CATE: Computation-aware Neural Architecture Encoding with Transformers

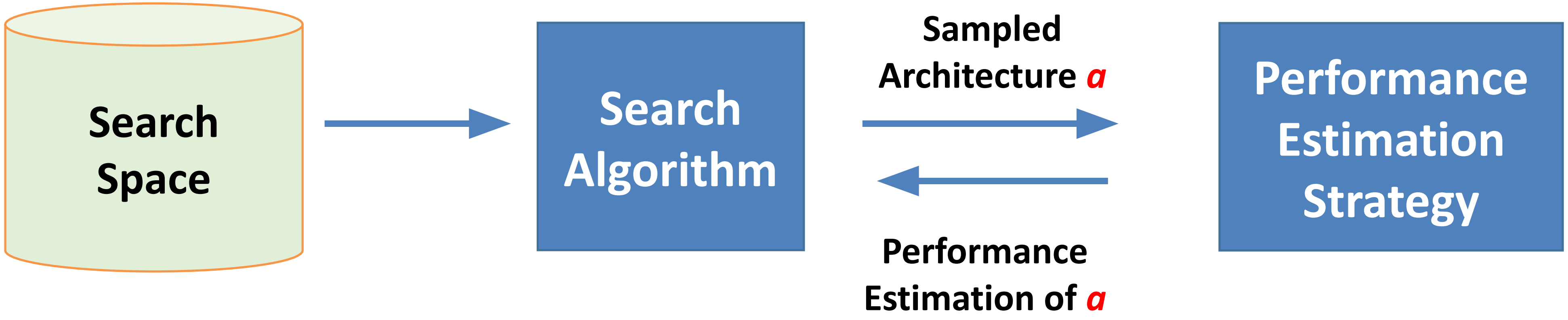
Shen Yan, Kaiqiang Song, Fei Liu, Mi Zhang

Michigan State University, Tencent AI Lab, University of Central Florida

https://arxiv.org/abs/2102.07108

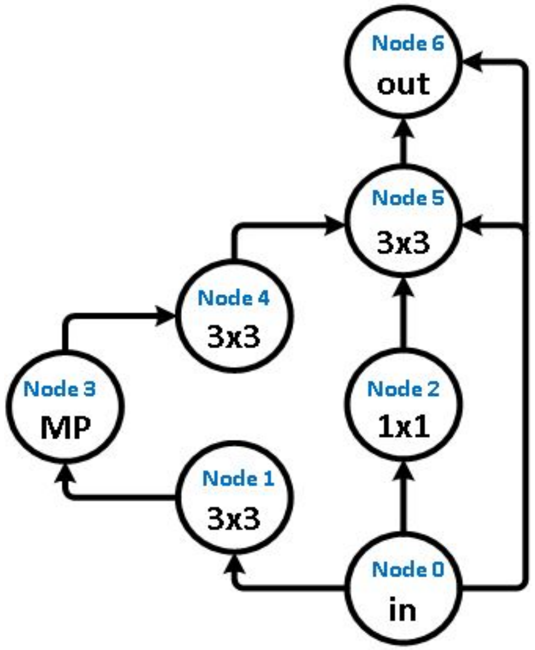
June 20th, 2021

Neural Architecture Search (NAS) Pipeline



Thomas Elsken, Jan Hendrik Metzen, Frank Hutter, Neural Architecture Search: A Survey, JMLR 2019

Structure-aware encodings: Adjacency matrix-based



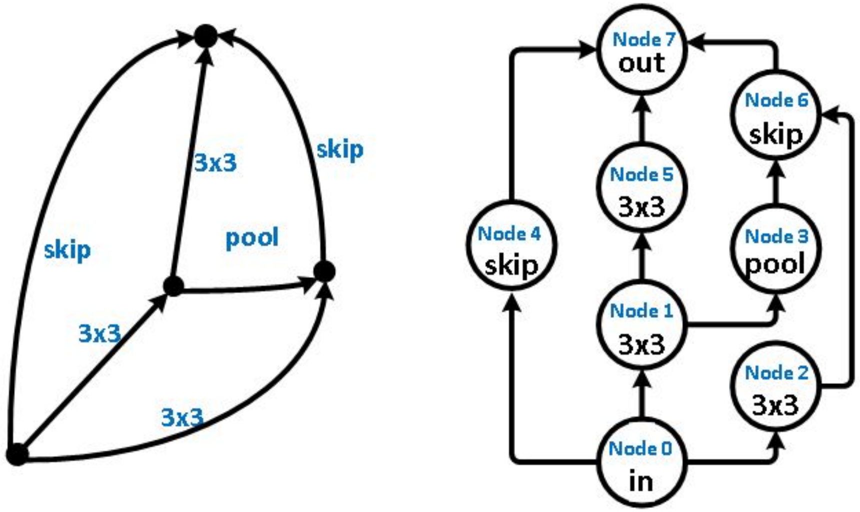
Node

$$\text{Node}\quad\begin{array}{c|c} & [0\ 1\ 2\ 3\ 4\ 5\ 6] \\ \hline \begin{matrix}0\\1\\2\\3\\4\\5\\6\end{matrix} & \begin{pmatrix} 0&1&1&0&0&1&1\\ 0&0&0&1&0&0&0\\ 0&0&0&0&0&1&0\\ 0&0&0&0&1&0&0\\ 0&0&0&0&0&1&0\\ 0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0 \end{pmatrix}\end{array}$$

Operation

$$\text{Node}\quad\begin{array}{c|c} & [\text{in}\ \ 1\text{x}1\ \ 3\text{x}3\ \ \text{MP}\ \ \text{out}] \\ \hline \begin{matrix}0\\1\\2\\3\\4\\5\\6\end{matrix} & \begin{pmatrix} 1&0&0&0&0\\ 0&0&1&0&0\\ 0&1&0&0&0\\ 0&0&0&1&0\\ 0&0&1&0&0\\ 0&0&1&0&0\\ 0&0&0&0&1 \end{pmatrix}\end{array}$$

One-hot adjacency encoding of NAS-Bench-101



Node

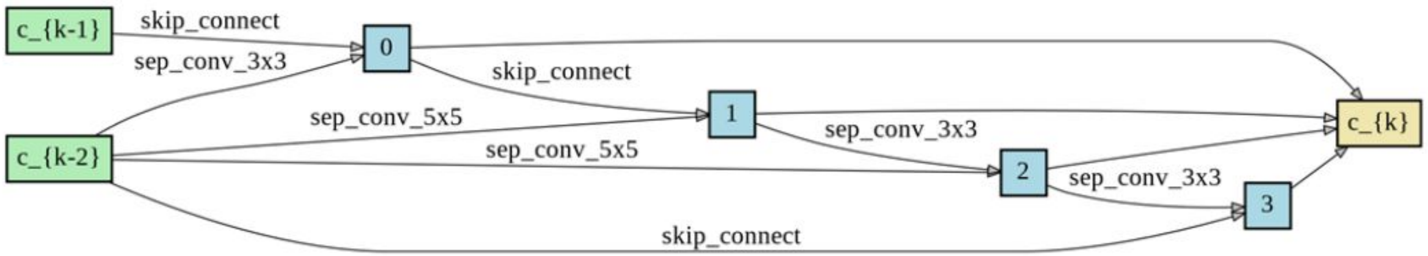
$$\text{Node}\quad\begin{array}{c|c} & [0\ 1\ 2\ 3\ 4\ 5\ 6\ 7] \\ \hline \begin{matrix}0\\1\\2\\3\\4\\5\\6\\7\end{matrix} & \begin{pmatrix} 0&1&1&0&1&0&0&0\\ 0&0&0&1&0&1&0&0\\ 0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&0 \end{pmatrix}\end{array}$$

Operation

$$\text{Node}\quad\begin{array}{c|c} & [\text{in}\ 1\text{x}1\ 3\text{x}3\ \text{pool}\ \text{skip}\ \text{none}\ \text{out}] \\ \hline \begin{matrix}0\\1\\2\\3\\4\\5\\6\\7\end{matrix} & \begin{pmatrix} 1&0&0&0&0&0&0\\ 0&0&1&0&0&0&0\\ 0&0&1&0&0&0&0\\ 0&0&0&1&0&0&0\\ 0&0&0&0&1&0&0\\ 0&0&1&0&0&0&0\\ 0&0&0&0&1&0&0\\ 0&0&0&0&0&0&1 \end{pmatrix}\end{array}$$

One-hot adjacency encoding of NAS-Bench-201

Structure-aware encodings: Adjacency matrix-based

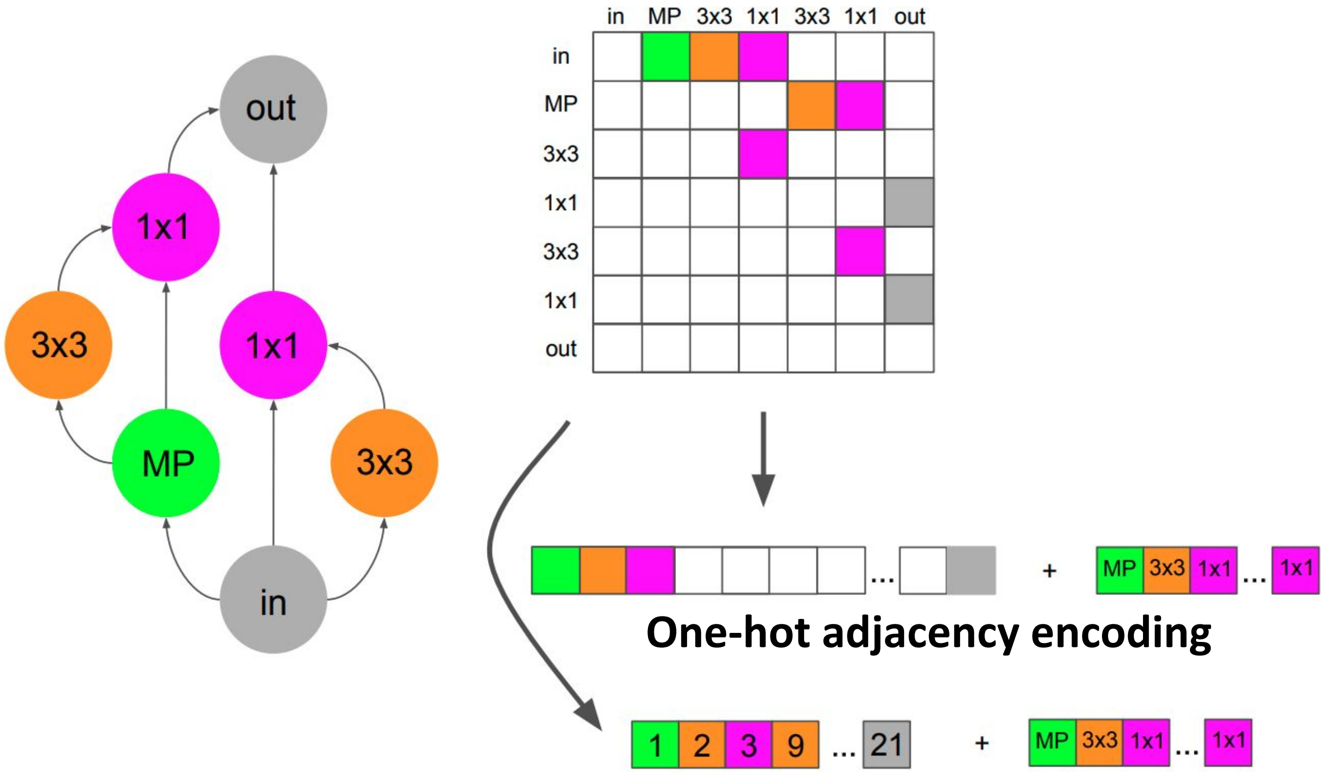


$$
\begin{array}{c|ccccccccccc|}
 & \text{c\_k-2} & \text{c\_k-1} & \multicolumn{2}{c}{\text{node0}} & \multicolumn{2}{c}{\text{node1}} & \multicolumn{2}{c}{\text{node2}} & \multicolumn{2}{c}{\text{node3}} & \text{c\_k} \\
\text{c\_k-2} & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\
\text{c\_k-1} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{node0} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
\text{node1} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
\text{node2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
\text{node3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\text{c\_k} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}
$$

$$
\begin{array}{ccccccccccc}
\text{c\_k-2} & \text{c\_k-1} & \text{zero} & \text{max\_pool\_3x3} & \text{avg\_pool\_3x3} & \text{skip\_connect} & \text{sep\_conv\_3x3} & \text{sep\_conv\_5x5} & \text{dil\_conv\_3x3} & \text{dil\_conv\_5x5} & \text{c\_k} \\
\end{array}
$$
$$
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
$$

One-hot adjacency encoding of a DARTS cell

Structure-aware encodings: Adjacency matrix-based

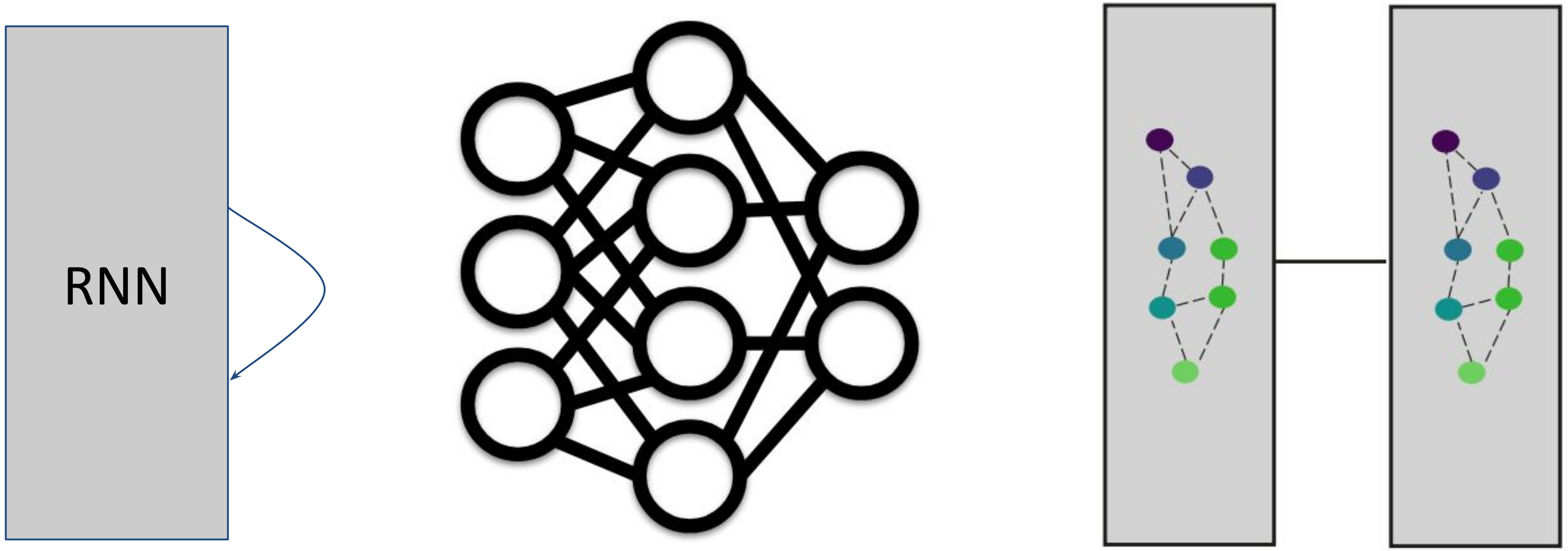


One-hot adjacency encoding

Categorical adjacency encoding

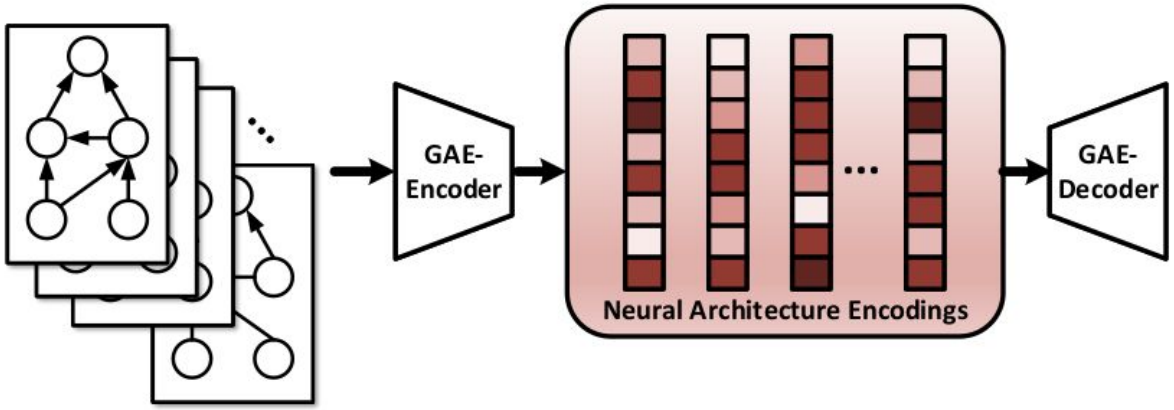
Colin White, Willie Neiswanger, Sam Nolen, Yash Savani., A Study on Encodings for Neural Architecture Search, NeurIPS 2020

Structure-aware encodings: LSTM/MLP/GCN

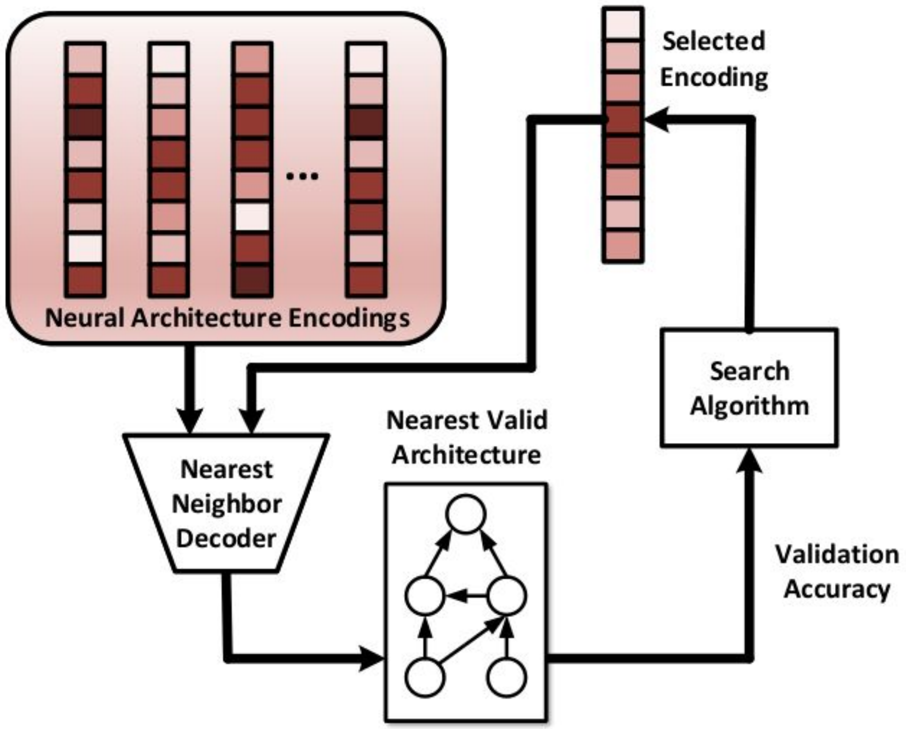


Different types of architecture encoders

Structure-aware encodings: Unsupervised Pre-training

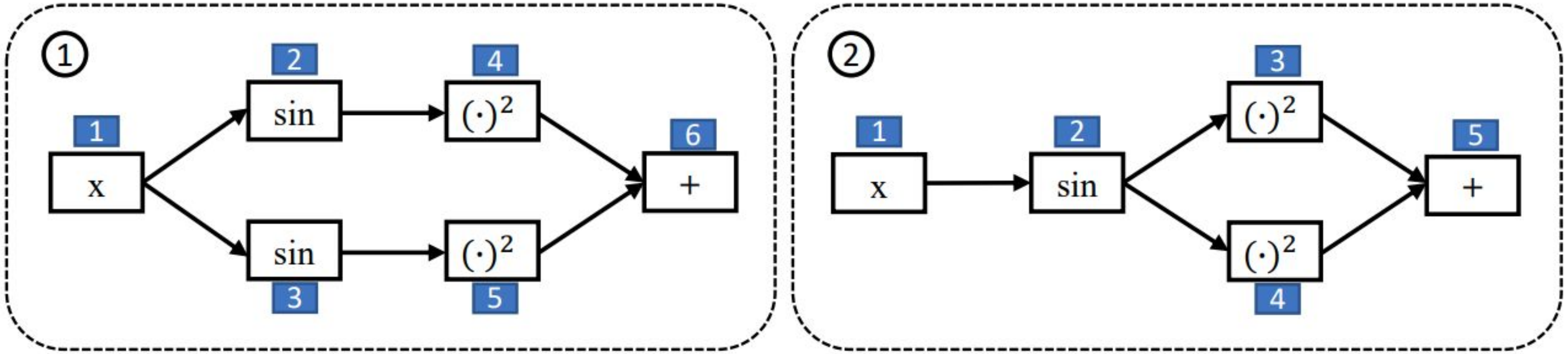


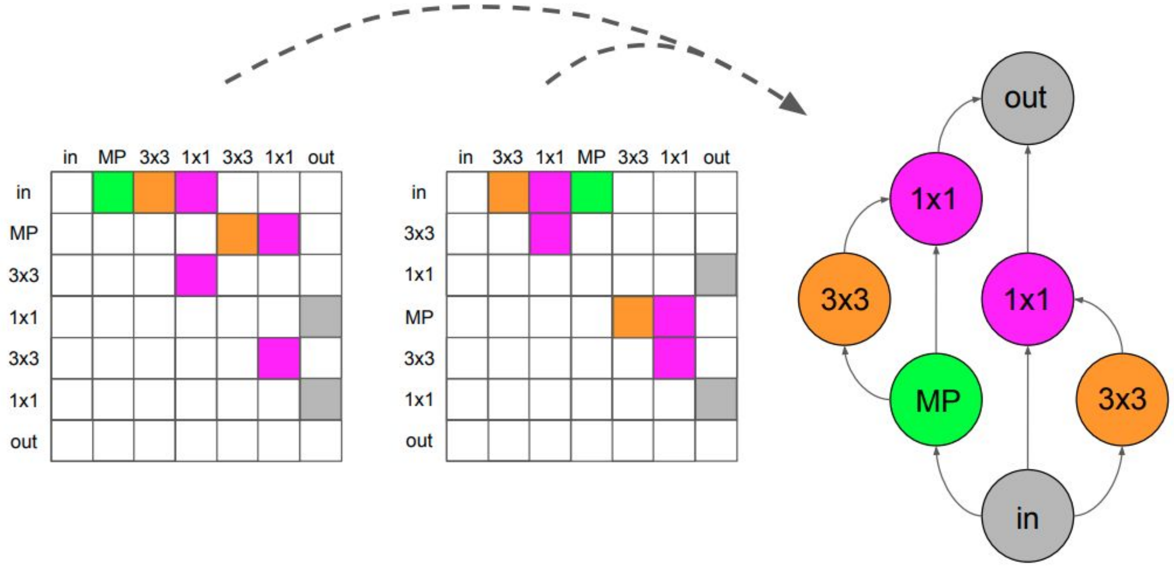
arch2vec: Unsupervised pretraining of neural architecture encodings



Neural architecture search using pretrained architecture encodings

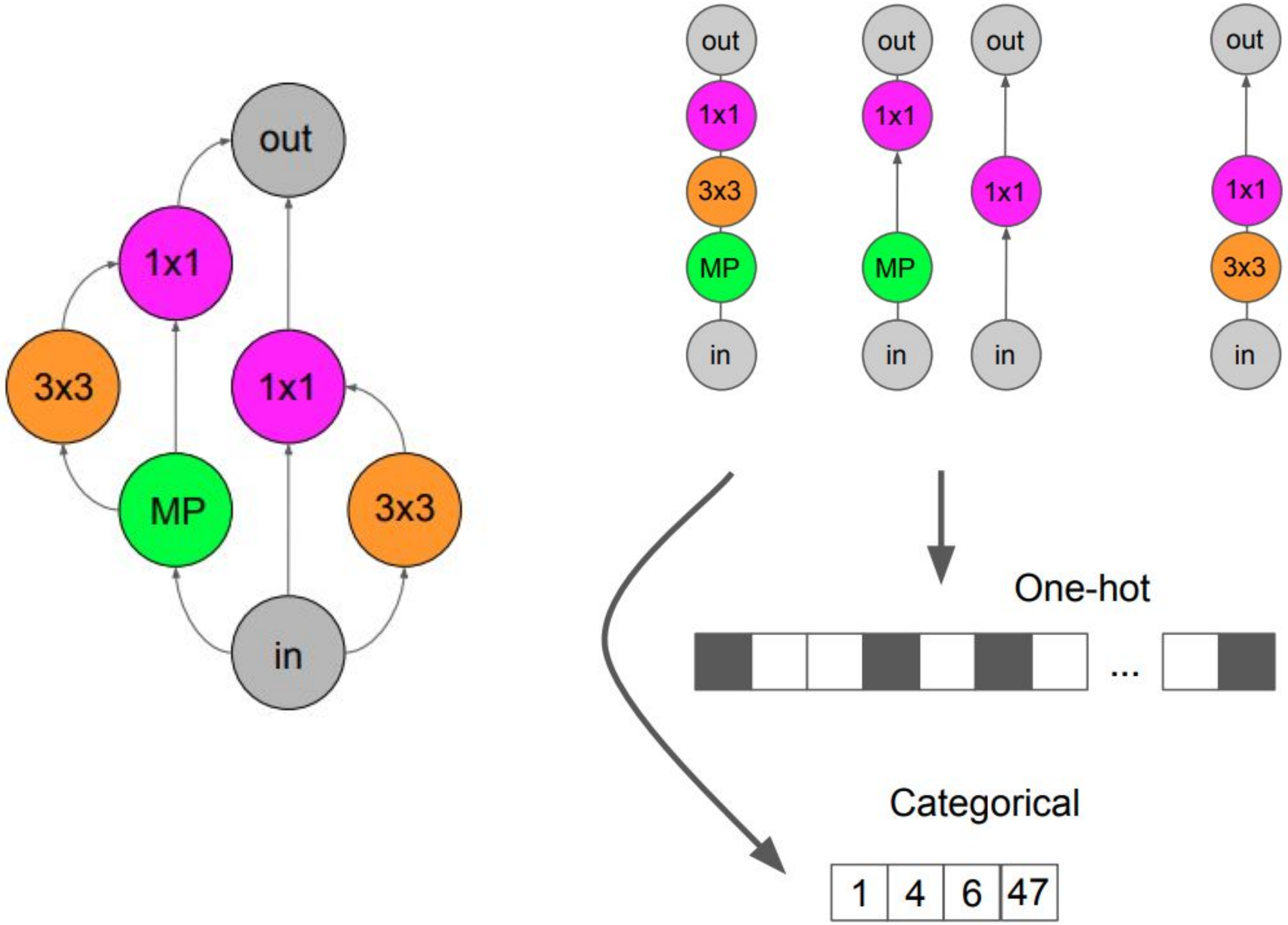
Drawbacks of Structure-aware encodings





Muhan Zhang, Shali Jiang, Zhicheng Cui, Roman Garnett, Yixin Chen., D-VAE: A Variational Autoencoder for Directed Acyclic Graphs, NeurIPS 2019
Colin White, Willie Neiswanger, Sam Nolen, Yash Savani., A Study on Encodings for Neural Architecture Search, NeurIPS 2020

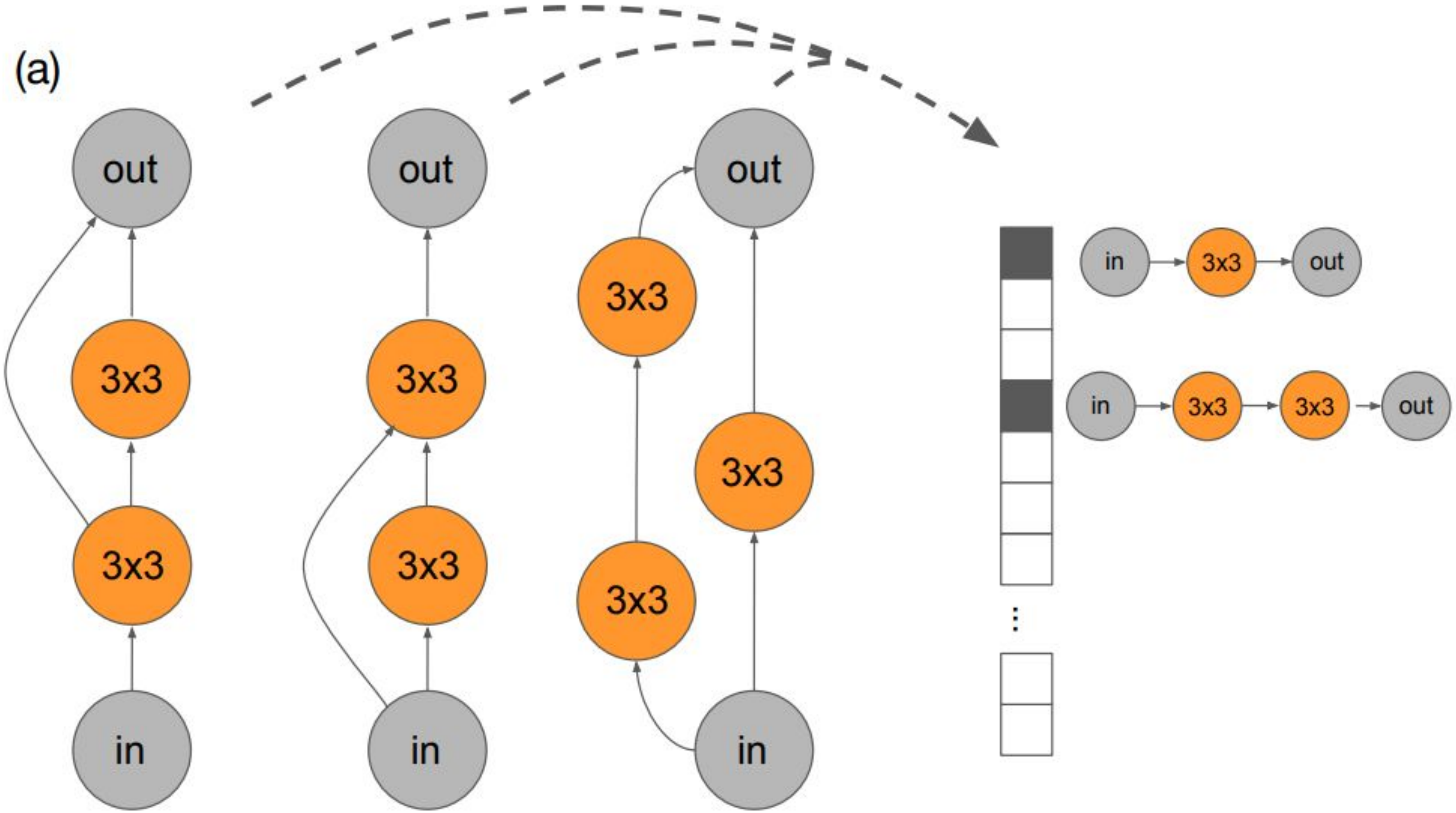
Computation-aware encodings: Path Encodings



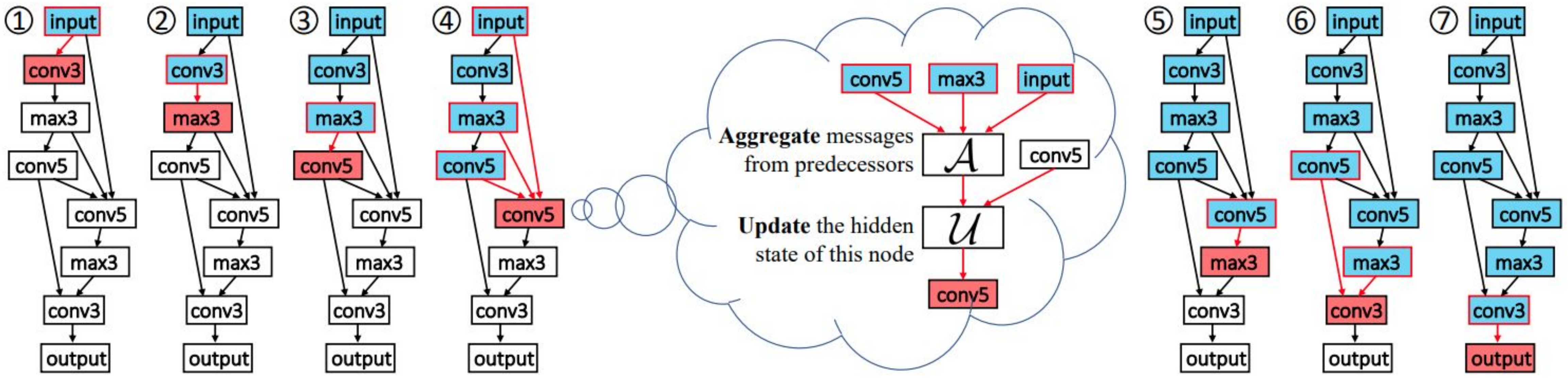
One-hot / Categorical path encoding

Colin White, Willie Neiswanger, Sam Nolen, Yash Savani., A Study on Encodings for Neural Architecture Search, NeurIPS 2020

Advantages of Computation-aware Encodings

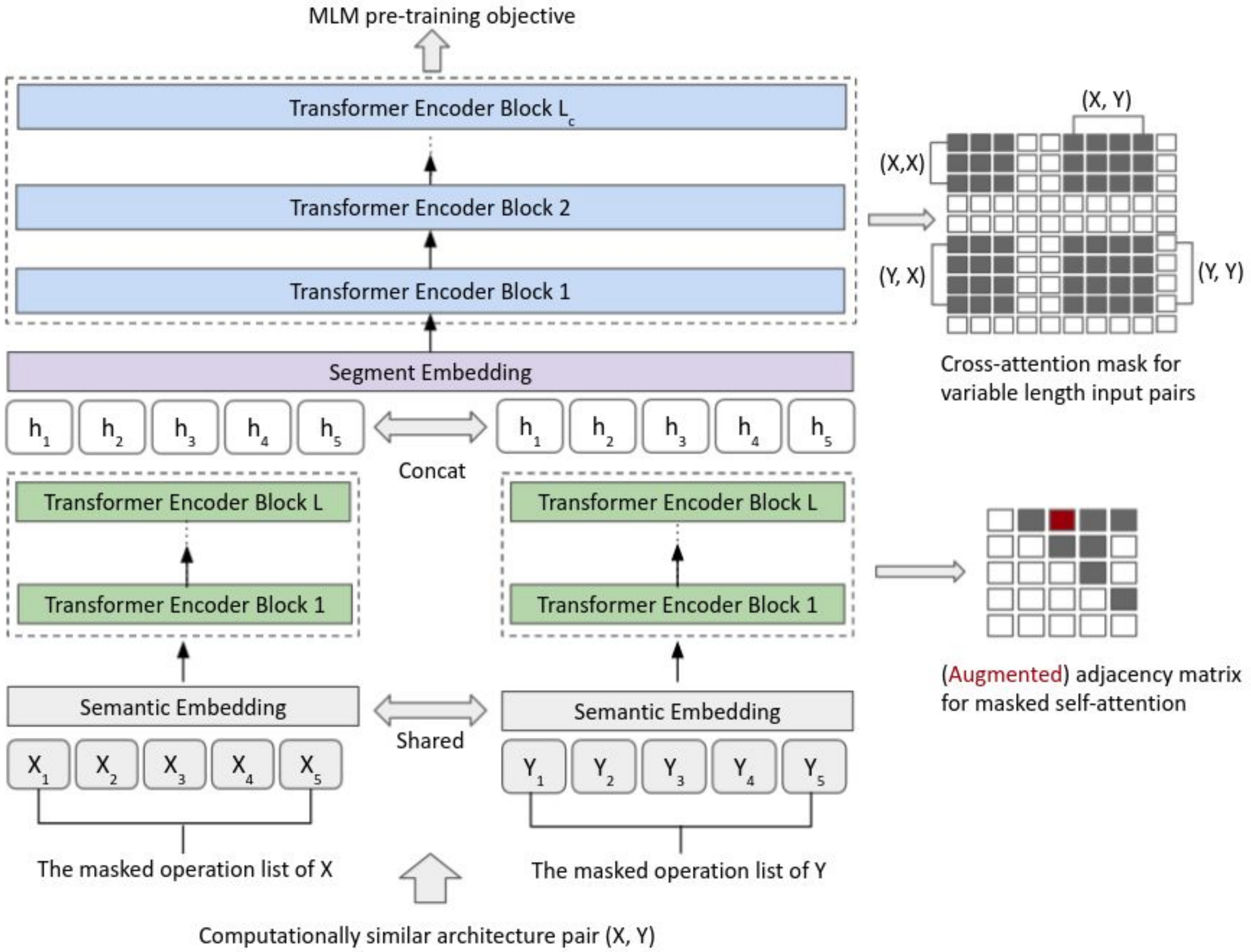


Computation-aware encodings: D-VAE



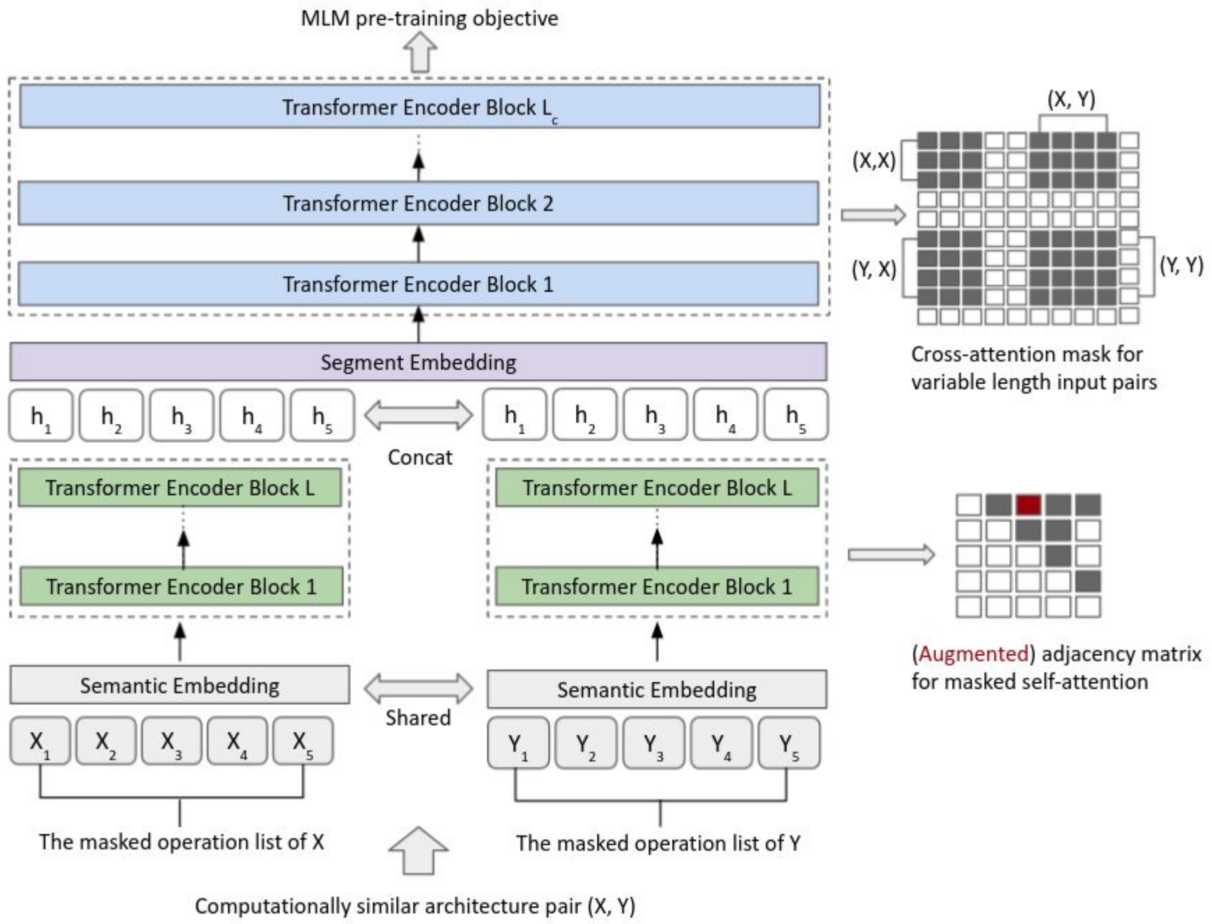
Muhan Zhang, Shali Jiang, Zhicheng Cui, Roman Garnett, Yixin Chen., D-VAE: A Variational Autoencoder for Directed Acyclic Graphs, NeurIPS 2019

Our Proposed Method: CATE



Why Pairwise Sampling

Conv 3x3 -> Conv 1x1 -> Conv 5x5 -> ~~Conv 1x1~~ -> Max Pool -> …
Conv 3x3 -> Conv 1x1 -> Conv 5x5 -> ? -> Max Pool -> …



Attention Mask

Algorithm 1 Floyd Algorithm

Input: the node set $\mathcal{V}$, the adjacent matrix $\mathbf{A}$

$\tilde{\mathbf{A}} \leftarrow \mathbf{A}$

for $k \in \mathcal{V}$ **do**

  for $i \in \mathcal{V}$ **do**

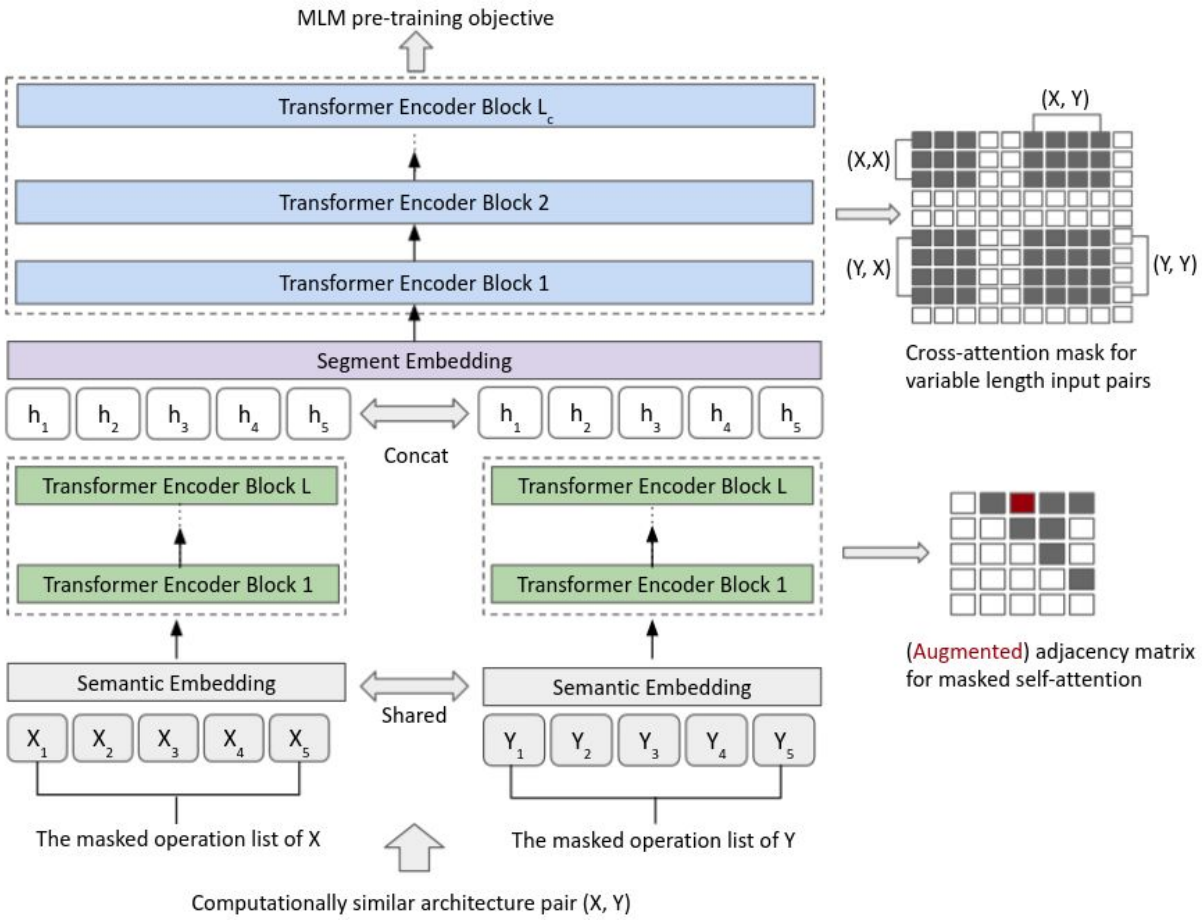
    for $j \in \mathcal{V}$ **do**

      $\tilde{\mathbf{A}}_{i,j} \mathrel{|}= \tilde{\mathbf{A}}_{i,k} \;\&\; \tilde{\mathbf{A}}_{k,j}$

Output: $\tilde{\mathbf{A}}$

$$\mathbf{M}^{Direct}_{i,j} = \begin{cases} 0, & if \quad A_{i,j} = 1 \\ -\infty, & if \quad A_{i,j} = 0 \end{cases}$$

$$\mathbf{M}^{Indirect}_{i,j} = \begin{cases} 0, & if \quad \tilde{A}_{i,j} = 1 \\ -\infty, & if \quad \tilde{A}_{i,j} = 0 \end{cases}$$



Encoding-dependent NAS Subroutines

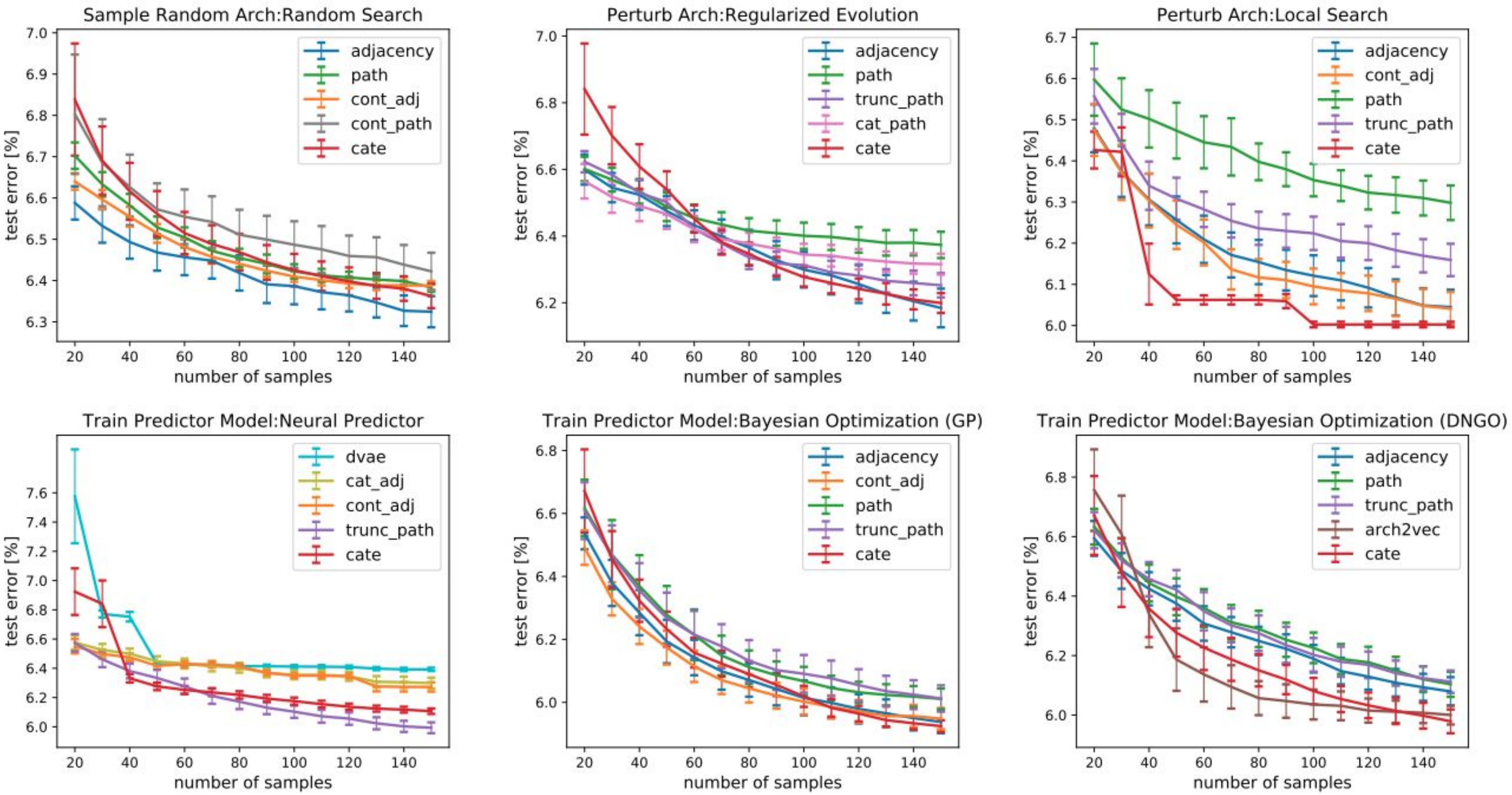
12 different encodings:

- One-hot/Categorical/Continuous adjacency matrix encoding (3)
- One-hot/Categorical/Continuous path encoding (3)
- The truncated counterparts (3)
- D-VAE
- arch2vec
- CATE

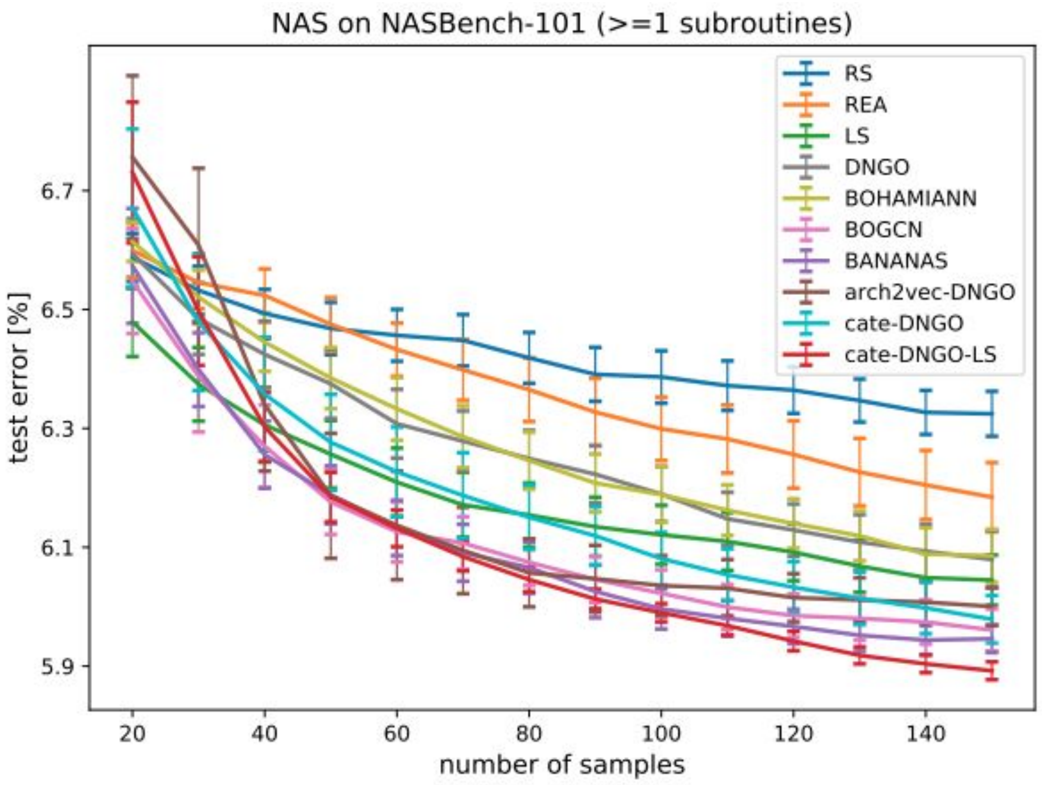
3 NAS subroutine:

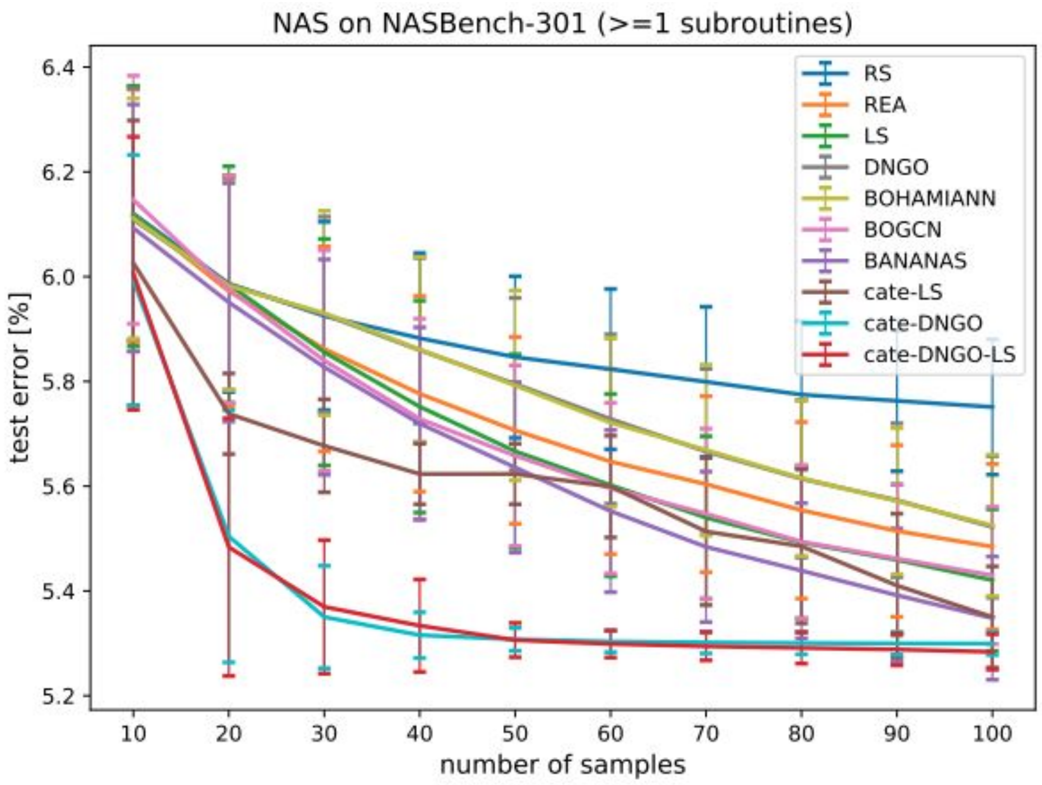
- Sample random architecture: random search
- Perturb architecture: regularized evolution, local search
- Train predictor: neural predictor, BO with GP, BO with DNGO

Comparison between CATE and other encoding schemes



Comparison between CATE and other NAS methods





Evaluation on DARTS without surrogate models

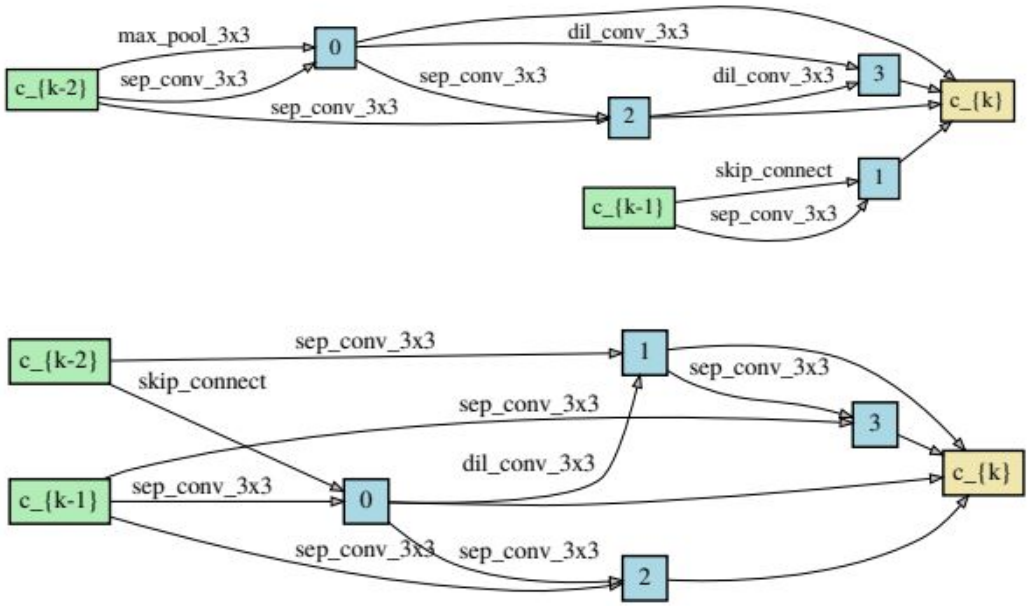


Figure 4. Top: Best found cell from CATE-DNGO-LS given the budget of 100 samples. Bottom: Best found cell from CATE-DNGO-LS given the budget of 300 samples.

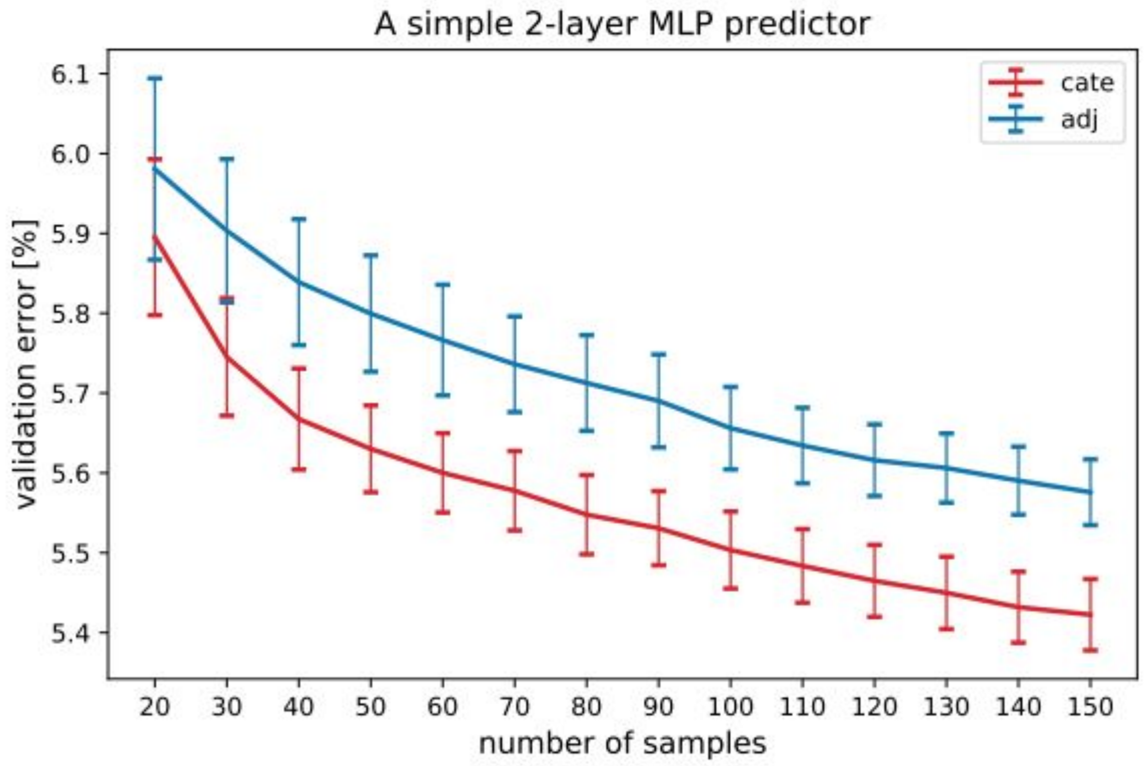
NAS Methods	Avg. Test Error (%)	Params (M)	Search Cost (GPU days)
RS (Li & Talwalkar, 2019)	3.29 ± 0.15	**3.2**	4
DARTS (Liu et al., 2019a)	2.76 ± 0.09	3.3	4
BANANAS (White et al., 2021)	2.67 ± 0.07	3.6	11.8
arch2vec-BO (Yan et al., 2020)	2.56 ± 0.05	3.6	9.2
CATE-DNGO-LS (small budget)	2.55 ± 0.08	3.5	**3.3**
CATE-DNGO-LS (large budget)	$\mathbf{2.46 \pm 0.05}$	4.1	10.3

Table 2. NAS results in DARTS search space using CIFAR-10.

NAS Methods	Params (M)	Mult-Adds (M)	Top-1 Test Error (%)
SNAS (Xie et al., 2019b)	4.3	522	27.3
DARTS (Liu et al., 2019a)	4.7	574	26.7
BayesNAS (Zhou et al., 2019)	**4.0**	**440**	26.5
arch2vec-BO (Yan et al., 2020)	5.2	580	25.5
BANANAS (ours)	5.1	576	26.3
CATE-DNGO-LS (small budget)	5.0	556	26.1
CATE-DNGO-LS (large budget)	5.8	642	**25.0**

Table 3. Transfer learning results on ImageNet.

Evaluation on Outside Search Space



Ablation Study

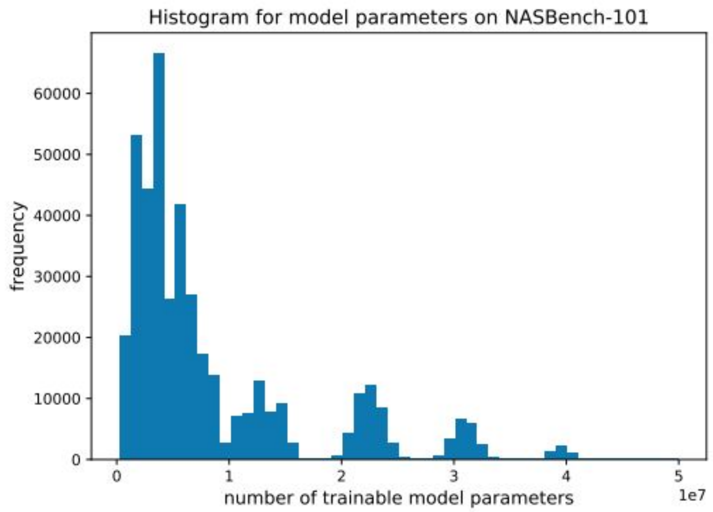


Figure 6. Histogram of model parameters on NAS-Bench-101.

d_{ff} \ L_c	6	12	24
64	6.07	5.99	5.95
128	6.01	**5.94**	5.95
256	5.97	**5.94**	**5.94**

Table 5. Effects of L_c and d_{ff} on pretraining CATE.

δ \ K	1	2	4	8
1×10^6	6.02	5.95	5.99	5.95
2×10^6	6.02	**5.94**	6.04	5.96
4×10^6	**5.94**	6.03	6.05	5.99
8×10^6	6.05	6.04	6.11	6.04

Table 4. Effects of δ and K on architecture pair sampling.

Mask type	**NAS-Bench-101**	**NAS-Bench-301**
Direct	6.03	5.35
Indirect	**5.94**	**5.30**

Table 6. Direct/Indirect dependency mask selection.

Conclusion

- **A non-contrastive, pairwise pre-training method to learn computation-aware encodings with cross-attention Transformers**
- **Competitive under all encoding-dependent NAS subroutines in both small and large search spaces**
- **Superior generalization ability beyond the search space on which it was trained**

For more detailed information and code, please refer to our paper:
https://arxiv.org/abs/2102.07108

Thank You