36TH AAAI CONFERENCE
ON ARTIFICIAL INTELLIGENCE

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

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

中国科学技术大学
University of Science and Technology of China

Menu

中国科学技术大学
University of Science and Technology of China

Main contents

- **Introduction**
- Problem Statement
- Black-box Model Linking
- Collaborative Multi-model Inference
- Evaluation
- Conclusion

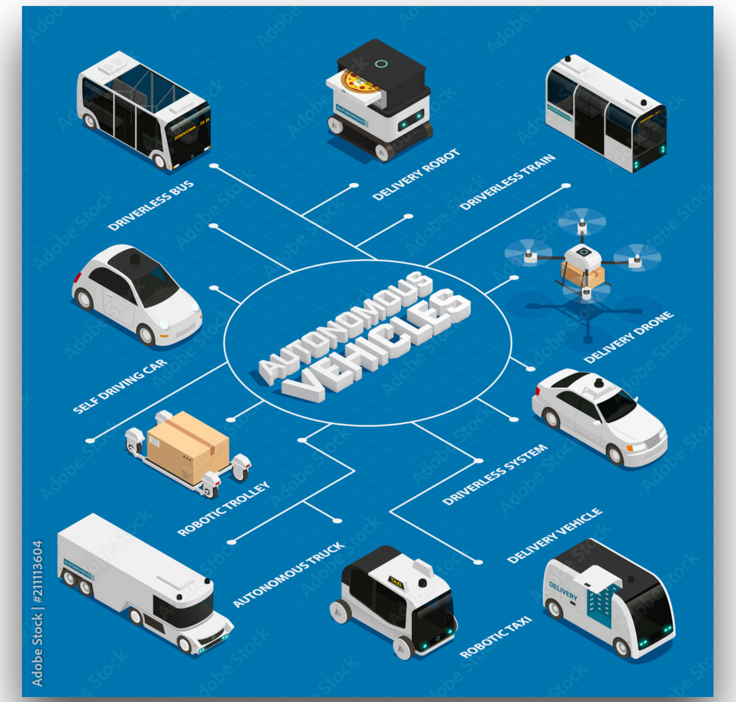
Introduction

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

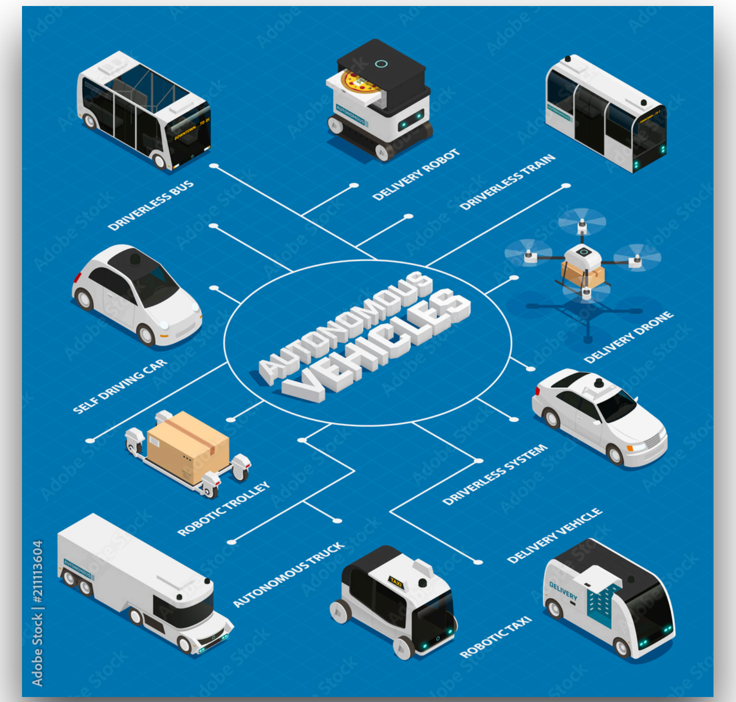
中国科学技术大学
University of Science and Technology of China

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

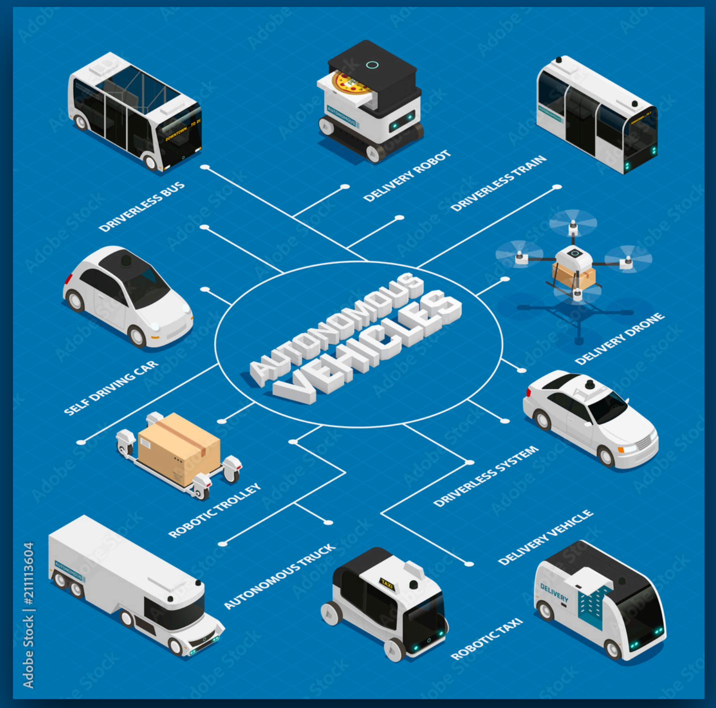
中国科学技术大学
University of Science and Technology of China

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

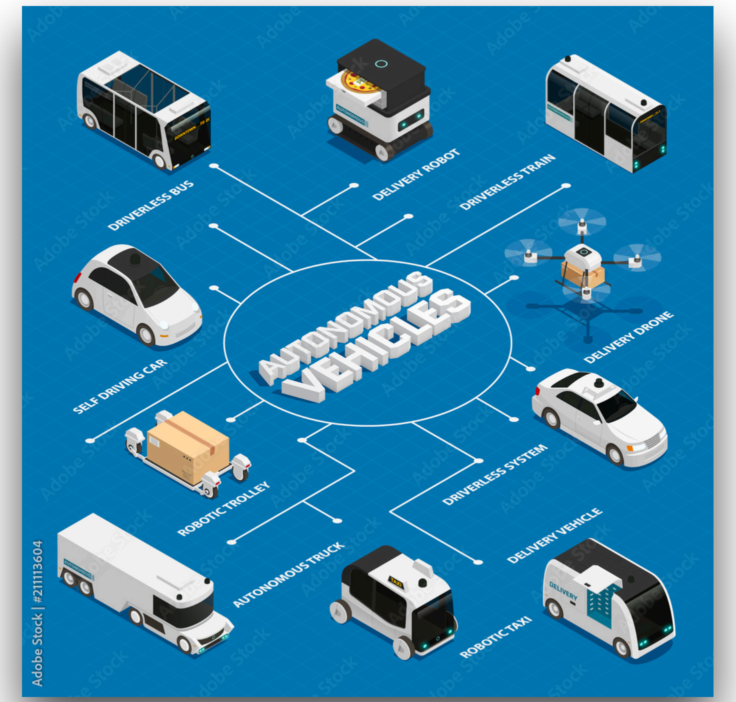
Introduction

Multi-model Inference Workloads

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

Smart Speaker

Intelligent Traffic



Autonomous Vehicles

Contextual Advertising

Introduction

中国科学技术大学
University of Science and Technology of China

Cost-effective Inference

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

Introduction

中国科学技术大学
University of Science and Technology of China

Cost-effective Inference

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

Cost-effective Inference

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

Cost-effective Inference

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

Introduction

中国科学技术大学
University of Science and Technology of China

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

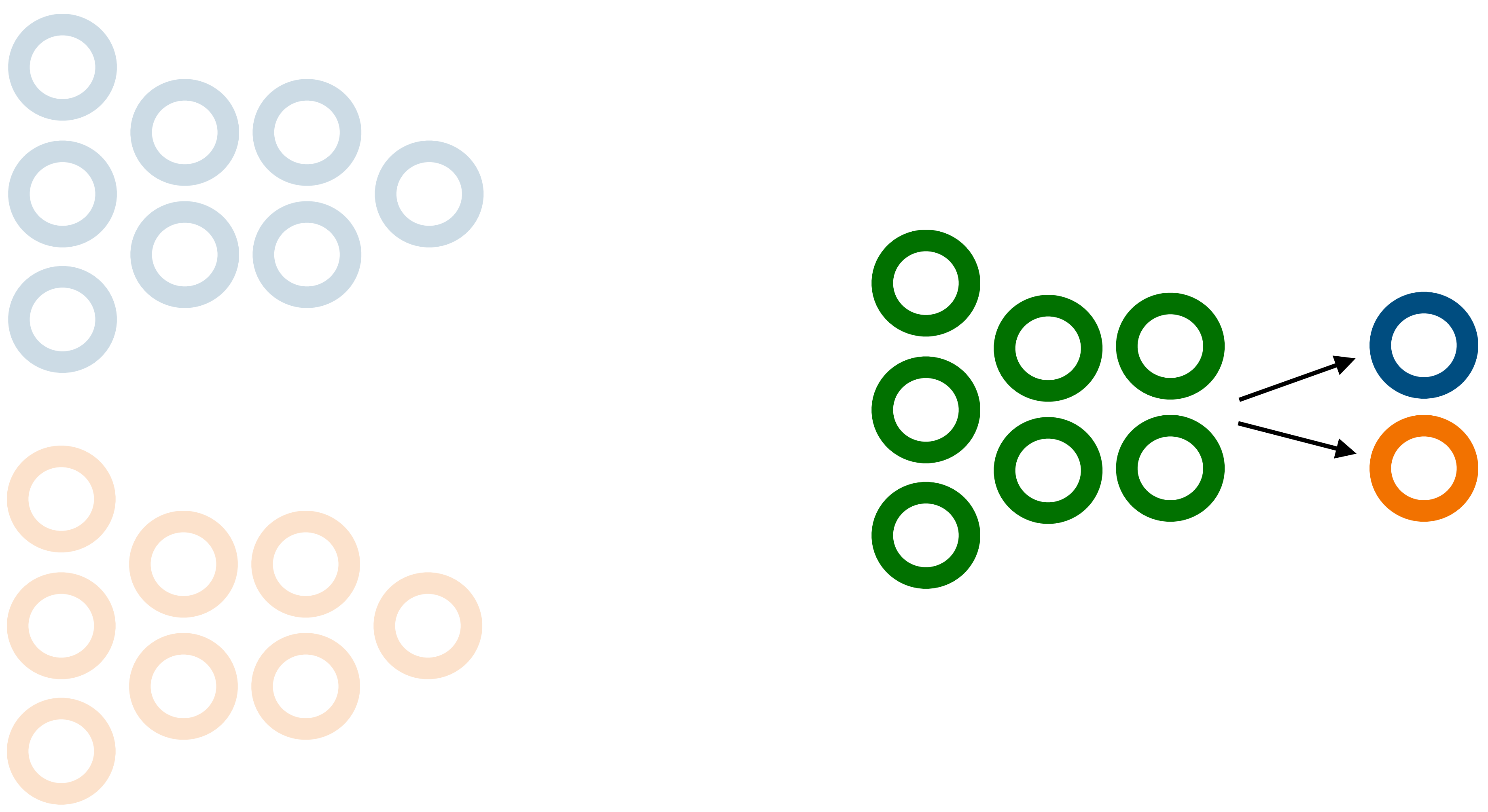
How to obtain as accurate inference results as possible without the exact execution of ML models?

Introduction

Cost-effective Inference

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

Exact Execution | Resulting Workload



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

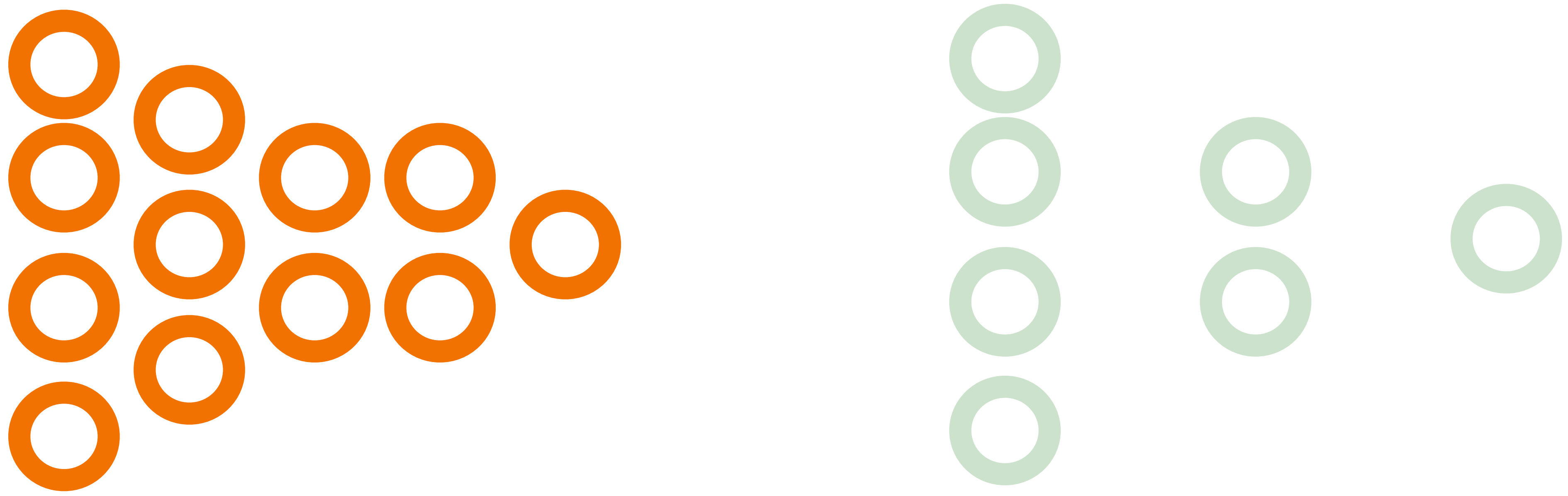
Introduction

中国科学技术大学
University of Science and Technology of China

Cost-effective Inference

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

Exact Execution　　Resulting Workload



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

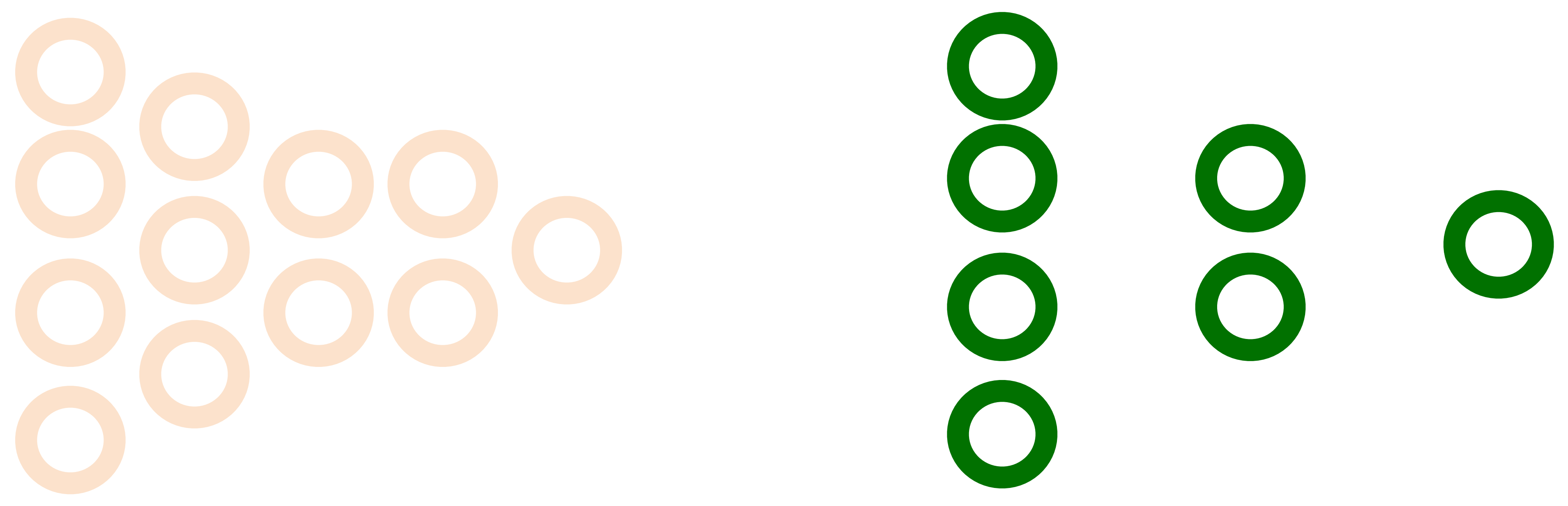
Introduction

中国科学技术大学
University of Science and Technology of China

Cost-effective Inference

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

Exact Execution | Resulting Workload



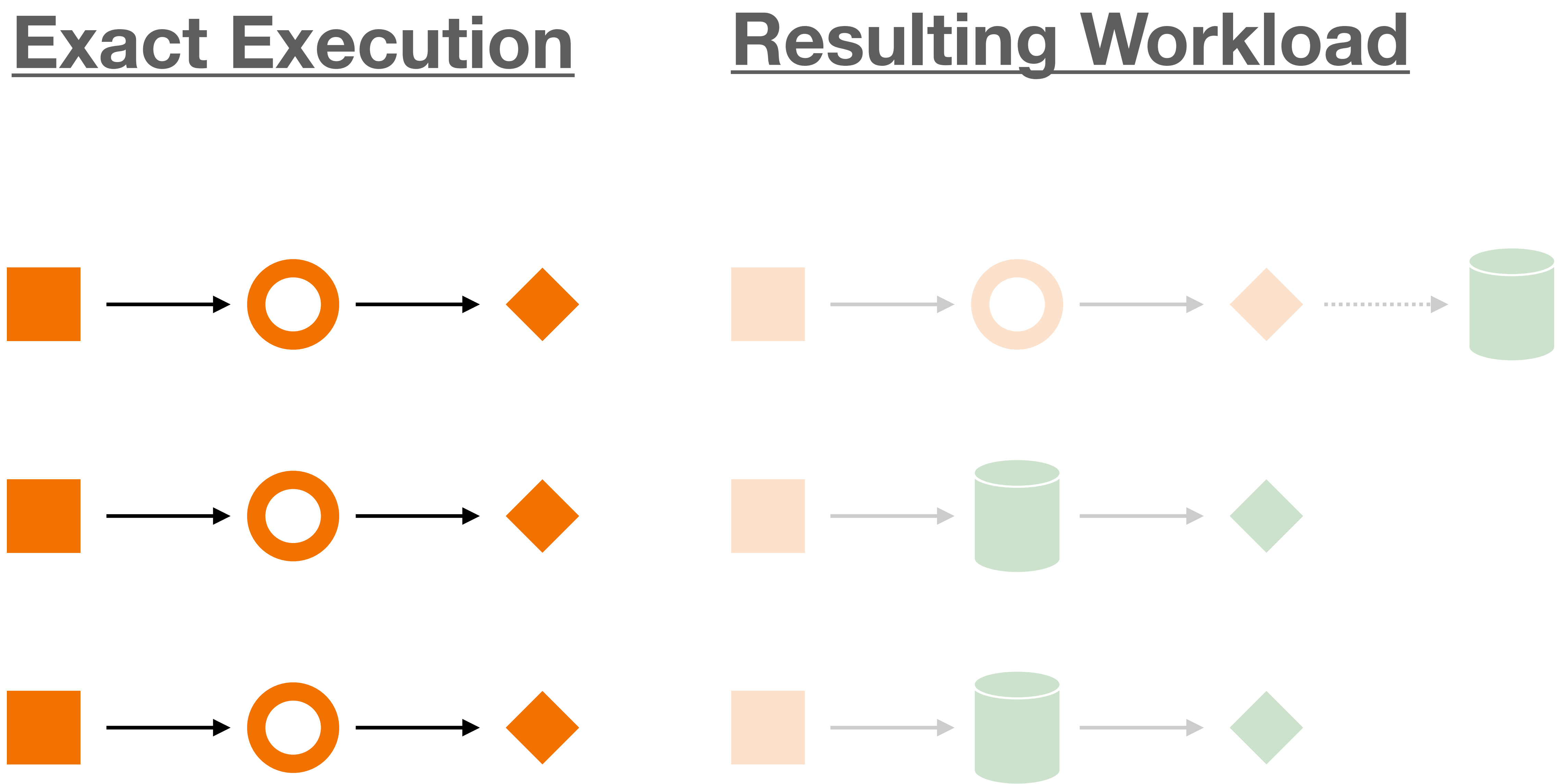
How to obtain as accurate inference results as possible without the exact execution of ML models?

Introduction

中国科学技术大学
University of Science and Technology of China

Cost-effective Inference

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



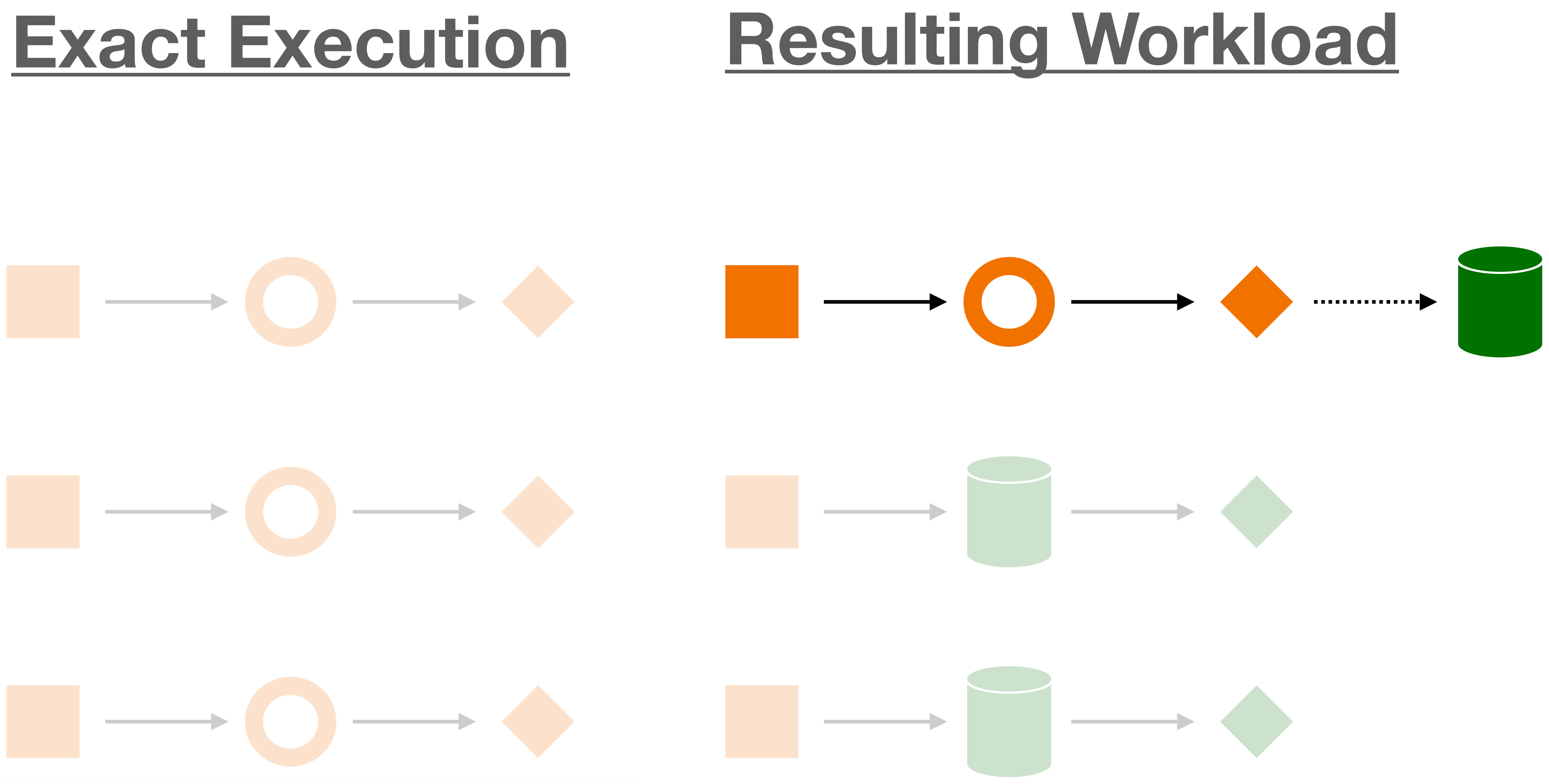
How to obtain as accurate inference results as possible without the exact execution of ML models?

Introduction

中国科学技术大学
University of Science and Technology of China

Cost-effective Inference

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

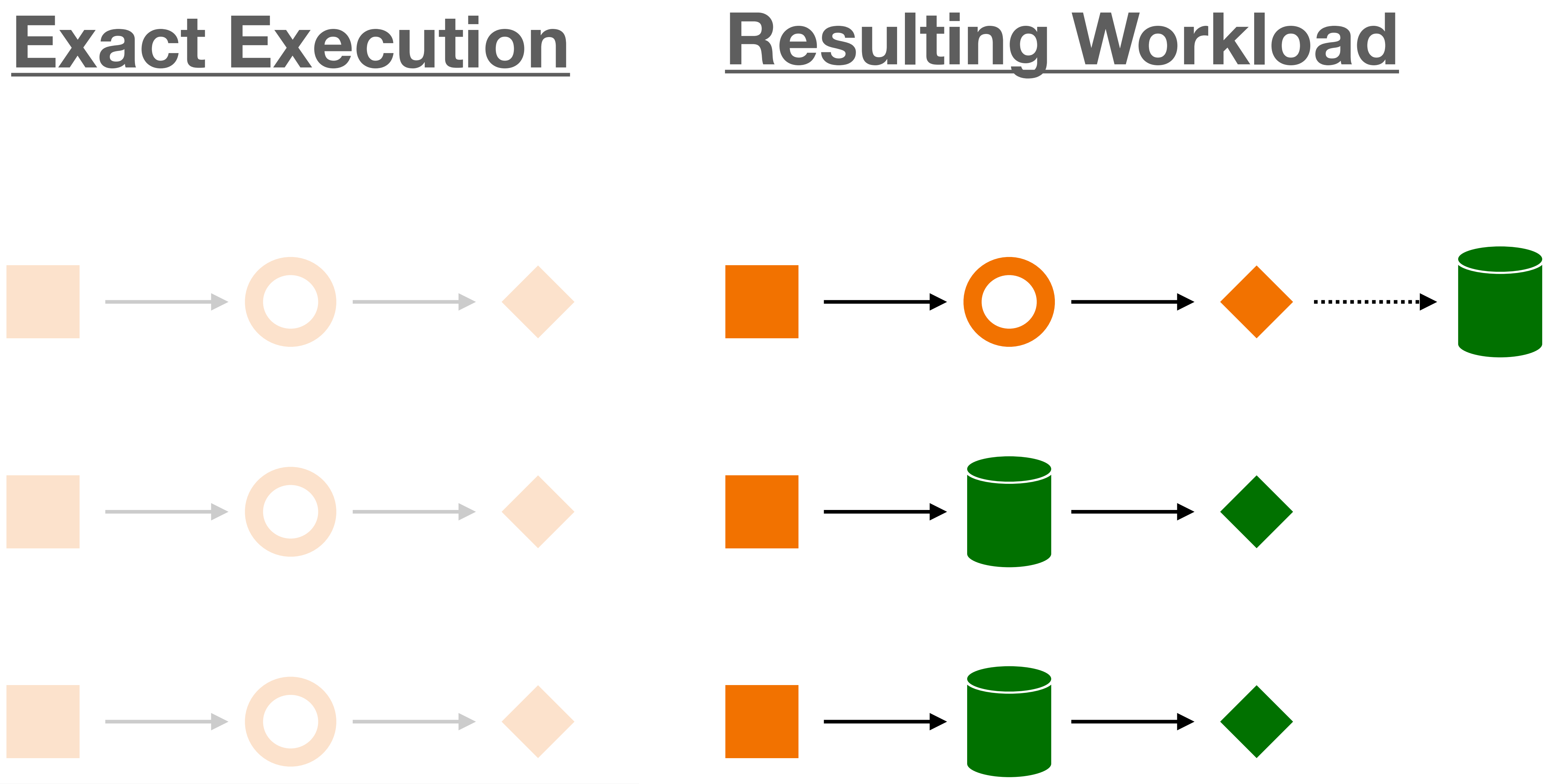


How to obtain as accurate inference results as possible without the exact execution of ML models?

Introduction

Cost-effective Inference

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



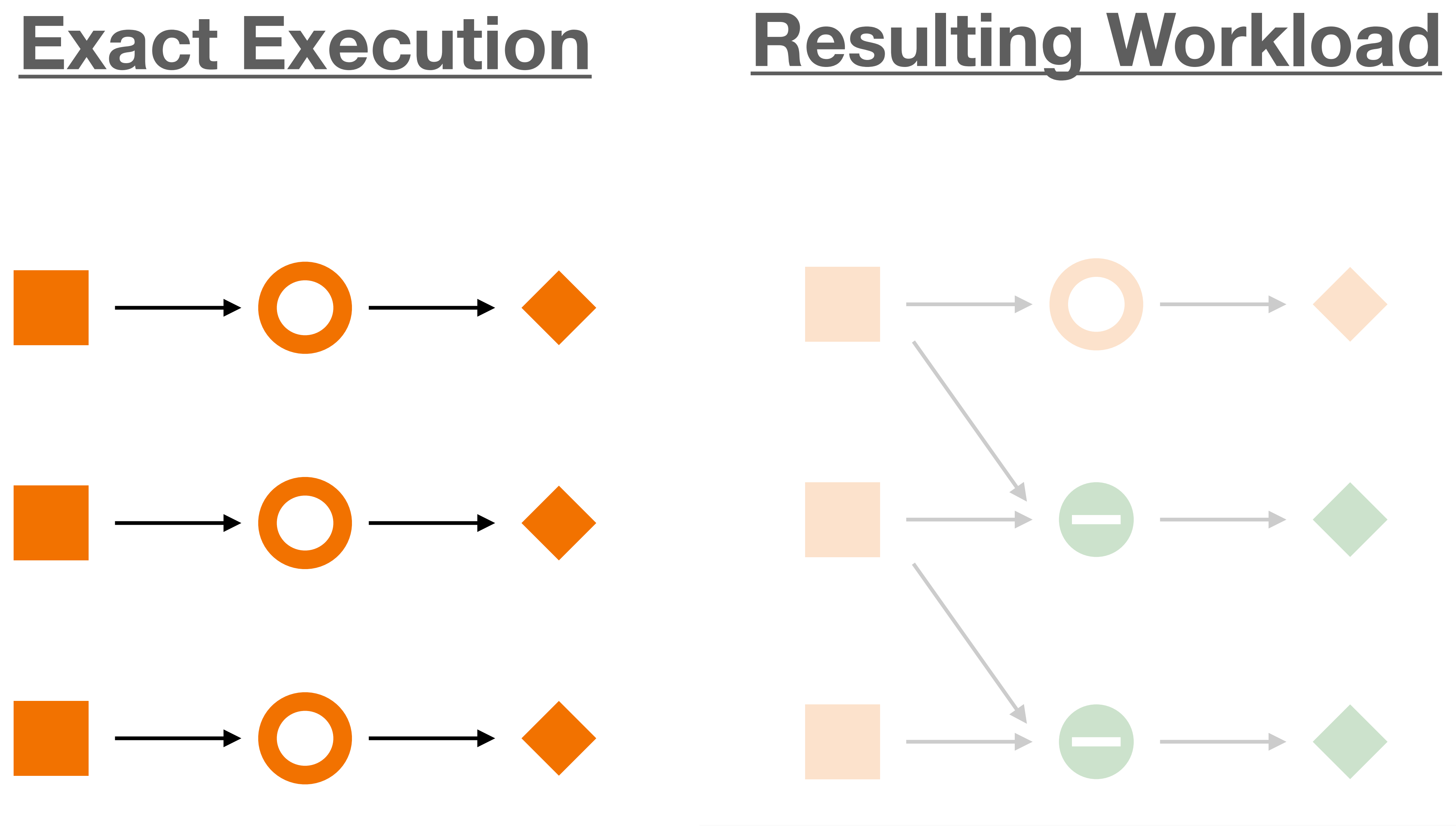
How to obtain as accurate inference results as possible without the exact execution of ML models?

Introduction

中国科学技术大学
University of Science and Technology of China

Cost-effective Inference

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



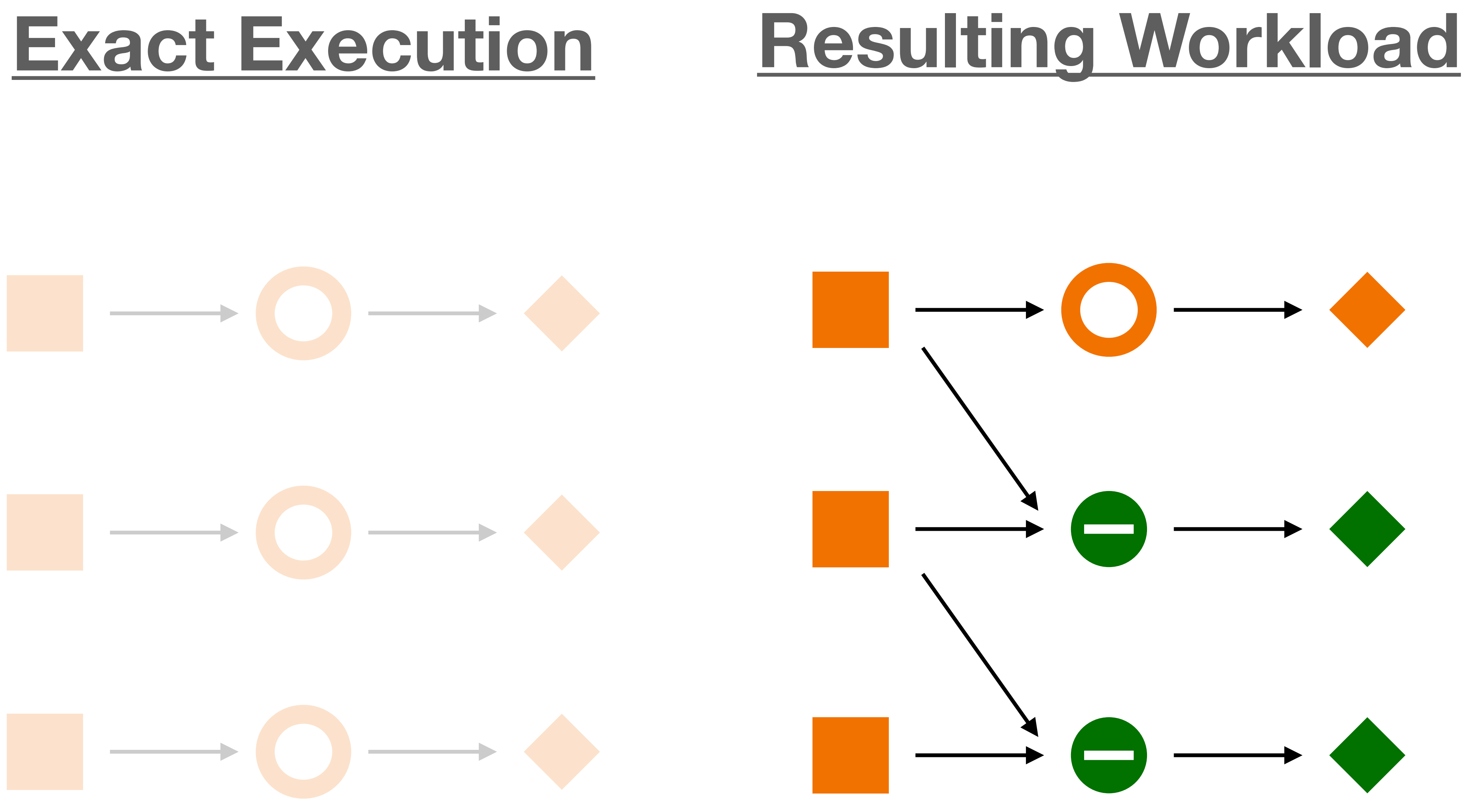
How to obtain as accurate inference results as possible without the exact execution of ML models?

Introduction

中国科学技术大学
University of Science and Technology of China

Cost-effective Inference

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



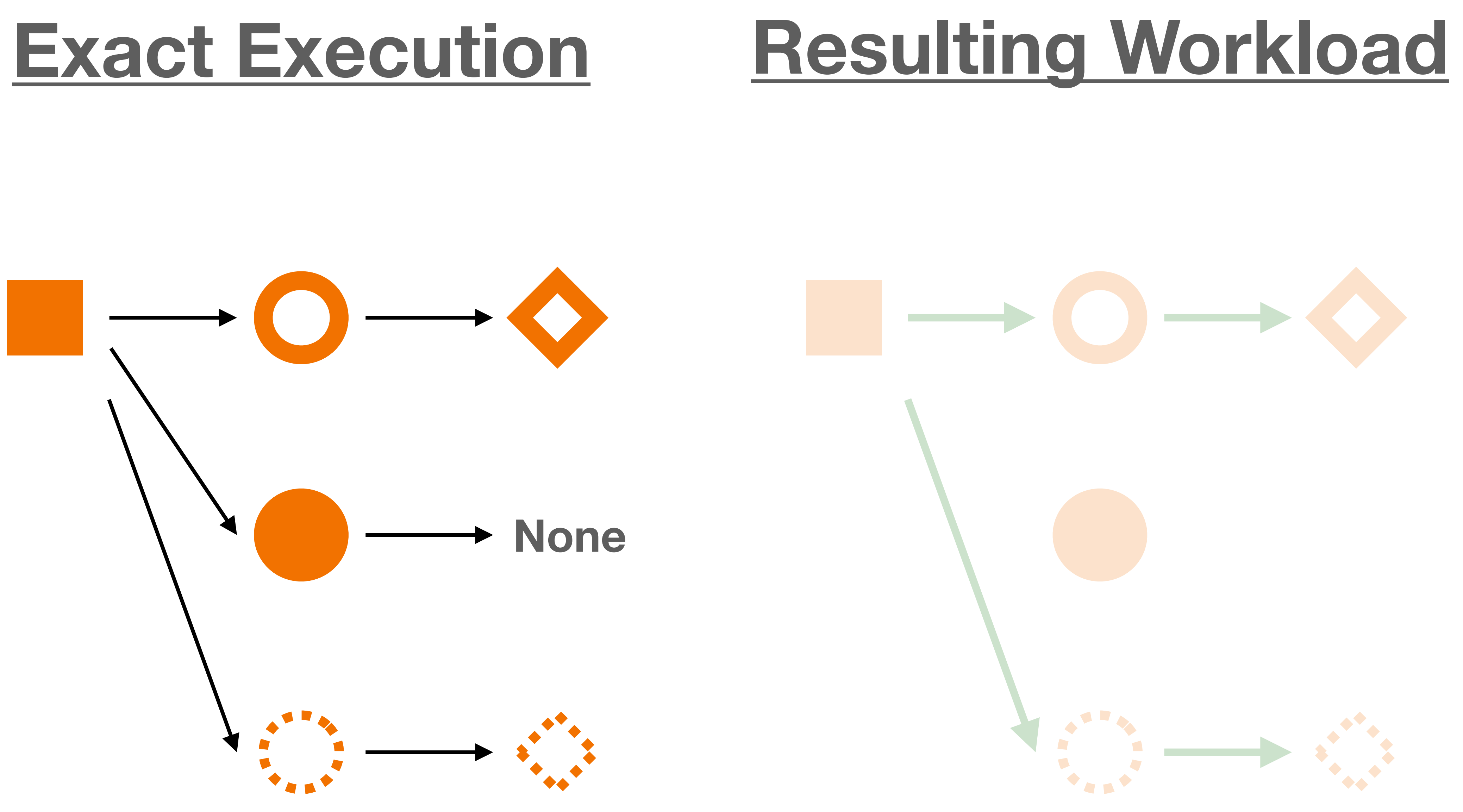
How to obtain as accurate inference results as possible without the exact execution of ML models?

Introduction

中国科学技术大学 University of Science and Technology of China

Cost-effective Inference

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



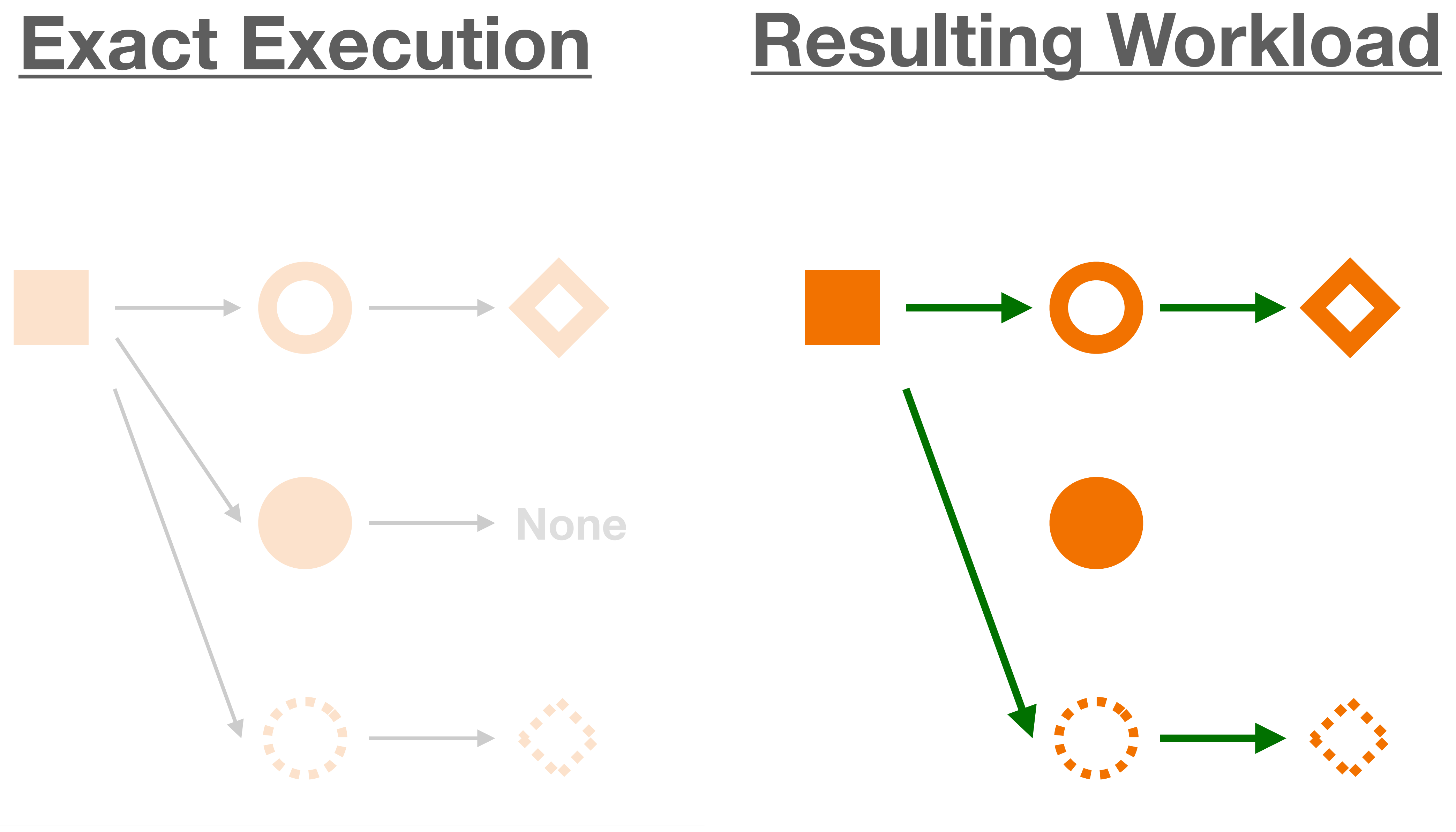
How to obtain as accurate inference results as possible without the exact execution of ML models?

Introduction

University of Science and Technology of China

Cost-effective Inference

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



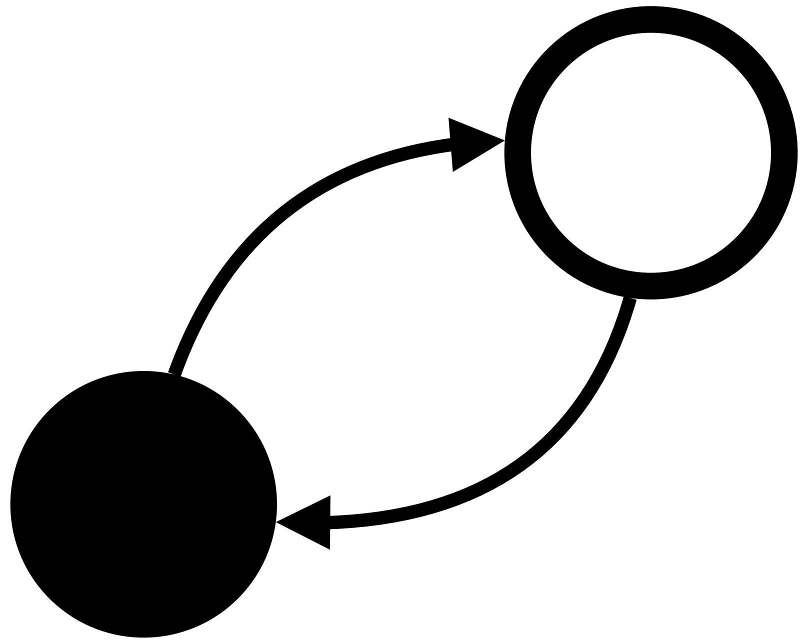
How to obtain as accurate inference results as possible without the exact execution of ML models?

Introduction

Linking Black-box Models

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

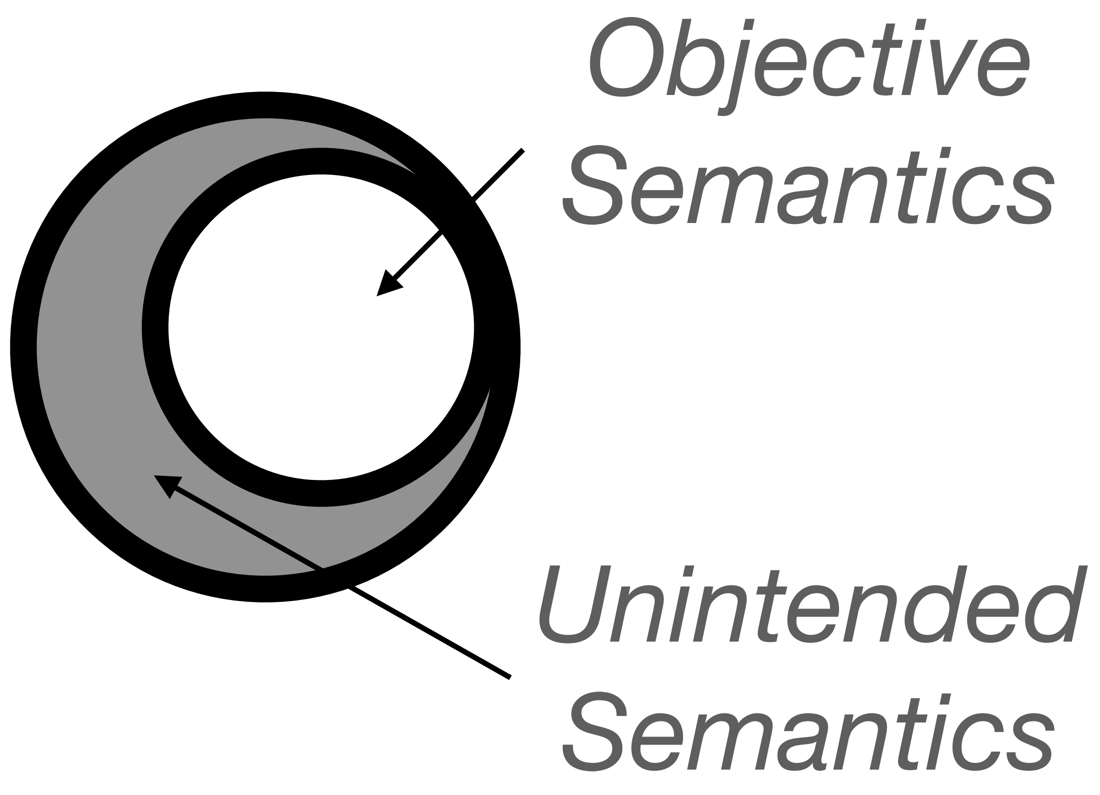
- **Model Linking**



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

Linking Black-box Models

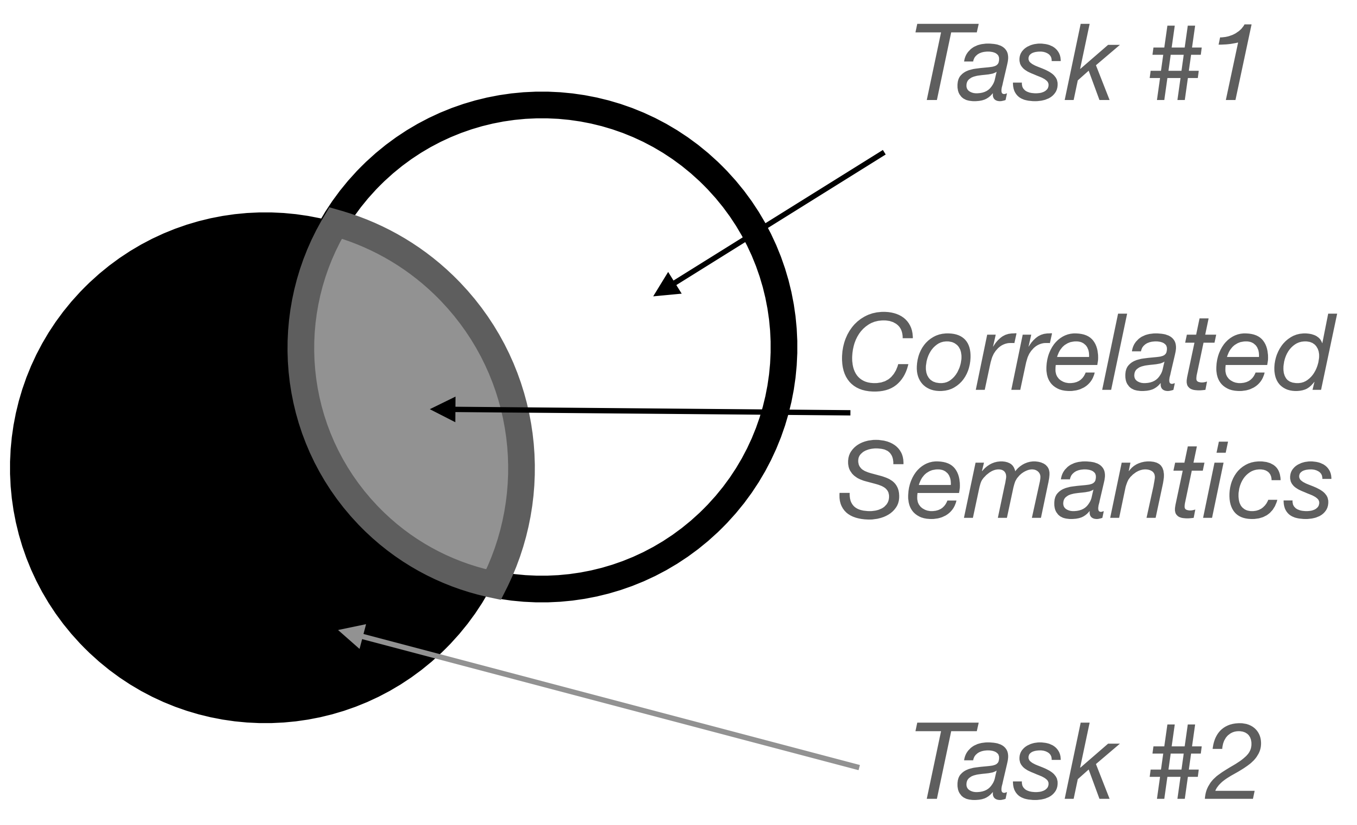
- **Model Linking**
  - machine over-learning



Introduction

Linking Black-box Models

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



Introduction

中国科学技术大学 University of Science and Technology of China

Linking Black-box Models

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

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

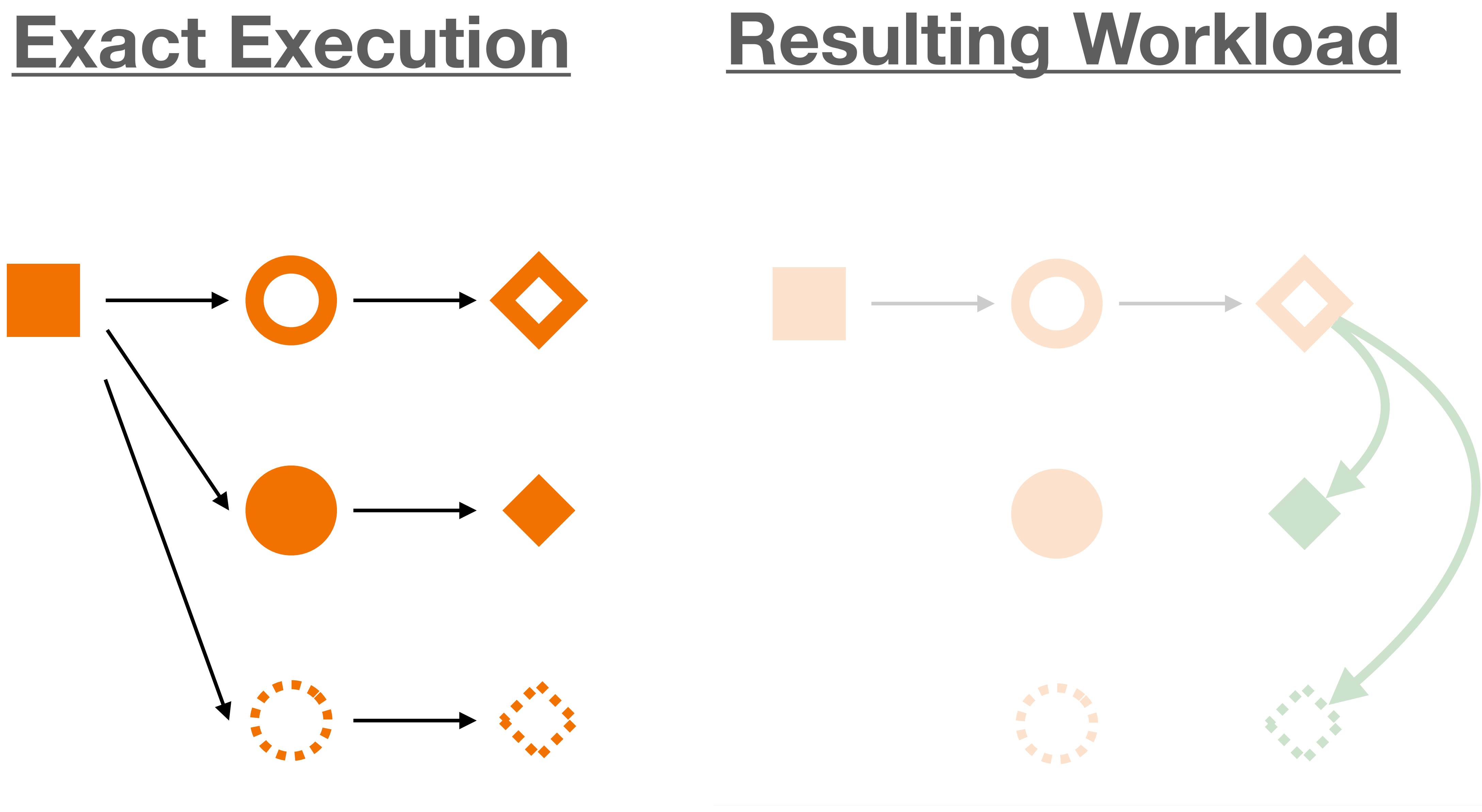
Introduction

中国科学技术大学
University of Science and Technology of China

Linking Black-box Models

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

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



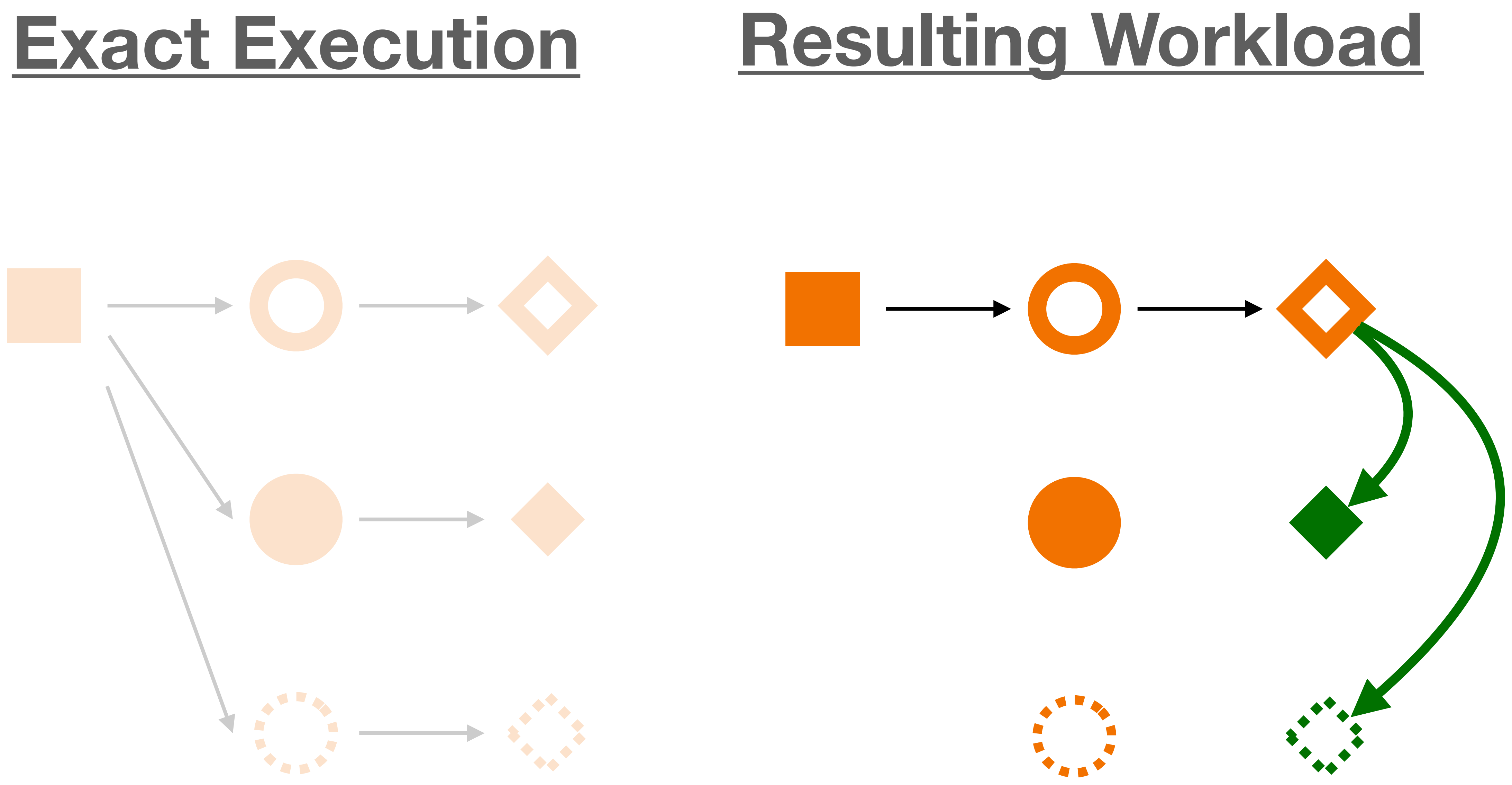
Introduction

中国科学技术大学
University of Science and Technology of China

Linking Black-box Models

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

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



Introduction

University of Science and Technology of China

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

CNNs, RNNs,
Auto-encoders,
Transformers …

Different DL frameworks

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

CNNs, RNNs, Auto-encoders, Transformers …

Different DL frameworks

non-intrusive design and implementation

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

University of Science and Technology of China

Main contents

- Introduction
- **Problem Statement**
- Black-box Model Linking
- Collaborative Multi-model Inference
- Evaluation
- Conclusion

Problem Statement

中国科学技术大学
University of Science and Technology of China

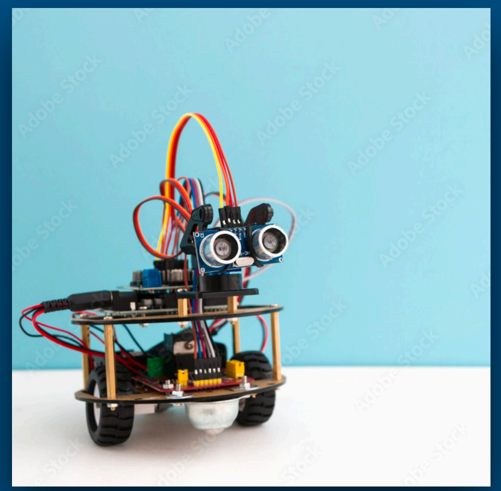
Model Linking

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

Problem Statement

Model Linking

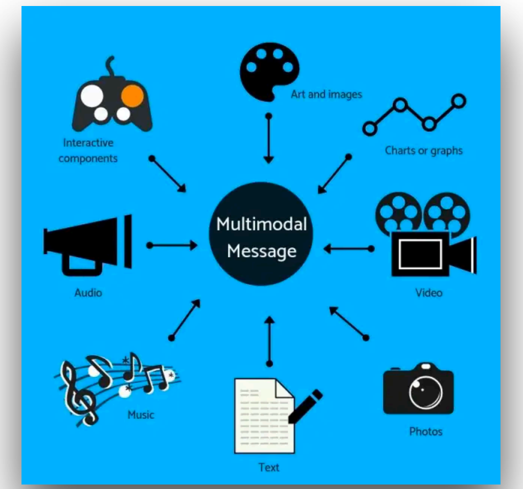
- black-box models $F = \{f_i\}_{i=1}^{k}$ where $f_i : X_i \to Y_i$
- **Assumption**: same or aligned input spaces $\{X_i\}_{i=1}^{k}$
  - common in multi-model applications



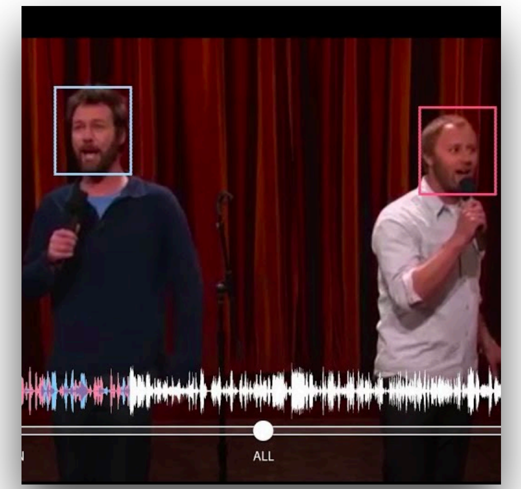
multi-task robotics

drone-based video monitoring

Same Input Spaces



multi-modal learning



audio-visual speech recognition

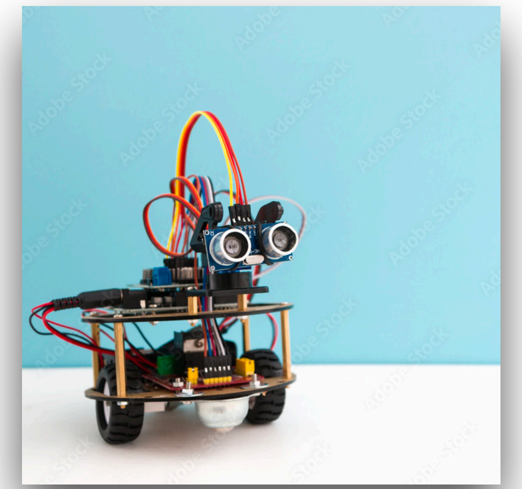
Aligned Input Spaces

Problem Statement

中国科学技术大学
University of Science and Technology of China

Model Linking

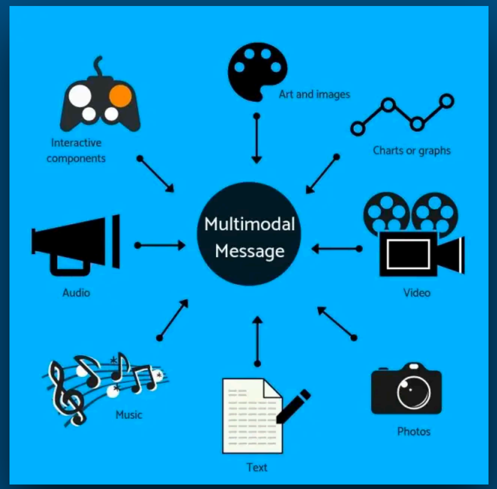
- black-box models $F = \{f_i\}_{i=1}^{k}$ where $f_i : X_i \rightarrow Y_i$
- **Assumption**: same or aligned input spaces $\{X_i\}_{i=1}^{k}$
  - common in multi-model applications



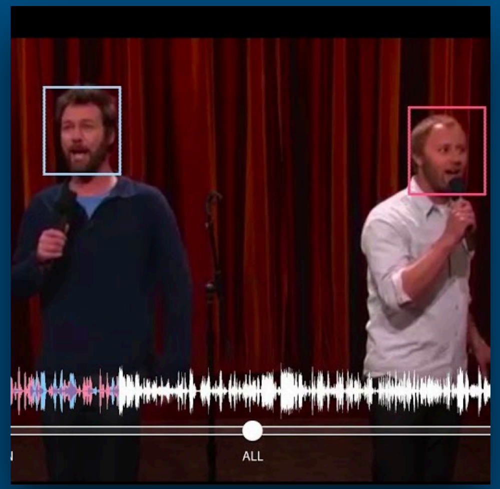
multi-task robotics

drone-based video monitoring

Same Input Spaces



multi-modal learning



audio-visual speech recognition

Aligned Input Spaces

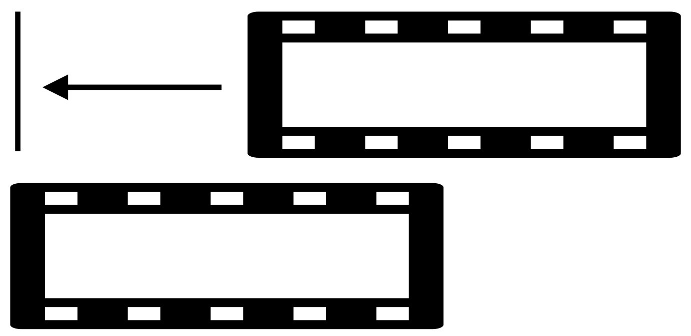
Problem Statement

中国科学技术大学 University of Science and Technology of China

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

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

semantic alignment

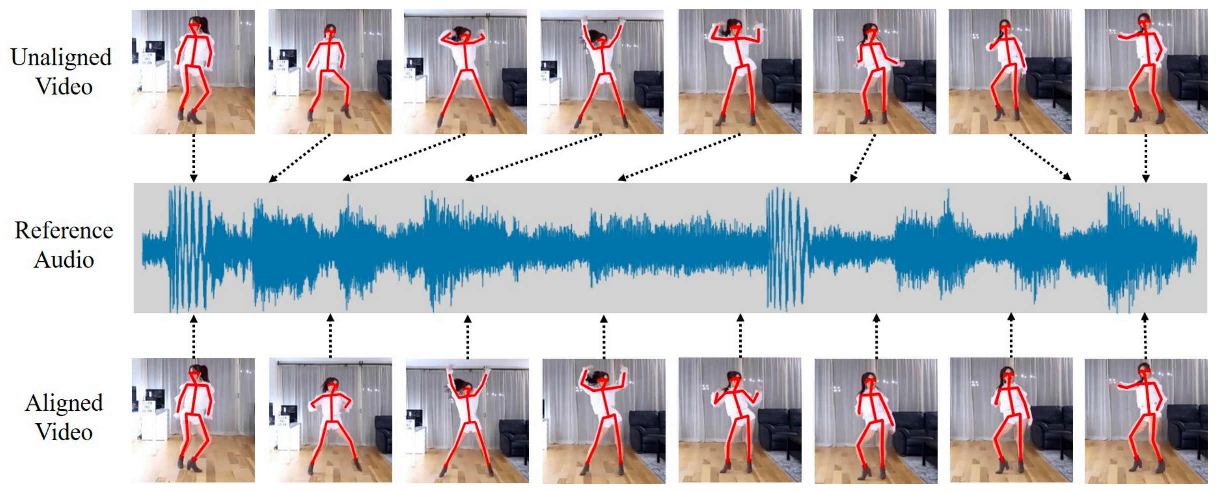


image from paper "AlignNet: A Unifying Approach to Audio-Visual Alignment"

Problem Statement

Model Linking

- black-box models $F = \{f_i\}_{i=1}^{k}$ where $f_i : X_i \to Y_i$
- model link $g_{i,j} : Y_i \to Y_j$
    - <u>*source*</u> model f_i
    - <u>*target*</u> model f_j

Problem Statement

中国科学技术大学 University of Science and Technology of China

Model Linking

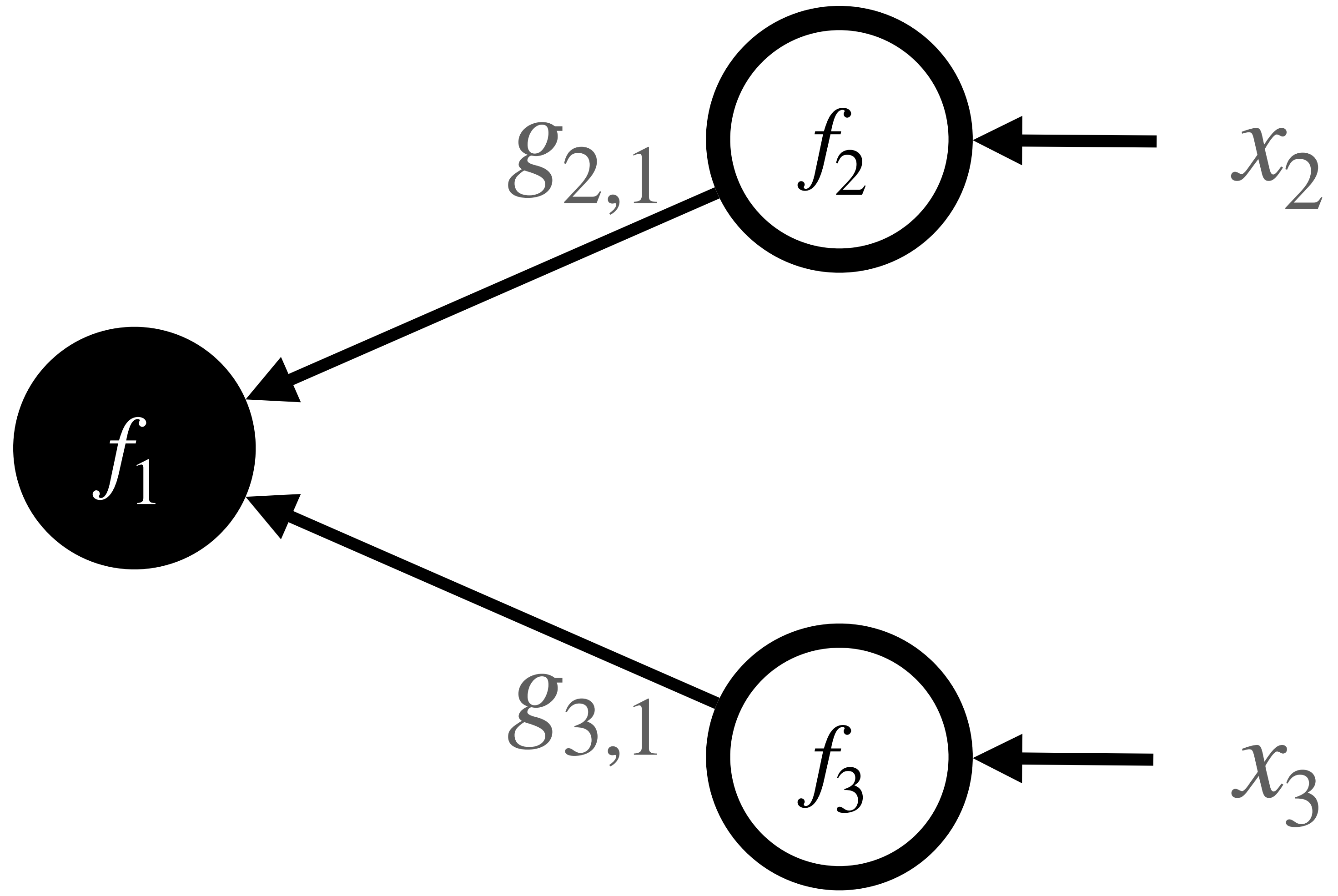
- black-box models $F = \{f_i\}_{i=1}^{k}$ where $f_i : X_i \to Y_i$
- model link $g_{i,j} : Y_i \to Y_j$
  - _source_ model f_i
  - _target_ model f_j
- composite function $g_{i,j} \circ f_i : X_i \to Y_j$

Problem Statement

University of Science and Technology of China

Multi-source Model Links Ensemble

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

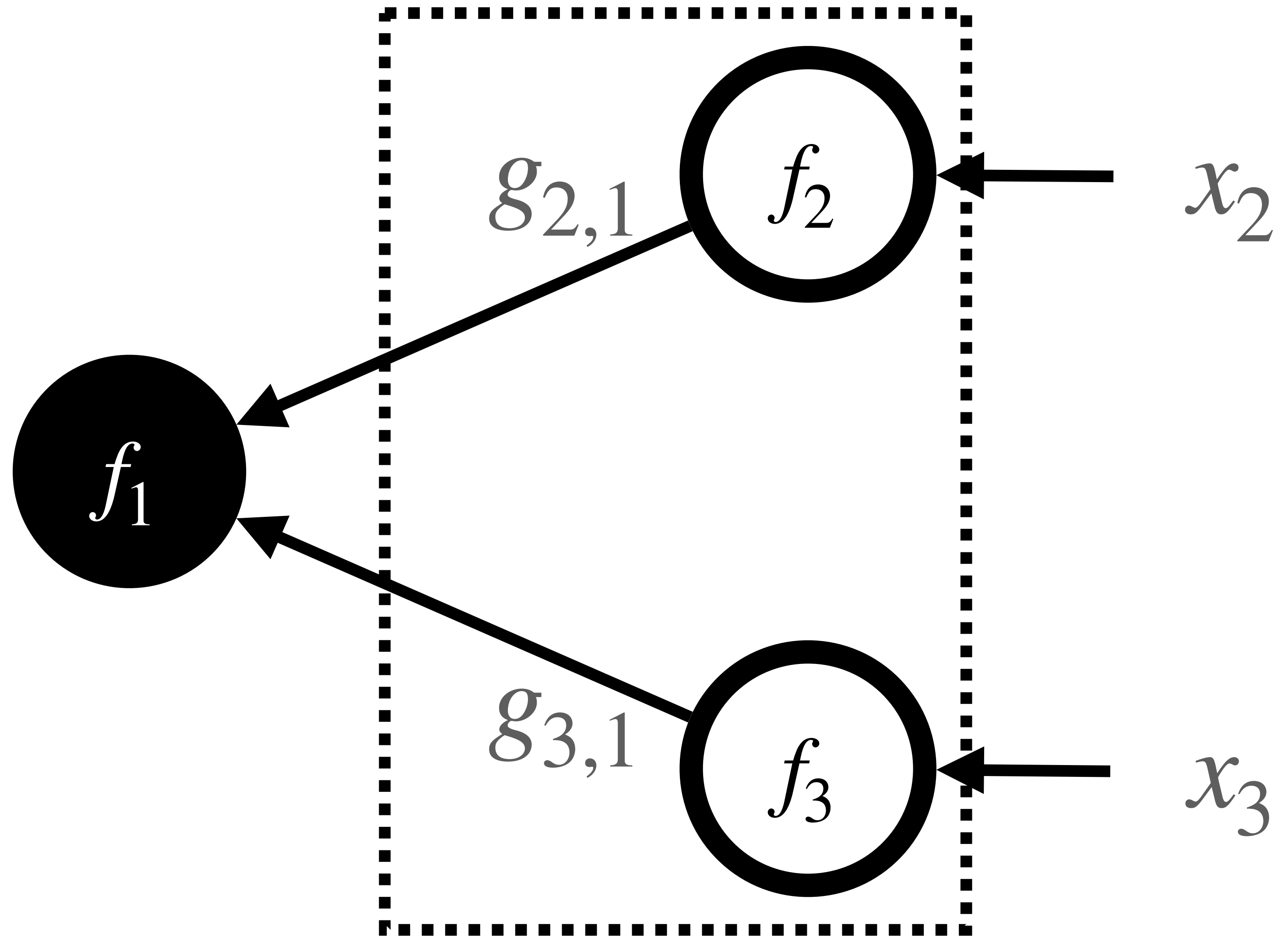


Problem Statement

中国科学技术大学 University of Science and Technology of China

Multi-source Model Links Ensemble

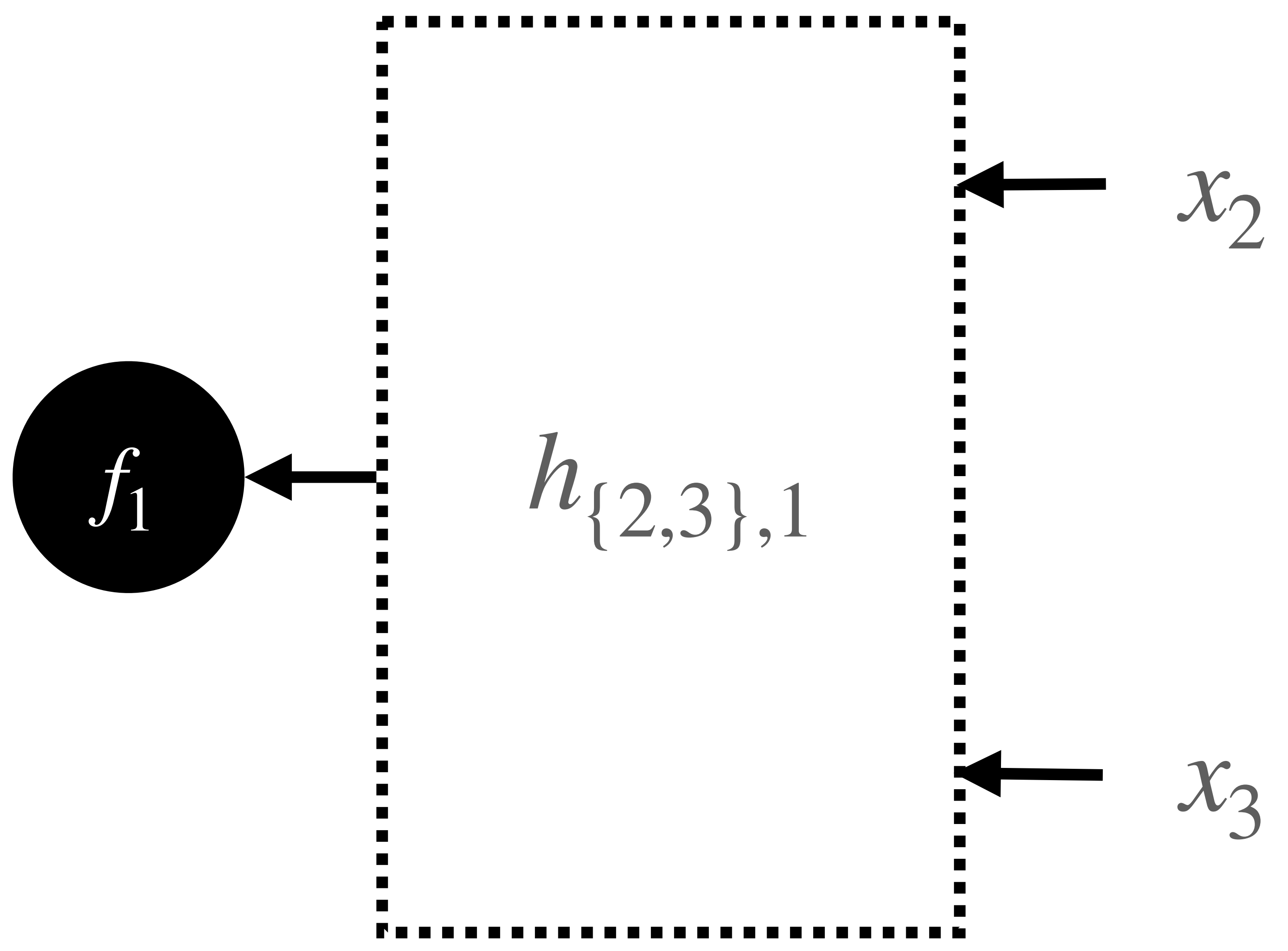
- when $k \geq 3$, there are multiple model links for one target model
- given a set of source models $A \subseteq F$ and a target model f_j, we have a multi-expert model $\{g_{i,j} \circ f_i\}_{f_i \in A}$



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

中国科学技术大学
University of Science and Technology of China

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

中国科学技术大学
University of Science and Technology of China

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

中国科学技术大学
University of Science and Technology of China

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

中国科学技术大学 University of Science and Technology of China

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

$$\max_{A \subseteq F} \overbrace{\left(\frac{1}{|F|}\left(\underbrace{\sum_{f_i \in A} 1}_{\text{activated}} + \underbrace{\sum_{f_j \in F \setminus A} p_j(h_{A,j})}_{\text{predicted}}\right)\right)}^{\text{average output accuracy}} \quad s.t. \quad \underbrace{\sum_{f_i \in A} c(f_i)}_{\text{exact inference}} + \underbrace{\sum_{f_j \in F \setminus A} c(h_{A,j})}_{\text{model links}} \leq B.$$

Menu

中国科学技术大学
University of Science and Technology of China

Main contents

Black-box Model Linking

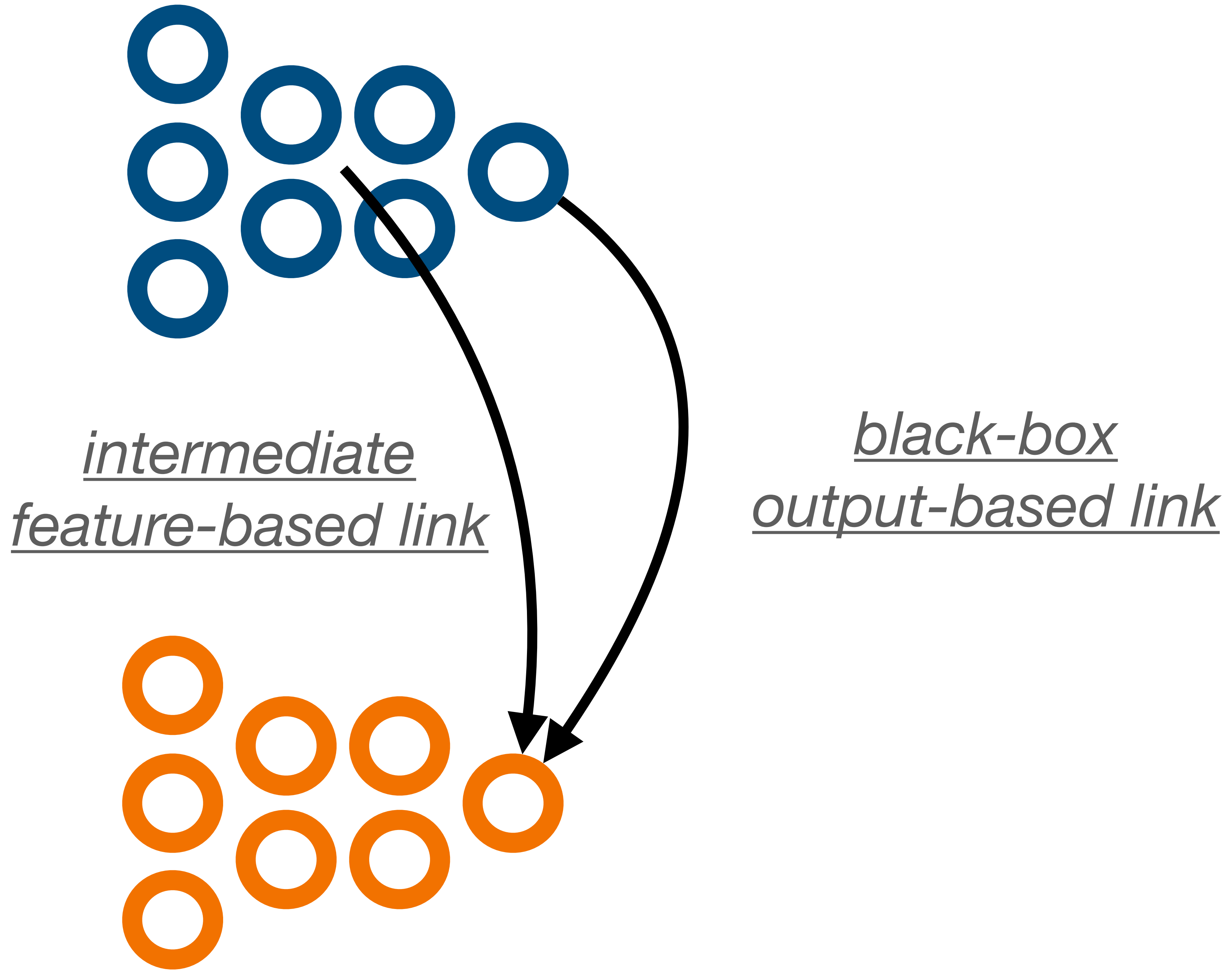
中国科学技术大学
University of Science and Technology of China

Black-box outputs or intermediate features?

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

VM

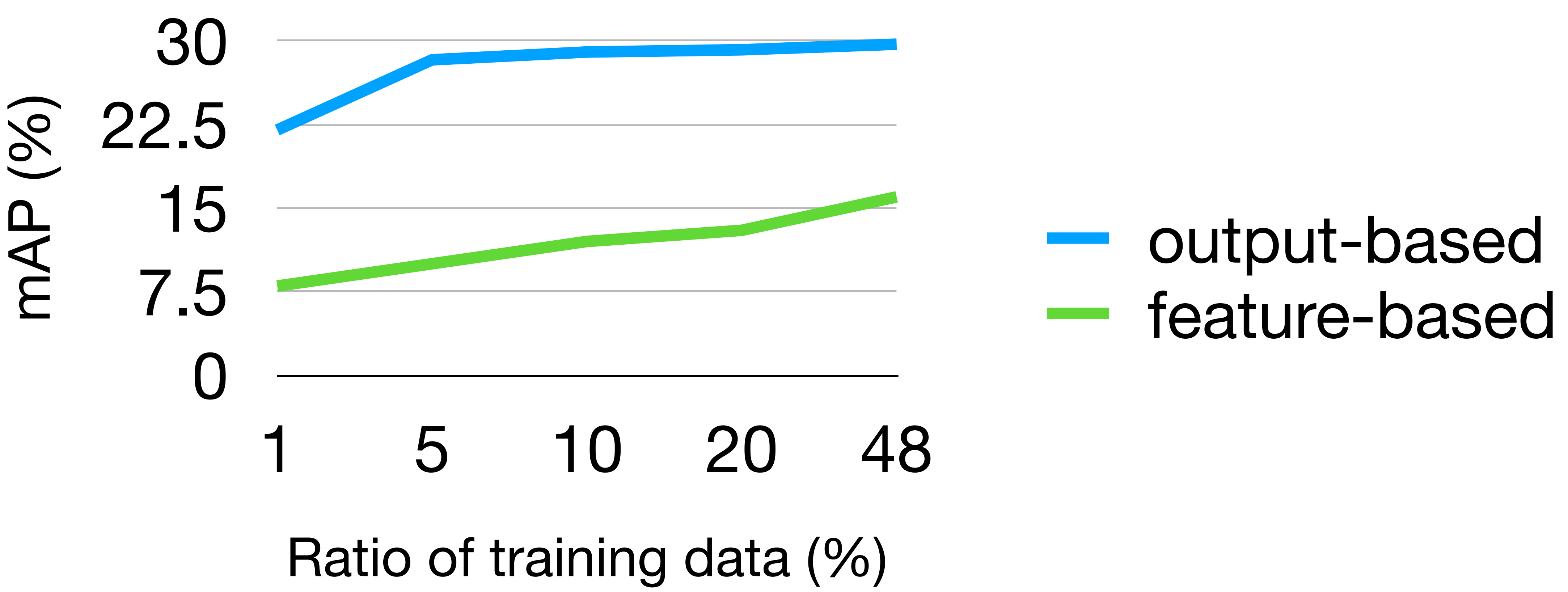
docker

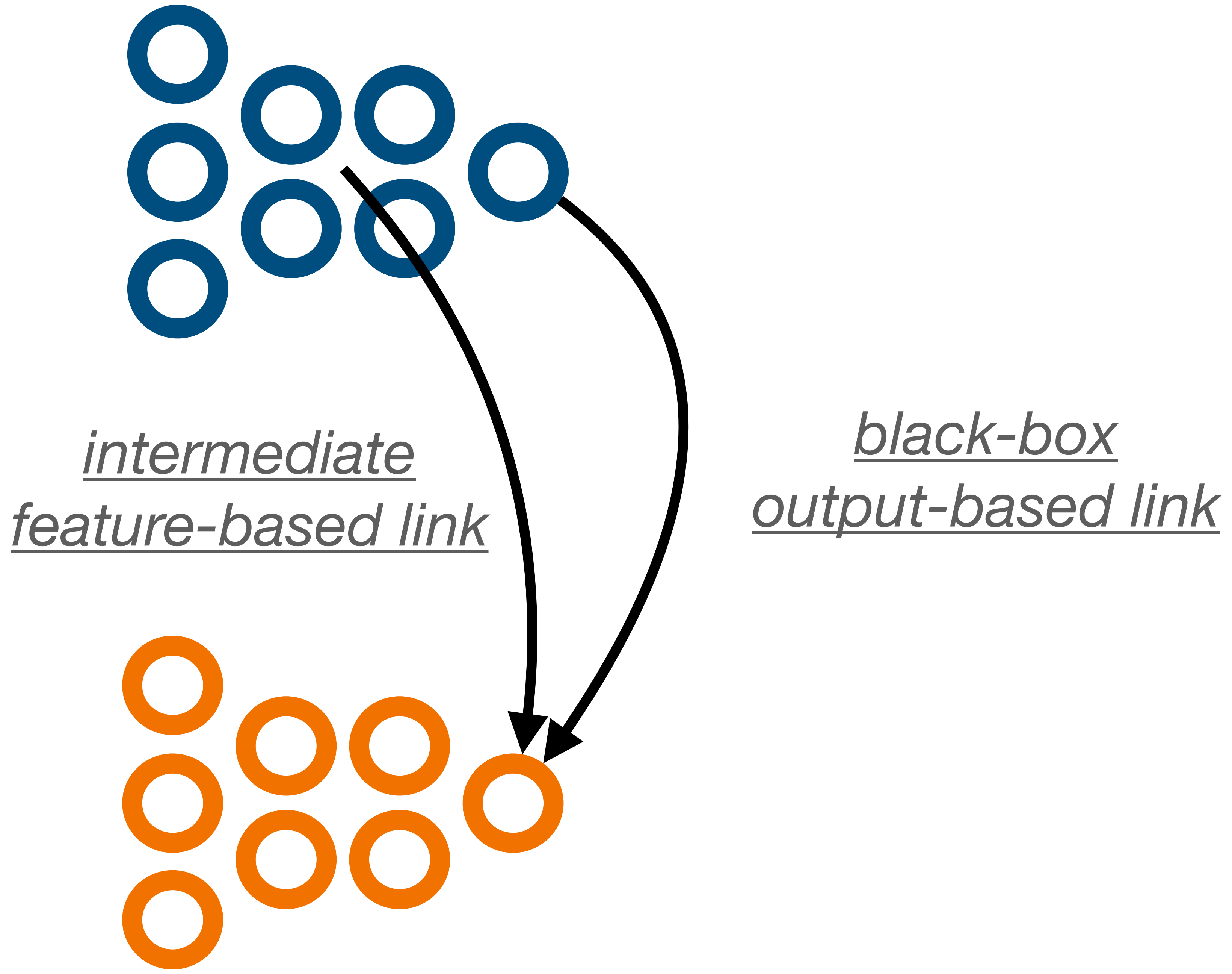


Black-box Model Linking

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



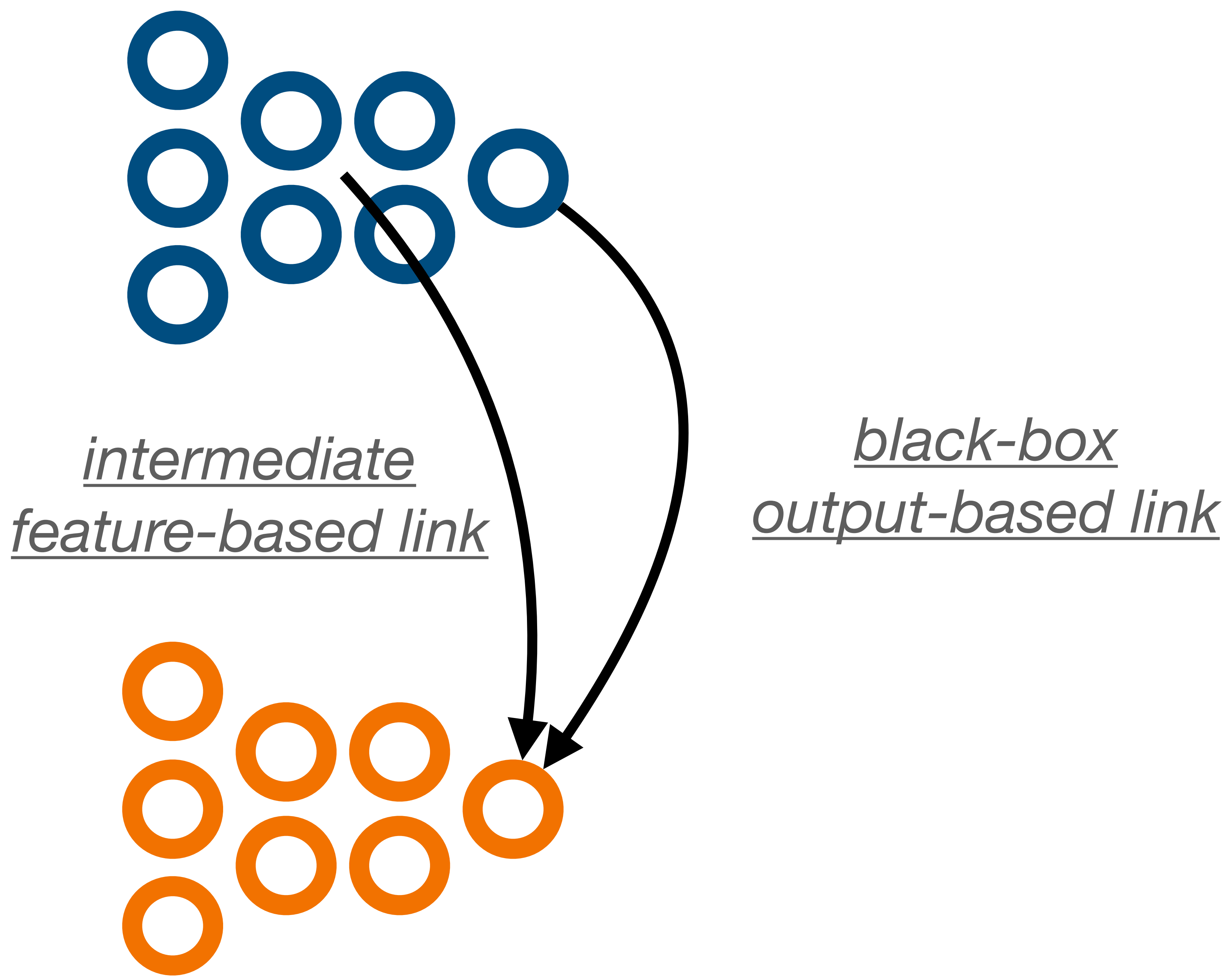


Black-box Model Linking

中国科学技术大学 University of Science and Technology of China

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

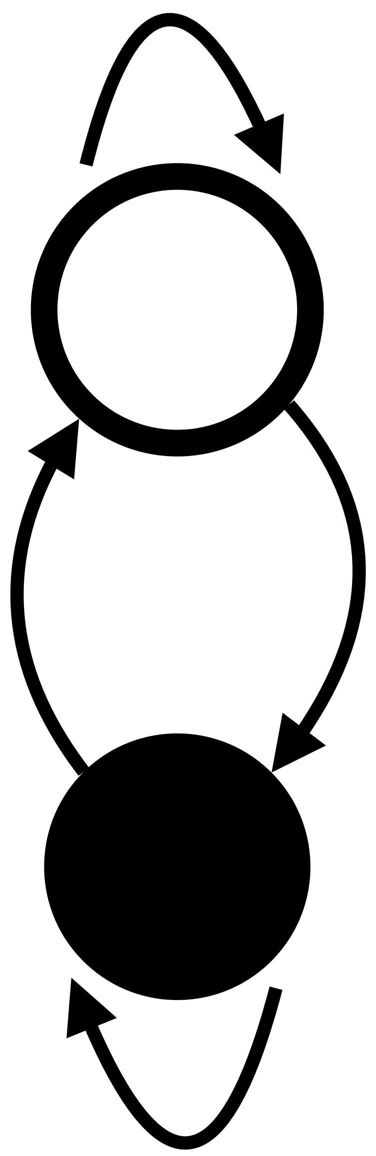


When the training data is abundant for the representation shared among tasks, learning a new task branch $f \in F$ requires $C(F)$ sample complexity, where $C(\cdot)$ measures the complexity of a hypothesis family.

Black-box Model Linking

Model link architecture

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

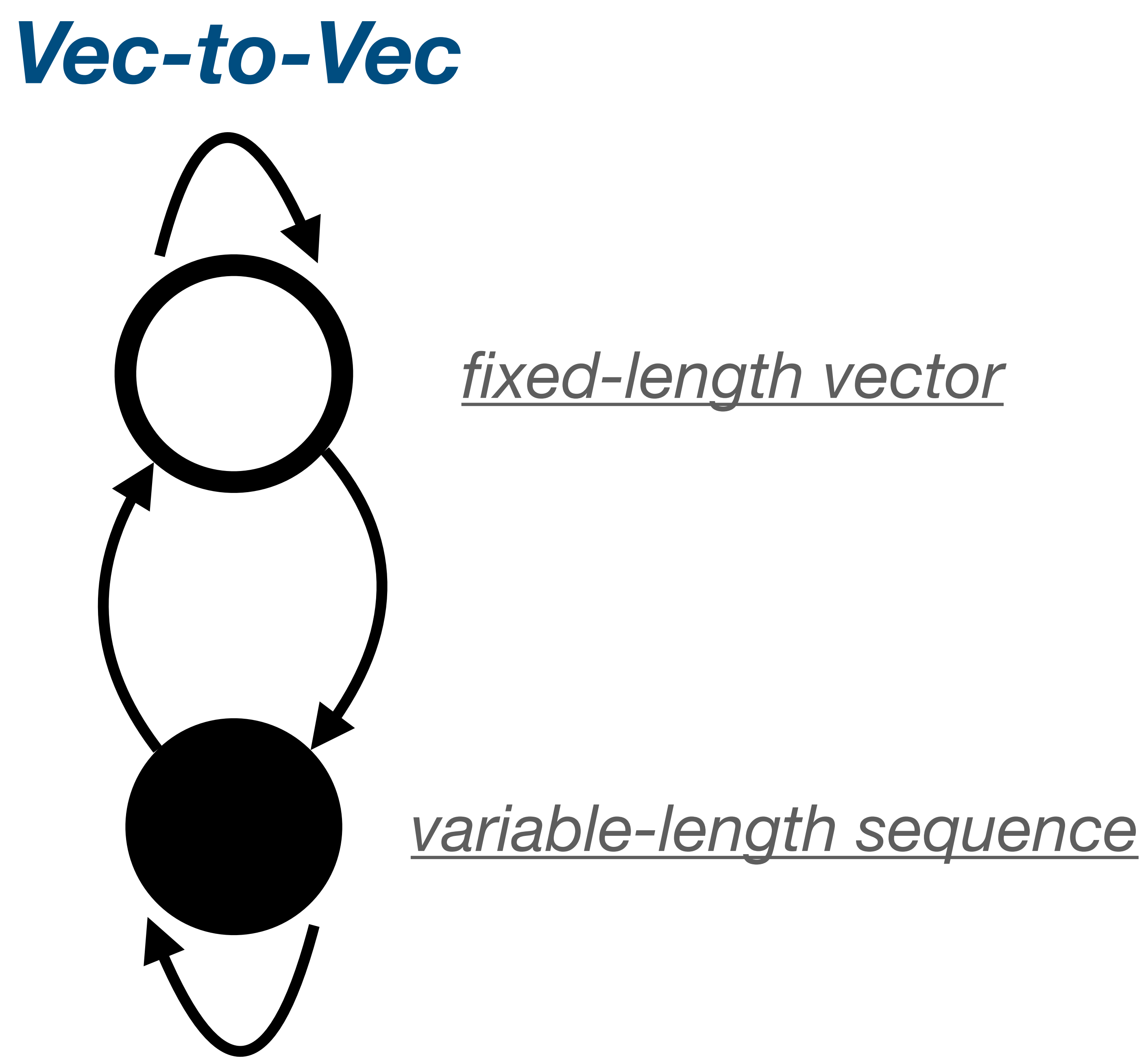


Black-box Model Linking

University of Science and Technology of China

Model link architecture

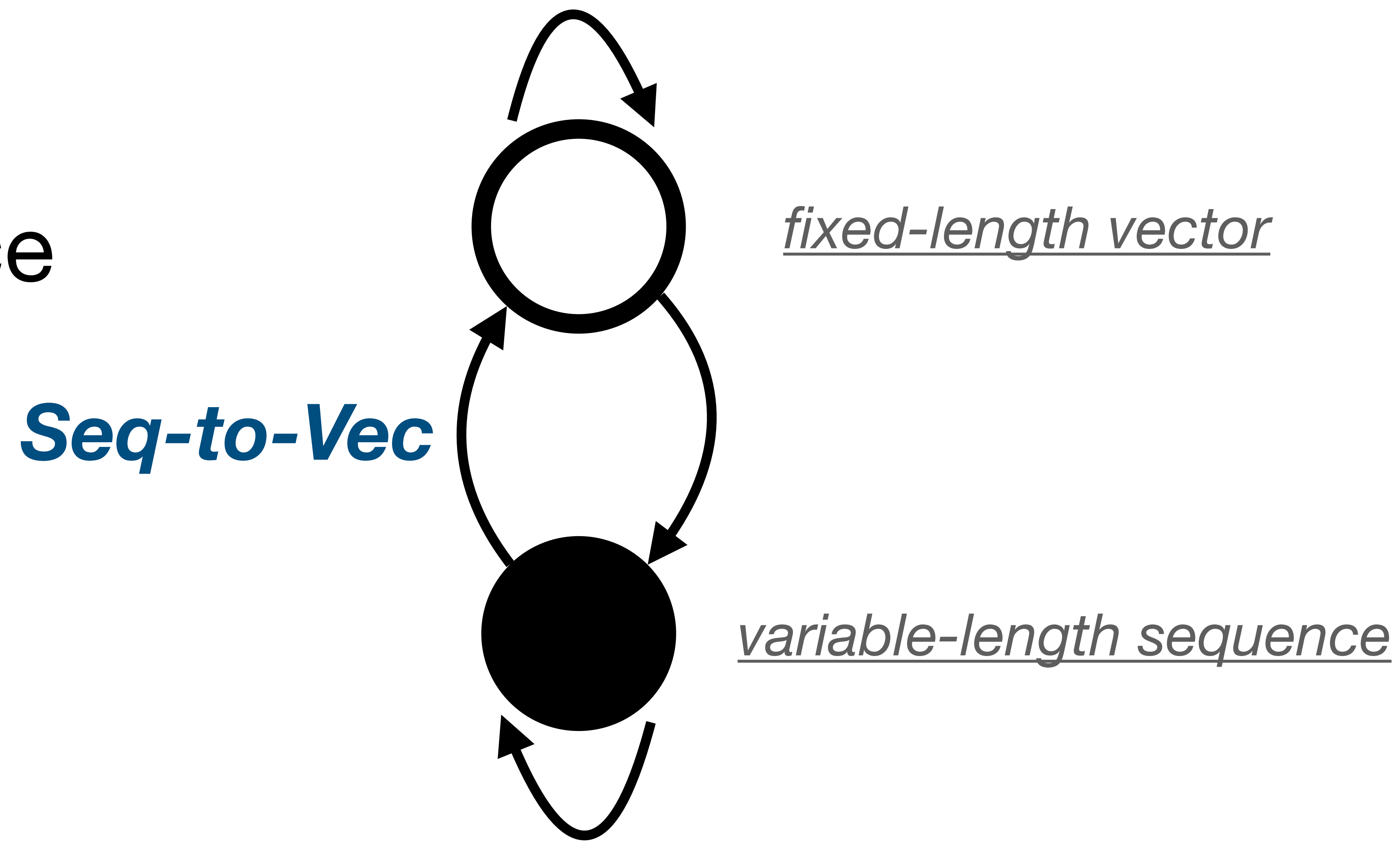
- output formats determine the model link's architecture
    - fixed-length vector & variable-length sequence
- **Vec-to-Vec**
    - ReLU-activated multilayer perception (MLP)



Black-box Model Linking

Model link architecture

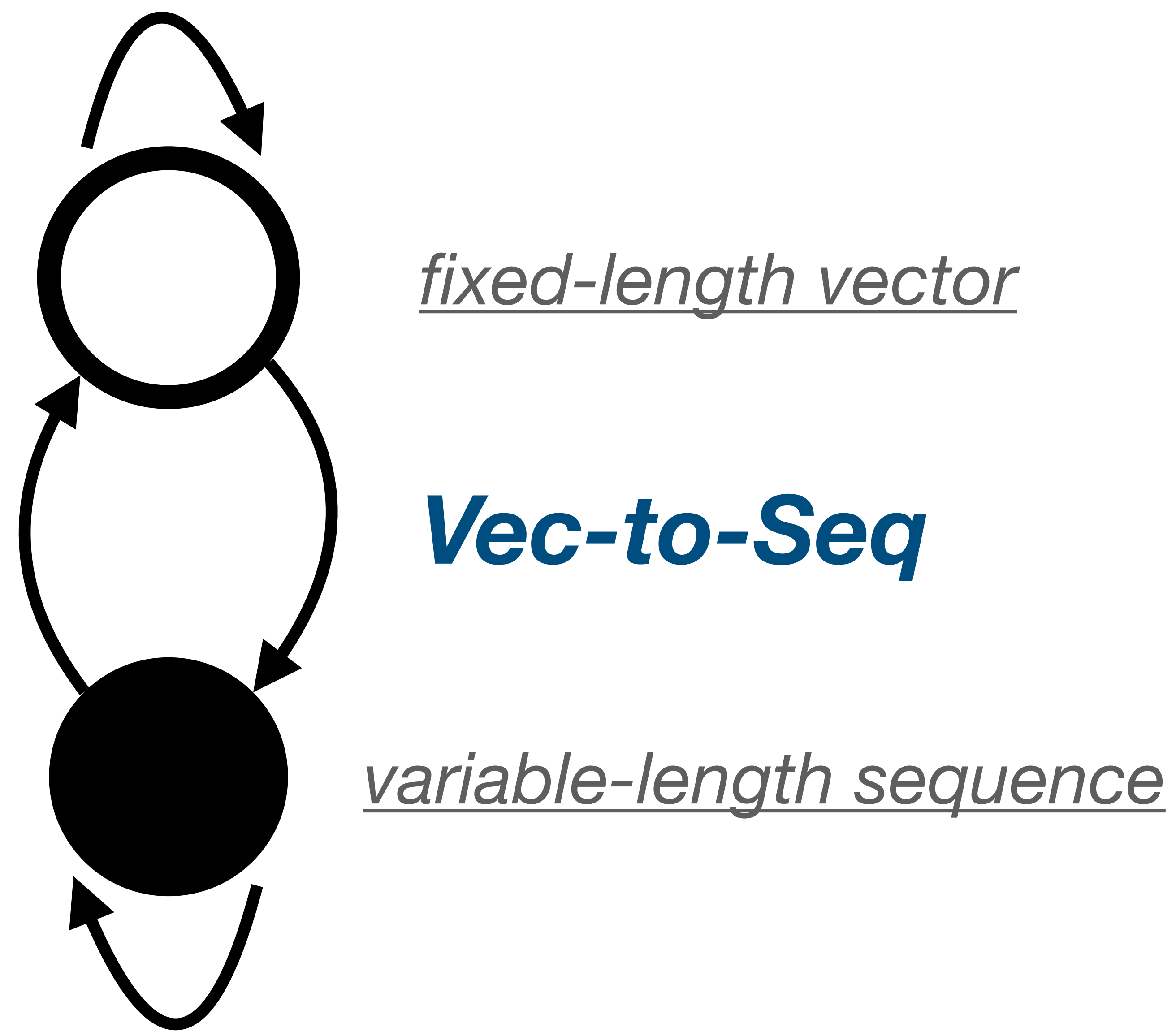
- output formats determine the model link's architecture
  - fixed-length vector & variable-length sequence
- **Seq-to-Vec**
  - Embedding - LSTM - MLP



Black-box Model Linking

Model link architecture

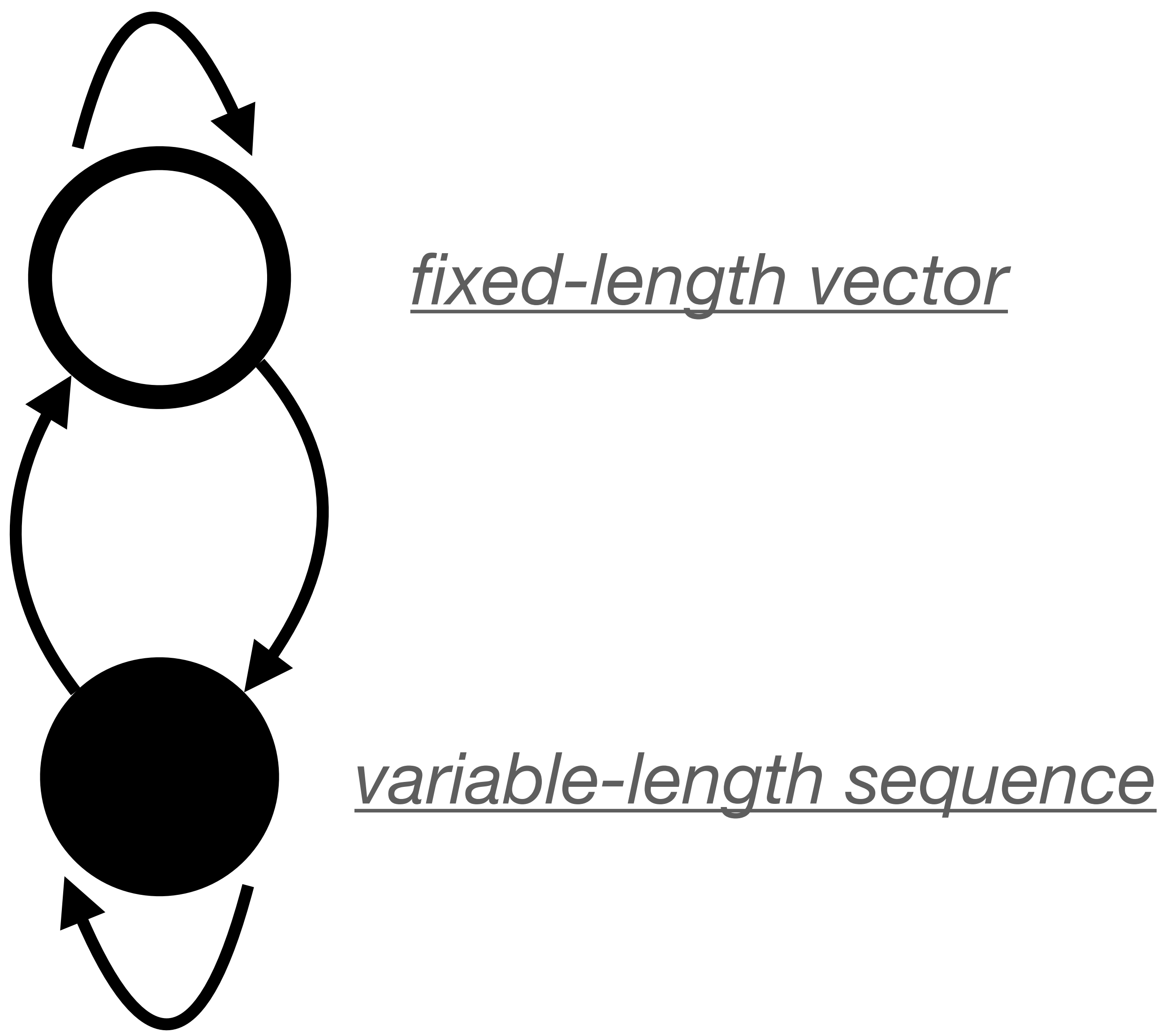
- output formats determine the model link's architecture
  - fixed-length vector & variable-length sequence
- **Vec-to-Seq**
  - MLP Encoder
  - Embedding - LSTM - Attention - MLP Decoder



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



Seq-to-Seq

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.

Black-box Model Linking

Ensemble

- weighted sum of model links

$$h_{A,j} = \sigma(\sum_{f_i \in A} g_{i,j} \circ f_i(x_i))$$

  - σ denotes the activation function

Black-box Model Linking

中国科学技术大学
University of Science and Technology of China

Training

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

$$\min \sum_{i=l}^{n} L_j(h_{A,j}(\{y_i^l\}_{f_i \in A}), y_j^l)$$

  - target model's task determines the loss function

Menu

中国科学技术大学
University of Science and Technology of China

Main contents

- Introduction
- Problem Statement
- Black-box Model Linking
- **Collaborative Multi-model Inference**
- Evaluation
- Conclusion

Assumptions and Observations

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

$$\Delta(A, f_i) = F(A \cup \{f_i\}) - F(A)$$

Optimization Problem

$$\max_{A \subseteq F} \overbrace{\left(\frac{1}{|F|}\left(\underbrace{\sum_{f_i \in A} 1}_{\text{activated}} + \underbrace{\sum_{f_j \in F \setminus A} p_j(h_{A,j})}_{\text{predicted}}\right)\right)}^{\text{average output accuracy}}$$

$$s.t. \underbrace{\sum_{f_i \in A} c(f_i)}_{\text{exact inference}} + \underbrace{\sum_{f_j \in F \setminus A} c(h_{A,j})}_{\text{model links}} \leq B.$$

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

$$\max_{A \subseteq F} \overbrace{\left(\frac{1}{|F|}\left(\underbrace{\sum_{f_i \in A} 1}_{\text{activated}} + \underbrace{\sum_{f_j \in F \setminus A} p_j(h_{A,j})}_{\text{predicted}}\right)\right)}^{\text{average output accuracy}}$$

$$s.t. \underbrace{\sum_{f_i \in A} c(f_i)}_{\text{exact inference}} + \underbrace{\sum_{f_j \in F \setminus A} c(h_{A,j})}_{\text{model links}} \leq B.$$

Assumptions and Observations

- two cases observed
  - <u>dominance:</u> the performance of the ensemble model approximately equals the best-performance source of model links.

$$f_{i*} = argmax_{f_i \in A} p_j(g_{ij})$$

$$p_j(h_{A,f_j}) \approx p_j(g_{i*,j})$$

Collaborative Multi-model Inference

中国科学技术大学 University of Science and Technology of China

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.

$$\forall f_i \in A, p_j(h_{A,f_j}) > p_j(g_{i,j})$$

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.

see paper: Sviridenko, M. 2004. A note on maximizing a submodular set function subject to a knapsack constraint. Operations Research Letters, 32(1): 41–43.

Optimization Problem

$$\max_{A \subseteq F} \overbrace{\left(\frac{1}{|F|}\left(\underbrace{\sum_{f_i \in A} 1}_{\text{activated}} + \underbrace{\sum_{f_j \in F \setminus A} p_j(h_{A,j})}_{\text{predicted}}\right)\right)}^{\text{average output accuracy}}$$

$$s.t. \quad \underbrace{\sum_{f_i \in A} c(f_i)}_{\text{exact inference}} + \underbrace{\sum_{f_j \in F \setminus A} c(h_{A,j})}_{\text{model links}} \le B.$$

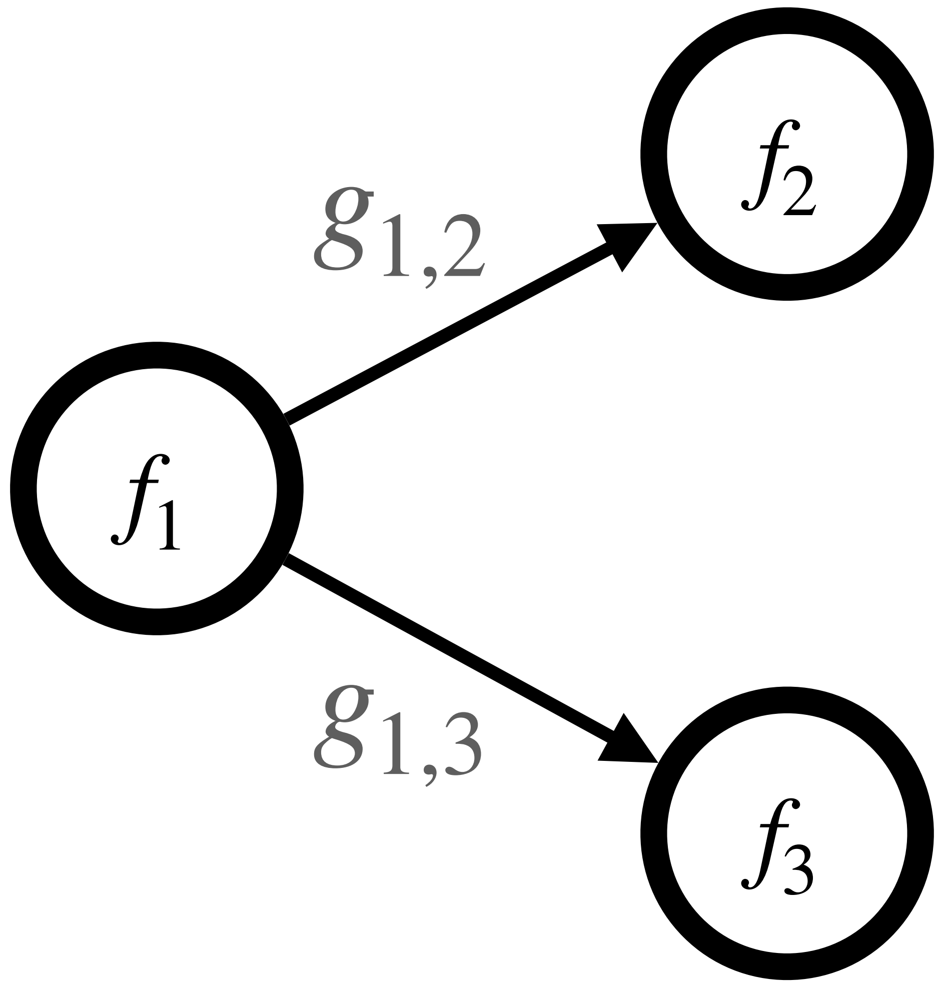
Collaborative Multi-model Inference

中国科学技术大学
University of Science and Technology of China

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

中国科学技术大学 University of Science and Technology of China

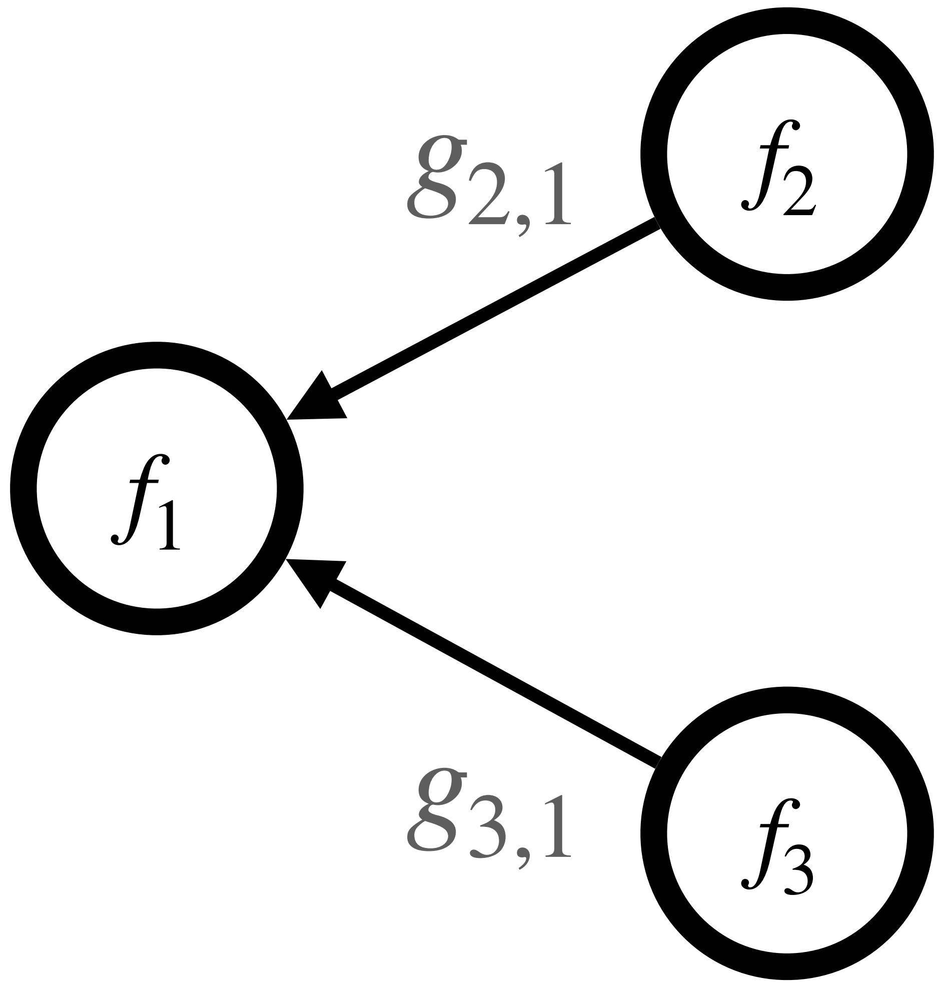
Activation Probability

- 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

中国科学技术大学 University of Science and Technology of China

Activation Probability

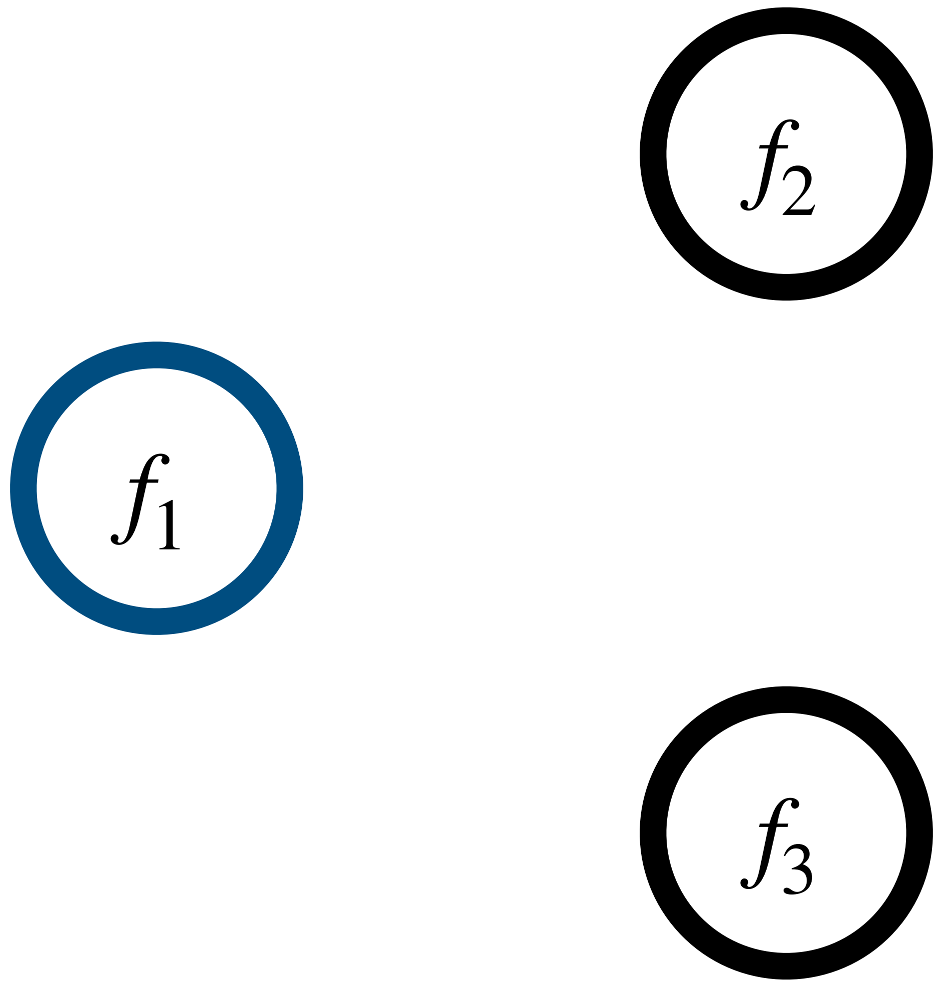
- three factors
  - the average performance of model links from f_i to all the others

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

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

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

  - the cost of f_i $\quad c(f_i)$



Collaborative Multi-model Inference

中国科学技术大学
University of Science and Technology of China

Activation Probability

- definition

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

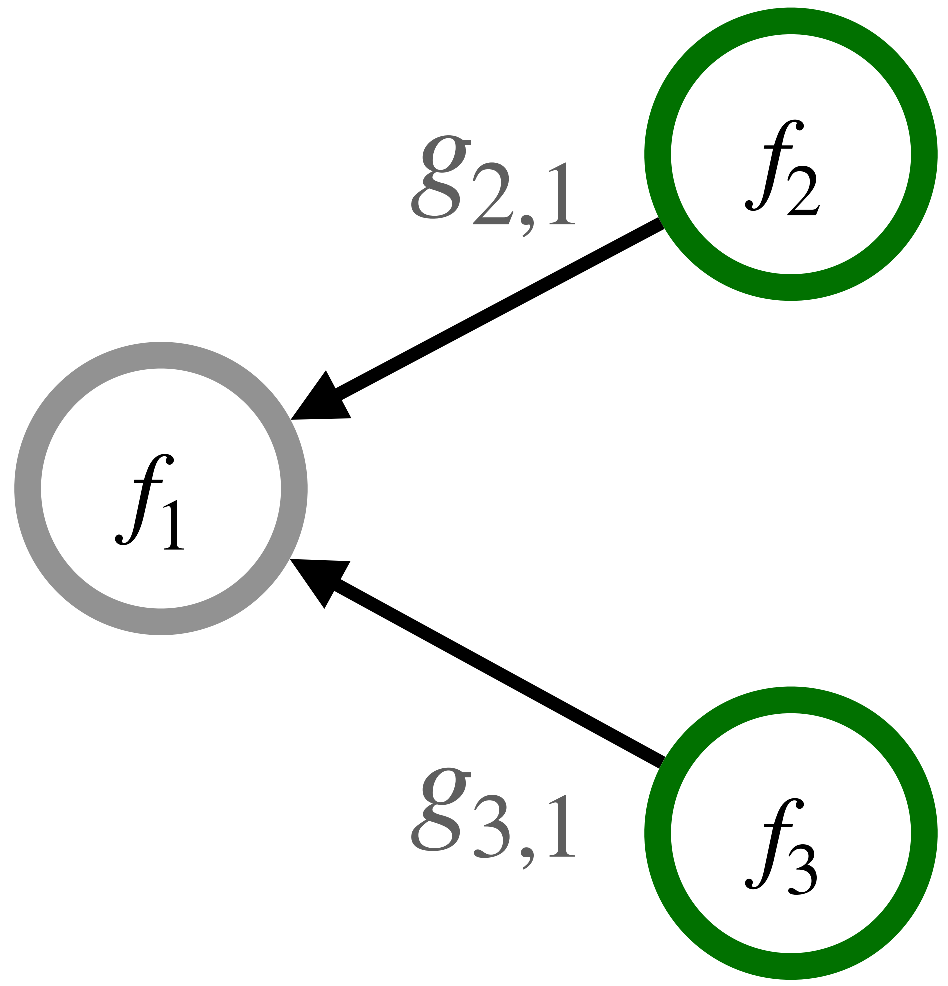
$w = 2/\min_i c(f_i)$ by normalization

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

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

Algorithm

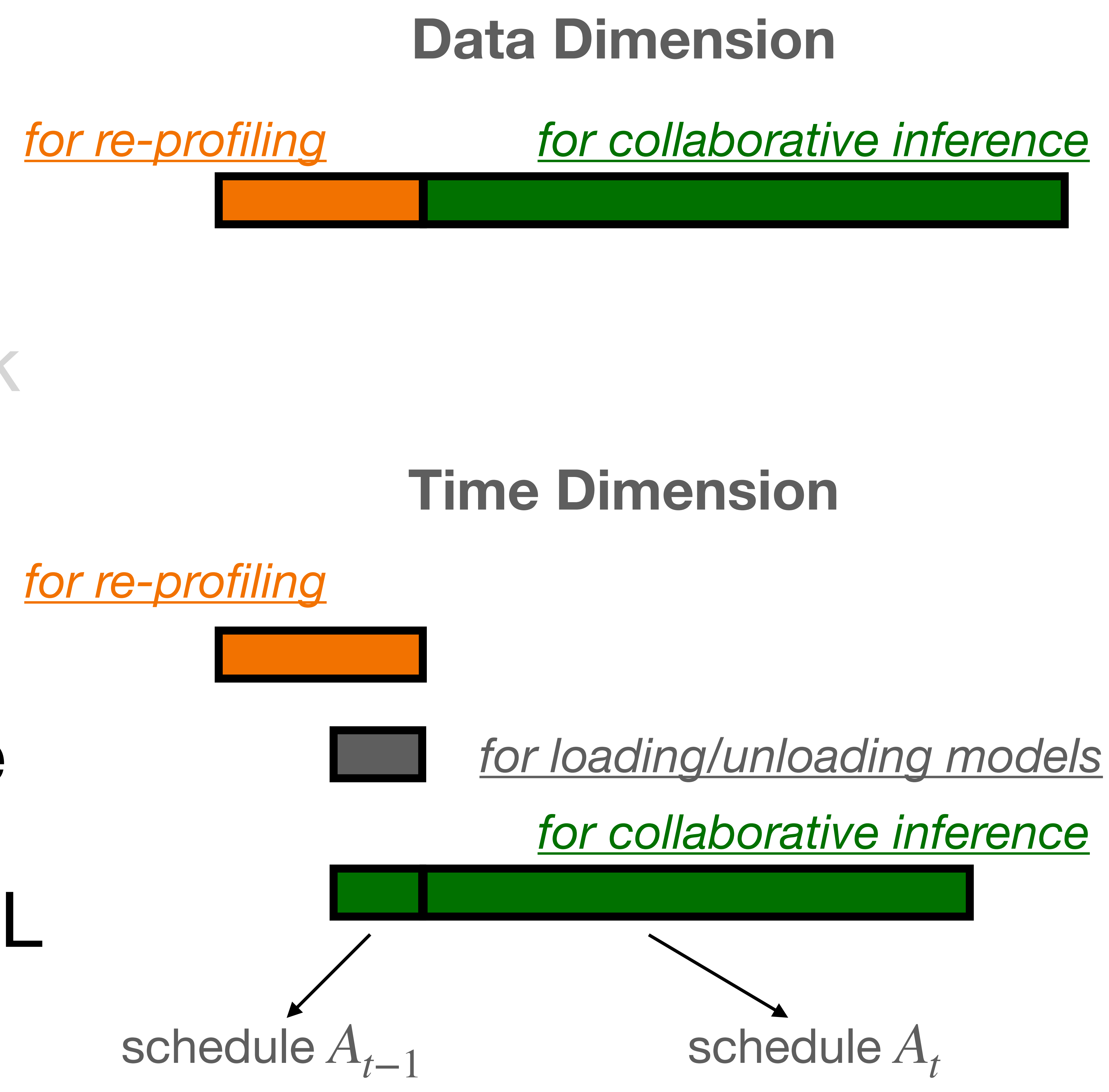
- select greedily w.r.t. activation probability under the cost budget
- activated models do exact inference while the others' outputs will be predicted by the model link ensemble of activated sources.



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

Evaluation

中国科学技术大学
University of Science and Technology of China

Implementation

- ***MLink*** implemented in Python based on TensorFlow 2.0
- We tested the integration on programs implemented with TensorFlow, PyTorch and MindSpore.

Datasets and Models

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

Table 1: ML Models on Hollywood2 Dataset

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

Evaluation

Datasets and Models

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

Table 1: ML Models on Hollywood2 Dataset

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

Evaluation

Datasets and Models

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

Table 1: ML Models on Hollywood2 Dataset

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

Evaluation

中国科学技术大学
University of Science and Technology of China

Datasets and Models

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

Table 1: ML Models on Hollywood2 Dataset

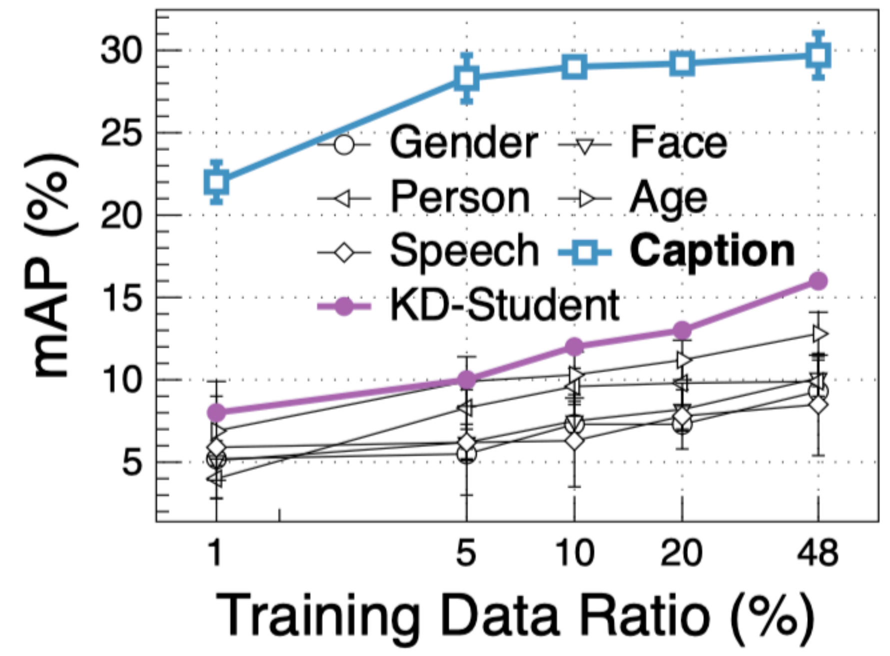
Task Class	ML Model	Input Modality	Output Format	Metric
Single-label Classification	Gender Classification	Audio	2-D Softmax Labels	Acc.
Multi-label Classification	Action Classification	Video	12-D Sigmoid Labels	mAP
Localization	Face Detection	Image	4-D Bounding Box	IoU
	Person Detection			
Regression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

Evaluation

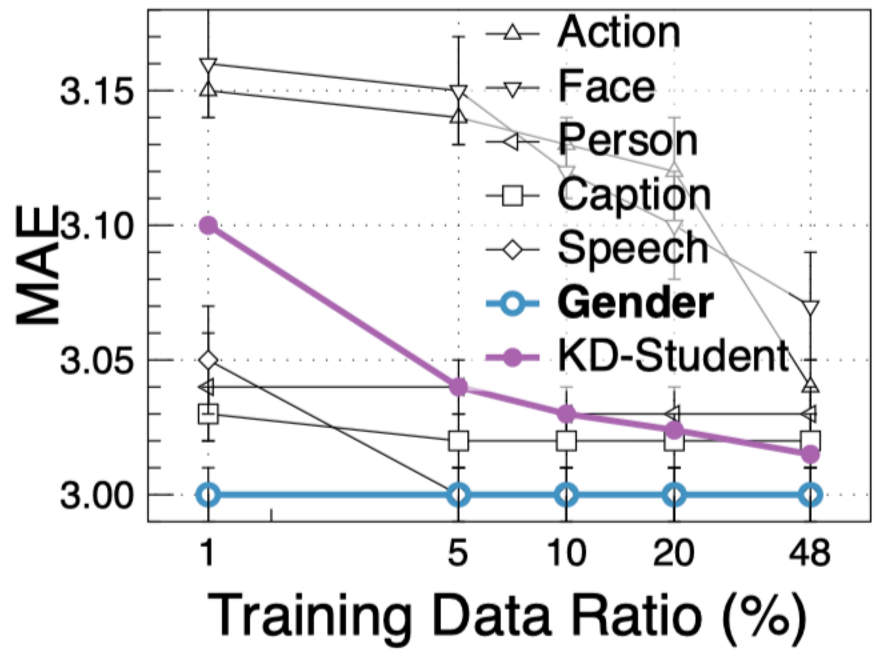
中国科学技术大学
University of Science and Technology of China

Model Links' Performance

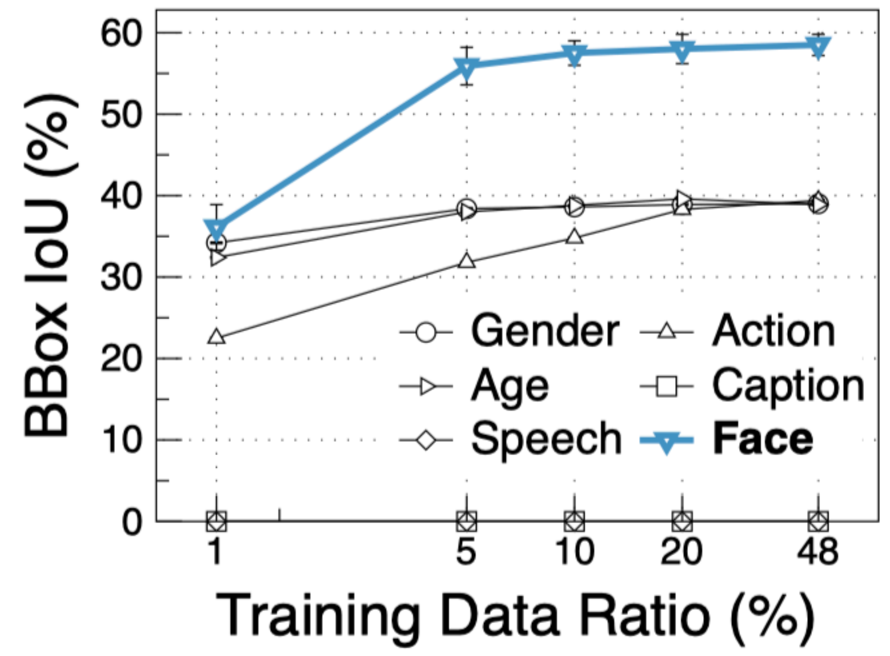
- pairwise model links are trained using 1%, 5%, 10%, 20%, 48% data
  - RMSprop optimizer with same hyper-parameters (0.01 learning rate, 100 epochs, 32 batch size)



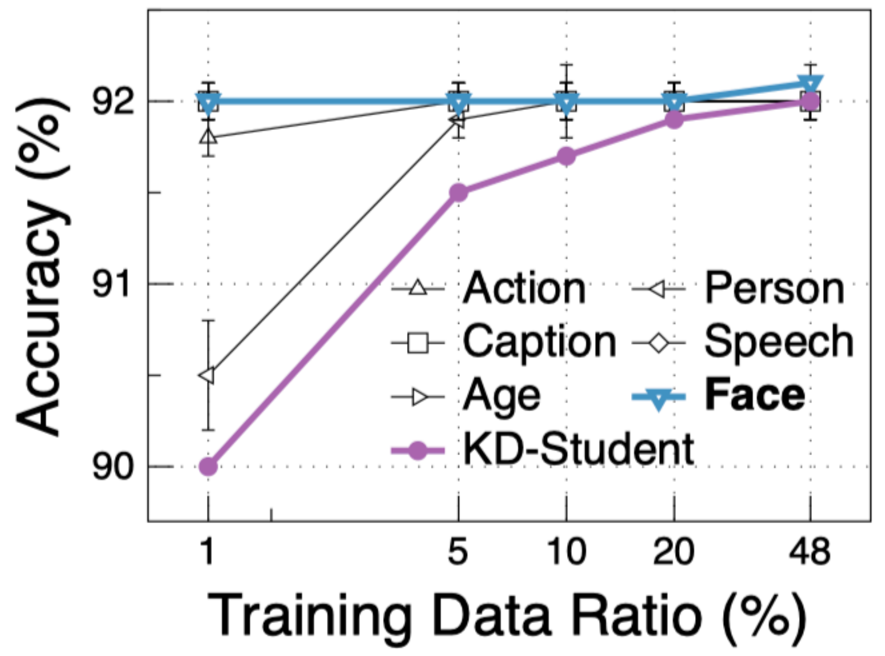
(a) Target: Action



(b) Target: Age



(c) Target: Person



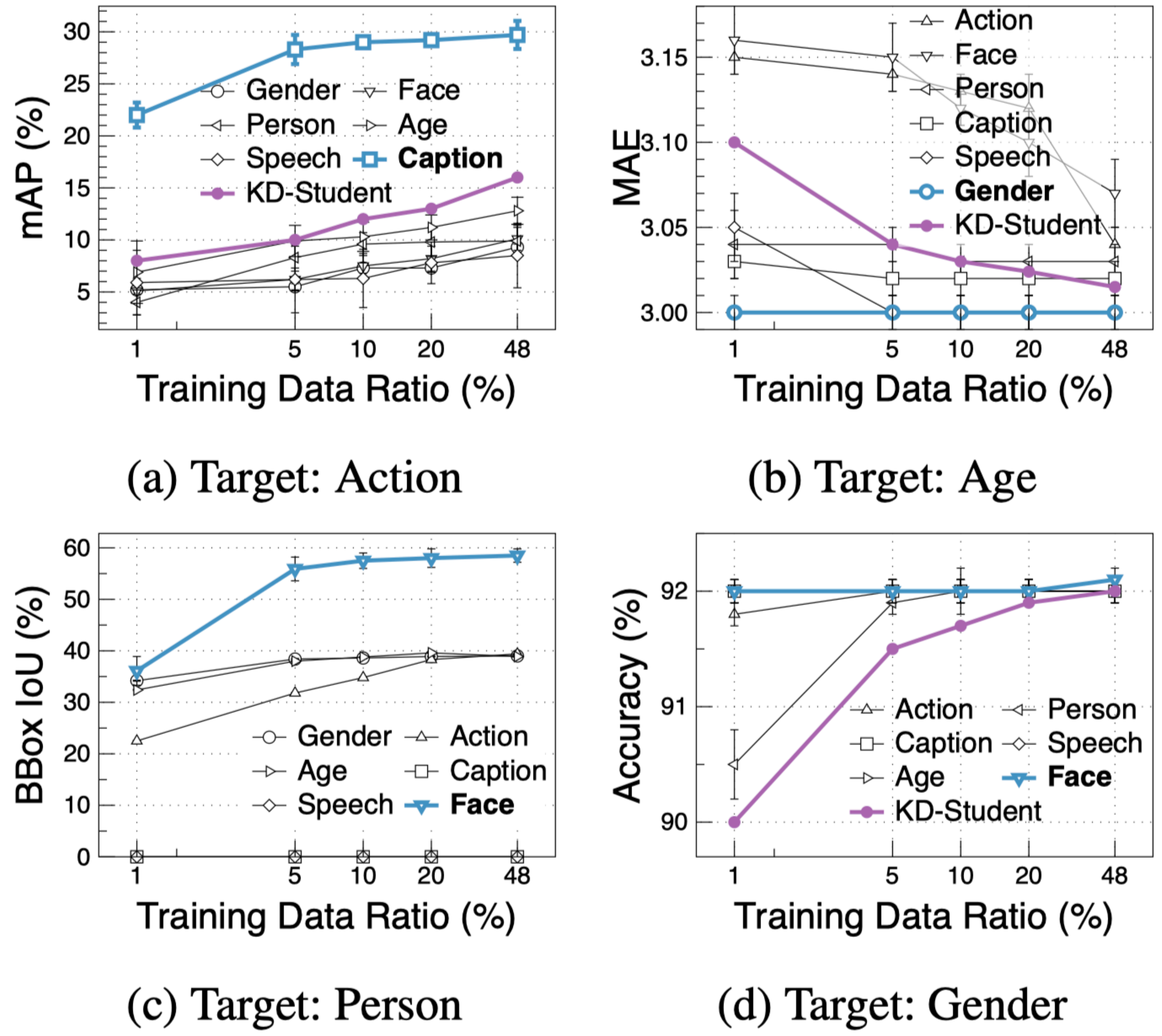
(d) Target: Gender

Evaluation

University of Science and Technology of China

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



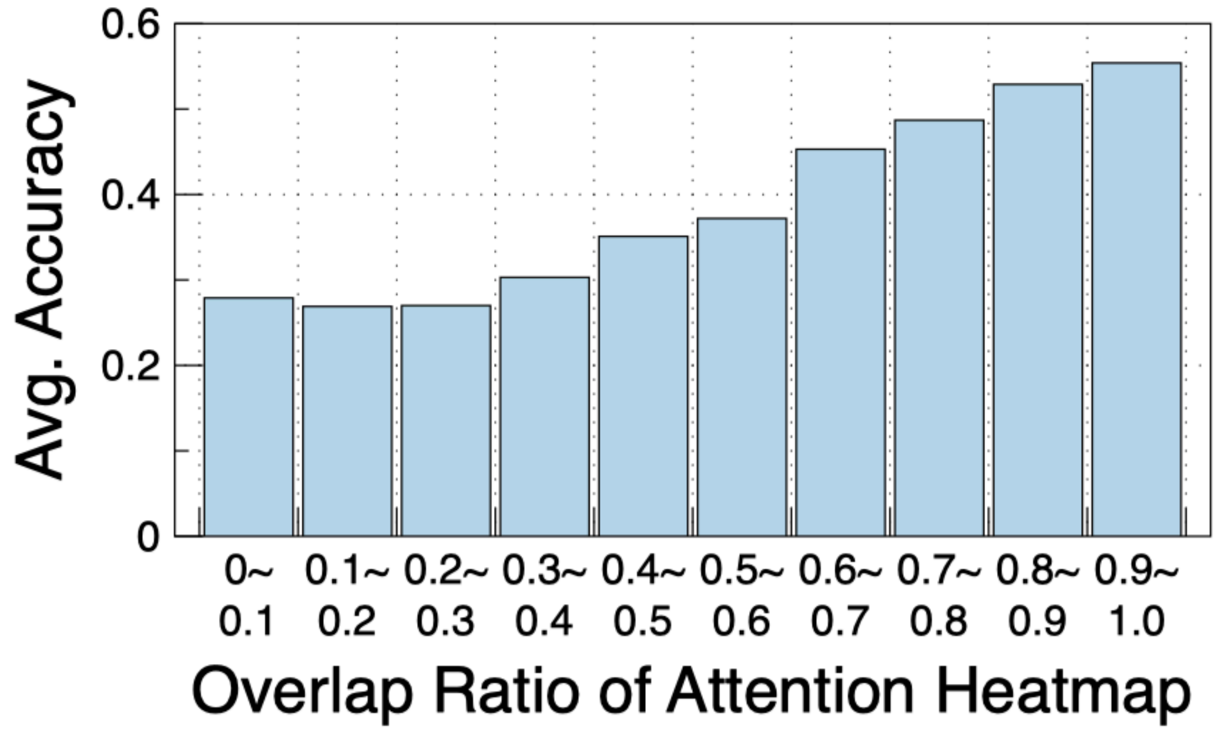
Evaluation

University of Science and Technology of China

Semantic Correlation

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

(a) Attention heatmaps of Object and Scene models.



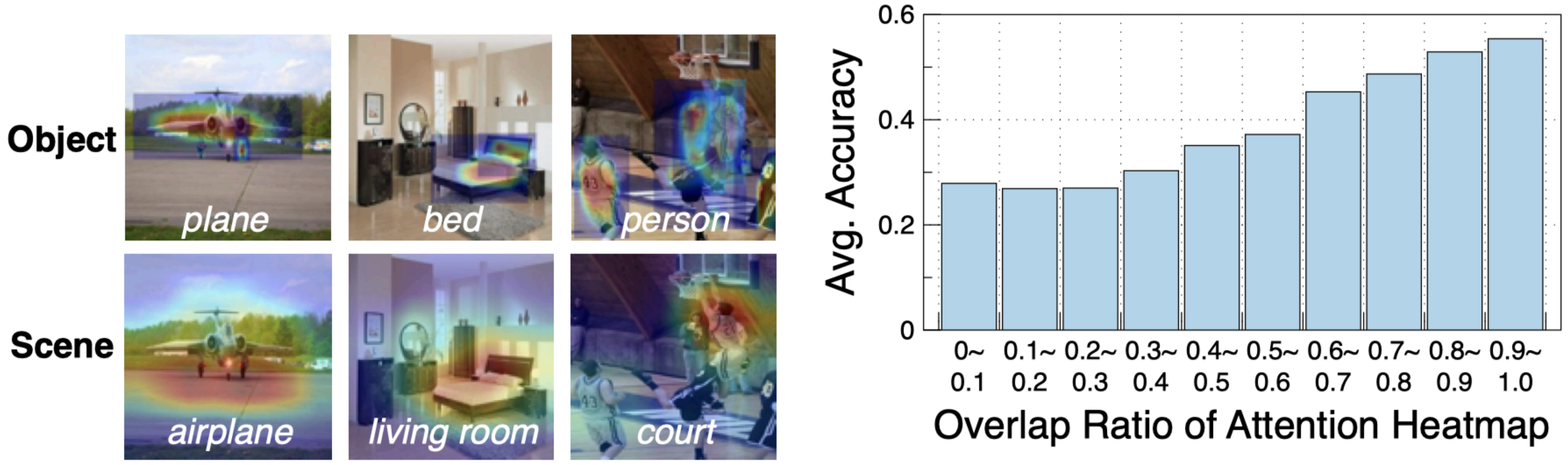
(b) Scene-to-Object MLink accuracy vs. attention overlaps.

Evaluation

University of Science and Technology of China

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



(a) Attention heatmaps of Object and Scene models.

(b) Scene-to-Object MLink accuracy vs. attention overlaps.

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

Source	Action	Age	Face	Gender
IoU (%)	39.4	38.9	58.5	39.0
Pearson Corr.	0.123	0.042	0.244	-0.053

MLink Ensemble

- dominance cases $p_j(h_{A,f_j}) \approx p_j(g_{i*,j})$

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.

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

Evaluation

中国科学技术大学
University of Science and Technology of China

MLink Ensemble

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

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.

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

Evaluation

University of Science and Technology of China

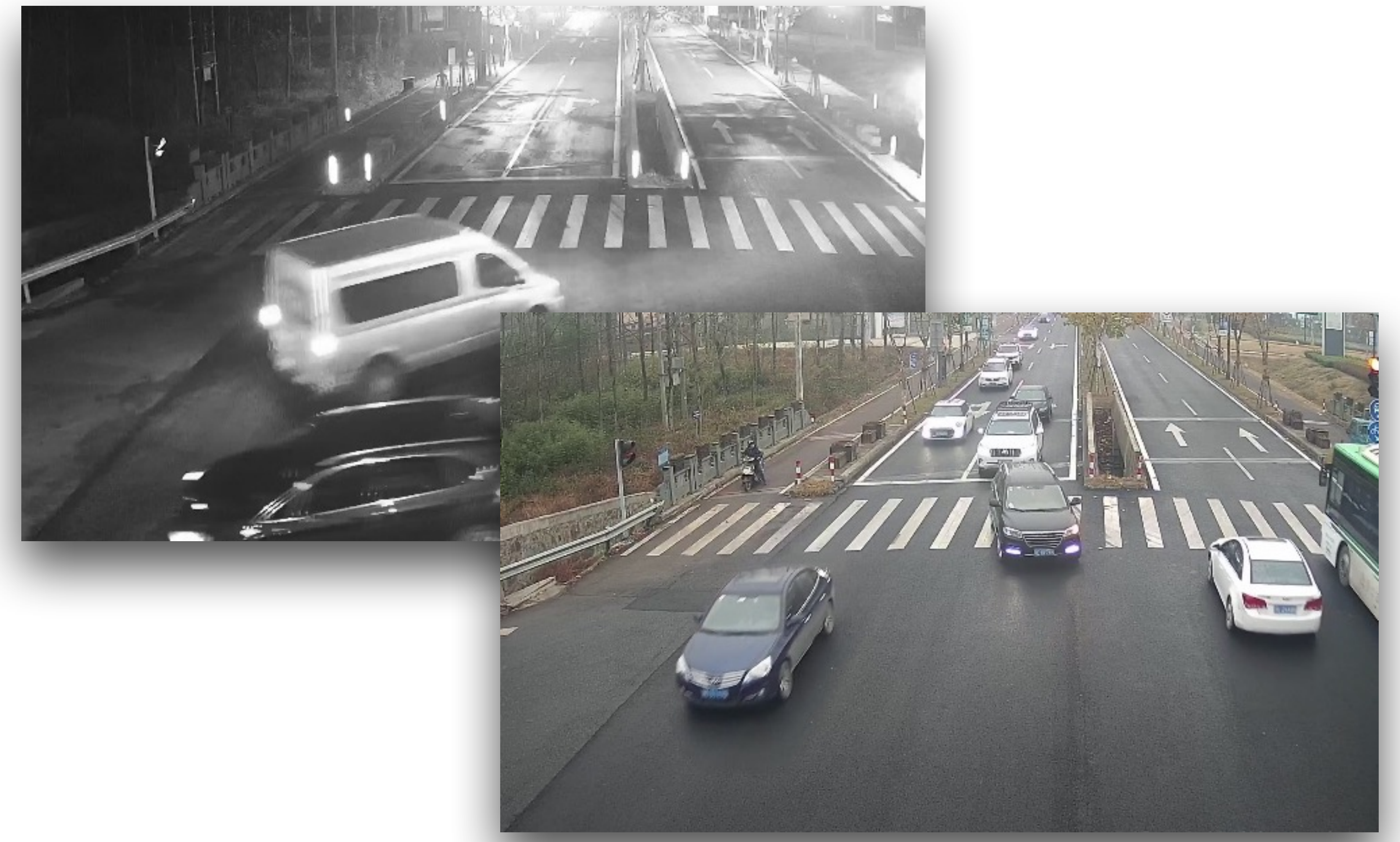
Real Systems

- 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

中国科学技术大学
University of Science and Technology of China

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

中国科学技术大学
University of Science and Technology of China

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

中国科学技术大学
University of Science and Technology of China

Baselines

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

<u>Target Application</u>

- 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

Method	Building (5/9 GB Mem.)		City (5/9 GB Mem.)	
	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)
Standalone	33.3/66.7	30/74	33.3/66.7	55/121
MTL	53.3	32.8	61.3	32.5
DRLS	45.7/81.3	58.7/107	39.5/77.6	102/188
Reducto	91.8/96.9	45.7/89	84.1/95.3	64/127
MLink	**94.1/97.9**	39.3/84	**94/97.4**	62/125

Evaluation

Video Analytics with Model Links

- GPU Memory as the cost budget

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

Method	Building (5/9 GB Mem.)		City (5/9 GB Mem.)	
	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)
Standalone	33.3/66.7	30/74	33.3/66.7	55/121
MTL	53.3	32.8	61.3	32.5
DRLS	45.7/81.3	58.7/107	39.5/77.6	102/188
Reducto	91.8/96.9	45.7/89	84.1/95.3	64/127
MLink	**94.1/97.9**	39.3/84	**94/97.4**	62/125

fast but accuracy is too low

Evaluation

中国科学技术大学
University of Science and Technology of China

Video Analytics with Model Links

- GPU Memory as the cost budget

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

Method	Building (5/9 GB Mem.)		City (5/9 GB Mem.)	
	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)
Standalone	33.3/66.7	30/74	33.3/66.7	55/121
MTL	53.3	32.8	61.3	32.5
DRLS	45.7/81.3	58.7/107	39.5/77.6	102/188
Reducto	91.8/96.9	45.7/89	84.1/95.3	64/127
MLink	**94.1/97.9**	39.3/84	**94/97.4**	62/125

improved accuracy
but too much overheads

Evaluation

University of Science and Technology of China

Video Analytics with Model Links

- GPU Memory as the cost budget

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

Method	Building (5/9 GB Mem.)		City (5/9 GB Mem.)	
	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)
Standalone	33.3/66.7	30/74	33.3/66.7	55/121
MTL	53.3	32.8	61.3	32.5
DRLS	45.7/81.3	58.7/107	39.5/77.6	102/188
Reducto	91.8/96.9	45.7/89	84.1/95.3	64/127
MLink	**94.1/97.9**	39.3/84	**94/97.4**	62/125

good trade-offs
but only applicable to video streams

Evaluation

University of Science and Technology of China

Video Analytics with Model Links

- GPU Memory as the cost budget

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

Method	Building (5/9 GB Mem.)		City (5/9 GB Mem.)	
	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)
Standalone	33.3/66.7	30/74	33.3/66.7	55/121
MTL	53.3	32.8	61.3	32.5
DRLS	45.7/81.3	58.7/107	39.5/77.6	102/188
Reducto	91.8/96.9	45.7/89	84.1/95.3	64/127
MLink	**94.1/97.9**	39.3/84	**94/97.4**	62/125

accurate, lightweight, and widely applicable

Menu

中国科学技术大学
University of Science and Technology of China

Main contents

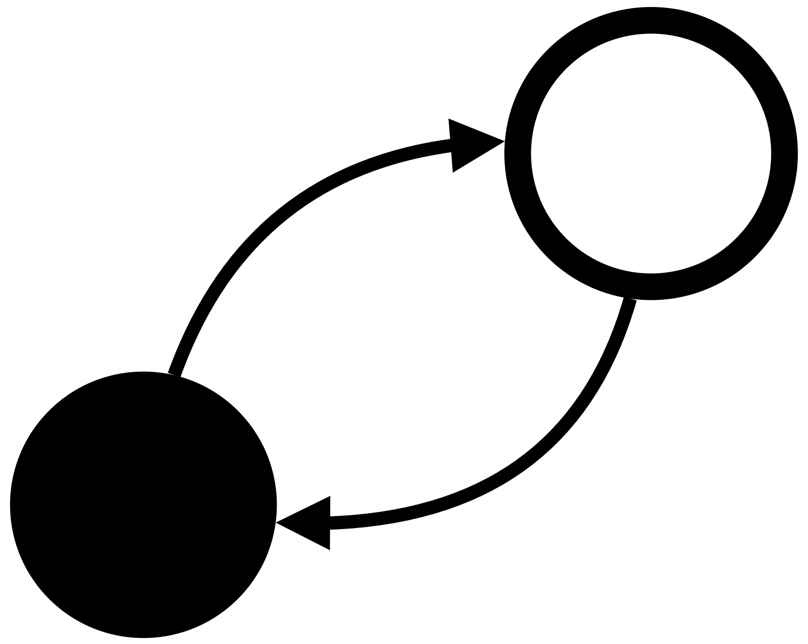
- Introduction
- Problem Statement
- Black-box Model Linking
- Collaborative Multi-model Inference
- Evaluation
- **Conclusion**

Conclusion

University of Science and Technology of China

Take-home Messages

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

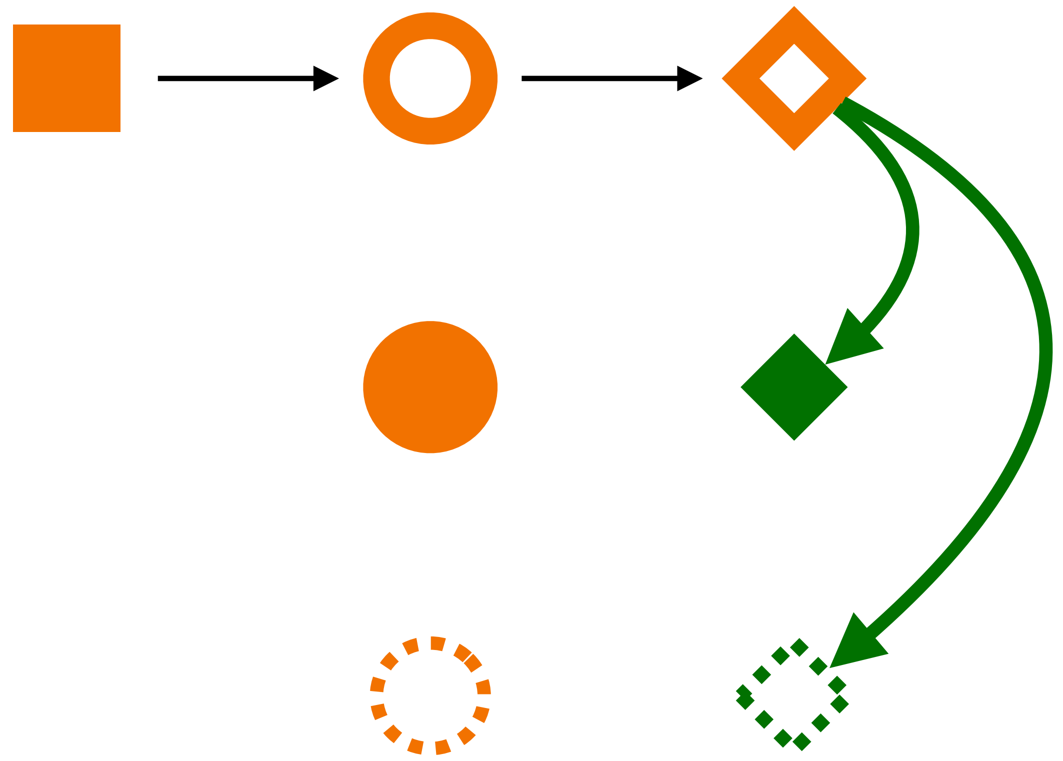


Conclusion

中国科学技术大学
University of Science and Technology of China

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



36TH AAAI CONFERENCE
ON ARTIFICIAL INTELLIGENCE

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

中国科学技术大学
University of Science and Technology of China