MLink: Linking Black-box Models for Collaborative Multi-model Inference

Mu Yuan, Lan Zhang, Xiang-Yang Li
University of Science and Technology of China
Main contents

• Introduction

• Problem Statement

• Black-box Model Linking

• Collaborative Multi-model Inference

• Evaluation

• Conclusion
Introduction

Multi-model Inference Workloads

- Complex intelligent services that are difficult (or even impossible) to develop with a single model.

- Smart Speaker
- Intelligent Traffic
- Autonomous Vehicles
- Contextual Advertising
Introduction

Multi-model Inference Workloads

- Complex intelligent services that are difficult (or even impossible) to develop with a single model.
Introduction

Multi-model Inference Workloads

- Complex intelligent services that are difficult (or even impossible) to develop with a single model.
Introduction

Multi-model Inference Workloads

• Complex intelligent services that are difficult (or even impossible) to develop with a single model.

Smart Speaker

Intelligent Traffic

Autonomous Vehicles

Contextual Advertising
Introduction
Cost-effective Inference

- Multi-task learning and zipping
- Model compression
- Inference reusing
- Source filtering
- Multi-model scheduling
Introduction
Cost-effective Inference

• Multi-task learning and zipping

**Model compression**

• Inference reusing
• Source filtering
• Multi-model scheduling
Introduction

Cost-effective Inference

- Multi-task learning and zipping
- Model compression
- Inference reusing
- Source filtering
- Multi-model scheduling
Introduction

Cost-effective Inference

• Multi-task learning and zipping
• Model compression
• Inference reusing
• **Source filtering**
• Multi-model scheduling
Introduction

Cost-effective Inference

• Multi-task learning and zipping
• Model compression
• Inference reusing
• Source filtering
• Multi-model scheduling
Introduction

Cost-effective Inference

- Multi-task learning and zipping
- Model compression
- Inference reusing
- Source filtering
- Multi-model scheduling

How to obtain as accurate inference results as possible without the exact execution of ML models?
Introduction
Cost-effective Inference

• Multi-task learning and zipping
• Model compression
• Inference reusing
• Source filtering
• Multi-model scheduling

How to obtain as accurate inference results as possible without the exact execution of ML models?
Introduction

Cost-effective Inference

• Multi-task learning and zipping
• Model compression
• Inference reusing
• Source filtering
• Multi-model scheduling

How to obtain as accurate inference results as possible without the exact execution of ML models?
How to obtain as accurate inference results as possible without the exact execution of ML models?
How to obtain as accurate inference results as possible without the exact execution of ML models?
How to obtain as accurate inference results as possible without the exact execution of ML models?
Introduction

Cost-effective Inference

- Multi-task learning and zipping
- Model compression
- Inference reusing
- Source filtering
- Multi-model scheduling

How to obtain as accurate inference results as possible without the exact execution of ML models?
Introduction

Cost-effective Inference

- Multi-task learning and zipping
- Model compression
- Inference reusing
- Source filtering
- Multi-model scheduling

How to obtain as accurate inference results as possible without the exact execution of ML models?
Introduction

Cost-effective Inference

- Multi-task learning and zipping
- Model compression
- Inference reusing
- Source filtering
- Multi-model scheduling

How to obtain as accurate inference results as possible without the exact execution of ML models?
Introduction
Cost-effective Inference

- Multi-task learning and zipping
- Model compression
- Inference reusing
- Source filtering
- Multi-model scheduling

How to obtain as accurate inference results as possible without the exact execution of ML models?
Introduction

Cost-effective Inference

- Multi-task learning and zipping
- Model compression
- Inference reusing
- Source filtering

- Multi-model scheduling

How to obtain as accurate inference results as possible without the exact execution of ML models?
Introduction

Cost-effective Inference

- Multi-task learning and zipping
- Model compression
- Inference reusing
- Source filtering
- Multi-model scheduling

How to obtain as accurate inference results as possible without the exact execution of ML models?
How to obtain as accurate inference results as possible without the exact execution of ML models?
Introduction

Linking Black-box Models

• Model Linking
  • machine over-learning
Introduction
Linking Black-box Models

• Model Linking
  • machine over-learning
  • cross-task semantic correlation
Introduction

Linking Black-box Models

• Model Linking
  • machine over-learning
  • cross-task semantic correlation

Predict un-executed models’ inference results based on executed models’?
Introduction

Linking Black-box Models

• Model Linking
  • machine over-learning
  • cross-task semantic correlation

Predict un-executed models’ inference results based on executed models’?
Introduction

Linking Black-box Models

- Model Linking
  - machine over-learning
  - cross-task semantic correlation

Predict un-executed models’ inference results based on executed models’?

Exact Execution

Resulting Workload
Introduction

Linking Black-box Models

- Model Linking
  - machine over-learning
  - cross-task semantic correlation

- Target application
  - inference results of multiple models are required
  - cost budget is too limited to run them all
Introduction

Challenges

- build lightweight and accurate links among heterogeneous models
- efficiently select models to execute and models to be predicted

Different input modalities

Different model architectures

Different DL frameworks

CNNs, RNNs, Auto-encoders, Transformers …
Introduction

Challenges

• build lightweight and accurate links among heterogeneous models

• efficiently select models to execute and models to be predicted

Different input modalities

Different model architectures

Different DL frameworks

non-intrusive design and implementation

CNNs, RNNs, Auto-encoders, Transformers …
Introduction

Challenges

• build lightweight and accurate links among heterogeneous models

• efficiently select models to execute and models to be predicted

dynamic re-selection

v.s.

NP-hard combinatory optimization problem
Menu

Main contents

• Introduction

• Problem Statement
  • Black-box Model Linking
  • Collaborative Multi-model Inference

• Evaluation

• Conclusion
Model Linking

- black-box models $F = \{f_i\}_{i=1}^k$ where $f_i : X_i \to Y_i$
Problem Statement

Model Linking

• black-box models $F = \{f_i\}_{i=1}^k$ where $f_i : X_i \rightarrow Y_i$

• **Assumption**: same or aligned input spaces $\{X_i\}_{i=1}^k$

  • common in multi-model applications

---

Multi-task robotics

Drone-based video monitoring

---

Same Input Spaces

Multi-modal learning

Audio-visual speech recognition

Aligned Input Spaces
Problem Statement

Model Linking

- Black-box models $F = \{f_i\}_{i=1}^{k}$ where $f_i : X_i \rightarrow Y_i$

- Assumption: same or aligned input spaces $\{X_i\}_{i=1}^{k}$
  - Common in multi-model applications

**Same Input Spaces**
- multi-task robotics
- drone-based video monitoring

**Aligned Input Spaces**
- multi-modal learning
- audio-visual speech recognition
Problem Statement

Model Linking

- black-box models $F = \{f_i\}_{i=1}^k$ where $f_i : X_i \rightarrow Y_i$

- **Assumption**: same or aligned input spaces $\{X_i\}_{i=1}^k$
  
  - common in multi-model applications
  
  - available alignment techniques

---

*time synchronization*  
*spatial alignment*  
*semantic alignment*

image from http://cvlab.cse.msu.edu/project-sequence-alignment.html

image from paper “AlignNet: A Unifying Approach to Audio-Visual Alignment”
Problem Statement

Model Linking

- black-box models $F = \{f_i\}_{i=1}^k$ where $f_i : X_i \rightarrow Y_i$

- model link $g_{i,j} : Y_i \rightarrow Y_j$
  - source model $f_i$
  - target model $f_j$
Problem Statement

Model Linking

- black-box models \( F = \{ f_i \}_{i=1}^{k} \) where \( f_i : X_i \to Y_i \)
- model link \( g_{i,j} : Y_i \to Y_j \)
  - \textit{source} model \( f_i \)
  - \textit{target} model \( f_j \)
- composite function \( g_{i,j} \circ f_i : X_i \to Y_j \)
Problem Statement

Multi-source Model Links Ensemble

- when $k \geq 3$, there are multiple model links for one target model
Problem Statement

Multi-source Model Links Ensemble

- when $k \geq 3$, there are multiple model links for one target model

- given a set of source models $A \subseteq F$ and a target model $f_j$, we have a multi-expert model $\{g_{i,j} \circ f_i\}_{f_i \in A}$
Problem Statement

Multi-source Model Links Ensemble

• when $k \geq 3$, there are multiple model links for one target model

• given a set of source models $A \subseteq F$ and a target model $f_j$, we have a multi-expert model $\{g_{i,j} \circ f_i\}_{f_i \in A}$

• $h_{A,j}$ as the ensemble model link
Problem Statement
Multi-model Inference under a Budget

Target Application
- inference results of multiple models are required
- cost budget is too limited to run them all
Problem Statement
Multi-model Inference under a Budget

- cost function \( c(\cdot) \)
  - e.g., GPU memory, inference delay

Target Application
- inference results of multiple models are required
- cost budget is too limited to run them all
Problem Statement

Multi-model Inference under a Budget

• cost function $c(\cdot)$
  • e.g., GPU memory, inference delay
• cost budget $B$

Target Application

• inference results of multiple models are required
• cost budget is too limited to run them all
Problem Statement

Multi-model Inference under a Budget

- cost function \( c(\cdot) \)
  - e.g., GPU memory, inference delay
- cost budget \( B \)
- performance measurement \( p_j(h_{A,j}) \)
  - normalized into \([0,1]\)
  - e.g., accuracy for classification, IoU for detection

Target Application

- inference results of multiple models are required
- cost budget is too limited to run them all
Problem Statement

Multi-model Inference under a Budget

• cost function \( c(\cdot) \)
• cost budget \( B \)
• performance measurement \( p_j(h_{A,j}) \)

Target Application

• inference results of multiple models are required
• cost budget is too limited to run them all

Optimization Problem

\[
\max_{A \subseteq F} \left( \frac{1}{|F|} \left( \sum_{f_i \in A} 1 + \sum_{f_j \notin A} p_j(h_{A,j}) \right) \right) \\
\text{s.t.} \quad \sum_{f_i \in A} c(f_i) + \sum_{f_j \notin A} c(h_{A,j}) \leq B.
\]
Main contents

• Introduction
• Problem Statement

• Black-box Model Linking
• Collaborative Multi-model Inference
• Evaluation
• Conclusion
Black-box Model Linking

Black-box outputs or intermediate features?

• real-world deployment typically provide only black-box inference API
  • virtual machine, container, …
Black-box Model Linking

Black-box outputs or intermediate features?

• real-world deployment typically provide only black-box inference API

• given the same (or aligned) inputs, correlations between black-box outputs are more explicit and easier to learn

• experimental evidences

![Graph showing mAP (%) against Ratio of training data (%) for output-based and feature-based links.](image)

- output-based
- feature-based

intermediate feature-based link

black-box output-based link
Black-box Model Linking

Black-box outputs or intermediate features?

- real-world deployment typically provide only black-box inference API

- given the same (or aligned) inputs, correlations between black-box outputs are more explicit and easier to learn
  - experimental evidences
  - theoretical evidences

When the training data is abundant for the representation shared among tasks, learning a new task branch \( f \in F \) requires \( C(F) \) sample complexity, where \( C(\cdot) \) measures the complexity of a hypothesis family.
Black-box Model Linking

Model link architecture

- output formats determine the model link’s architecture
  - fixed-length vector & variable-length sequence
Black-box Model Linking

Model link architecture

- output formats determine the model link’s architecture
  - fixed-length vector & variable-length sequence
- Vec-to-Vec
  - ReLU-activated multilayer perception (MLP)
Black-box Model Linking

Model link architecture

• output formats determine the model link’s architecture
  • fixed-length vector & variable-length sequence

• Seq-to-Vec
  • Embedding - LSTM - MLP
Black-box Model Linking

Model link architecture

- output formats determine the model link’s architecture
  - fixed-length vector & variable-length sequence

- Vec-to-Seq
  - MLP Encoder
  - Embedding - LSTM - Attention - MLP Decoder
Black-box Model Linking

Model link architecture

• output formats determine the model link’s architecture
  • fixed-length vector & variable-length sequence

• Seq-to-Seq
  • Embedding - LSTM Encoder
  • Embedding - LSTM - Attention - MLP Decoder
Black-box Model Linking

Model link architecture

- output formats determine the model link’s architecture
  - fixed-length vector & variable-length sequence
- target model’s task determines output activation
  - softmax for single-label classification, linear for regression and localization, etc.
• weighted sum of model links

\[ h_{A,j} = \sigma \left( \sum_{f_i \in A} g_{i,j} \circ f_i(x_i) \right) \]

• \( \sigma \) denotes the activation function
Black-box Model Linking

Training

- soft-label supervision

- knowledge distillation methods show that the teacher model’s outputs augment the hard-label space with relations among different classes

\[
\min \sum_{i=l}^{n} L_f(h_{A,f}(\{y_i^l\}_{f\in A}), y_j^l)
\]

- target model’s task determines the loss function
Main contents

• Introduction
• Problem Statement
• Black-box Model Linking

• Collaborative Multi-model Inference
• Evaluation
• Conclusion
Collaborative Multi-model Inference

Assumptions and Observations

- $F(A)$ as the objective function to optimize
- Gain of selecting one more model $f_i$

\[
\Delta(A, f_i) = F(A \cup \{f_i\}) - F(A)
\]
Assumptions and Observations

- \( F(A) \) as the objective function to optimize
  - gain of selecting one more model \( f_i \)
    \[ \Delta(A, f_i) = F(A \cup \{ f_i \}) - F(A) \]
  - Assume that adding a source of model link into the ensemble model will not decrease the performance:
    \[ p(A \cup \{ f_i \}, f_j) \geq p(A, f_j) \]
  - Then \( \Delta(A, f_i) \geq 0 \), i.e., the objective function is nondecreasing.

Optimization Problem

\[
\max_{A \subseteq F} \left( \frac{1}{|F|} \left( \sum_{f_j \in A} 1 + \sum_{f_j \in F \setminus A} p_j(h_{A,j}) \right) \right) \\
\text{s.t.} \quad \sum_{f_i \in A} c(f_i) + \sum_{f_j \in F \setminus A} c(h_{A,j}) \leq B.
\]
Collaborative Multi-model Inference

Assumptions and Observations

- two cases observed

\[ f_{i*} = \arg\max_{f_i \in A} p_j(g_{ij}) \]
\[ p_j(h_{A,f_j}) \approx p_j(g_{i*}, j) \]
Collaborative Multi-model Inference

Assumptions and Observations

• two cases observed

  • **dominance:** the performance of the ensemble model approximately equals the best-performance source of model links.

  • **mutual assistance:** the multi-source model links ensemble outperforms any single source.

\[
\forall f_i \in A, p_j(h_{A,f_i}) > p_j(g_{i,j})
\]
Collaborative Multi-model Inference

Activation Probability

- solving the optimization problem is NP-hard and the existing $(1 - 1/e)$-approximation algorithm needs partial-enumeration and requires $O(n^5)$ computations of the objective function.

Collaborative Multi-model Inference

Activation Probability

- three factors

- the average performance of model links from $f_i$ to all the others

$$P_i^1 = \frac{\sum_{j \neq i} p_j(g_{i,j})}{|F| - 1}$$
Collaborative Multi-model Inference

Activation Probability

• three factors

• the average performance of model links from $f_i$ to all the others

$$P_i^1 = \frac{\sum_{j \neq i} p_j(g_{i,j})}{|F| - 1}$$

• the average performance of model links targeted to $f_i$ from all the others

$$P_i^2 = \frac{\sum_{j \neq i} p_j(g_{j,i})}{|F| - 1}$$
Collaborative Multi-model Inference

Activation Probability

- three factors
  - the average performance of model links from $f_i$ to all the others
    \[
    P^1_i = \frac{\sum_{j \neq i} p_j(g_{i,j})}{|F| - 1}
    \]
  - the average performance of model links targeted to $f_i$ from all the others
    \[
    P^2_i = \frac{\sum_{j \neq i} p_j(g_{j,i})}{|F| - 1}
    \]
  - the cost of $f_i$ $c(f_i)$
Collaborative Multi-model Inference

Activation Probability

- definition

\[ P_i = \frac{1 + P_i^1 - P_i^2}{wc(f_i)} \]

\[ w = 2 / \min_i c(f_i) \text{ by normalization} \]

- This activation probability can be regarded as a coefficient that is positively correlated with the gain when selecting a model.

\[ P_i^1 = \frac{\sum_{j \neq i} p_j(g_{i,j})}{|F| - 1} \quad P_i^2 = \frac{\sum_{j \neq i} p_j(g_{j,i})}{|F| - 1} \]
Collaborative Multi-model Inference

Algorithm

• select greedily w.r.t. activation probability under the cost budget

• activated models do exact inference while the others’ outputs will be predicted by the model link ensemble of activated sources.
Collaborative Multi-model Inference

Algorithm

• select greedily w.r.t. activation probability under the cost budget

• activated models do exact inference while the others’ outputs will be predicted by the model link ensemble of activated sources.

• periodic re-profiling and re-selection

  • By reasonably setting the period length and the proportion of data used for profiling, we can amortize the overheads of loading/unloading ML models to negligible.
Main contents

• Introduction

• Problem Statement

• Black-box Model Linking

• Collaborative Multi-model Inference

• Evaluation

• Conclusion
• **MLink** implemented in Python based on TensorFlow 2.0

• We tested the integration on programs implemented with TensorFlow, PyTorch and MindSpore.
Evaluation

Datasets and Models

- Hollywood2
  - reprocess original videos to obtain a multi-modality dataset
- 7 models deployed

<table>
<thead>
<tr>
<th>Task Class</th>
<th>ML Model</th>
<th>Input Modality</th>
<th>Output Format</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-label Classification</td>
<td>Gender Classification</td>
<td>Audio</td>
<td>2-D Softmax Labels</td>
<td>Acc.</td>
</tr>
<tr>
<td>Multi-label Classification</td>
<td>Action Classification</td>
<td>Video</td>
<td>12-D Sigmoid Labels</td>
<td>mAP</td>
</tr>
<tr>
<td>Localization</td>
<td>Face Detection</td>
<td>Image</td>
<td>4-D Bounding Box</td>
<td>IoU</td>
</tr>
<tr>
<td></td>
<td>Person Detection</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>Age Prediction</td>
<td>Image</td>
<td>1-D Scalar</td>
<td>MAE</td>
</tr>
<tr>
<td>Sequence Generation</td>
<td>Image Captioning</td>
<td>Image</td>
<td>Variable-length Text</td>
<td>WER</td>
</tr>
<tr>
<td></td>
<td>Speech Recognition</td>
<td>Audio</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation
Datasets and Models

- Hollywood2
  - reprocess original videos to obtain a multi-modality dataset
- 7 models deployed

<table>
<thead>
<tr>
<th>Task Class</th>
<th>ML Model</th>
<th>Input Modality</th>
<th>Output Format</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-label Classification</td>
<td>Gender Classification</td>
<td>Audio</td>
<td>2-D Softmax Labels</td>
<td>Acc.</td>
</tr>
<tr>
<td>Multi-label Classification</td>
<td>Action Classification</td>
<td>Video</td>
<td>12-D Sigmoid Labels</td>
<td>mAP</td>
</tr>
<tr>
<td>Localization</td>
<td>Face Detection</td>
<td>Image</td>
<td>4-D Bounding Box</td>
<td>IoU</td>
</tr>
<tr>
<td></td>
<td>Person Detection</td>
<td>Image</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>Age Prediction</td>
<td>Image</td>
<td>1-D Scalar</td>
<td>MAE</td>
</tr>
<tr>
<td>Sequence Generation</td>
<td>Image Captioning</td>
<td>Image</td>
<td>Variable-length Text</td>
<td>WER</td>
</tr>
<tr>
<td></td>
<td>Speech Recognition</td>
<td>Audio</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation
Datasets and Models

- Hollywood2
  - reprocess original videos to obtain a mutli-modality dataset
  - 7 models deployed

<table>
<thead>
<tr>
<th>Task Class</th>
<th>ML Model</th>
<th>Input Modality</th>
<th>Output Format</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-label Classification</td>
<td>Gender Classification</td>
<td>Audio</td>
<td>2-D Softmax Labels</td>
<td>Acc.</td>
</tr>
<tr>
<td>Multi-label Classification</td>
<td>Action Classification</td>
<td>Video</td>
<td>12-D Sigmoid Labels</td>
<td>mAP</td>
</tr>
<tr>
<td>Localization</td>
<td></td>
<td>Image</td>
<td>4-D Bounding Box</td>
<td>IoU</td>
</tr>
<tr>
<td>Regression</td>
<td>Age Prediction</td>
<td>Image</td>
<td>1-D Scalar</td>
<td>MAE</td>
</tr>
<tr>
<td>Sequence Generation</td>
<td>Image Captioning</td>
<td>Image</td>
<td>Variable-length Text</td>
<td>WER</td>
</tr>
<tr>
<td></td>
<td>Speech Recognition</td>
<td>Audio</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation
Datasets and Models

- Hollywood2
- reprocess original videos to obtain a multi-modality dataset
- 7 models deployed

Table 1: ML Models on Hollywood2 Dataset

<table>
<thead>
<tr>
<th>Task Class</th>
<th>ML Model</th>
<th>Input Modality</th>
<th>Output Format</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-label Classification</td>
<td>Gender Classification</td>
<td>Audio</td>
<td>2-D Softmax Labels</td>
<td>Acc.</td>
</tr>
<tr>
<td>Multi-label Classification</td>
<td>Action Classification</td>
<td>Video</td>
<td>12-D Sigmoid Labels</td>
<td>mAP</td>
</tr>
<tr>
<td>Localization</td>
<td>Face Detection</td>
<td>Image</td>
<td>4-D Bounding Box</td>
<td>IoU</td>
</tr>
<tr>
<td></td>
<td>Person Detection</td>
<td>Image</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>Age Prediction</td>
<td>Image</td>
<td>1-D Scalar</td>
<td>MAE</td>
</tr>
<tr>
<td>Sequence Generation</td>
<td>Image Captioning</td>
<td>Image</td>
<td>Variable-length Text</td>
<td>WER</td>
</tr>
<tr>
<td></td>
<td>Speech Recognition</td>
<td>Audio</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation

Model Links’ Performance

• pairwise model links are trained using 1%, 5%, 10%, 20%, 48% data

• RMSprop optimizer with same hyper-parameters (0.01 learning rate, 100 epochs, 32 batch size)
Evaluation

Model Links’ Performance

• pairwise model links are trained using 1%, 5%, 10%, 20%, 48% data

• RMSprop optimizer with same hyper-parameters (0.01 learning rate, 100 epochs, 32 batch size)

• model links significantly outperform knowledge distillation-based student models
Evaluation

Semantic Correlation

- attention coverage has a positive correlation with the model linking performance

(a) Attention heatmaps of Object and Scene models.
(b) Scene-to-Object MLink accuracy vs. attention overlaps.
Evaluation

Semantic Correlation

- attention coverage has a positive correlation with the model linking performance
- Pearson correlation coefficients between outputs also show a positive correlation with the performance

Table 2: IoU scores of model links targeted to the Pearson model and the Pearson correlations.

<table>
<thead>
<tr>
<th>Source</th>
<th>Action</th>
<th>Age</th>
<th>Face</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>IoU (%)</td>
<td>39.4</td>
<td>38.9</td>
<td>58.5</td>
<td>39.0</td>
</tr>
<tr>
<td>Pearson Corr.</td>
<td>0.123</td>
<td>0.042</td>
<td><strong>0.244</strong></td>
<td>-0.053</td>
</tr>
</tbody>
</table>

(a) Attention heatmaps of Object and Scene models.
(b) Scene-to-Object MLink accuracy vs. attention overlaps.
Evaluation

MLink Ensemble

• dominance cases

\[ p_j(h_{A,j}) \approx p_j(g_{i*,j}) \]

Table 3: Dominance and mutual assistance cases in model link ensemble. Column titles are source models and row titles are target models. The dominant source’s performance is in bold.

<table>
<thead>
<tr>
<th>Target \ Source</th>
<th>Action</th>
<th>Age</th>
<th>Caption</th>
<th>Face</th>
<th>Gender</th>
<th>Person</th>
<th>Speech</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action mAP (%)</td>
<td>-</td>
<td>12.8</td>
<td>29.7</td>
<td>10.1</td>
<td>9.3</td>
<td>9.9</td>
<td>8.5</td>
<td>30.8</td>
</tr>
<tr>
<td>Face IoU (%)</td>
<td>11</td>
<td>11.2</td>
<td>0</td>
<td>-</td>
<td>10.3</td>
<td>31.9</td>
<td>0</td>
<td>32.2</td>
</tr>
<tr>
<td>Person IoU (%)</td>
<td>39.4</td>
<td>38.9</td>
<td>0</td>
<td>58.5</td>
<td>39.0</td>
<td>-</td>
<td>0</td>
<td>59.2</td>
</tr>
<tr>
<td>Age MAE</td>
<td>3.04</td>
<td>-</td>
<td>3.02</td>
<td>3.07</td>
<td>3.0</td>
<td>3.03</td>
<td>3.0</td>
<td>2.98</td>
</tr>
<tr>
<td>Gender Acc. (%)</td>
<td>92</td>
<td>92.1</td>
<td>92</td>
<td>92.1</td>
<td>-</td>
<td>92</td>
<td>92</td>
<td>92.3</td>
</tr>
</tbody>
</table>
Evaluation

MLink Ensemble

- dominance cases
- mutual assistance cases \( \quad \forall f_i \in A, p_j(h_{A,f_j}) > p_j(g_{i,j}) \)

Table 3: Dominance and mutual assistance cases in model link ensemble. Column titles are source models and row titles are target models. The dominant source’s performance is in bold.

<table>
<thead>
<tr>
<th>Target \ Source</th>
<th>Action</th>
<th>Age</th>
<th>Caption</th>
<th>Face</th>
<th>Gender</th>
<th>Person</th>
<th>Speech</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action mAP (%)</td>
<td>-</td>
<td>12.8</td>
<td>29.7</td>
<td>10.1</td>
<td>9.3</td>
<td>9.9</td>
<td>8.5</td>
<td>30.8</td>
</tr>
<tr>
<td>Face IoU (%)</td>
<td>11</td>
<td>11.2</td>
<td>0</td>
<td>-</td>
<td>10.3</td>
<td>31.9</td>
<td>0</td>
<td>32.2</td>
</tr>
<tr>
<td>Person IoU (%)</td>
<td>39.4</td>
<td>38.9</td>
<td>0</td>
<td>58.5</td>
<td>39.0</td>
<td>-</td>
<td>0</td>
<td>59.2</td>
</tr>
<tr>
<td>Age MAE</td>
<td>3.04</td>
<td>-</td>
<td>3.02</td>
<td>3.07</td>
<td>3.0</td>
<td>3.03</td>
<td>3.0</td>
<td>2.98</td>
</tr>
<tr>
<td>Gender Acc. (%)</td>
<td>92</td>
<td>92.1</td>
<td>92</td>
<td>92.1</td>
<td>-</td>
<td>92</td>
<td>92</td>
<td>92.3</td>
</tr>
</tbody>
</table>
Evaluation

Real Systems

- Smart Building
  - two days (one weekday & one weekend) of videos (1 frame per minute) from 58 cameras
  - 3 models deployed
    - person counting, action classification, object counting
Evaluation

Real Systems

• City Traffic
  • two days (one weekday & one weekend) of videos (1 FPS) from 10 cameras at road intersections
  • 3 models deployed
    • person counting, traffic condition classification, vehicle counting
Evaluation

Baselines

• **Standalone**: selects models in ascending order of delay and runs models independently

Target Application

• inference results of multiple models are required
• cost budget is too limited to run them all
Evaluation

Baselines

• **Standalone**: selects models in ascending order of delay and runs models independently

• **MTL**: a multi-task learning approach

**Target Application**

• inference results of multiple models are required
• cost budget is too limited to run them all
Evaluation

Baselines

- **Standalone**: selects models in ascending order of delay and runs models independently
- **MTL**: a multi-task learning approach
- **DRLS**: a deep reinforcement learning-based scheduling approach

Target Application

- inference results of multiple models are required
- cost budget is too limited to run them all
Evaluation

Baselines

- **Standalone**: selects models in ascending order of delay and runs models independently
- **MTL**: a multi-task learning approach
- **DRLS**: a deep reinforcement learning-based scheduling approach
- **Reducto**: a low-level feature difference-based frame filtering approach

Target Application

- inference results of multiple models are required
- cost budget is too limited to run them all
Evaluation

Video Analytics with Model Links

• GPU Memory as the cost budget

Table 4: Comparisons of MLink, MTL, Reducto, DRLS, and Standalone

<table>
<thead>
<tr>
<th>Method</th>
<th>Building (5/9 GB Mem.)</th>
<th>City (5/9 GB Mem.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc. (%)</td>
<td>Time (ms)</td>
</tr>
<tr>
<td>Standalone</td>
<td>33.3/66.7</td>
<td>30/74</td>
</tr>
<tr>
<td>MTL</td>
<td>53.3</td>
<td>32.8</td>
</tr>
<tr>
<td>DRLS</td>
<td>45.7/81.3</td>
<td>58.7/107</td>
</tr>
<tr>
<td>Reducto</td>
<td>91.8/96.9</td>
<td>45.7/89</td>
</tr>
<tr>
<td><strong>MLink</strong></td>
<td><strong>94.1/97.9</strong></td>
<td><strong>39.3/84</strong></td>
</tr>
</tbody>
</table>
Evaluation

Video Analytics with Model Links

- GPU Memory as the cost budget

Table 4: Comparisons of MLink, MTL, Reducto, DRLS, and Standalone

<table>
<thead>
<tr>
<th>Method</th>
<th>Building (5/9 GB Mem.)</th>
<th>City (5/9 GB Mem.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc. (%)</td>
<td>Time (ms)</td>
</tr>
<tr>
<td>Standalone</td>
<td>33.3/66.7</td>
<td>30/74</td>
</tr>
<tr>
<td>MTL</td>
<td>53.3</td>
<td>32.8</td>
</tr>
<tr>
<td>DRLS</td>
<td>45.7/81.3</td>
<td>58.7/107</td>
</tr>
<tr>
<td>Reducto</td>
<td>91.8/96.9</td>
<td>45.7/89</td>
</tr>
<tr>
<td><strong>MLink</strong></td>
<td><strong>94.1/97.9</strong></td>
<td><strong>39.3/84</strong></td>
</tr>
</tbody>
</table>

*fast but accuracy is too low*
Evaluation

Video Analytics with Model Links

- GPU Memory as the cost budget

Table 4: Comparisons of MLink, MTL, Reducto, DRLS, and Standalone

<table>
<thead>
<tr>
<th>Method</th>
<th>Building (5/9 GB Mem.)</th>
<th>City (5/9 GB Mem.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc. (%)</td>
<td>Time (ms)</td>
</tr>
<tr>
<td>Standalone</td>
<td>33.3/66.7</td>
<td>30/74</td>
</tr>
<tr>
<td>MTL</td>
<td>53.3</td>
<td>32.8</td>
</tr>
<tr>
<td>DRLS</td>
<td>45.7/81.3</td>
<td>58.7/107</td>
</tr>
<tr>
<td>Reducto</td>
<td>91.8/96.9</td>
<td>45.7/89</td>
</tr>
<tr>
<td><strong>MLink</strong></td>
<td><strong>94.1/97.9</strong></td>
<td><strong>39.3/84</strong></td>
</tr>
</tbody>
</table>

*improved accuracy but too much overheads*
Evaluation

Video Analytics with Model Links

• GPU Memory as the cost budget

Table 4: Comparisons of MLink, MTL, Reducto, DRLS, and Standalone

<table>
<thead>
<tr>
<th>Method</th>
<th>Building (5/9 GB Mem.)</th>
<th>City (5/9 GB Mem.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc. (%)</td>
<td>Time (ms)</td>
</tr>
<tr>
<td>Standalone</td>
<td>33.3/66.7</td>
<td>30/74</td>
</tr>
<tr>
<td>MTL</td>
<td>53.3</td>
<td>32.8</td>
</tr>
<tr>
<td>DRLS</td>
<td>45.7/81.3</td>
<td>58.7/107</td>
</tr>
<tr>
<td>Reducto</td>
<td>91.8/96.9</td>
<td>45.7/89</td>
</tr>
<tr>
<td><strong>MLink</strong></td>
<td><strong>94.1/97.9</strong></td>
<td><strong>39.3/84</strong></td>
</tr>
</tbody>
</table>

*good trade-offs but only applicable to video streams*
## Evaluation

### Video Analytics with Model Links

- GPU Memory as the cost budget

Table 4: Comparisons of MLink, MTL, Reducto, DRLS, and Standalone

<table>
<thead>
<tr>
<th>Method</th>
<th>Building (5/9 GB Mem.)</th>
<th>City (5/9 GB Mem.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc. (%)</td>
<td>Time (ms)</td>
</tr>
<tr>
<td>Standalone</td>
<td>33.3/66.7</td>
<td>30/74</td>
</tr>
<tr>
<td>MTL</td>
<td>53.3</td>
<td>32.8</td>
</tr>
<tr>
<td>DRLS</td>
<td>45.7/81.3</td>
<td>58.7/107</td>
</tr>
<tr>
<td>Reducto</td>
<td>91.8/96.9</td>
<td>45.7/89</td>
</tr>
<tr>
<td><strong>MLink</strong></td>
<td><strong>94.1/97.9</strong></td>
<td><strong>39.3/84</strong></td>
</tr>
</tbody>
</table>

accurate, lightweight, and widely applicable
Introduction

Problem Statement

Black-box Model Linking

Collaborative Multi-model Inference

Evaluation

Conclusion
Take-home Messages

• effective connections between black-box outputs of models can be built via our model linking approach
Conclusion

Take-home Messages

• effective connections between black-box outputs of models can be built via our model linking approach

• model link-based scheduling is a promising way towards cost-performance trade-off of multi-model inference
MLink: Linking Black-box Models for Collaborative Multi-model Inference

Thanks for your listening.

Mu Yuan (ym0813@mail.ustc.edu.cn), Lan Zhang, Xiang-Yang Li
University of Science and Technology of China