Graph Structure of Neural Networks

报告人：陈宇航

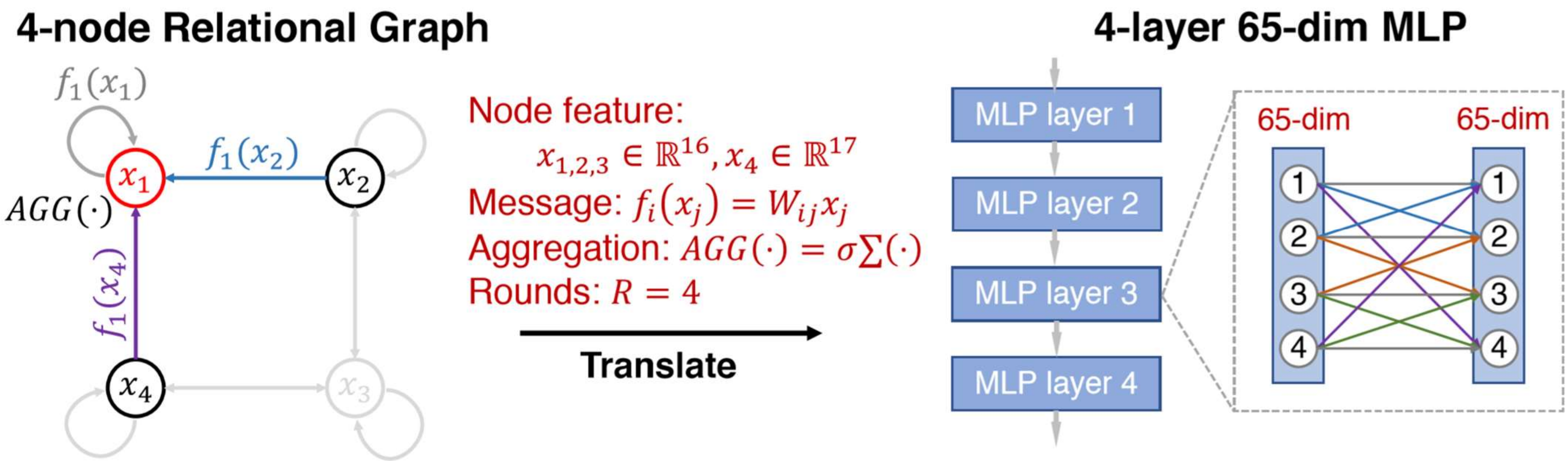
时间：2020年9月3日

Outline

- Definitions
- Search space
- Experiments

Method

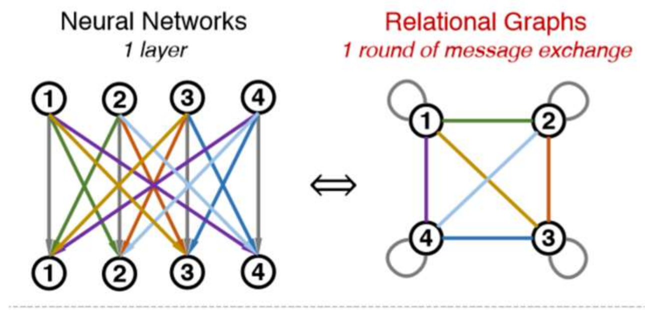
relational graph



1 message function f()

2 aggregated function AGG()

3 $$\mathbf{x}_v^{(r+1)} = \mathrm{AGG}^{(r)}(\{f_v^{(r)}(\mathbf{x}_u^{(r)}), \forall u \in N(v)\}) \qquad (1)$$

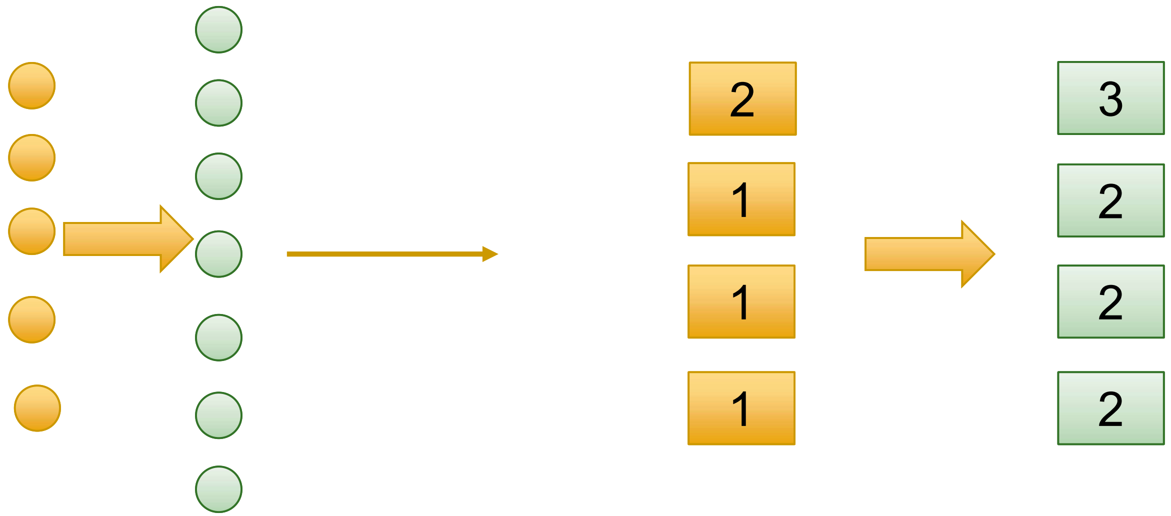


Definitions

Graph Structure of Neural Networks

	Fixed-width MLP	**Variable-width MLP**	**ResNet-34**	**ResNet-34-sep**	**ResNet-50**
Node feature $\mathbf{x}_i$	Scalar: 1 dimension of data	Vector: multiple dimensions of data	Tensor: multiple channels of data	Tensor: multiple channels of data	Tensor: multiple channels of data
Message function $f_i(\cdot)$	Scalar multiplication	(Non-square) matrix multiplication	3×3 Conv	3×3 depth-wise and 1×1 Conv	3×3 and 1×1 Conv
Aggregation function $\textsc{Agg}(\cdot)$	$\sigma(\sum(\cdot))$	$\sigma(\sum(\cdot))$	$\sigma(\sum(\cdot))$	$\sigma(\sum(\cdot))$	$\sigma(\sum(\cdot))$
Number of rounds R	1 round per layer	1 round per layer	34 rounds with residual connections	34 rounds with residual connections	50 rounds with residual connections

Table 1: **Diverse neural architectures expressed in the language of relational graphs.** These architectures are usually implemented as complete relational graphs, while we systematically explore more graph structures for these architectures.



$$\mathbf{x}_i^{(r+1)} = \sigma\Big(\sum_{j\in N(i)} \mathbf{W}_{ij}^{(r)}\mathbf{x}_j^{(r)}\Big) \tag{4}$$

Definitions

Graph Structure of Neural Networks

	Fixed-width MLP	**Variable-width MLP**	**ResNet-34**	**ResNet-34-sep**	**ResNet-50**
Node feature $\mathbf{x}_i$	Scalar: 1 dimension of data	Vector: multiple dimensions of data	Tensor: multiple channels of data	Tensor: multiple channels of data	Tensor: multiple channels of data
Message function $f_i(\cdot)$	Scalar multiplication	(Non-square) matrix multiplication	3×3 Conv	3×3 depth-wise and 1×1 Conv	3×3 and 1×1 Conv
Aggregation function AGG$(\cdot)$	$\sigma(\sum(\cdot))$	$\sigma(\sum(\cdot))$	$\sigma(\sum(\cdot))$	$\sigma(\sum(\cdot))$	$\sigma(\sum(\cdot))$
Number of rounds R	1 round per layer	1 round per layer	34 rounds with residual connections	34 rounds with residual connections	50 rounds with residual connections

Table 1: **Diverse neural architectures expressed in the language of relational graphs.** These architectures are usually implemented as complete relational graphs, while we systematically explore more graph structures for these architectures

$$\mathbf{X}_i^{(r+1)} = \sigma\Big(\sum_{j \in N(i)} \mathbf{W}_{ij}^{(r)} * \mathbf{X}_j^{(r)}\Big) \qquad (5)$$

Outline

- Definitions
- Measures
- Experiments

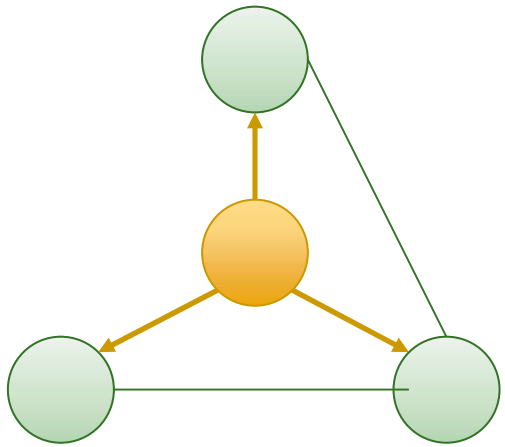
Measures

Average path length :

average shortest path distance between any pair of nodes

clustering coefficient :

measures the proportion of edges between the nodes within a given node's neighborhood, divided by the number of edges that could possibly exist between them, averaged over all the nodes



2/3

Measures

WS
1 the N nodes are regularly placed in a ring
2 each node is connected to its K/2 neighbors on both sides(K is an even number)
3 rewire the edge iteratively with a probability P

WS-flex
1 The N nodes are regularly placed in a ring
2 The number of edges is determined as $e = \lfloor n * k/2 \rfloor$
3 each node connect $\lfloor e/n \rfloor$ edges, picks e mode n nodes connect to one closest neighboring node
4 rewire the edge iteratively with a probability P

WS-flex relaxing the constraint that all the nodes have the same degree before random rewiring.

Measures

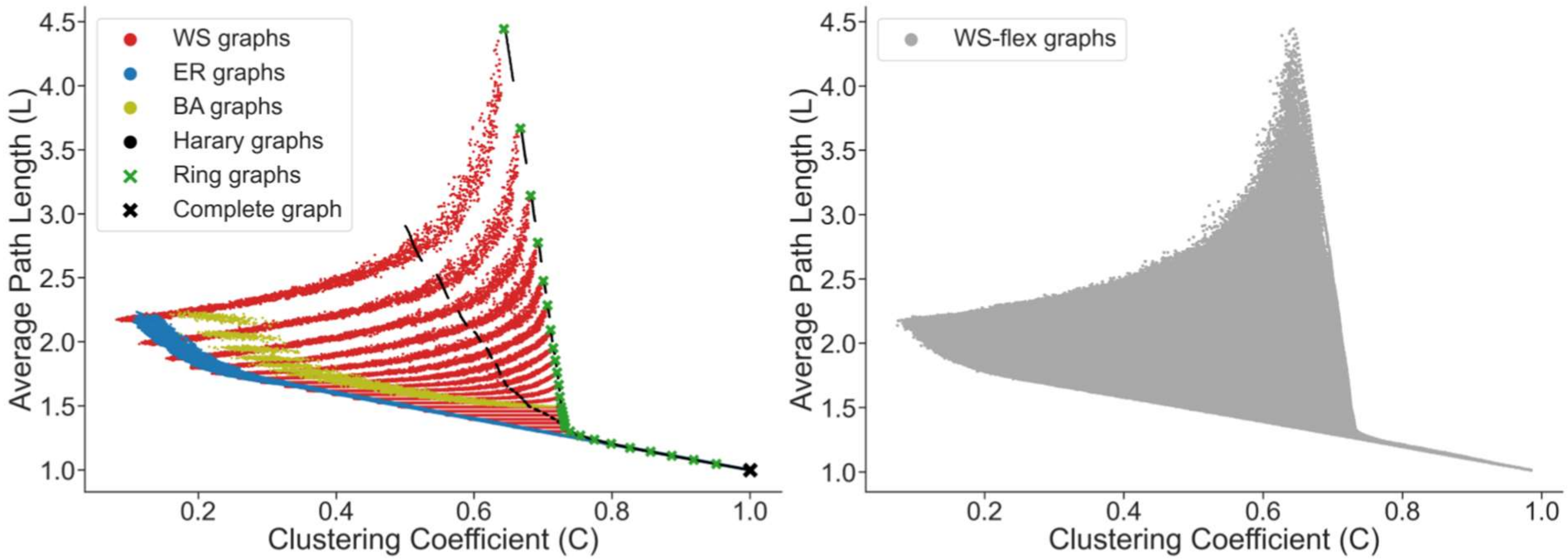
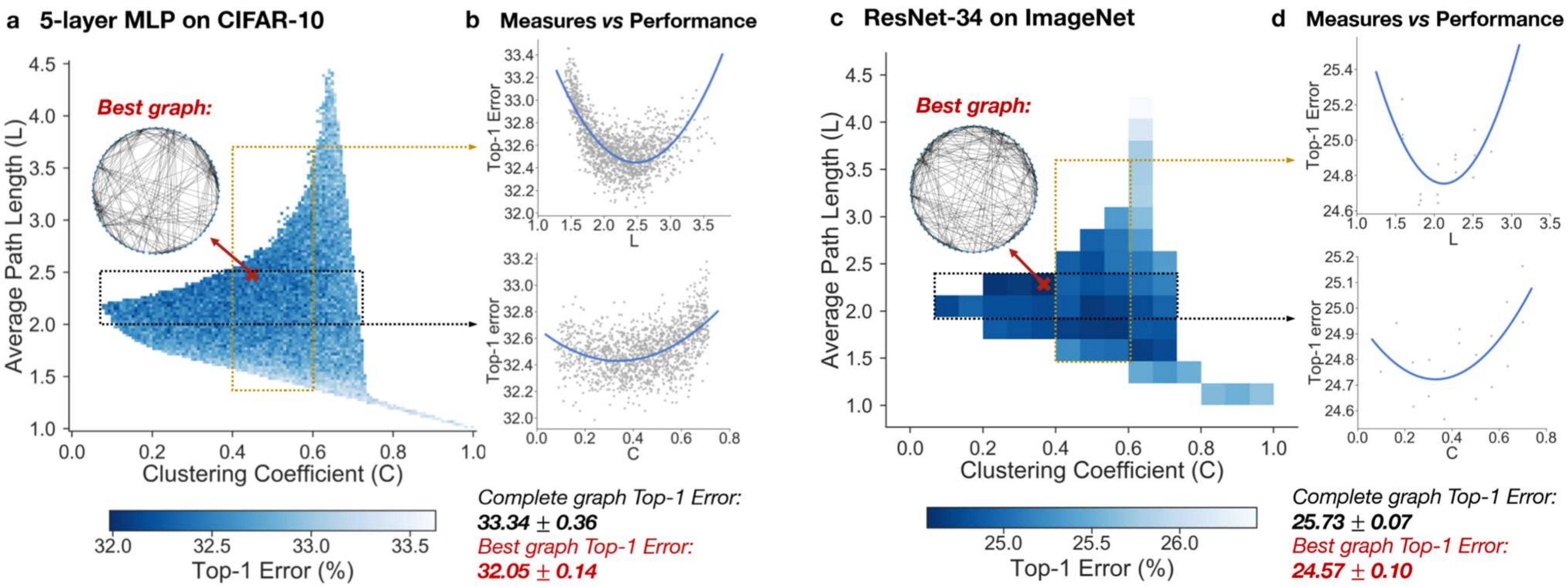


Figure 3: **Graphs generated by different graph generators.** The proposed graph generator WS-flex can cover a much larger region of graph design space. WS (Watts-Strogatz), BA (Barabási-Albert), ER (Erdős-Rényi).

Outline

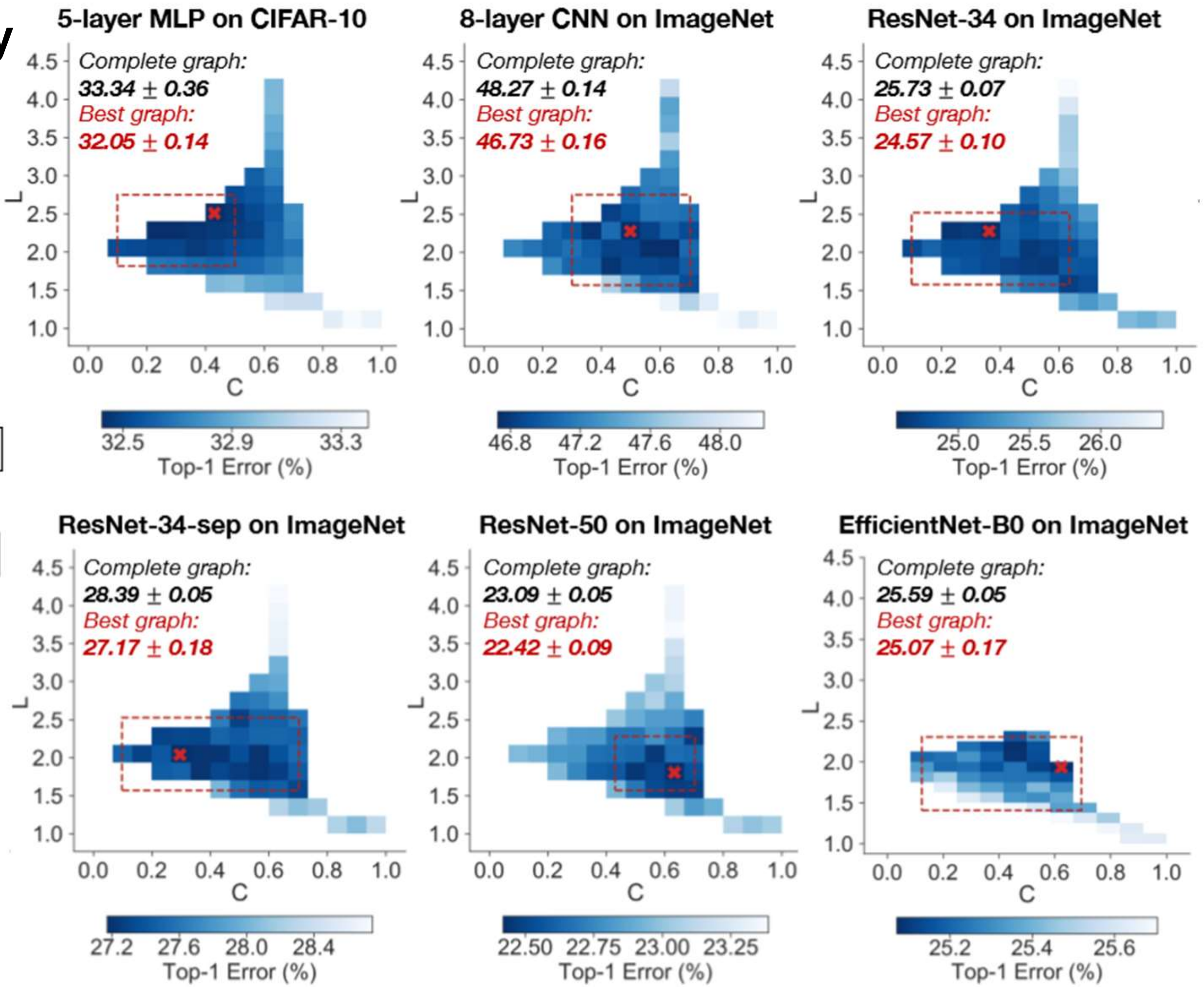
- Definitions
- Measures
- Experiments

Experiment


a 5-layer MLP on CIFAR-10
b Measures vs Performance
c ResNet-34 on ImageNet
d Measures vs Performance
Best graph:
Best graph:
Average Path Length (L)
Clustering Coefficient (C)
Top-1 Error (%)
Top-1 Error
Top-1 error
L
C
Complete graph Top-1 Error:
33.34 ± 0.36
Best graph Top-1 Error:
32.05 ± 0.14
Complete graph Top-1 Error:
25.73 ± 0.07
Best graph Top-1 Error:
24.57 ± 0.10

Experiment

Consistency



Sweet Pot

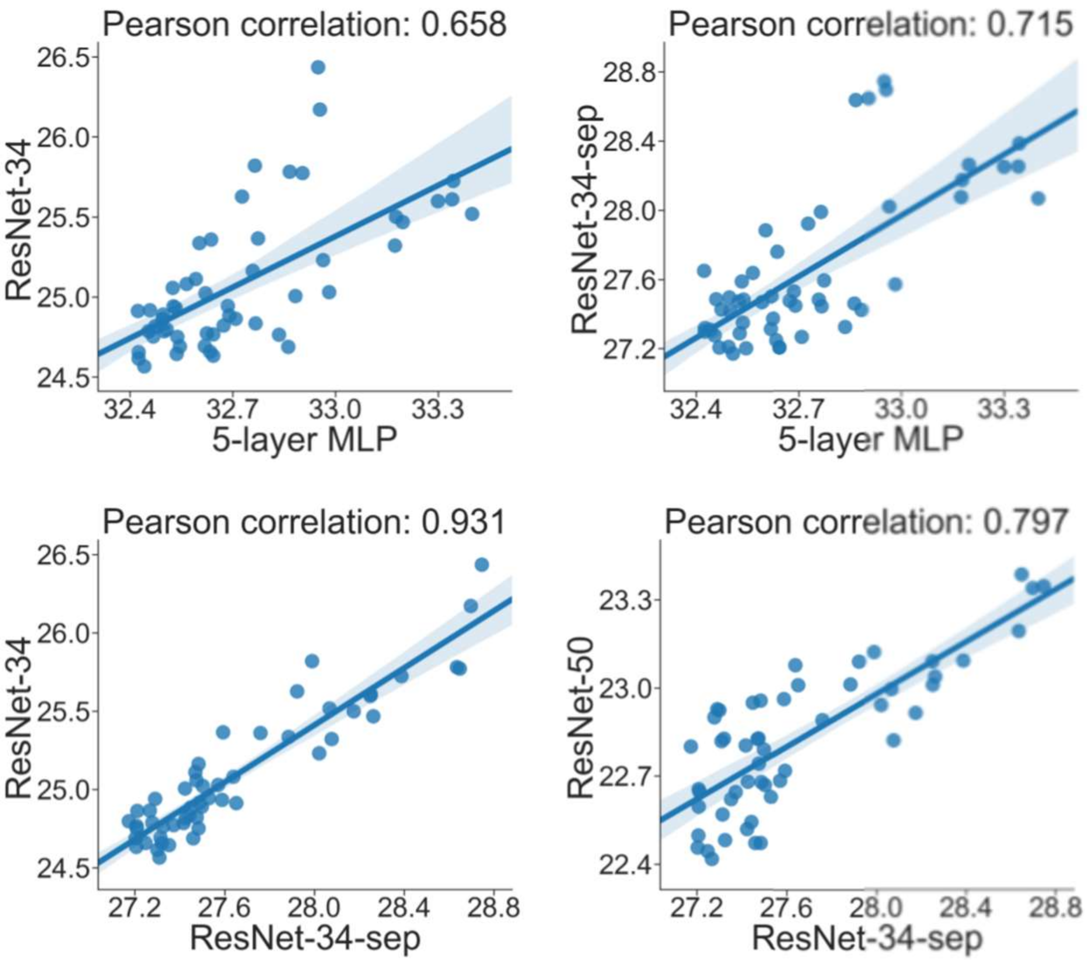
$C \in [0.43, 0.50]$

$L \in [1.82, 2.28]$

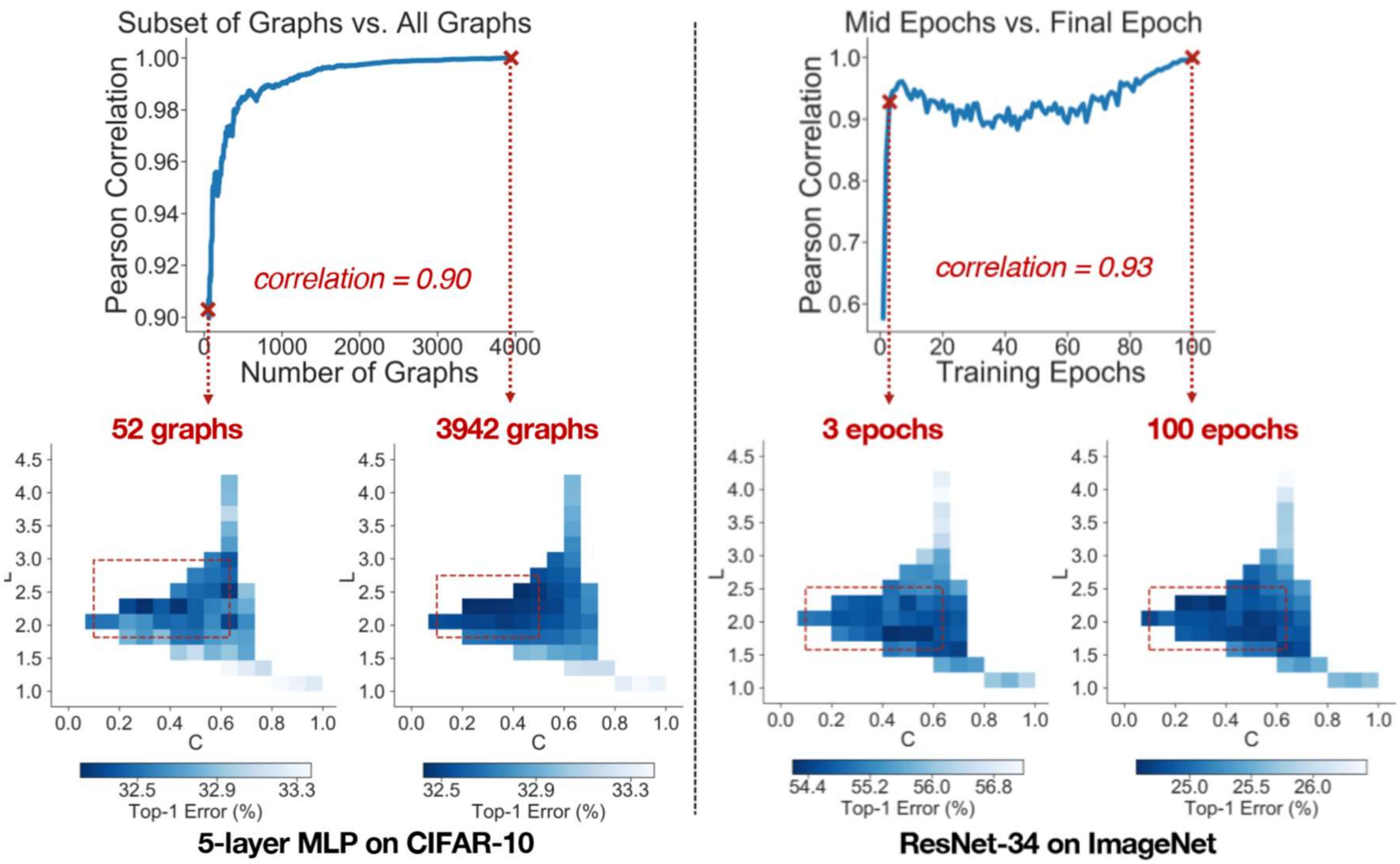
Experiment

Consistency

e Correlation across neural architectures



Experiment



Subset of Graphs vs. All Graphs
Pearson Correlation
1.00
0.98
0.96
0.94
0.92
0.90
correlation = 0.90
0
1000
2000
3000
4000
Number of Graphs
52 graphs
3942 graphs
L
C
Top-1 Error (%)
32.5
32.9
33.3
5-layer MLP on CIFAR-10
Mid Epochs vs. Final Epoch
Pearson Correlation
1.0
0.9
0.8
0.7
0.6
correlation = 0.93
0
20
40
60
80
100
Training Epochs
3 epochs
100 epochs
54.4
55.2
56.0
56.8
25.0
25.5
26.0
ResNet-34 on ImageNet

Experiment

Graph	Path (L)	Clustering (C)	CIFAR-10 Error (%)
Complete graph	1.00	1.00	33.34 ± 0.36
Cat cortex	1.81	0.55	33.01 ± 0.22
Macaque visual cortex	1.73	0.53	32.78 ± 0.21
Macaque whole cortex	**2.38**	**0.46**	32.77 ± 0.14
Consistent sweet spot across neural architectures	**1.82-2.28**	**0.43-0.50**	32.50 ± 0.33
Best 5-layer MLP	**2.48**	**0.45**	32.05 ± 0.14

Table 2: Top artificial neural networks can be similar to biological neural networks (Bassett & Bullmore, 2006).

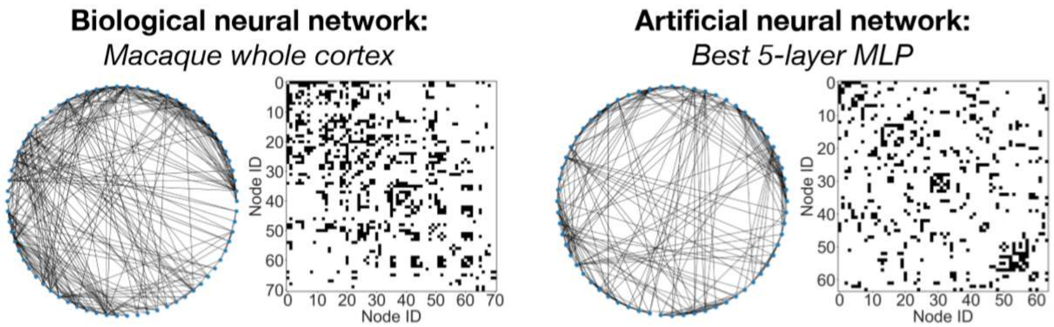
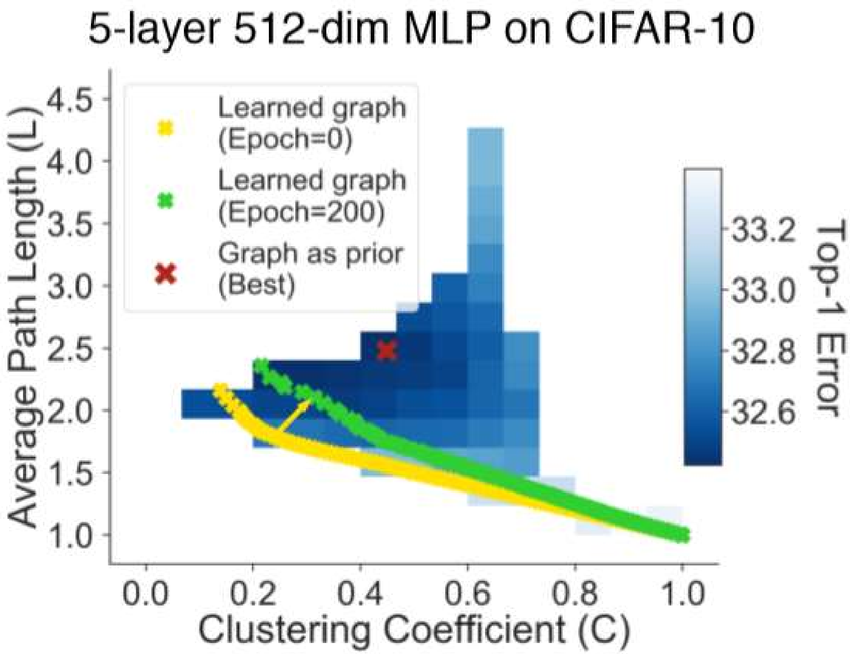


Figure 6: Visualizations of graph structure of biological (**left**) and artificial (**right**) neural networks.

Experiment



$$\mathbf{x}_v^{(r+1)} = \mathrm{AGG}^{(r)}(\{f_v^{(r)}(\mathbf{x}_u^{(r)}), \forall u \in N(v)\})$$

The End!