FlexFlow
Distributed DNN Training on Legion

- Memory Optimization
- Generalized Pipeline Parallelism
- MoE Inference
- JAX Interface
- ECP-CANDLE
- Cryo-EM
- Hugging Face
FlexFlow

Distributed DNN Training
Data Parallel
Data Parallel
Data Parallel

Input → MatMul 1 → MatMul 2 → Output

Weight 1

Weight 2
Data Parallel

Input

MatMul 1

Weight 1

MatMul 2

Weight 2

Output

0 3/3 9
Data Parallel
Data Parallel

Input

MatMul 1

Weight 1

Weight 2

MatMul 2

Output

0

3

/3

9
Data Parallel
Data Parallel
Data Parallel
Data Parallel

```
Data Parallel
Input
MatMul 1
MatMul 2
Weight 1
Weight 2
MatMul 1
MatMul 2
Weight 1
Weight 2
MatMul 1
MatMul 2
Output
0 3
3
9
```
Data Parallel

Input \rightarrow \text{MatMul 1} \rightarrow \text{Weight 1} \rightarrow \text{MatMul 2} \rightarrow \text{Output}

Input \rightarrow \text{MatMul 1} \rightarrow \text{Weight 1} \rightarrow \text{MatMul 2} \rightarrow \text{Output}

Input \rightarrow \text{MatMul 1} \rightarrow \text{Weight 1} \rightarrow \text{MatMul 2} \rightarrow \text{Output}
Data Parallel
Data Parallel

Input → MatMul 1 → Weight 1 → MatMul 2 → Output

Weight 2 → MatMul 2
Data Parallel

Input → MatMul 1 → MatMul 2 → Output

Input → MatMul 1 → MatMul 2

Weight 1 → MatMul 1 → MatMul 2

Weight 2 → MatMul 1 → MatMul 2

Output

0 / 3

3 / 9

9
Data Parallel
Data Parallel

Input → MatMul 1 → Weight 1 → MatMul 2 → Output

Input → MatMul 1 → Weight 1 → MatMul 2 → Output

Input → MatMul 1 → Weight 1 → MatMul 2 → Output

Input → MatMul 1 → Weight 1 → MatMul 2 → Output

Input → MatMul 1 → Weight 1 → MatMul 2 → Output

Input → MatMul 1 → Weight 1 → MatMul 2 → Output

Input → MatMul 1 → Weight 1 → MatMul 2 → Output

Input → MatMul 1 → Weight 1 → MatMul 2 → Output
Data Parallel

Input → MatMul 1 → MatMul 2 → Weight 1 → MatMul 1 → MatMul 2 → Weight 1 → Output

Output: 0

Input: 3

3 / 3 = 1

Output: 9
Data Parallel
Data Parallel

```
Data Parallel
Input
MatMul 1
MatMul 2
Weight 1
Weight 2
MatMul 1
MatMul 2
Weight 1
Weight 2
MatMul 1
MatMul 2
Weight 1
Weight 2
Input
Output
0
3/3
9
```

```
Input
MatMul 1
MatMul 2
Weight 1
Weight 2
MatMul 1
MatMul 2
Weight 1
Weight 2
Input
Output
0
3/3
9
```
Data Parallel

Input → MatMul 1 → Weight 1 → MatMul 2 → Weight 2 → Output

Input → MatMul 1 → Weight 1 → MatMul 2

Input → MatMul 1 → Weight 1 → MatMul 2

Input → MatMul 1 → Weight 1 → MatMul 2

Input → MatMul 1 → Weight 1 → MatMul 2

Input → MatMul 1 → Weight 1 → MatMul 2

Input → MatMul 1 → Weight 1 → MatMul 2

Output = 03/39
Data Parallel

Input
MatMul 1
Weight 1
MatMul 2
Weight 2
Output

Input
MatMul 1
Weight 1
MatMul 2
Weight 2
Output

Input
MatMul 1
Weight 1
MatMul 2
Weight 2
Output

0 3 3
9/3 9
5 0

0 3 3
9
3 9
Data Parallel

```
MatMul 1
<table>
<thead>
<tr>
<th>Weight 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MatMul 1</td>
</tr>
<tr>
<td>Weight 2</td>
</tr>
<tr>
<td>MatMul 2</td>
</tr>
<tr>
<td>Output</td>
</tr>
</tbody>
</table>

MatMul 2

Output

```

```
MatMul 1
<table>
<thead>
<tr>
<th>Weight 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MatMul 1</td>
</tr>
<tr>
<td>Weight 2</td>
</tr>
<tr>
<td>MatMul 2</td>
</tr>
<tr>
<td>Output</td>
</tr>
</tbody>
</table>

MatMul 2

Output

```

```
Input

```

```
Input

```

```
Output

```

```
Output

```

50

3/3

9

0

3

3

9

3

9
Data Parallel

Input

MatMul 1

Weight 1

MatMul 2

Weight 2

Output

MatMul 1

Weight 1

MatMul 2

Weight 2

Output

Output

Output

Output

Output

Output

Output

0

3/3

9

3

5

0
Data Parallel

Input → MatMul 1 → MatMul 2 → Weight 1 → Weight 2 → Output

Input → MatMul 1 → MatMul 2 → Weight 1 → Weight 2 → Output

Input → MatMul 1 → MatMul 2 → Weight 1 → Weight 2 → Output

Input → MatMul 1 → MatMul 2 → Weight 1 → Weight 2 → Output

Output

Output

Output

Output

0 3 / 3 9
Data Parallel
Data Parallel
Data Parallel
Model Parallel
Data Parallel
Model Parallel
Attribute Parallel
Data Parallel
Model Parallel
Attribute Parallel
Reduction Parallel
<table>
<thead>
<tr>
<th>Parameter Parallel</th>
<th>Data Parallel</th>
<th>Model Parallel</th>
<th>Attribute Parallel</th>
<th>Reduction Parallel</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4/3</td>
<td>9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Parallel
Model Parallel
Attribute Parallel
Reduction Parallel
Parameter Parallel
Pipeline Parallel
Data Parallel
Model Parallel
Attribute Parallel
Reduction Parallel
Parameter Parallel
Pipeline Parallel

Parallelization...
FlexFlow

Distributed DNN Training
Distributed DNN Training Runtime
Distributed DNN Training Runtime

FlexFlow

Distributed DNN Training Runtime
Distributed DNN Training

Runtime

FlexFlow

Distributed DNN Training Runtime

Keras

PyTorch

ONNX
Distributed DNN Training

FlexFlow

Runtime
Distributed DNN Training Runtime

FlexFlow
Distributed DNN Training Runtime

FlexFlow
Distributed DNN Training

FlexFlow

Distributed DNN Training Runtime
Distributed DNN Training
Auto-Parallelizer

FlexFlow

Distributed DNN Training Auto-Parallelizer
Distributed DNN Training
Auto-Parallelizer

FlexFlow

Distributed DNN Training Auto-Parallelizer
Distributed DNN Training
Search-Based Auto-Parallelizer

FlexFlow
Distributed DNN Training
Search-Based Auto-Parallelizer
Optimizer

FlexFlow
Accelerating DNN Training Through Joint Optimization of Algebraic Transformations and Parallelization

Colin Unger*, Zhihao Jia*, Wei Wu, Sina Lin, Mandeep Baines, Carlos Efrain Quintero Narvaez, Vinay Ramakrishnaiah, Nirmal Prajapati, Pat McCormick, Jamaludin Mohd-Yusof, Xi Luo, Dheevatsa Mudigere, Jongsoo Park, Misha Smelyanskiy, Alex Aiken
FlexFlow

TASO/MetaFlow

Unity

Memory Optimization
Generalized Pipeline Parallelism
MoE Inference
JAX Interface
ECP-CANDLE
Cryo-EM
Hugging Face
Input → ReLU → MatMul → ReLU → MatMul → Output

Weight 1 → MatMul → Weight 2 → ReLU → MatMul → Output
Memory
ReLU
MatMul

MatMul + ReLU

Output
ReLU
MatMul

MatMul + ReLU

MatMul

Weight 2

Weight 1

Input

Memory
MatMul + ReLU

Operator Fusion
Operator Fusion

Operator Splitting
MatMul + ReLU

Output

ReLU

MatMul

Weight 2

MatMul + ReLU

Input

Weight 1

Operator Fusion
Operator Splitting
Operator Reordering
Operator Fusion
Operator Splitting
Operator Reordering
...

MatMul + ReLU
Operator Fusion
Operator Splitting
Operator Reordering
...
“Algebraic Transformations”

Operator Fusion
Operator Splitting
Operator Reordering
...

Operator Fusion
Operator Splitting
Operator Reordering
...
### Auto-Parallelization

- FlexFlow [MLSys 19]
- Tofu [EuroSys 19]
- PipeDream [SOSP 19]
- automap [arXiv 19]
- Whale [arXiv 21]
- Alpa [OSDI 22]

### Algebraic Optimizers

- MetaFlow [MLSys 19]
- TASO [SOSP 19]
- PET [OSDI 21]
- Tensat [MLSys 21]

...
Auto-Parallelization
Auto-Parallelization
Algebraic Optimizer
Auto-Parallelization

Algebraic Optimizer
Joint Optimization
Joint Optimization

1. 

2. 
1. Representation
1. Representation
2. Scalability
Unity
Representation
Parallel Computation Graph (PCG)

Unity
Unity

- Representation
  - Parallel Computation Graph (PCG)
- Scalability
  - Hierarchical Search Algorithm
Representation
Parallel Computation Graph (PCG)
Parallel Computation Graph (PCG)

annotated computation graph
Parallel Computation Graph (PCG)
Parallel Computation Graph (PCG)

- **Data Parallelism**
  - **MatMul**
  - **A**
  - **B**
  - **C**

- **Reduction Parallelism**
  - **MatMul**
  - **Reduce**
  - **MatMul**
  - **Replicate**
  - **Partition**
  - **A**
  - **B**
  - **C**

**annotated computation graph**

**parallel computation graph (PCG)**
Parallel Computation Graph (PCG)

annotated computation graph

parallel computation graph (PCG)
MatMul

Output

MatMul

MatMul

C

A

B

PCG

1/8/39
MatMul

Output

MatMul

MatMul

A

B

C

MatMul

Partition (c)

Partition (r)

Reduce

MatMul

Substitution

PCG

\[
\begin{align*}
18/39
\end{align*}
\]
MatMul
Reduce
MatMul
Partition (c)
Partition (r)
A
B
C
Output

PCG

Substitution

\[ \frac{18}{39} \]
MatMul

Output

Reduce

MatMul

Replicate

C

Partition (c)

Partition (r)

A

B

MatMul

Reduce

C

MatMul

Reduce

C

Replicate

Substitution

PCG

1

8/3

9
PCG

Output
  └── Reduce
      └── MatMul
          └── Replicate
              └── C

MatMul
  ├── Partition (c)
  │   └── A
  └── Partition (r)
    └── B

18/39
Hierarchical Search Algorithm
Hierarchical Search Algorithm

Algebraic Transformation
Hierarchical Search Algorithm

Algebraic Transformation

Parallelism Type
Hierarchical Search Algorithm

Algebraic Transformation

Parallelism Type

Parallelism Degree
Hierarchical Search Algorithm

Algebraic Transformation

Parallelism Type

Parallelism Degree

Device Mapping
Hierarchical Search Algorithm

Algebraic Transformation

Parallelism Type

Parallelism Degree

Device Mapping
Hierarchical Search Algorithm

Algebraic Transformation

Parallelism Type

Parallelism Degree

Device Mapping

Backtracking Search
<table>
<thead>
<tr>
<th>Models</th>
<th>Precision Medicine</th>
<th>Regression</th>
<th>Recommendation</th>
<th>Computer Vision</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-Large (Language Modeling)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candle-UNO (Precision Medicine)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP (Regression)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DLRM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNeXt-50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inception-v3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
BERT-Large

Overall Throughput (samples/second)

- DeepSpeed
- Megatron
- TASO+FlexFlow

Number of GPUs (Number of nodes):
- 6 (1)
- 12 (2)
- 24 (4)
- 48 (8)
- 96 (16)
- 192 (32)
Distributed DNN Training
Search-Based Auto-Parallelizer
Optimizer
Machine-Tailored Optimizations

FlexFlow

Distributed DNN Training
Search-Based Auto-Parallelizer
Distributed DNN Training
Search-Based Optimizer
Optimizer
Machine-Tailored Optimizations

FlexFlow

Distributed DNN Training
Search-Based Optimizer
New Features

- FlexFlow
- TASO/MetaFlow
- Unity
- Memory Optimization
- Generalized Pipeline Parallelism
- MoE Inference
- JAX Interface
- ECP-CANDLE
- Cryo-EM
- Hugging Face
Extending Unity

- Algebraic Transformations
- Parallelization
- Memory Constraints

Memory Optimization
Generalized Pipeline Parallelism
MoE Inference
JAX Interface
ECP-CANDLE
Cryo-EM
Hugging Face
Maximizes throughput *but can exceed memory capacity*
Extending Unity

Algebraic Transformations
Parallelization
Memory Optimizations

Maximizes throughput but can exceed memory capacity
Extending Unity

- Algebraic Transformations
- Parallelization
- Memory Optimizations
  - Recomputation
  - Offloading

Maximizes throughput but can exceed memory capacity

- Memory Optimization
- Generalized Pipeline Parallelism
- MoE Inference
- JAX Interface
- ECP-CANDLE
- Cryo-EM
- Hugging Face
FlexFlow
Unity
TASO/MetaFlow
Memory Optimization
Generalized Pipeline
Parallelism
MoE
Inference
JAX Interface
ECP-CANDLE
Cryo-EM
Hugging Face
Extending Unity
Memory Optimization
Generalized Pipeline Parallelism

MoE Inference
JAX Interface
ECP-CANDLE
Cryo-EM
Hugging Face

Extending Unity
Memory Optimization
Generalized Pipeline Parallelism
MoE Inference
JAX Interface
ECP-CANDLE
Cryo-EM
Hugging Face

Extending Unity
Extending Unity

Memory Optimization

Generalized Pipeline Parallelism

MoE Inference

JAX Interface

ECP-CANDLE

Cryo-EM

Hugging Face
“sparsely-gated MoE” (2017)
GShard (2020)
GLaM (2021)
Switch Transformer (2021)
Expert Choice (2022)

SpeechMoE (2021)
SpeechMoE2 (2021)

DeepSpeed-MoE (2022)
New Features

- Switch Transformer (2021)
- GShard (2020)
- Expert Choice (2022)
- GLaM (2021)
- ST-MoE (2022)
- DeepSpeed-MoE (2022)
- V-MoE (2022)
- SpeechMoE (2021)
- “sparsely-gated MoE” (2017)
- SpeechMoE2 (2021)

- Memory Optimization
- Generalized Pipeline Parallelism
- MoE Inference
- JAX Interface
- ECP-CANDLE
- Cryo-EM
- Hugging Face

Mixture of Experts → data-dependent control flow + distributed execution + improved inference support!
A functional deep learning framework

immutability, compositionality, modularity

New Features
A functional deep learning framework

immutability, compositionality, modularity

New Features

- Memory Optimization
- Generalized Pipeline Parallelism
- MoE Inference
- JAX Interface
- ECP-CANDLE
- Cryo-EM
- Hugging Face
A functional deep learning framework

immutability, compositionality, modularity

New Features

Memory Optimization
Generalized Pipeline Parallelism
MoE Inference
JAX Interface
ECP-CANDLE
Cryo-EM
Hugging Face
Set of Deep Neural Networks for Precision Medicine

Applications

- Memory Optimization
- Generalized Pipeline Parallelism
- MoE Inference
- JAX Interface
- ECP-CANDLE
- Cryo-EM
- Hugging Face
Set of Deep Neural Networks for Precision Medicine

Difficult to scale w/ existing frameworks

- Memory Optimization
- Generalized Pipeline Parallelism
- MoE Inference
- JAX Interface
- ECP-CANDLE
- Cryo-EM
- Hugging Face

Applications
Set of Deep Neural Networks for Precision Medicine

Difficult to scale with existing frameworks → Hybrid Parallelism

Applications

- Memory Optimization
- Generalized Pipeline Parallelism
- MoE Inference
- JAX Interface
- ECP-CANDLE
- Cryo-EM
- Hugging Face
Set of Deep Neural Networks for Precision Medicine

Difficult to scale w/ existing frameworks

Hybrid Parallelism

Memory Optimization
Generalized Pipeline Parallelism
MoE Inference
JAX Interface

Applications

ECP-CANDLE
Cryo-EM
Hugging Face
Cryo-EM

electron microscopy

Applications

- Memory Optimization
- Generalized Pipeline Parallelism
- MoE Inference
- JAX Interface

ECP-CANDLE  Cryo-EM  Hugging Face

Cryo-EM: electron microscopy
FlexFlow
Unity
TASO/MetaFlow
Memory Optimization
Generalized Pipeline Parallelism
MoE
Inference
JAX Interface
ECP-CANDLE
Cryo-EM
Hugging Face

Raw EM Data (slices)

Cryo-EM

electron microscopy

Applications
FlexFlow | Unity | TASO/MetaFlow
---|---|---
Memory Optimization | Generalized Pipeline | Parallelism
MoE Inference | JAX Interface

Applications

Cryo-EM

electron microscopy

Raw EM Data (slices)

ECP-CANDLE | Cryo-EM | Hugging Face
Cryo-EM

\textit{electron microscopy}

Raw EM Data (slices) \rightarrow \text{Memory Optimization} \rightarrow \text{Generalized Pipeline Parallelism} \rightarrow \text{MoE Inference} \rightarrow \text{JAX Interface} \rightarrow 3D Protein Structure

Applications

- ECP-CANDLE
- Cryo-EM
- Hugging Face
Cryo-EM (electron microscopy) takes raw EM data (slices) and processes it through a series of steps involving Memory Optimization, Generalized Pipeline Parallelism, MoE Inference, JAX Interface, ECP-CANDLE, Cryo-EM, and Hugging Face. The result is a 3D Protein Structure.
Cryo-EM (electron microscopy) → Raw EM Data (slices) → GNN → 3D Protein Structure

Applications:
- Memory Optimization
- Generalized Pipeline Parallelism
- MoE Inference
- JAX Interface
- ECP-CANDLE
- Cryo-EM
- Hugging Face
FlexFlow
Unity
TASO/MetaFlow
Memory Optimization
Generalized Pipeline Parallelism
MoE Inference
JAX Interface
ECP-CANDLE
Cryo-EM
Hugging Face
Applications
Library for Transformer-based Models
Ease-of-use
Performance
Ease-of-use
Performance
Memory Optimization
Generalized Pipeline Parallelism
MoE Inference
JAX Interface
ECP-CANDLE
Cryo-EM
Hugging Face
Applications
Library for Transformer-based Models

Ease-of-use

FlexFlow

Applications

Memory Optimization
Generalized Pipeline Parallelism
MoE Inference
JAX Interface

ECP-CANDLE
Cryo-EM
Hugging Face
Questions?

https://github.com/flexflow/FlexFlow