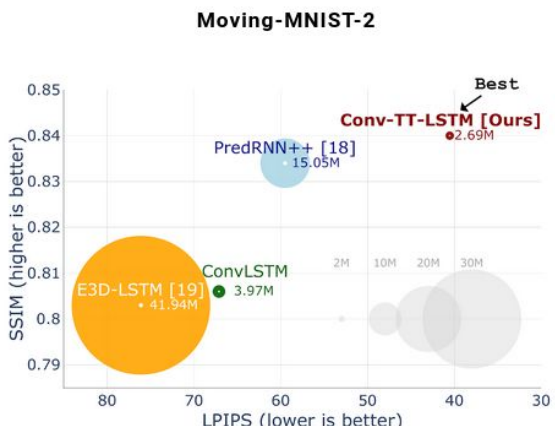
MIXED PRECISION TRAINING FOR CONVOLUTIONAL TENSOR-TRAIN LSTM

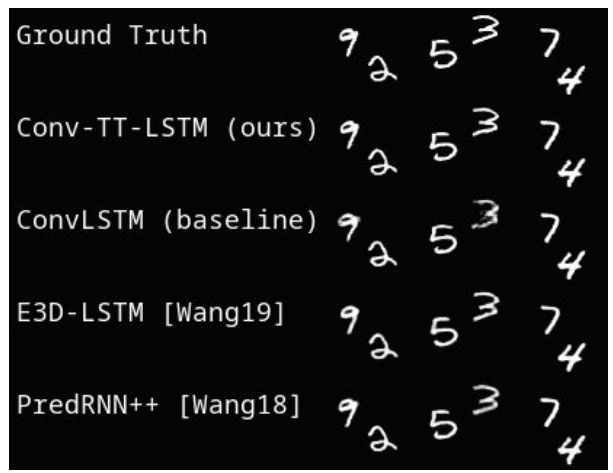
Wonmin Byeon, August 23, 2020

APPLICATIONS

Spatio-Temporal Learning

Video Prediction

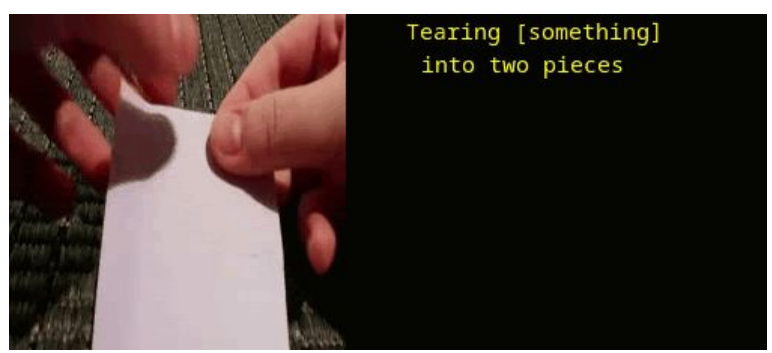




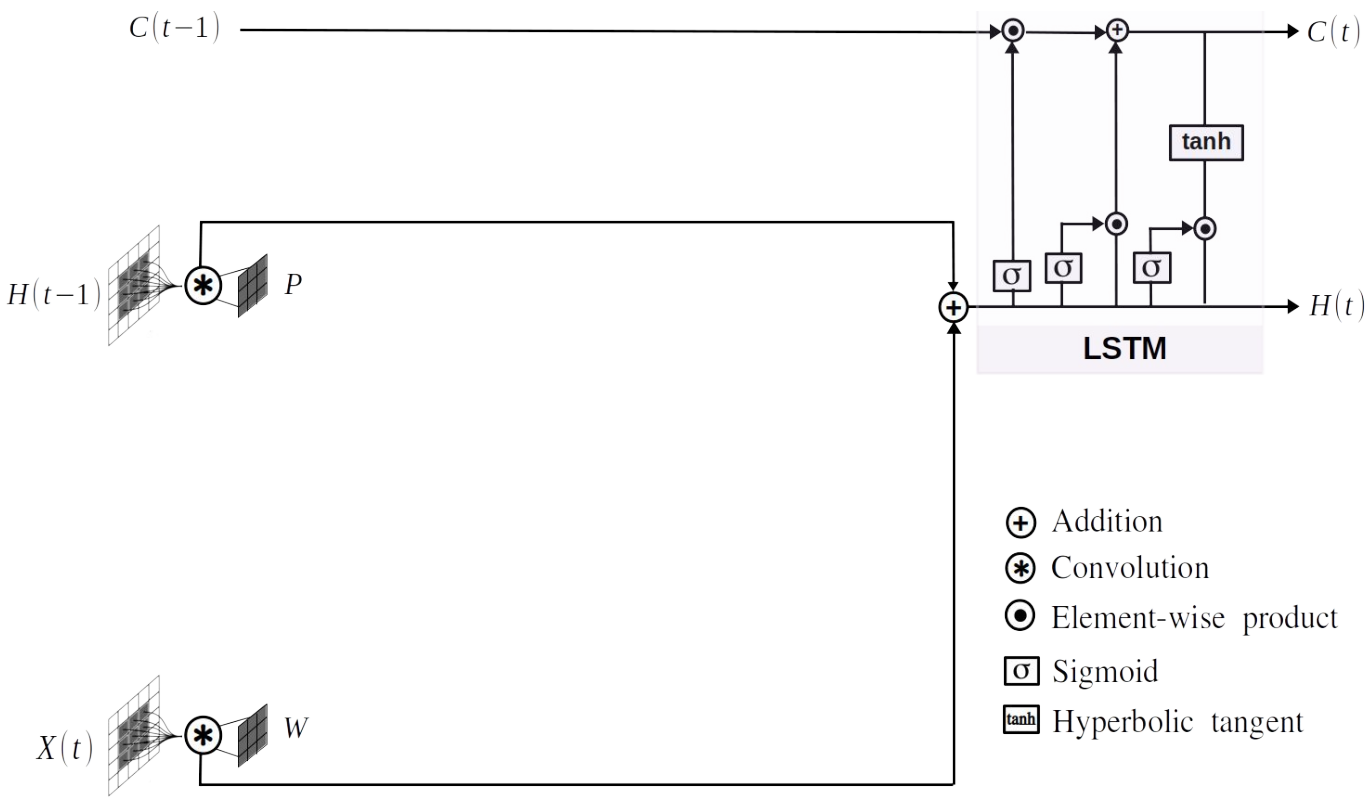
Early Activity Recognition

Model	Input Ratio	
	Front 25%	Front 50%
3D-CNN*	9.11	10.30
E3D-LSTM* [7]	14.59	22.73
3D-CNN	13.26	20.72
ConvLSTM	15.46	21.97
Conv-TT-LSTM (ours)	**19.53**	**30.05**

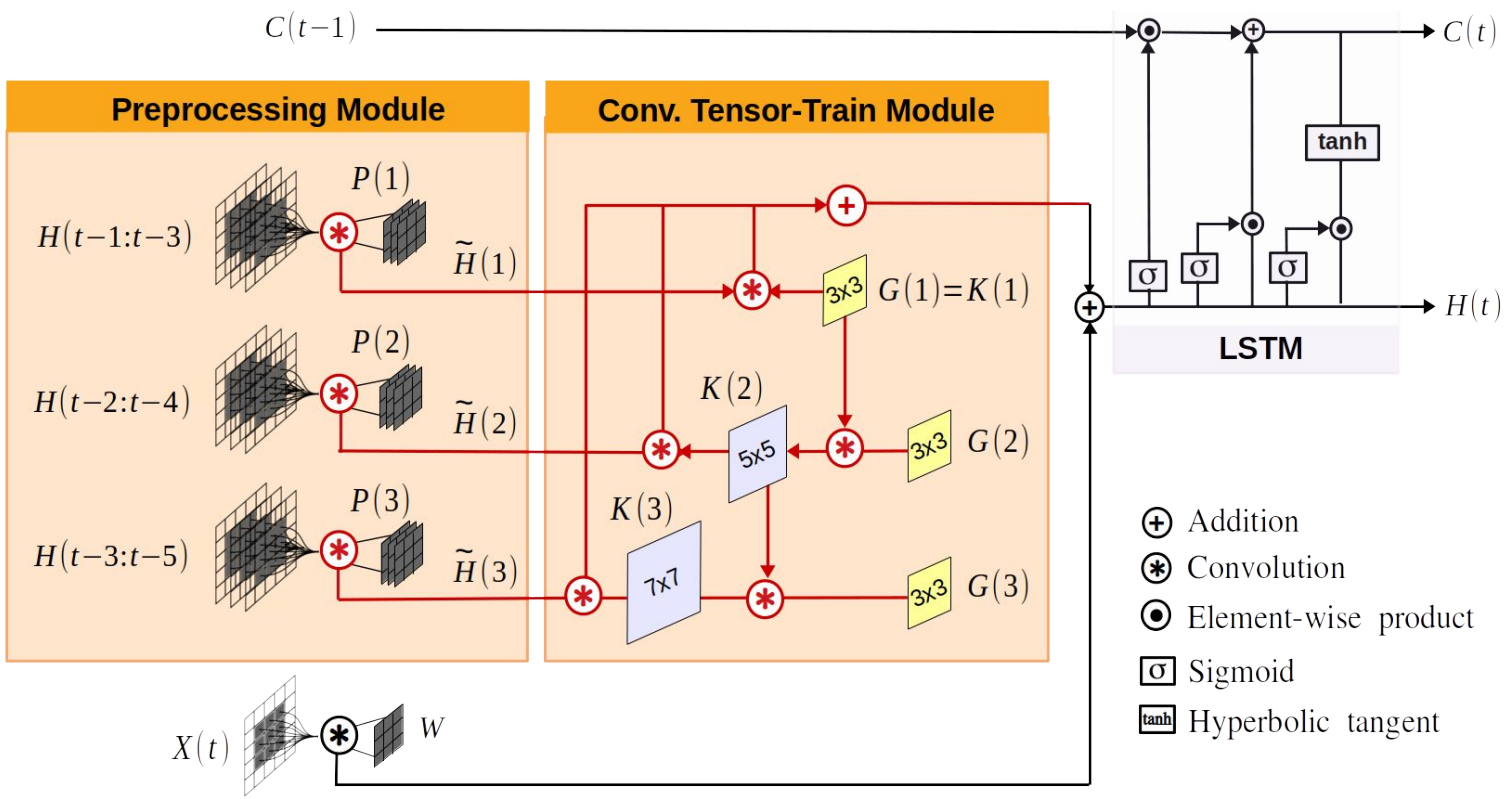
Table 2: **Early activity recognition on the Something-Something V2 dataset** using 41 categories as [7]. (*) indicates the result by [7].



CONVOLUTIONAL LSTM



CONVOLUTIONAL TENSOR-TRAIN LSTM



OPTIMIZATION TRICKS

- Speed-up tricks to train Convolutional LSTM/Convolutional Tensor-Train LSTM

  #1. **Enabling mixed precision training using NVIDIA AMP**

  #2. **GPU Optimization:** Fused Adam, Fused kernels in LSTM cell, Thread affinity binding

  #3. LSTM **Activation Checkpointing**

- Fast convergence without performance change
- **Model parallelism** using multi-streams

#1. ENABLING AMP

```
from apex import amp

model, optimizer = amp.initialize(model, optimizer, opt_level="O1")


with amp.scale_loss(loss, optimizer) as scaled_loss:
    scaled_loss.backward()
if gradient_clipping:
    grad_norm = torch.nn.utils.clip_grad_norm_(
        amp.master_params(optimizer), clipping_threshold)
```

```
# save the checkpoint
checkpoint_info['amp'] = amp.state_dict()

# load the checkpoint
amp.load_state_dict(checkpoint['amp'])
```

#2. GPU Optimization

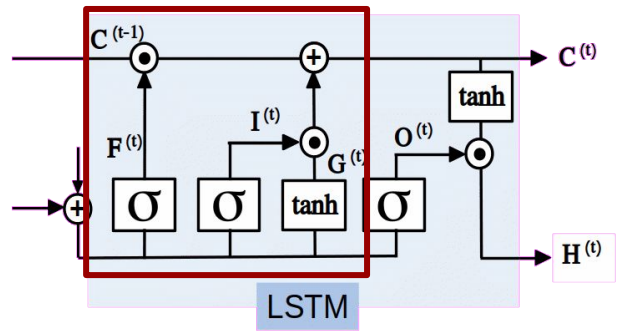
- **Fused Adam: Adam optimizer**

```
from apex.optimizers import FusedAdam

optimizer = FusedAdam(model.parameters(), lr = learning_rate, eps = eps)
```

#2. GPU Optimization

- **Fused optimizer: Adam optimizer**

```
from apex.optimizers import FusedAdam

optimizer = FusedAdam(model.parameters(), lr = learning_rate, eps = eps)
```



- **Fused kernel with JIT: LSTM cell**

```
@torch.jit.script
def fuse_mul_add_mul(f, cell_states, i, g):
    return f * cell_states + i * g
```

```
self.cell_states  = f * self.cell_states + i * g
```

```
self.cell_states = fuse_mul_add_mul(f, self.cell_states, i, g)
```

#2. GPU Optimization

- **Fused optimizer: Adam optimizer**

```
from apex.optimizers import FusedAdam

optimizer = FusedAdam(model.parameters(), lr = learning_rate, eps = eps)
```

- **Fused kernel with JIT: LSTM cell**

```
@torch.jit.script
def fuse_mul_add_mul(f, cell_states, i, g):
    return f * cell_states + i * g
```

```
self.cell_states  = f * self.cell_states + i * g
```

```
self.cell_states = fuse_mul_add_mul(f, self.cell_states, i, g)
```

- **Thread affinity binding: distributed computing**

```
from utils.gpu_affinity import set_affinity

set_affinity(args.local_rank)
```

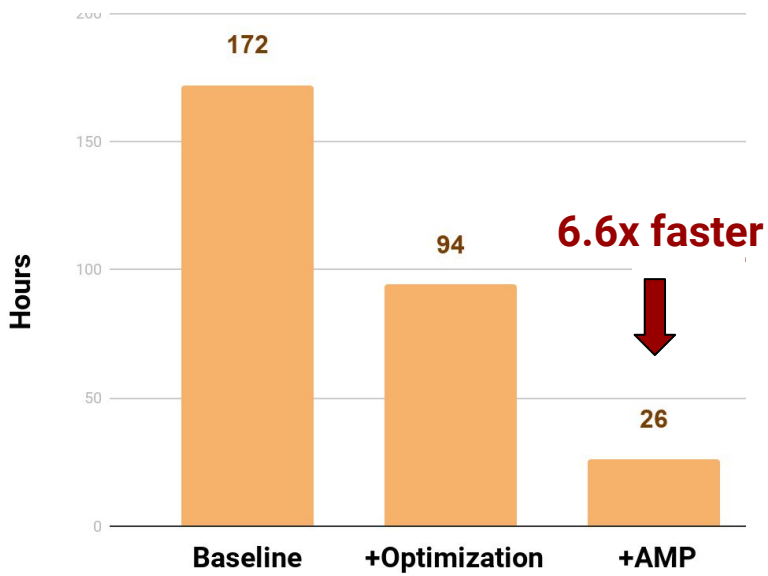
Application: video prediction
Machine: V100 x 8, 16GB
Batch Size: 16 videos
12 Conv. LSTM layers
Input/output image resolution: 128x128

TRAINING

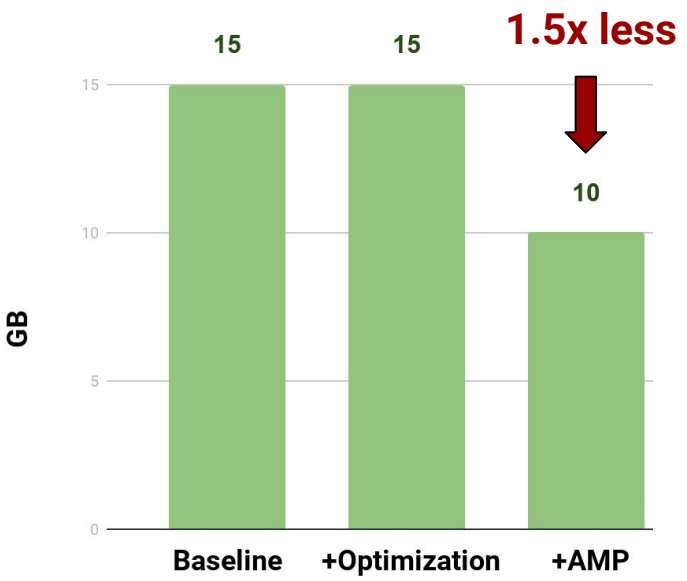
Convolutional LSTM

No performance change!

Convergence time



GPU Memory



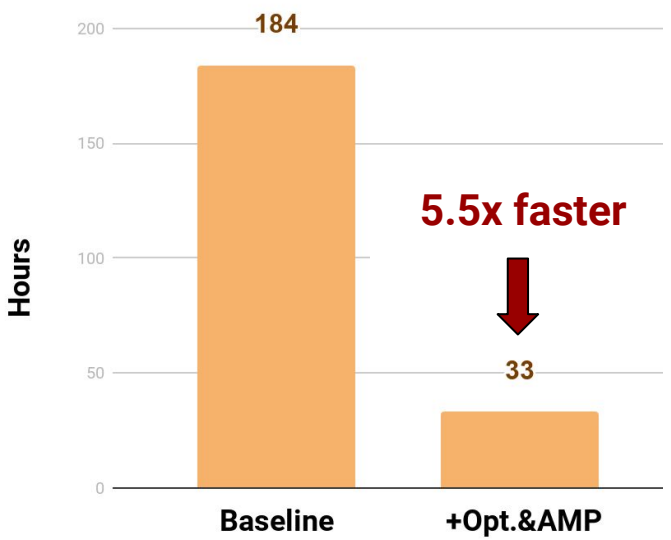
Application: video prediction
Machine: V100 x 8, 16GB
Batch Size: 16 videos
12 Conv. LSTM layers
Input/output image resolution: 128x128

TRAINING

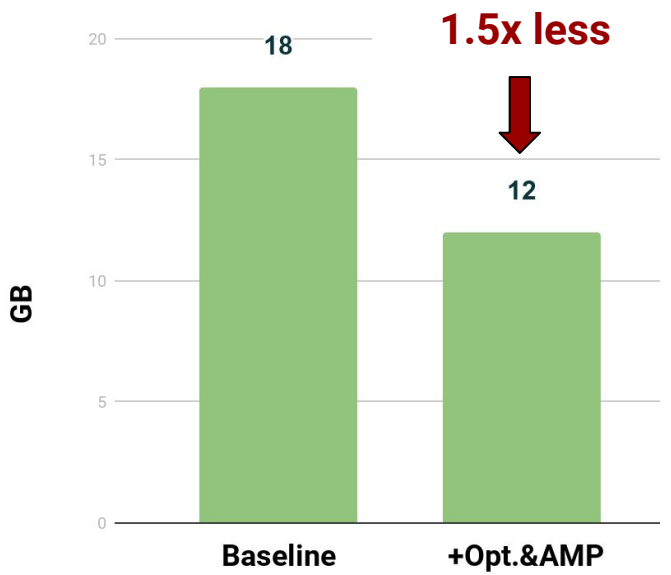
Convolutional Tensor-Train LSTM

No performance change!

Convergence time

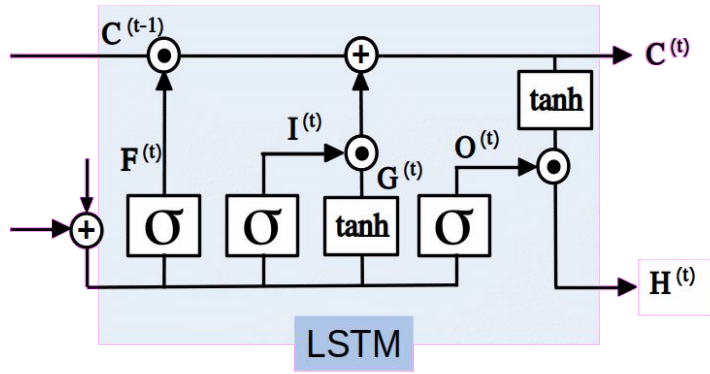


GPU Memory



#3. ACTIVATION CHECKPOINTING

```
from torch.utils.checkpoint import checkpoint

def chkpt_blk(cc_i, cc_f, cc_o, cc_g, cell_states):

    i = torch.sigmoid(cc_i)
    f = torch.sigmoid(cc_f)
    o = torch.sigmoid(cc_o)
    g = torch.tanh(cc_g)

    cell_states = fuse_mul_add_mul(f, cell_states, i, g)
    outputs = o * torch.tanh(cell_states)

    return outputs, cell_states
```



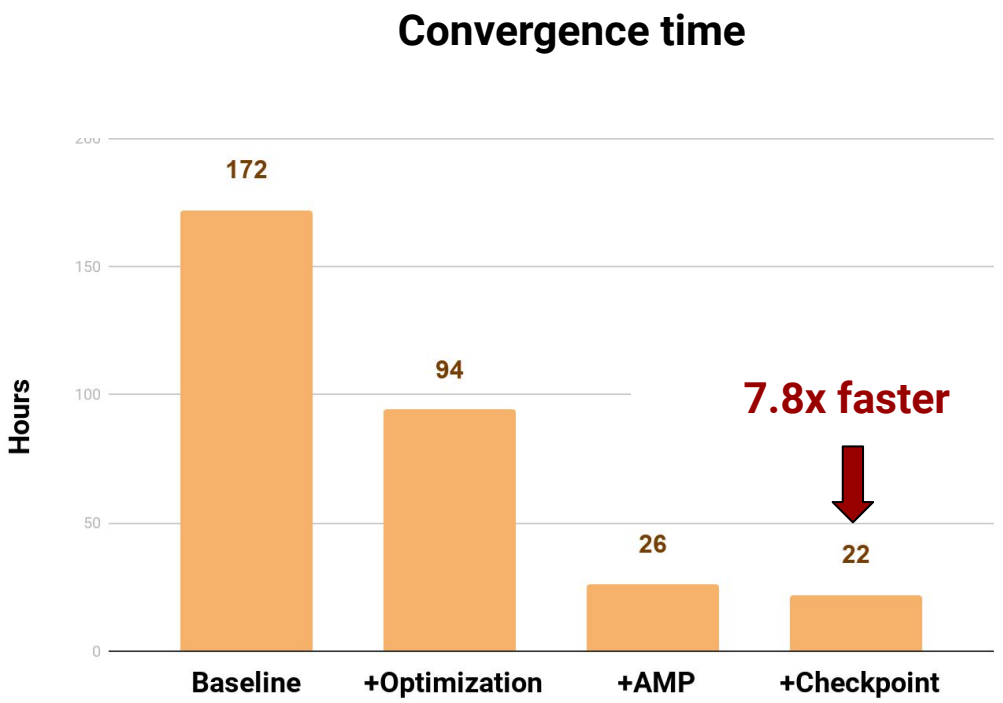

```
outputs, self.cell_states = checkpoint(chkpt_blk, cc_i, cc_f, cc_o, cc_g, self.cell_states)
```

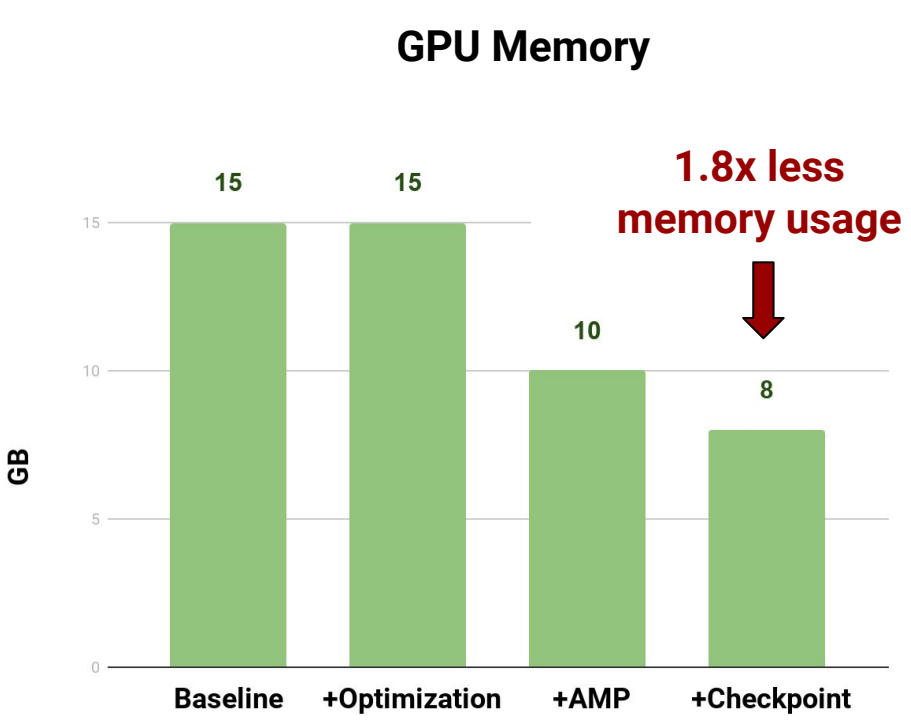
Application: video prediction
Machine: V100 x 8, 16GB
Batch Size: 16 videos
12 Conv. LSTM layers
Input/output image resolution: 128x128

TRAINING

Convolutional LSTM

No performance change!





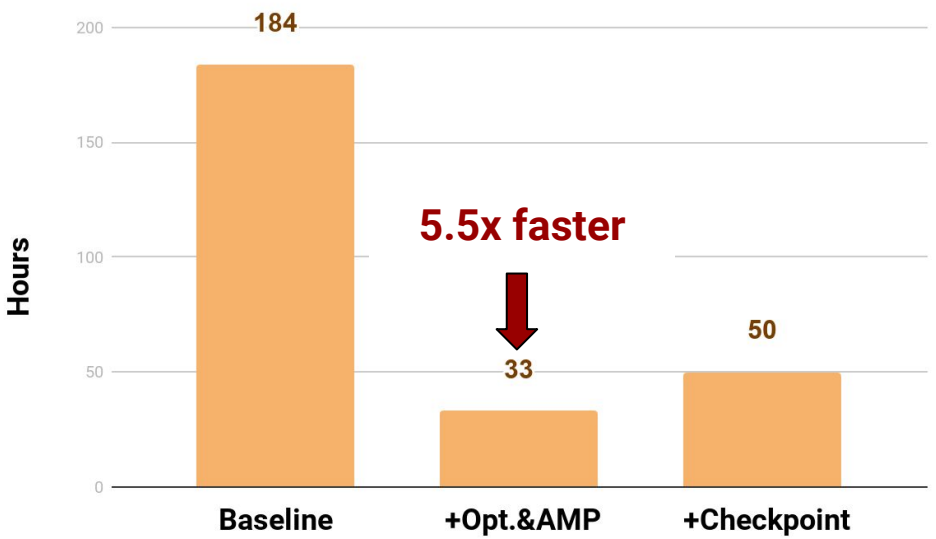
Application: video prediction
Machine: V100 x 8, 16GB
Batch Size: 16 videos
12 Conv. LSTM layers
Input/output image resolution: 128x128

TRAINING

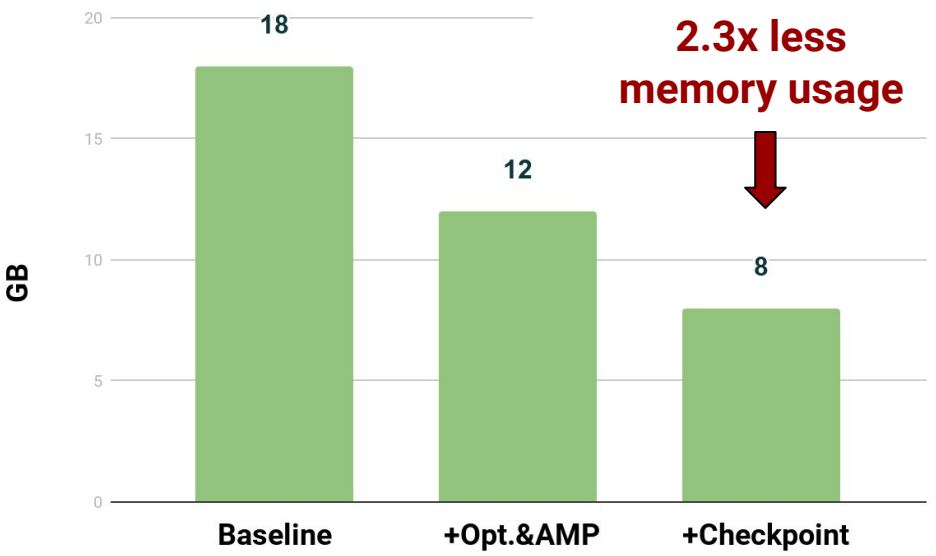
Conv. Tensor-Train LSTM

No performance change!

Convergence time



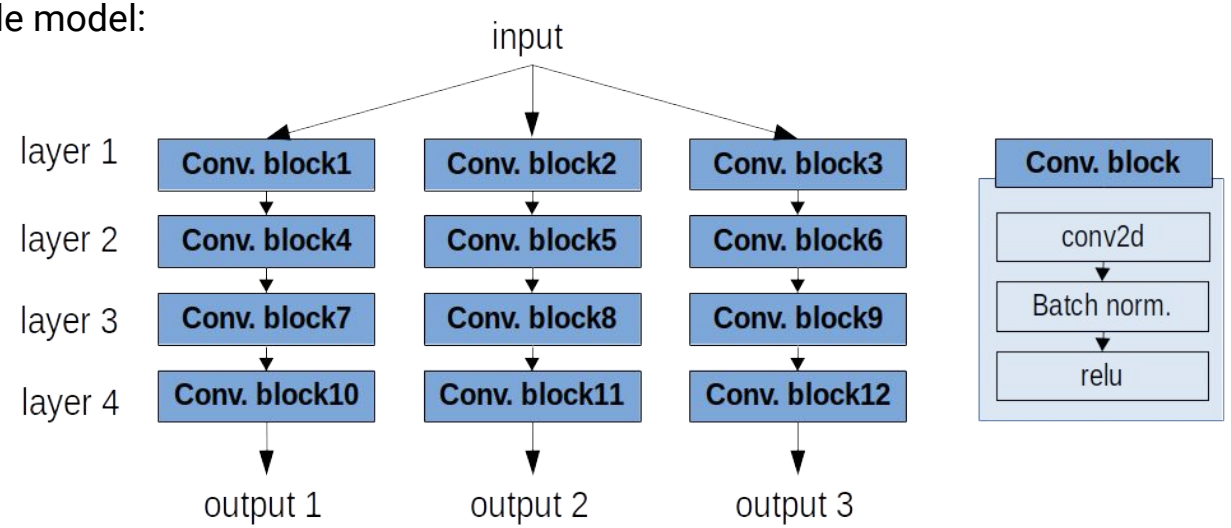
GPU Memory



MODEL PARALLEL: MULTI-STREAMS

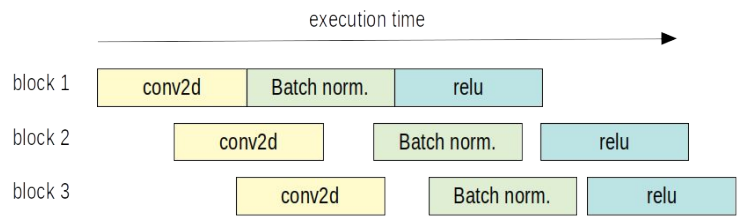
- Use multiple cuda-streams
- Execute multiple kernels that do not have data dependency in parallel

example model:



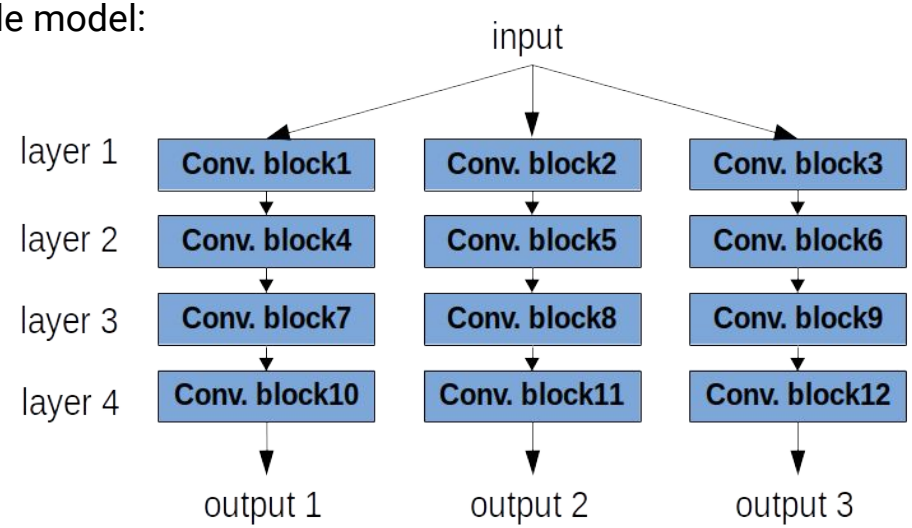
MODEL PARALLEL: MULTI-STREAMS

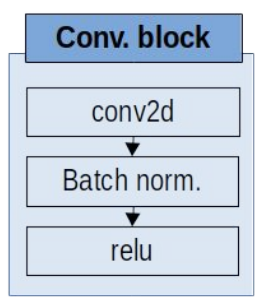
- Use multiple cuda-streams
- Execute multiple kernels that do not have data dependency in parallel



Operations of layers at different branches are overlapped.

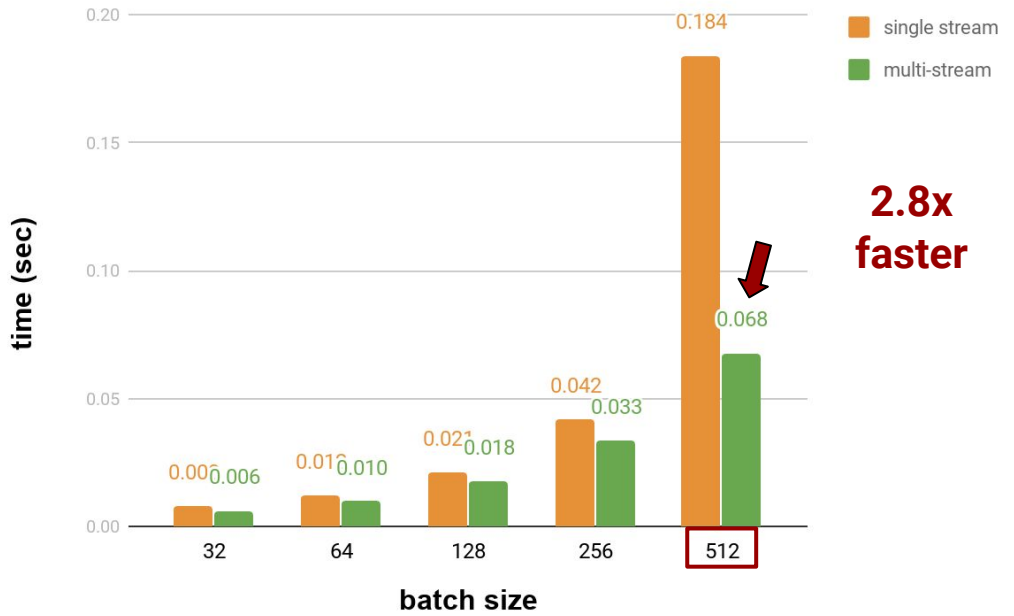
example model:





MODEL PARALLEL: MULTI-STREAMS

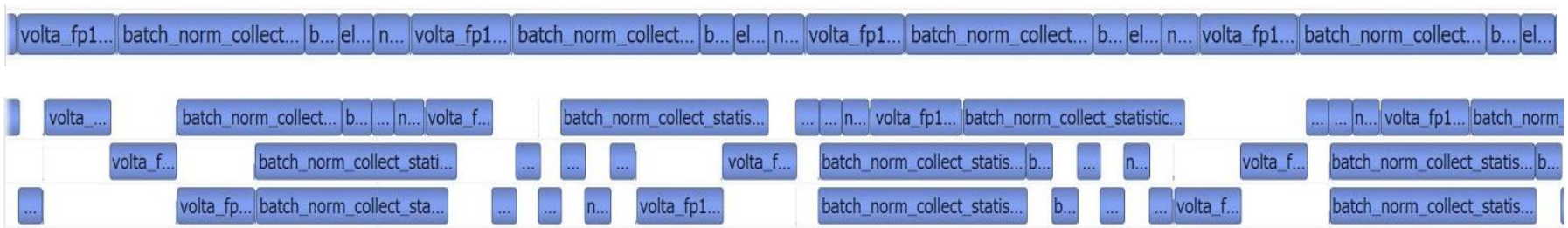
Computation time comparison



2.8x faster

single-stream

multi-stream



NVIDIA

SUMMARY

Speed-up summary to train Convolutional Tensor-Train LSTM

- **Enabling AMP**
- **GPU Optimization**
- **Activation Checkpointing**
- ConvLSTM: **7.8x** speed-up, **1.8x** less memory usage
- Conv. Tensor-Train LSTM: **5.5x** speed-up, **2.3x** less memory usage.
- Fast convergence without performance change

Model parallelism using multi-streams

CONCLUSION

- Paper: Jiahao Su* (UMD), Wonmin Byeon* (NVIDIA), Jean Kossaifi (NVIDIA), Furong Huang (UMD), Jan Kautz (NVIDIA), Animashree Anandkumar (NVIDIA), '*Convolutional Tensor-Train LSTM for Spatio-Temporal Learning*', under submission, 2020.
- Project page: https://sites.google.com/nvidia.com/conv-tt-lstm/home
- Code Optimization: Sangkug Lym (NVIDIA)
- The code with the optimization tricks will be available soon: https://github.com/NVlabs/conv-tt-lstm