# ShiftAddNAS: Hardware-Inspired Search for More Accurate and Efficient Neural Networks

**Haoran You**[1], Baopu Li[2], Huihong Shi[1], Yonggan Fu[1], Yingyan Lin[1]

***ICML 2022***

[1]**Rice University**, [2]Oracle Corporation

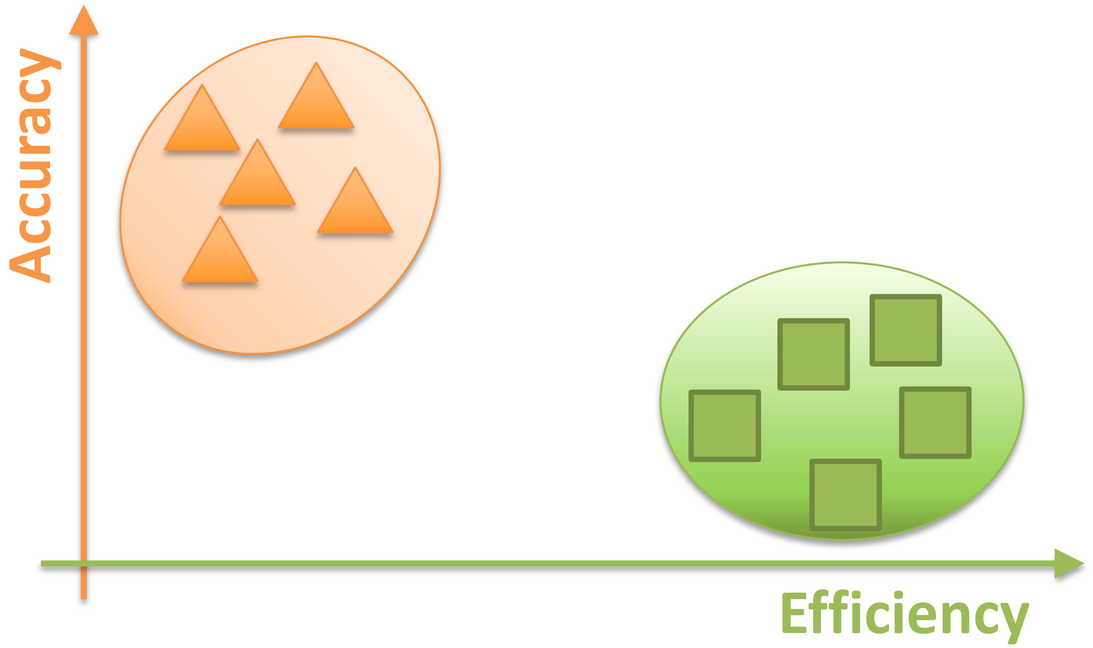
Efficient and Intelligent Computing Lab

# ShiftAddNAS: Background and Motivation

- **Two branches of SOTA DNN design: Trade off accuracy and efficiency**
  - ***Multiplication-based* DNNs, e.g., CNNs, Transformers**
    - ☺ **Achieve unprecedented task accuracy**
    - ☹ **Power hungry → Challenge their deployment to edge devices**

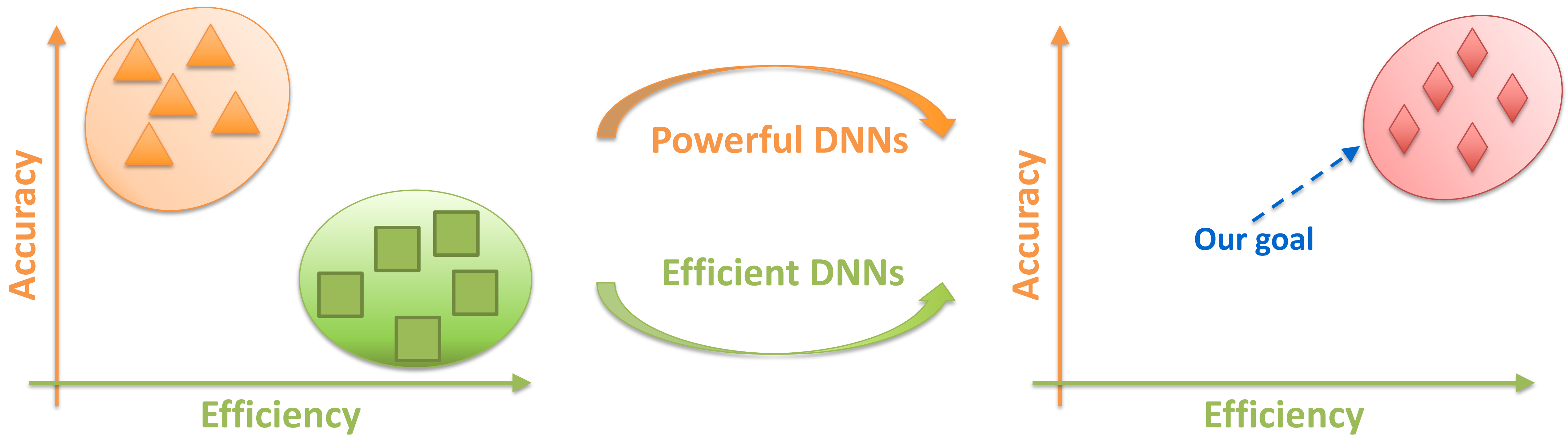
# ShiftAddNAS: Background and Motivation

- Two branches of SOTA DNN design: Trade off accuracy and efficiency
  - *Multiplication-based DNNs*, e.g., CNNs, Transformers
    - ☺ Achieve unprecedented task accuracy
    - ☹ Power hungry → Challenge their deployment to edge devices
  - *Multiplication-free DNNs*, e.g., ShiftNet, AdderNet, ShiftAddNet
    - ☺ Efficient and favor their deployment to edge devices
    - ☹ Under-perform their multiplication-based counterparts in terms of task accuracy



# ShiftAddNAS: Background and Motivation

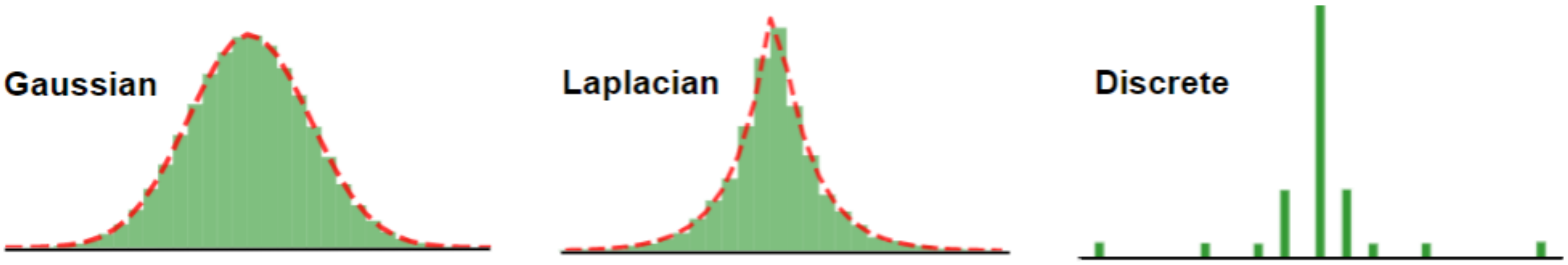
- **Motivation of ShiftAddNAS**
  - **Enable automated search for hybrid network architecture to marry the best of both worlds**
    - ☺ **Multiplication-based operators (e.g., Conv & Attention) → High accuracy**
    - ☺ **Multiplication-free operators (e.g., Shift & Add) → High efficiency**



# ShiftAddNAS: Tackled Challenges

- Motivation of ShiftAddNAS
  - Enable automated search for hybrid network architecture to marry the best of both worlds
    - Multiplication-based operators (e.g., Conv & Attention) → High accuracy
    - Multiplication-free operators (e.g., Shift & Add) → High efficiency

- Associated Challenges
  - How to construct an effective hybrid search space?
  - More operators → larger SuperNets, but SOTA weight sharing strategy is not applicable




(a) Weights in Conv (b) Weights in Add (c) Weights in Shift

# ShiftAddNAS: Our Contributions

**For the first time, we**

- **Develop ShiftAddNAS, featuring a hybrid search space that incorporates both *multiplication-based* and *multiplication-free* operators**

- **Propose a new heterogeneous weight sharing strategy that enables automated search for hybrid operators with heterogeneous weight distributions**

- **Conduct extensive experiments on both CV and NLP tasks to validate the effectiveness of our proposed ShiftAddNAS framework**

# Contribution 1: Hybrid Search Space and SuperNet

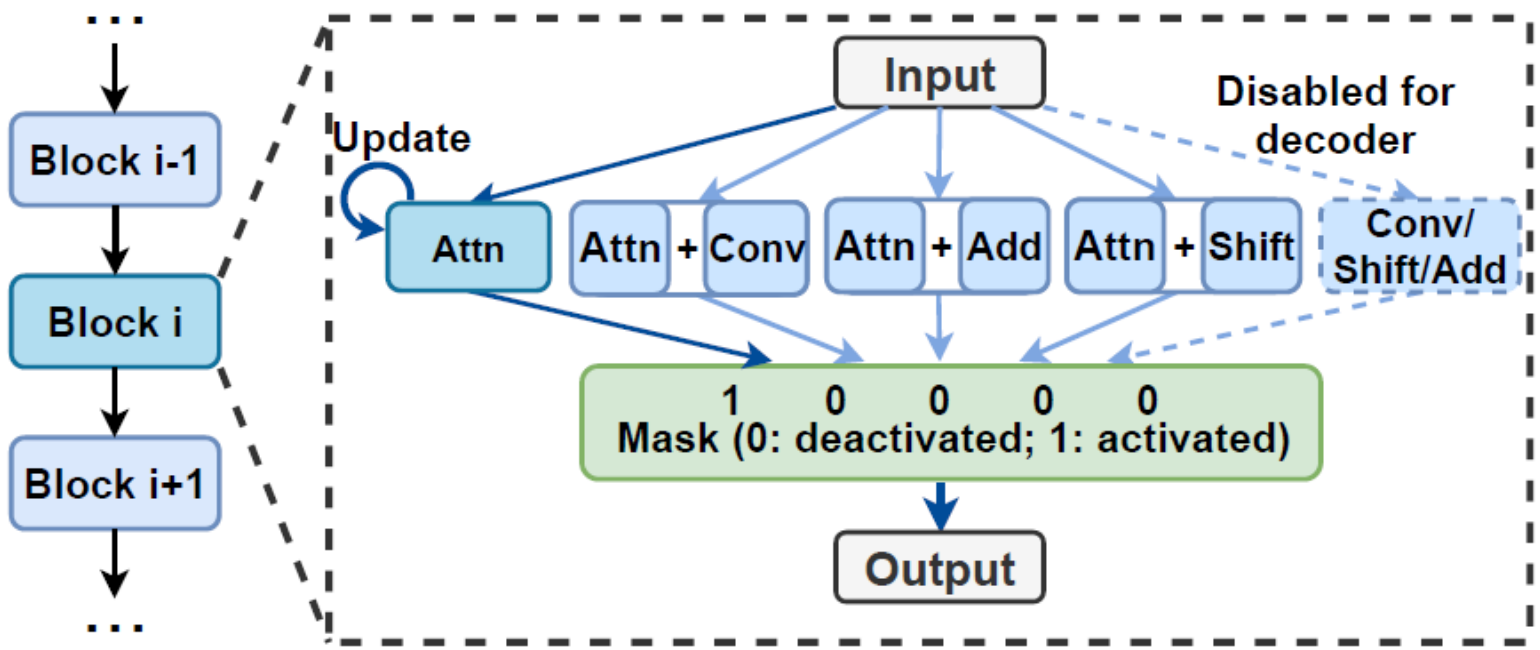
- Search space for NLP tasks
  - Seven different blocks
    - **Attn, Conv, Shift, and Add**
    - **Attn+Conv, Attn+Add, and Attn+Shift**
  - Elastic dimensions **for MLPs, embeddings, and heads**

| | |
|---|---|
| Encoder block types | [Attn, Attn+Conv, Attn+Shift]<br>[Attn+Add, Conv, Shift, Add] |
| Decoder block types | [Attn, Attn+Conv]<br>[Attn+Shift, Attn+Add] |
| Num. of decoder blocks | [6, 5, 4, 3, 2, 1] |
| Elastic embed. Dim. | [1024, 768, 512] |
| Elastic head number | [16, 8, 4] |
| Elastic MLP dim. | [4096, 3072, 2048, 1024] |
| Arbitrary Attn | [3, 2, 1] |

**The Search Space for NLP Tasks**

# Contribution 1: Hybrid Search Space and SuperNet

- Search space for NLP tasks
  - Seven different blocks
    - Attn, Conv, Shift,  and Add
    - Attn+Conv, Attn+Add, and Attn+Shift
  - Elastic dimensions for MLPs, embeddings, and heads




The SuperNet for NLP Tasks

# Contribution 1: Hybrid Search Space and SuperNet

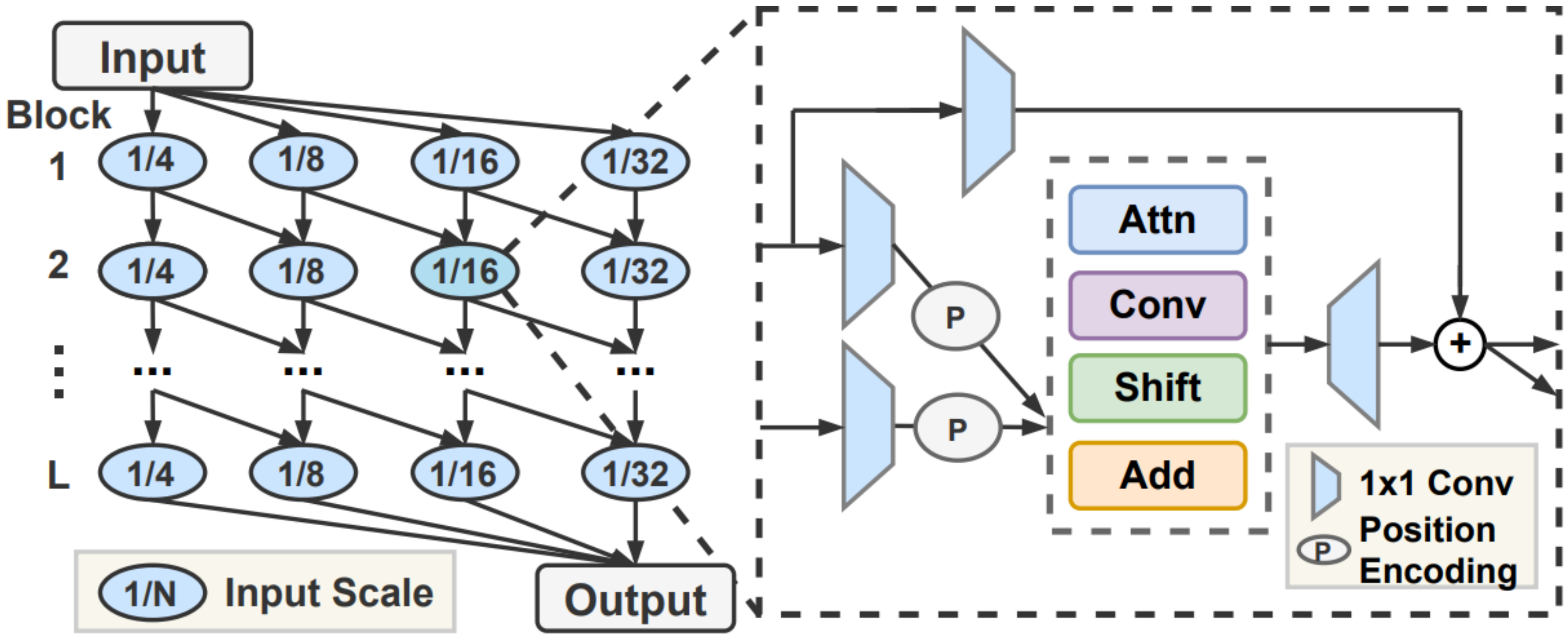
- **Search space for NLP tasks**
- **Search space for CV tasks**
  - **Multi-resolution**
    - **Various spatial resolutions or scales are essential for CV tasks**

| | |
|---|---|
| Block types | [Attn, Conv, Shift, Add] |
| Num. of $56^2 \times 128$ blocks | [1, 2, 3, 4] |
| Num. of $28^2 \times 256$ blocks | [1, 2, 3, 4] |
| Num. of $14^2 \times 512$ blocks | [3, 4, 5, 6, 7] |
| Num. of $7^2 \times 1024$ blocks | [4, 5, 6, 7, 8, 9] |

**The Search Space for CV Tasks**

# Contribution 1: Hybrid Search Space and SuperNet

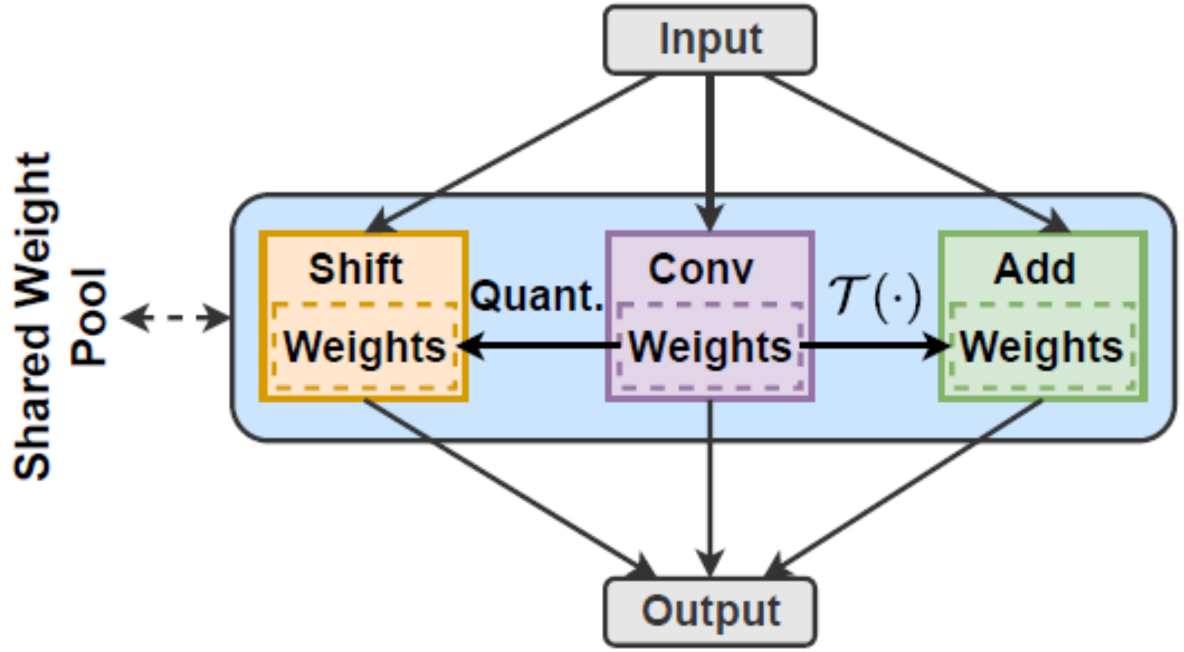
- Search space for NLP tasks
- Search space for CV tasks
  - Multi-resolution
    - Various spatial resolutions or scales are essential for CV tasks



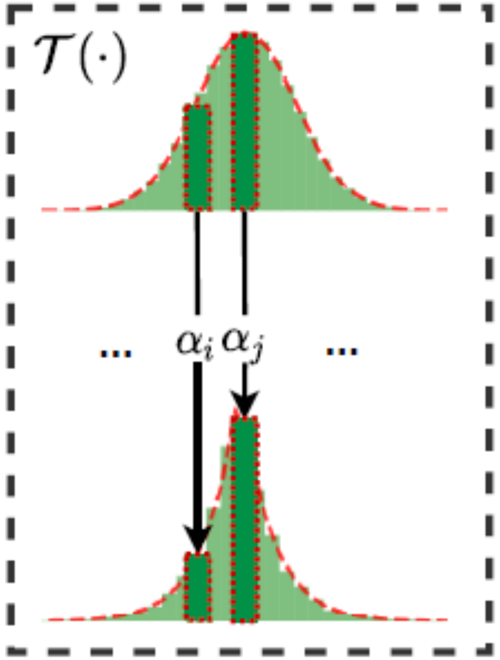

The SuperNet for CV Tasks

# Contribution 2: Heterogenous Weight Sharing Strategy

- **One-shot NAS with heterogeneous weight sharing**
  - **Weight sharing among Conv, Add, and Shift blocks**




(a) Heterogenous Weight Sharing Strategy



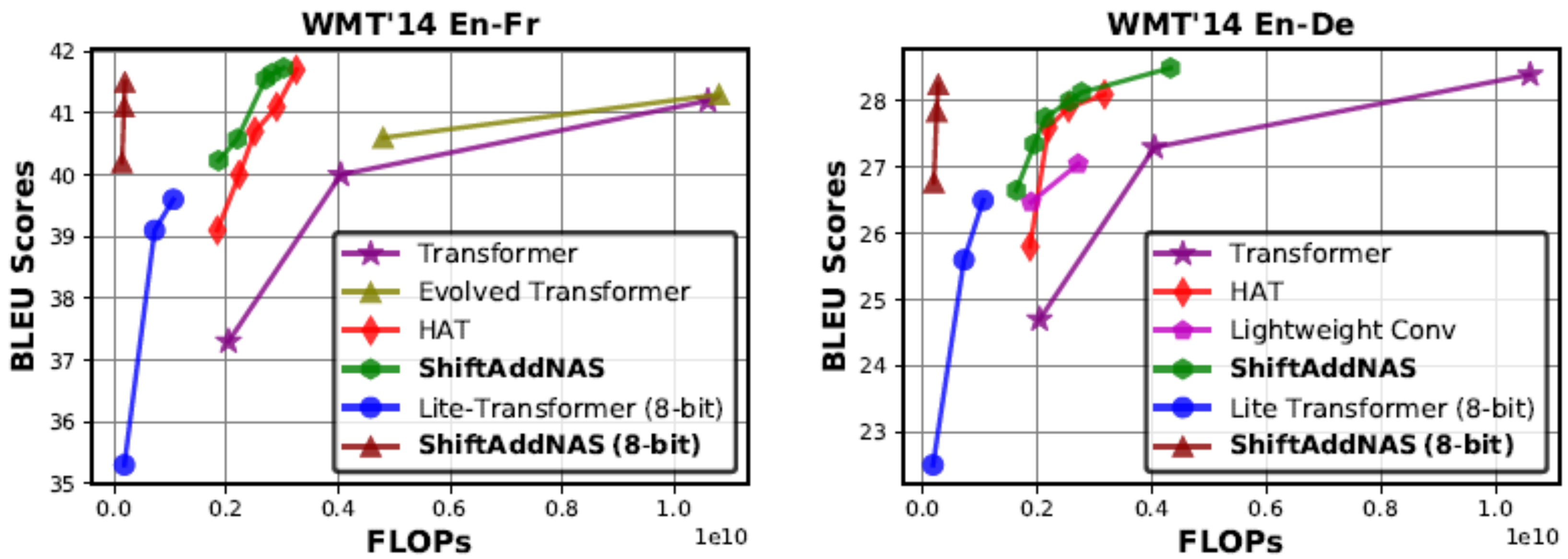

(b) Transformation Kernel

$$\mathcal{L}_{\mathcal{S}} = \mathcal{L}_{CE} + \mathcal{L}_{KL} = -\frac{1}{N}\sum_{i=1}^{N} P(y_i|x_i)\log(P(\hat{y}_i|x_i))$$
$$+ \mathcal{D}_{KL}(P_{\text{Conv}}(\mathbf{W}_{\mathcal{S}}) \;||\; \mathcal{N}(0, I)) + \mathcal{D}_{KL}(P_{\text{Add}}(\mathcal{T}(\mathbf{W}_{\mathcal{S}})) \;||\; \mathcal{L}_p(0, \lambda)),$$

# ShiftAddNAS: Experimental Setting

- NLP tasks
  - Two datasets
    - WMT’14 English to French (En-Fr)
    - WMT’14 English to German (En-De)
  - Five evaluation metrics
    - BLEU score
    - Number of parameters/FLOPs
    - Hardware energy and latency
  - Four baselines
    - Transformer
    - Lightweight Conv
    - Lite Transformer
    - HAT
- CV tasks
  - One dataset: ImageNet
  - Five evaluation metrics
    - Accuracy
    - Number of parameters/MACs
    - Hardware energy and latency
  - Four categories of baselines
    - Multiplication-free NNs
      - AdderNet, DeepShift, BNN
    - CNNs
      - ResNet, SENet
    - Transformer
      - ViT, DeiT, VITAS, Autoformer
    - CNN-Transformer
      - BoT, HR-NAS, BossNAS

# ShiftAddNAS: Experimental Results for NLP Tasks




BLEU scores vs. FLOPs of ShiftAddNAS over SOTA baselines on NLP tasks.

ShiftAddNAS vs. SOTA baselines in terms of accuracy and efficiency on NLP tasks.

| | WMT'14 En-Fr | | | | | WMT'14 En-De | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Params | FLOPs | BLEU | Latency | Energy | Params | FLOPs | BLEU | Latency | Energy |
| Transformer | 176M | 10.6G | 41.2 | 130ms | 214mJ | 176M | 10.6G | 28.4 | 130ms | 214mJ |
| Evolved Trans. | 175M | 10.8G | 41.3 | - | - | 47M | 2.9G | 28.2 | - | - |
| HAT | 48M | 3.4G | 41.4 | 49ms | 81mJ | 44M | 2.7G | 28.2 | 42ms | 69mJ |
| **ShiftAddNAS** | **46M** | **3.0G** | **41.8** | **43ms** | **71mJ** | **43M** | **2.7G** | **28.2** | **40ms** | **66mJ** |
| HAT | 46M | 2.9G | 41.1 | 42ms | 69mJ | 36M | 2.2G | 27.6 | 34ms | 56mJ |
| **ShiftAddNAS** | **41M** | **2.7G** | **41.6** | **39ms** | **64mJ** | **33M** | **2.1G** | **27.8** | **31ms** | **52mJ** |
| HAT | 30M | 1.8G | 39.1 | 29ms | 48mJ | 25M | **1.5G** | 25.8 | 24ms | 40mJ |
| **ShiftAddNAS** | **29M** | **1.8G** | **40.2** | **16ms** | **45mJ** | **25M** | 1.6G | **26.7** | **24ms** | **40mJ** |
| Lite Trans. (8-bit) | 17M | 1G | 39.6 | 19ms | 31mJ | 17M | 1G | 26.5 | 19ms | 31mJ |
| **ShiftAddNAS (8-bit)** | **11M** | **0.2G** | **41.5** | **11ms** | **16mJ** | **17M** | **0.3G** | **28.3** | **16ms** | **24mJ** |
| Lite Trans. (8-bit) | 12M | 0.7G | 39.1 | 14ms | 24mJ | 12M | 0.7G | 25.6 | 14ms | 24mJ |
| **ShiftAddNAS (8-bit)** | **10M** | **0.2G** | **41.1** | **10ms** | **15mJ** | **12M** | **0.2G** | **26.8** | **9.2ms** | **14mJ** |

- **Overall Improvement on NLP**
  - **ShiftAddNAS achieves up to +2 BLEU scores improvement and 69.1% and 69.2% energy and latency savings**

# ShiftAddNAS: Experimental Results for CV Tasks

Comparison with SOTA baselines on ImageNet classification task.

| Model | Top-1 Acc. | Top-5 Acc. | Params | Res. | MACs | #Mult. | #Add | #Shift | Model Type |
|---|---|---|---|---|---|---|---|---|---|
| BNN | 55.8% | 78.4% | 26M | $224^2$ | 3.9G | 0.1G | 3.9G | 3.8G | Mult.-free |
| AdderNet | 74.9% | 91.7% | 26M | $224^2$ | 3.9G | 0.1G | 7.6G | 0 | Mult.-free |
| AdderNet-PKKD | 76.8% | 93.3% | 26M | $224^2$ | 3.9G | 0.1G | 7.6G | 0 | Mult.-free |
| DeepShift-Q | 70.7% | 90.2% | 26M | $224^2$ | 3.9G | 0.1G | 3.9G | 3.8G | Mult.-free |
| DeepShift-PS | 71.9% | 90.2% | 52M | $224^2$ | 3.9G | 0.1G | 3.9G | 3.8G | Mult.-free |
| ResNet-50 | 76.1% | 92.9% | 26M | $224^2$ | 3.9G | 3.9G | 3.9G | 0 | CNN |
| ResNet-101 | 77.4% | 94.2% | 45M | $224^2$ | 7.6G | 7.6G | 7.6G | 0 | CNN |
| SENet-50 | 79.4% | 94.6% | 26M | $224^2$ | 3.9G | 3.9G | 3.9G | 0 | CNN |
| SENet-101 | 81.4% | 95.7% | 45M | $224^2$ | 7.6G | 7.6G | 7.6G | 0 | CNN |
| ViT-B/16 | 77.9% | - | 86M | $384^2$ | 18G | 18G | 17G | 0 | Transformer |
| ViT-L/16 | 76.5% | - | 304M | $384^2$ | 64G | 64G | 63G | 0 | Transformer |
| DeiT-T | 74.5% | - | 6M | $224^2$ | 1.3G | 1.3G | 1.3G | 0 | Transformer |
| DeiT-S | 81.2% | - | 22M | $224^2$ | 4.6G | 4.6G | 4.6G | 0 | Transformer |
| VITAS | 77.4% | 93.8% | 13M | $224^2$ | 2.7G | 2.7G | 2.7G | 0 | Transformer |
| Autoformer-S | 81.7% | 95.7% | 23M | $224^2$ | 5.1G | 5.1G | 5.1G | 0 | Transformer |
| BoT-50 | 78.3% | 94.2% | 21M | $224^2$ | 4.0G | 4.0G | 4.0G | 0 | CNN + Trans. |
| BoT-50 + SE | 79.6% | 94.6% | 21M | $224^2$ | 4.0G | 4.0G | 4.0G | 0 | CNN + Trans. |
| HR-NAS | 77.3% | - | 6.4M | $224^2$ | 0.4G | 0.4G | 0.4G | 0 | CNN + Trans. |
| BossNAS-T0 | 80.5% | 95.0% | 38M | $224^2$ | 3.5G | 3.5G | 3.5G | 0 | CNN + Trans. |
| BossNAS-T0 + SE | 80.8% | 95.2% | 38M | $224^2$ | 3.5G | 3.5G | 3.5G | 0 | CNN + Trans. |
| **ShiftAddNAS-T0** | **82.1%** | **95.8%** | **30M** | $224^2$ | **3.7G** | **2.7G** | **3.8G** | **1.0G** | Hybrid |
| **ShiftAddNAS-T0↑** | **82.6%** | **96.2%** | **30M** | $256^2$ | **4.9G** | **3.6G** | **4.9G** | **1.4G** | Hybrid |
| T2T-ViT-19 | 81.9% | - | 39M | $224^2$ | 8.9G | 8.9G | 8.9G | 0 | Transformer |
| TNT-S | 81.3% | 95.6% | 24M | $224^2$ | 5.2G | 5.2G | 5.2G | 0 | Transformer |
| Autoformer-B | 82.4% | 95.7% | 54M | $224^2$ | 11G | 11G | 11G | 0 | Transformer |
| BoTNet-S1-59 | 81.7% | 95.8% | 28M | $224^2$ | 7.3G | 7.3G | 7.3G | 0 | CNN + Trans. |
| BossNAS-T1 | 82.2% | 95.8% | 38M | $224^2$ | 8.0G | 8.0G | 8.0G | 0 | CNN + Trans. |
| **ShiftAddNAS-T1** | **82.7%** | **96.1%** | **30M** | $224^2$ | **6.4G** | **5.4G** | **6.4G** | **1.0G** | Hybrid |
| **ShiftAddNAS-T1↑** | **83.0%** | **96.4%** | **30M** | $256^2$ | **8.5G** | **7.1G** | **8.5G** | **1.4G** | Hybrid |

- **Overall Improvement on CV**
  - **ShiftAddNAS on average offers a +0.8% ~ +7.7% higher accuracy and 24% ~ 93% energy savings**

# Summary

**For the first time, we**

- **Develop ShiftAddNAS, featuring a hybrid search space that incorporates both *multiplication-based* and *multiplication-free* operators**
- **Propose a new heterogeneous weight sharing strategy that enables automated search for hybrid operators with heterogeneous weight distributions**
- **Conduct extensive experiments on both CV and NLP tasks to validate the effectiveness of our proposed ShiftAddNAS framework**

**Open-source Code:**
https://github.com/RICE-EIC/ShiftAddNAS

NSF
NIH National Institute of Mental Health