CATE: Computation-aware Neural Architecture Encoding with Transformers
Shen Yan¹, Kaiqiang Song²,³, Fei Liu³, Mi Zhang¹
Michigan State University, Tencent AI Lab, University of Central Florida

Introduction
Neural Architecture Search has recently drawn considerable attention. While majority of the prior work focuses on either constructing new search spaces or designing efficient search and evaluation methods, some of the most recent work [1,2] sheds light on the importance of architecture encoding on the subroutines in the NAS pipeline as well as overall performance of NAS.

Background
• Structure-aware encodings such as adjacency matrix-based encodings may not be computationally unique. A same encoding can have many different representations using adjacency matrix.

• Computation-aware encodings [2,3] are proposed to tackle the above drawbacks. However, it suffers scalability and generalization issues.

Our proposed method: CATE
• Use computationally similar architecture pairs as input. The model is trained to predict masked operators given the joint information. Extracted architecture encodings are used for the downstream encoding-dependent NAS subroutines.

Results
• Comparison between CATE and other encodings under different NAS subroutines on NAS-Bench-101.

• Comparison between CATE with other NAS methods on NAS-Bench-101 (left) and NAS-Bench-301 (right):

• Generalization performance on out-of-training search space:

Background

Key difference with BERT:
● No contextual bias in the prediction as long as it is a valid graph
Conv 3x3 -> Conv 5x5 -> Connected -> Max Pool -> ...
Conv 3x3 -> Conv 5x5 -> ... -> Max Pool -> ...

● Cannot use the fully-visible attention mask as it does not reflect the single directional flow (e.g. direct, acyclic, single-in-single-out)