Parameter Prediction for Unseen Deep Architectures

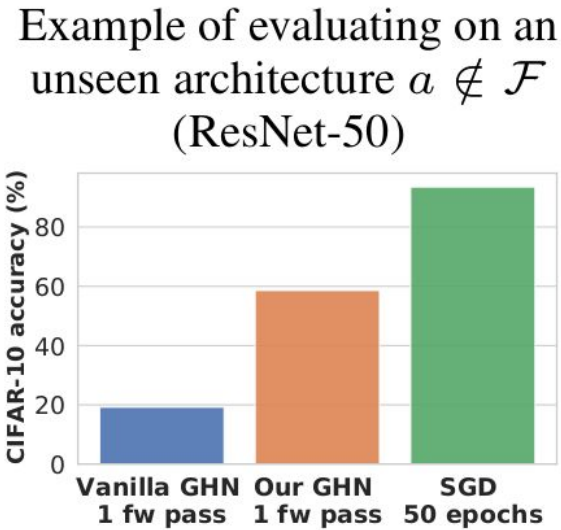
Докладчик: Анастасия Дроздова
Рецензент: Айнур Нуриев
Практик-исследователь: Полина Гусева
Хакер: Алексей Цеховой

https://openreview.net/pdf?id=vqHak8NLk25

Идея

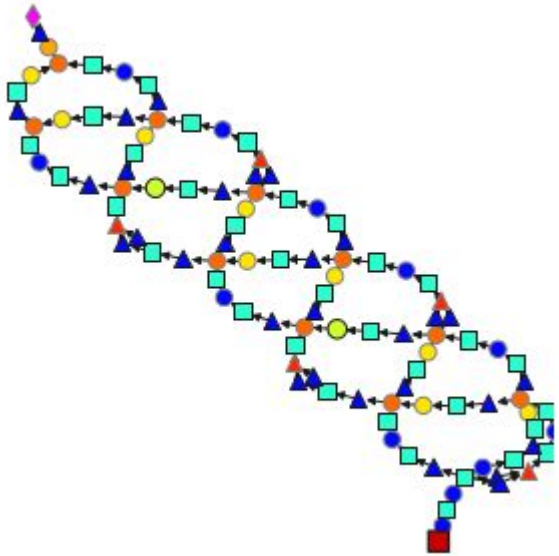
- предсказываем веса для архитектуры за один проход
- набор данных и задача фиксированы

$$\arg\min_{\theta} \sum_{j=1}^{N} \sum_{i=1}^{M} \mathcal{L}\Big(f\Big(\mathbf{x}_j; a_i, H_{\mathcal{D}}(a_i; \theta)\Big), y_j\Big)$$



Graph HyperNetwork GHN-1

- архитектура - направленный ациклический граф вычислений
- вершины $V = \{v_i\}_{i=1}^{|V|}$
- матрица смежности $\mathbf{A} \in \{0,1\}^{|V|\times|V|}$
- матрица операций $\mathbf{H}^0 = [\mathbf{h}_1^0, \mathbf{h}_2^0, ..., \mathbf{h}_{|V|}^0]$
- кодируем размерности параметров



Graph HyperNetwork GHN-1

1. Эмбеддинг $\mathbf{H}^0$ размерности d: $\mathbf{H}^1 \in \mathbb{R}^{|V| \times d}$

2. GatedGNN имитирует проходы по графу

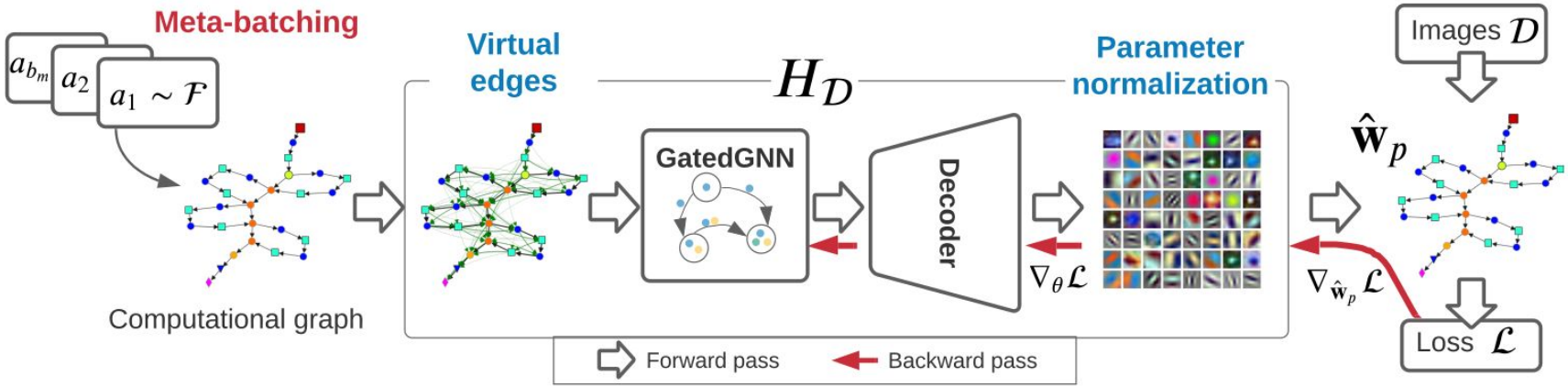
$$\forall t \in [1, ..., T] : \Big[\forall \pi \in [\text{fw}, \text{bw}] : \Big(\forall v \in \pi : \mathbf{m}_v^t = \sum_{u \in \mathcal{N}_v^\pi} \text{MLP}(\mathbf{h}_u^t), \;\; \mathbf{h}_v^t = \text{GRU}(\mathbf{h}_v^t, \mathbf{m}_v^t)\Big)\Big]$$

3. Декодер использует $\mathbf{h}_v^T$, чтобы получить параметры $\hat{\mathbf{w}}_p^v$

GHN-2

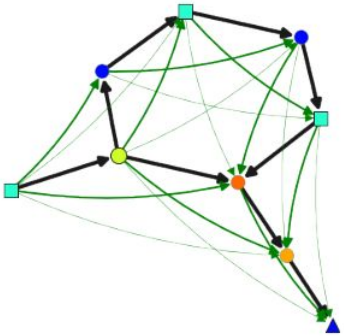
- нормализация предсказанных параметров
- виртуальные ребра
- мета-батчи

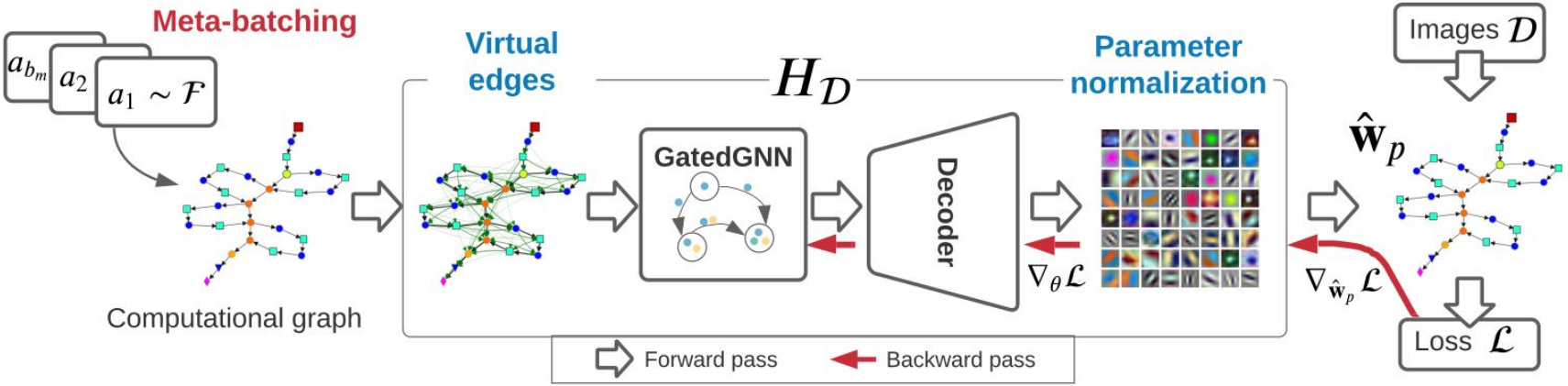
Type of node v	Normalization
Conv./fully-conn.	$\hat{\mathbf{w}}_p^v \sqrt{\beta/(C_{in}\mathcal{H}\mathcal{W})}$
Norm. weights	$2 \times \text{sigmoid}(\hat{\mathbf{w}}_p^v/T)$
Biases	$\tanh(\hat{\mathbf{w}}_p^v/T)$



GHN-2

- нормализация предсказанных параметров
- виртуальные ребра
- мета-батчи



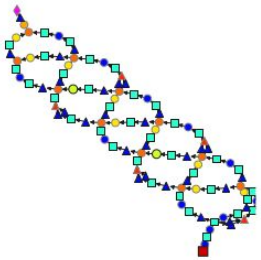
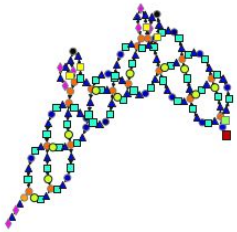
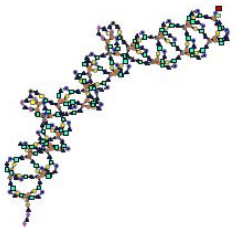
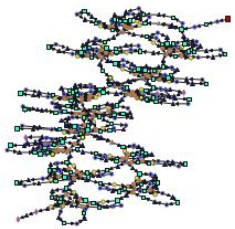
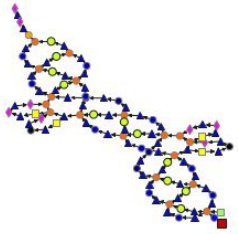
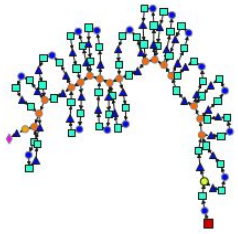


DeepNets-1M

- Было в DARTS
  - stems
  - normal cells
  - reduction cells
  - classification heads

- Добавили:
  - non-separable 2D convolutions
  - Squeeze&Excite
  - multihead self-attention
  - positional encoding
  - layer norm

- Генерация архитектур
  - in-distribution
  - out-of-distribution (wide, deep, dense, BN-free, ResNet, ViT)

	IN-DISTRIBUTION		OUT-OF-DISTRIBUTION				
							
	TRAIN	VAL/TEST	WIDE	DEEP	DENSE	BN-FREE	RESNET/VIT
#graphs	10^6	500/500	100	100	100	100	1/1
#cells	4-18	4-18	4-18	**10-36**	4-18	4-18	16/12
#channels	16-128	32-128	**128-1216**	32-208	32-240	32-336	64/128
#nodes ($\|V\|$)	21-827	33-579	33-579	**74-1017**	**57-993**	33-503	161/114
% w/o BN	3.5%	4.1%	4.1%	2.0%	5.0%	**100%**	0%/**100%**
#params(M)*	0.01-3.1	2.5-35	**39-101**	2.5-15.3	2.5-8.8	2.5-7.7	**23.5**/1.0
avg degree	2.3±0.1	2.3±0.1	2.3±0.1	2.3±0.1	**2.4**±0.1	**2.4**±0.1	2.2/2.3
avg path	14.5±4.8	14.5±4.9	14.7±4.9	**26.2**±9.3	15.1±4.1	10.0±2.8	11.2/10.7

marker / primitive	conv	BN	sum	bias	group conv	concat	dilated gr. conv	LN	max pool	avg pool	MSA	SE	input	glob avg	pos enc
fraction in TRAIN (%)	36.3	25.5	11.1	6.5	5.1	3.8	2.5	2.5	1.8	1.7	1.2	1.0	0.5	0.5	0.2

Эксперименты: предсказание параметров

Table 4: ImageNet results on DEEPNETS-1M. Mean (±standard error of the mean) top-5 accuracies are reported (random chance ≈0.5%). *Estimated on ResNet-50 with batch size 128.

METHOD	#upd	GPU sec.	CPU sec.	ID-TEST		OOD-TEST				
		avg	avg	avg	max	WIDE	DEEP	DENSE	BN-FREE	RESNET/VIT
GHN-1	1	0.3	0.5	17.2±0.4	32.1	15.8±0.9	15.9±0.8	15.1±0.7	0.5±0.0	**6.9**/0.9
GHN-2	1	0.3	0.7	**27.2**±0.6	**48.3**	**19.4**±1.4	**24.7**±1.4	**26.4**±1.2	**7.2**±0.6	5.3/**4.4**
Iterative optimizers (all architectures are ID in this case)										
SGD (1 step)	1	0.4	6.0	0.5±0.0	0.7	0.5±0.0	0.5±0.0	0.5±0.0	0.5±0.0	0.5/0.5
SGD (5000 steps)	5k	2×10^3	3×10^4	25.6±0.3	50.7	26.2±1.4	13.2±1.1	25.4±1.1	4.8±0.8	34.8/24.3
SGD (10000 steps)	10k	4×10^3	6×10^4	37.7±0.6	62.0	38.7±1.6	22.1±1.4	36.3±1.2	8.0±1.2	49.0/33.4
SGD (100 epochs)	1000k	$6\times10^{5*}$	$6\times10^{7*}$	–		–	–	–	–	92.9/72.2

Эксперименты: предсказание метрик

- accuracy on clean set
- accuracy on corrupted set
- inference speed
- convergence speed (SGD)

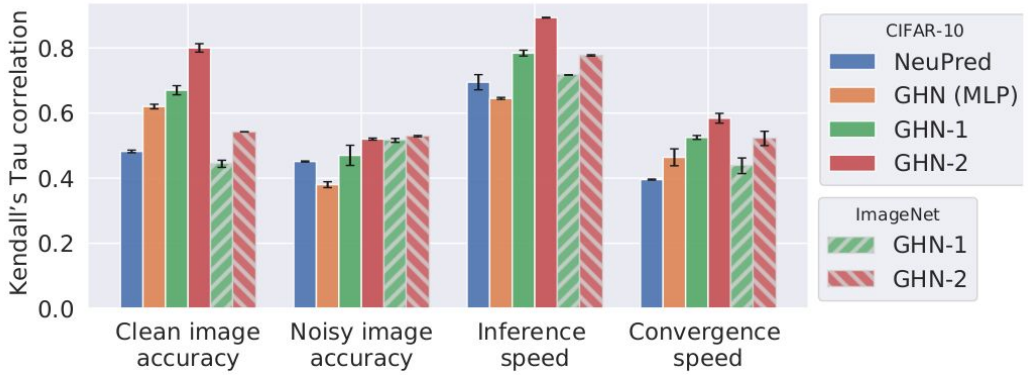


Figure 4: Property prediction of neural networks in terms of correlation (higher is better). Error bars denote the standard deviation across 5 runs.

Эксперименты: transfer learning

Initialization Method	GPU sec. to init.*	100-Shot CIFAR-10			Penn-Fudan Object Detection		
		ResNet-50	ViT	DARTS	ResNet-50	ViT	DARTS
He's [57]	0.003	41.0±0.4	33.2±0.3	45.4±0.4	0.197±0.042	0.144±0.010	0.486±0.035
GHN-1 (trained on ImageNet)	0.6	46.6±0.0	23.3±0.1	49.2±0.1	0.433±0.013	0.0±0.0	0.468±0.024
GHN-2 (trained on ImageNet)	0.7	**56.4**±0.1	**41.4**±0.6	**60.7**±0.3	**0.560**±0.019	**0.436**±0.032	**0.785**±0.032
ImageNet (1k pretraining steps)	6×10^2	45.4±0.3	**44.3**±0.1	**62.4**±0.3	0.302±0.022	0.182±0.046	**0.814**±0.033
ImageNet (2.5k pretraining steps)	1.5×10^3	**55.4**±0.2	50.4±0.3	70.4±0.2	**0.571**±0.056	0.322±0.073	0.823±0.022
ImageNet (5 pretraining epochs)	3×10^4	84.6±0.2	70.2±0.5	83.9±0.1	0.723±0.045	0.391±0.024	0.827±0.053
ImageNet (final epoch)	6×10^5	89.2±0.2	74.5±0.2	85.6±0.2	0.876±0.011	**0.468**±0.023	0.881±0.023

Рецензент

Сильные стороны

- Значительно ускоряет fine-tuning (100 -> 5 эпох) благодаря хорошей инициализации весов
- Может также предсказывать получаемое accuracy, inference speed, convergence speed

Слабые стороны

- Проигрывает SGD и Adam
- Не можем предсказывать параметры для новых задач
- Не может обобщаться на разные датасеты (на вход не подаются входные данные исходной нейронной сети на которой обучаемся)

Оценки

Понятность

Написано хорошо, но для полного и безболезненного понимания статьи нужны знания в графовых нейронных сетях.

Воспроизводимость

Несмотря на отсутствие практика, качественное описания всех необходимых деталей для запуска дает уверенность в хорошей воспроизводимости. Есть хорошо проработанный авторский репозиторий с моделью.

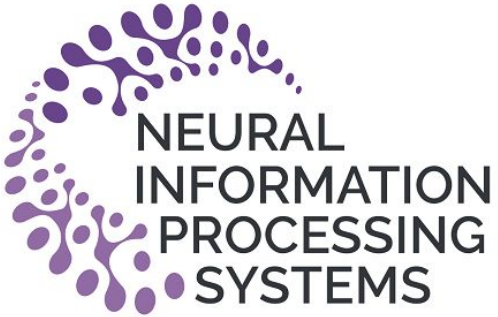
Вывод

Громкая статья с большим количеством подробностей, интересными идеями и впечатляющими результатами.

Оценка по НИПС: 8 (уверенность 4)

Практик-исследователь

О статье



- Первая версия написана весной 2021 года
- Постер на NeurIPS'21
- Статья активно обсуждалась ML сообществом:
  - Анонс от авторов в Twitter[1] набрал 1.5k+ лайков и 400 ретвитов
  - Статье посвящен выпуск на подкасте Янника Килчера
  - Статья освещалась в неформальных научных изданиях

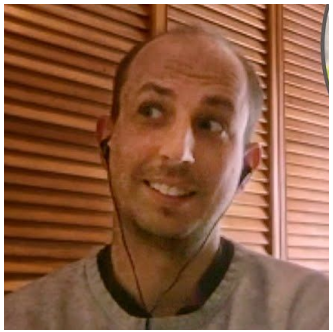
1. Запрещенная на территории РФ организация

Авторы

Boris Knyazev

PhD студент,
University of Guelph

GNN и CV



Michal Drozdal

Research Scientist,
Facebook AI Research

Medical AI и CV

Graham W. Taylor

Профессор, University
of Guelph

GNN и многое другое...

Adriana Romero-Soriano

Research Scientist,
Facebook AI Research

Multimodal data

Смежные статьи

Базовые статьи

- Graph HyperNetworks for Neural Architecture Search (2019)

  Предложенная сеть может принимать другие сети на вход.

- DARTS: Differentiable Architecture Search (2018)

  Предложили базовые блоки для нейросетевых архитектур.

Конкуренты

Нет (пока)

Цитирования

На статью опираются 2 работы:

- Teaching Networks to Solve Optimization Problems
- Tutorial on amortized optimization for learning to optimize over continuous domains

Еще 2 работы упоминают статью в обзорах литературы.

Итого 4 цитирования.

Что дальше?

Практика

- Использовать модель для предсказания статистик — финальной метрики, скорости инференса и скорости сходимости.
- Использовать предсказанные веса в качестве инициализации.

Эксперименты

- Сравниться с методами для инициализации сетей.
- Изучить метод на задачах отличных от классификации изображений.