

# Parameter Prediction for Unseen Deep Architectures

Докладчик: Медведев Антон  
Рецензент: Семенова Елена  
Исследователь: Морозов Никита  
Хакер: Гришанин Виктор

# Мотивация

Рассмотрим набор данных  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ .

$$\arg \min_{\mathbf{w}} \sum_{i=1}^N \mathcal{L}(f(\mathbf{x}_i; a, \mathbf{w}), y_i)$$

С постоянно растущим размером нейронных сетей и необходимостью их многократного обучения этот способ становится вычислительно неустойчивым.

При оптимизации параметров для новой архитектуры хотим использовать опыт, полученный при оптимизации других нейронных сетей.
















# Предлагаемый метод

Пусть  $\mathcal{F} = \{a_i\}_{i=1}^M$ .

Основная идея: использовать парадигму мета-обучения.

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

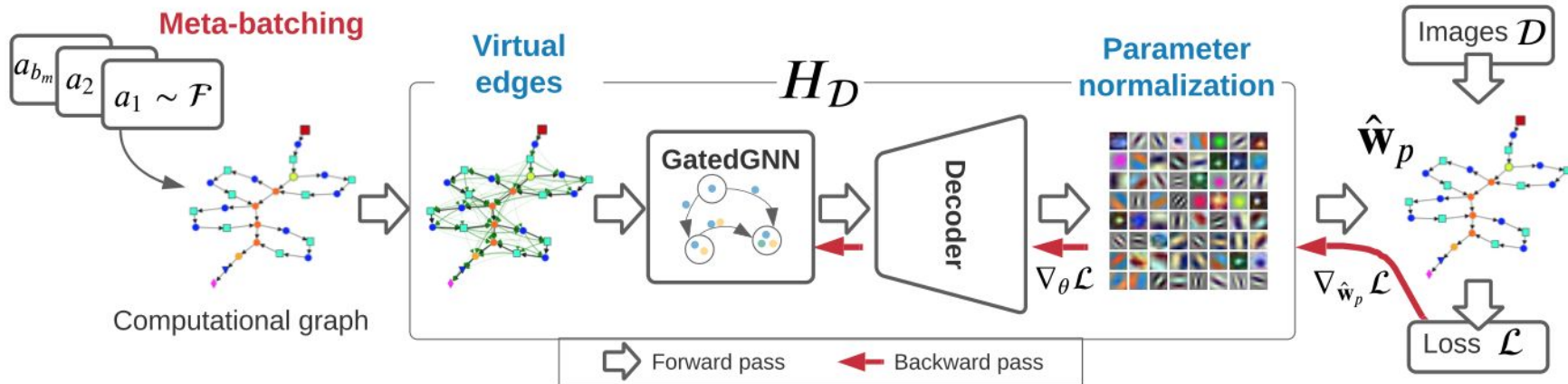
# DeepNets-1M

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

ID & OOD architectures

# Graph Hyper Network: GHN-1

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



## GHN-2: Normalization

Table 2: Parameter normalizations.

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: Virtual edges

$$\mathbf{m}_v^t = \sum_{u \in \mathcal{N}_v^\pi} \text{MLP}(\mathbf{h}_u^t) + \sum_{u \in \mathcal{N}_v^{(\text{sp})}} \frac{1}{s_{vu}} \text{MLP}_{\text{sp}}(\mathbf{h}_u^t)$$

$$1 < s_{vu} \leq s^{(\text{max})}$$

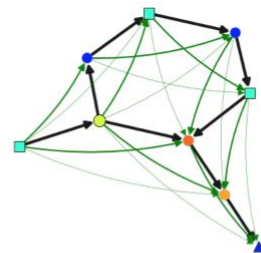


Figure 2: Virtual edges (in green) allow for better capture of global context.

## GHM-2: Meta-batching

Пусть  $b_m$  — количество архитектур для каждого набора изображений.

$$\nabla_{\theta} \mathcal{L} = 1/b_m \sum_{i=1}^{b_m} \nabla_{\theta} \mathcal{L}_i$$