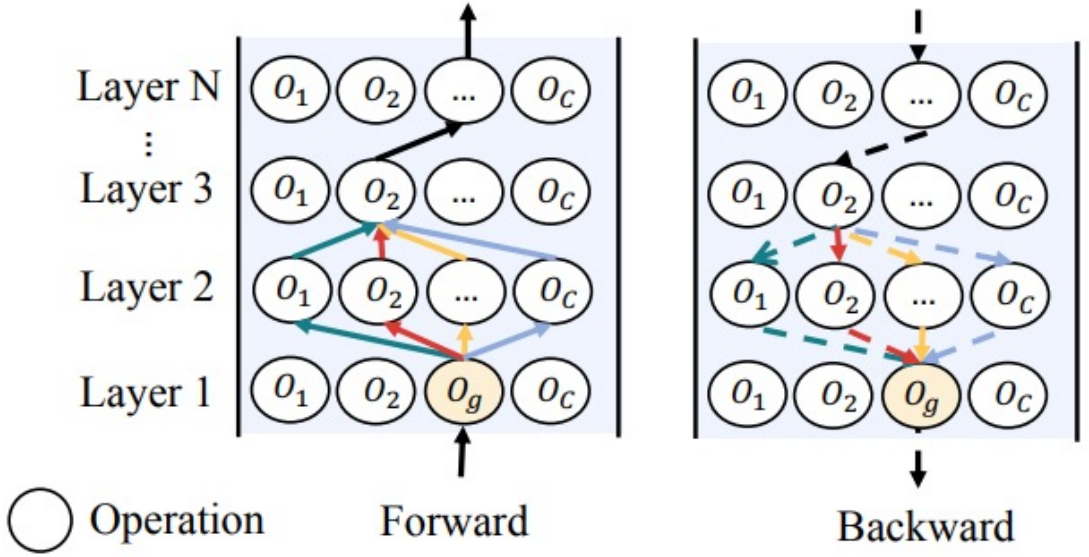


Figure 1: The illustration of the forward and backward process regarding the operator o_g in layer 1 that is shared by child models that differ in layer 2.

- Little has been discussed about the **causes of the interference** and how to mitigate it
- This paper focuses on the interference issue of chain-styled search space in sampled single path one-shot NAS

Analyses

- Interference: Gradient interference on shared operators

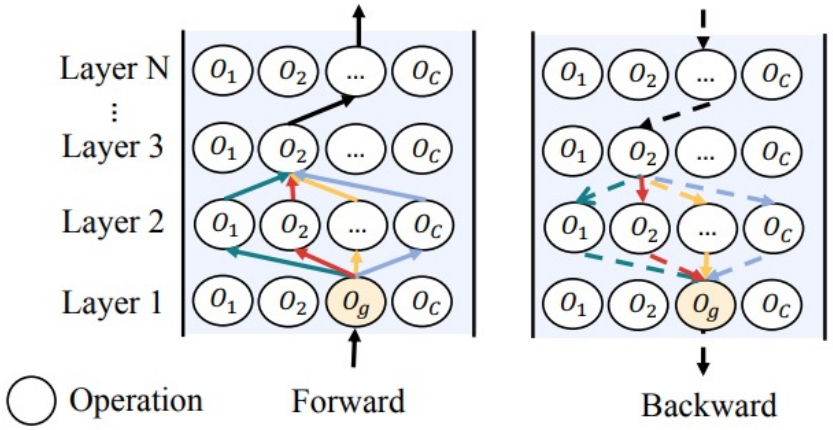
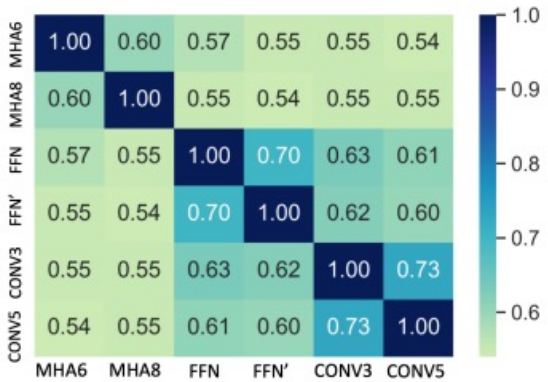


Figure 1: The illustration of the forward and backward process regarding the operator o_g in layer 1 that is shared by child models that differ in layer 2.

- Analyze the gradient similarity on shared operators between different candidate models

(a) Cosine similarity matrix of the super-net trained by single path one-shot



(b) Cosine similarity matrix of the super-net trained by single path one-shot with alignment

- We find
  - By aligning the inputs and outputs of the shared operators to be similar to the average inputs and outputs, the gradient interference can be reduced
  - The interference on a shared operator between two child models is positively correlated to the number of different operators between them.

Analyses

- Interference: Gradient interference on shared operators

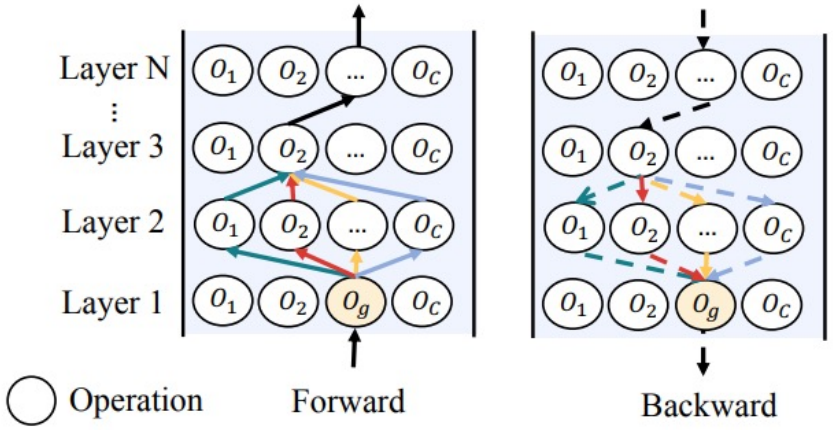
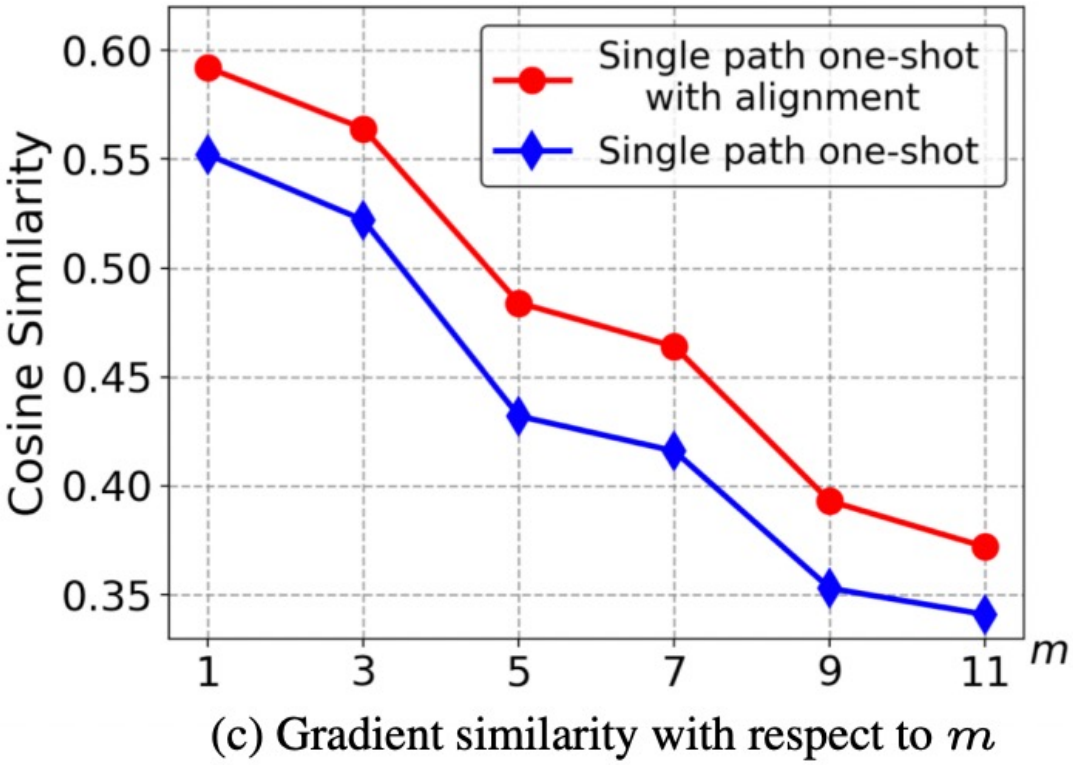


Figure 1: The illustration of the forward and backward process regarding the operator o_g in layer 1 that is shared by child models that differ in layer 2.

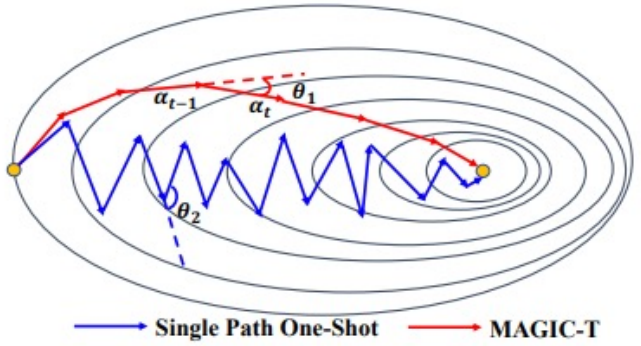


(c) Gradient similarity with respect to m

- We find
  - By aligning the inputs and outputs of the shared operators to be similar to the average inputs and outputs, the gradient interference can be reduced
  - The interference on a shared operator between two child models is positively correlated to the number of different operators between them.

Methods

- Approach 1: MitigAtinG InTerferenCe (**MAGIC-T**) from the perspective of Topological environment
  - Gradually change the topological environment for the shared operators
  - Samples a child model by randomly substituting one operator in the child model sampled at the last step with another operator fo weights updating at each training step



- Approach 2: MitigAtinG InTerferenCe (**MAGIC-A**) from the perspective of inputs and outputs Alignment
  - Pick a top-performing anchor child model from the search space to align other child models
  - The anchor model can be replaced when the performance of another child model outperforms it

Experiments

Model	Params	MNLI	QQP	QNLI	CoLA	SST-2	STS-B	RTE	MRPC	AVG
dev set										
DistilBERT (Sanh et al., 2019)	66M	82.2	88.5	89.2	51.3	91.3	86.9	59.9	87.5	79.6
MiniLM (Wang et al., 2020)	66M	84.0	91.0	91.0	49.2	92.0	-	71.5	88.4	-
BERT-of-Theseus (Xu et al., 2020)	66M	82.3	89.6	89.5	51.1	91.5	88.7	68.2	-	-
PD-BERT (Turc et al., 2019)	66M	82.5	90.7	89.4	-	91.1	-	66.7	84.9	-
DynaBERT* (Hou et al., 2020)	60M	84.2	91.2	91.5	56.8	92.7	89.2	72.2	84.1	82.7
NAS-BERT (Xu et al., 2021)	60M	84.1	91.0	91.3	58.1	92.1	89.4	79.2	88.5	84.2
SPOS (Guo et al., 2020)	60M	84.0	90.7	**91.1**	57.1	91.6	88.2	75.9	86.5	83.1
MAGIC-AT	60M	84.5	90.9	91.1	61.8	92.8	89.0	78.9	89.2	**84.8**
test set										
BERT-of-Theseus (Xu et al., 2020)	66M	82.4	89.3	89.6	47.8	92.2	84.1	66.2	83.2	79.4
PD-BERT (Turc et al., 2019)	66M	82.8	88.5	88.9	-	91.8	-	65.3	81.7	-
BERT-PKD (Sun et al., 2019)	66M	81.5	88.9	89.0	-	92.0	-	65.5	79.9	-
TinyBERT* (Jiao et al., 2020)	66M	84.6	89.1	90.4	51.1	93.1	83.7	70.0	82.6	80.6
NAS-BERT (Xu et al., 2021)	60M	83.5	88.9	90.9	48.4	92.9	86.1	73.7	84.5	81.1
SPOS (Guo et al., 2020)	60M	83.5	88.5	**90.6**	52.4	91.7	86.5	74.2	83.6	81.4
MAGIC-AT	60M	84.2	88.8	90.6	53.6	92.1	86.8	75.6	84.3	**82.0**

Experiments

Model	Params	FLOPs	MNLI	QQP	QNLI	CoLA	SST-2	STS-B	RTE	MRPC	AVG
dev set											
$\text{BERT}_{\text{base}}$ (Devlin et al., 2019)	110M	2.9e10	84.4	89.9	88.4	54.3	92.7	88.9	71.1	86.7	82.1
$\text{RoBERTa}_{\text{base}}$ (Liu et al., 2019)	125M	3.3e10	85.3	91.1	91.1	61.0	92.7	90.0	77.5	87.9	84.6
$\text{ELECTRA}_{\text{base}}$ (Clark et al., 2020)	110M	2.9e10	-	-	-	-	-	-	-	-	85.1
$\text{MPNet}_{\text{base}}$ (Song et al., 2020)	110M	2.9e10	85.2	-	-	-	93.4	-	-	-	-
SPOS (Guo et al., 2020)	114M	3.3e10	84.7	91.4	91.4	59.6	92.1	89.7	80.9	86.3	84.4
MAGIC-AT	113M	3.3e10	85.6	91.3	91.8	61.1	**93.5**	90.3	80.9	**90.9**	85.7
E-MAGIC-AT	110M	2.9e10	**86.3**	**91.7**	**92.5**	**65.8**	92.5	**91.0**	**84.0**	89.7	**86.7**
test set											
$\text{BERT}_{\text{base}}$ (Devlin et al., 2019)	110M	2.9e10	84.6	89.2	90.5	52.1	93.5	85.8	66.4	84.8	80.9
$\text{RoBERTa}_{\text{base}}$ (Liu et al., 2019)	125M	3.3e10	84.8	89.0	91.7	57.1	93.3	88.0	74.1	84.1	82.8
$\text{ELECTRA}_{\text{base}}$ (Clark et al., 2020)	110M	2.9e10	85.8	89.1	**92.7**	59.7	93.4	87.7	73.1	86.7	83.5
SPOS (Guo et al., 2020)	114M	3.3e10	84.3	88.6	91.0	56.1	92.8	88.1	74.9	83.4	82.4
MAGIC-AT	113M	3.3e10	84.9	89.1	92.0	57.0	**94.1**	**87.8**	77.4	85.2	83.4
E-MAGIC-AT	110M	2.9e10	**85.9**	**89.6**	92.4	**60.3**	93.4	87.3	**80.4**	**87.4**	**84.6**

Experiments

Table 5: Comparison of models on ImageNet.

Model	Top1/Top5 Err.	Params	FLOPS
MobileNetV2 (Sandler et al., 2018)	25.3/-	6.9M	585M
ShuffleNetV2 (Zhang et al., 2018b)	25.1/-	~5M	591M
DARTS (Liu et al., 2018)	26.9/9.0	4.9M	595M
PC-DARTS (Xu et al., 2019)	24.2/7.3	5.3M	597M
CARS (Yang et al., 2020)	24.8/7.5	5.1M	591M
PC-NAS (Li et al., 2020)	23.9/-	5.1M	-
EnTranNAS-DST (Yang et al., 2021)	23.8/7.0	5.2M	594M
Models searched on the MobileNetV2 search space			
NAO (Luo et al., 2018)	24.5/7.8	6.5M	590M
LaNAS (Wang et al., 2021a)	25.0/7.7	5.1M	570M
BN-NAS (Chen et al., 2021)	24.3/-	4.4M	470M
ProxelessNAS (Cai et al., 2018)	24.0/7.1	5.8M	595M
RLNAS (Zhang et al., 2021)	24.4/7.4	5.3M	473M
SemiNAS (Luo et al., 2020)	23.5/6.8	6.3M	599M
MAGIC-AT	**23.2/6.7**	6.0M	598M

Thanks!