Deep Geometric Knowledge Distillation with Graphs

Carlos Lassance, Myriam Bontonou, Ghouthi Boukli Hacene, Vincent Gripon, Jian Tang, Antonio Ortega

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Mila

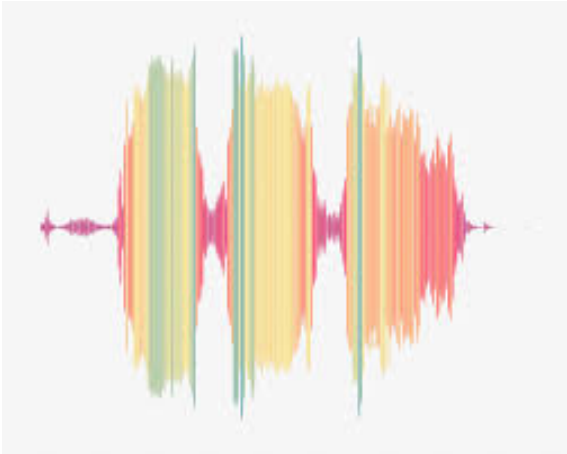
USC University of Southern California

ICASSP 2020

Outline

1. *Define* and motivate knowledge distillation;
2. *Introduce* the concept of Graph Knowledge Distillation (GKD);
3. *Present* empirical evaluation and analysis.

Motivation



Motivation

1T FLOPs for one decision

1024 V100 during 1 day for training

100M parameters to learn

4 TPUs during 1 month for training

Knowledge distillation

Goal

Neural network compression:

- Teacher transfers knowledge to student ;
- Student has less parameters than teacher;
- Student decisions consistent with teacher leads to
  - Student's accuracy teacher's accuracy;

Distilling the Knowledge in a Neural Network, Hinton et al., 2014

- Student mimicks the teacher's output;
  - Form of pseudo-labeling;
  - Uses teacher understanding of classes;

Knowledge distillation

Goal

Neural network compression:

- Teacher transfers knowledge to student ;
- Student has less parameters than teacher;
- Student decisions consistent with teacher leads to
  - Student's accuracy teacher's accuracy;

Distilling the Knowledge in a Neural Network, Hinton et al., 2014

- Student mimicks the teacher's output;
  - Form of pseudo-labeling;
  - Uses teacher understanding of classes;

Distillation

Layer/Block-wise distillation

Modern neural networks tend to be very deep;

Distilling only the output does not guarantee in uencing all layers;

Solution :

Enforce [student latent space = teacher latent space];

Drawback : intermediate representation dimensions may not match.

Distillation

Layer/Block-wise distillation

Modern neural networks tend to be very deep;

Distilling only the output does not guarantee in uencing all layers;

Solution :

Enforce [student latent space = teacher latent space];

Drawback : intermediate representation dimensions may not match.

Fitnets, Romero et al., 2015

Solution : add linear transformations so that dimensions match;

Drawback : the linear transformations are removed after training, jointly with part of the distilled knowledge.

Modern neural networks tend to be very deep;

Distilling only the output does not guarantee in uencing all layers;

Solution :

Enforce [student latent space = teacher latent space];

Drawback : intermediate representation dimensions may not match.

LIT, Koratana et al., 2019

Solution : perform the distillation block-wise and ensure that the outputs of each block have the same size;

Drawback : limits the architecture choice.

Distillation

IKD vs RKD

Individual Knowledge Distillation (IKD)

Methods we presented perform IKD;

Consider each example separately;

Either need transformations or same size representations.

Relational Knowledge Distillation (RKD)

Formalized in Park et al., 2019.

Goal: Transfer higher order knowledge to the student, e.g.:

- Distance between pairs of examples;
- Angles between triplets of examples.

Distillation

IKD vs RKD

Individual Knowledge Distillation (IKD)

Methods we presented perform IKD;

Consider each example separately;

Either need transformations or same size representations.

Relational Knowledge Distillation (RKD)

Formalized in Park et al., 2019.

Goal: Transfer higher order knowledge to the student, e.g.:

- Distance between pairs of examples;
- Angles between triplets of examples.

Training NN with KD

$$L = L_{task} + \quad _{KD} \quad L_{KD} \tag{1}$$

Individual Knowledge Distillation (IKD)

$$L_{IKD} = \sum_{`2} \sum_{x2X} L_d(x_{S`}; x_{T`}) \tag{2}$$

Relational Knowledge Distillation (RKD) - distance between pairs of examples

$$L_{RKD\text{-}D} = \sum_{`2} \sum_{(x;x^0)2X^2} L_d \left( \frac{kx_{S`} \quad x^0_{S`} k_2}{S`}; \frac{kx_{T`} \quad x^0_{T`} k_2}{T`} \right) \tag{3}$$

Distillation

IKD vs RKD

Training NN with KD

$$L = L_{task} + \quad _{KD} \quad L_{KD} \tag{1}$$

Individual Knowledge Distillation (IKD)

$$L_{IKD} = \sum_{`2} \sum_{x2X} L_d(x_{S`}; x_{T`}) \tag{2}$$

Relational Knowledge Distillation (RKD) - distance between pairs of examples

$$L_{RKD\text{-}D} = \sum_{`2} \sum_{(x;x^0)2X^2} L_d \left( \frac{kx_{S`} \quad x^0_{S`}\, k_2}{S`}; \frac{kx_{T`} \quad x^0_{T`}\, k_2}{T`} \right) \tag{3}$$

Distillation

IKD vs RKD

Training NN with KD

$$L = L_{task} + \quad _{KD} \quad L_{KD} \tag{1}$$

Individual Knowledge Distillation (IKD)

$$L_{IKD} = \sum_{`2} \sum_{x2X} L_d(x_{S`}; x_{T`}) \tag{2}$$

Relational Knowledge Distillation (RKD) - distance between pairs of examples

$$L_{RKD\text{-}D} = \sum_{`2} \sum_{(x;x^0)2X^2} L_d \left( \frac{kx_{S`} \quad x^0_{S`}k_2}{S`}; \frac{kx_{T`} \quad x^0_{T`}k_2}{T`} \right) \tag{3}$$

Graph Knowledge Distillation

We propose to use graphs to distillate knowledge:

- Use graphs to represent latent spaces;
- Student should mimick the teacher's graphs;
- Introducing a graph formalism opens research directions:
  - Graph Signal Processing (GSP) analysis of the results;
  - Better normalization ! easier to compare;
  - More meaningful relational distances;
  - Graph variations:
    - Task speci c graphs (inter/intra-class graphs);
    - Localized graphs (k-neighbors graphs);
    - Smoothed graphs (adjacency matrix to power p).
- Form of RKD.
- Concurrently proposed by Liu et al., 2019; Lee et al., 2019; and this work.

Graph Knowledge Distillation

We propose to use graphs to distillate knowledge:

- Use graphs to represent latent spaces;
- Student should mimick the teacher's graphs;
- Introducing a graph formalism opens research directions:
  - Graph Signal Processing (GSP) analysis of the results;
  - Better normalization ! easier to compare;
  - More meaningful relational distances;
  - Graph variations:
    - Task speci c graphs (inter/intra-class graphs);
    - Localized graphs (k-neighbors graphs);
    - Smoothed graphs (adjacency matrix to power p).
- Form of RKD.
- Concurrently proposed by Liu et al., 2019; Lee et al., 2019; and this work.

Graph Knowledge Distillation

We propose to use graphs to distillate knowledge:

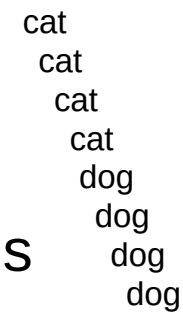
- Use graphs to represent latent spaces;
- Student should mimick the teacher's graphs;
- Introducing a graph formalism opens research directions:
  - Graph Signal Processing (GSP) analysis of the results;
  - Better normalization ! easier to compare;
  - More meaningful relational distances;
  - Graph variations:
    - Task speci c graphs (inter/intra-class graphs);
    - Localized graphs (k-neighbors graphs);
    - Smoothed graphs (adjacency matrix to power p).
- Form of RKD.
- Concurrently proposed by Liu et al., 2019; Lee et al., 2019; and this work.

Graph representation of latent spaces

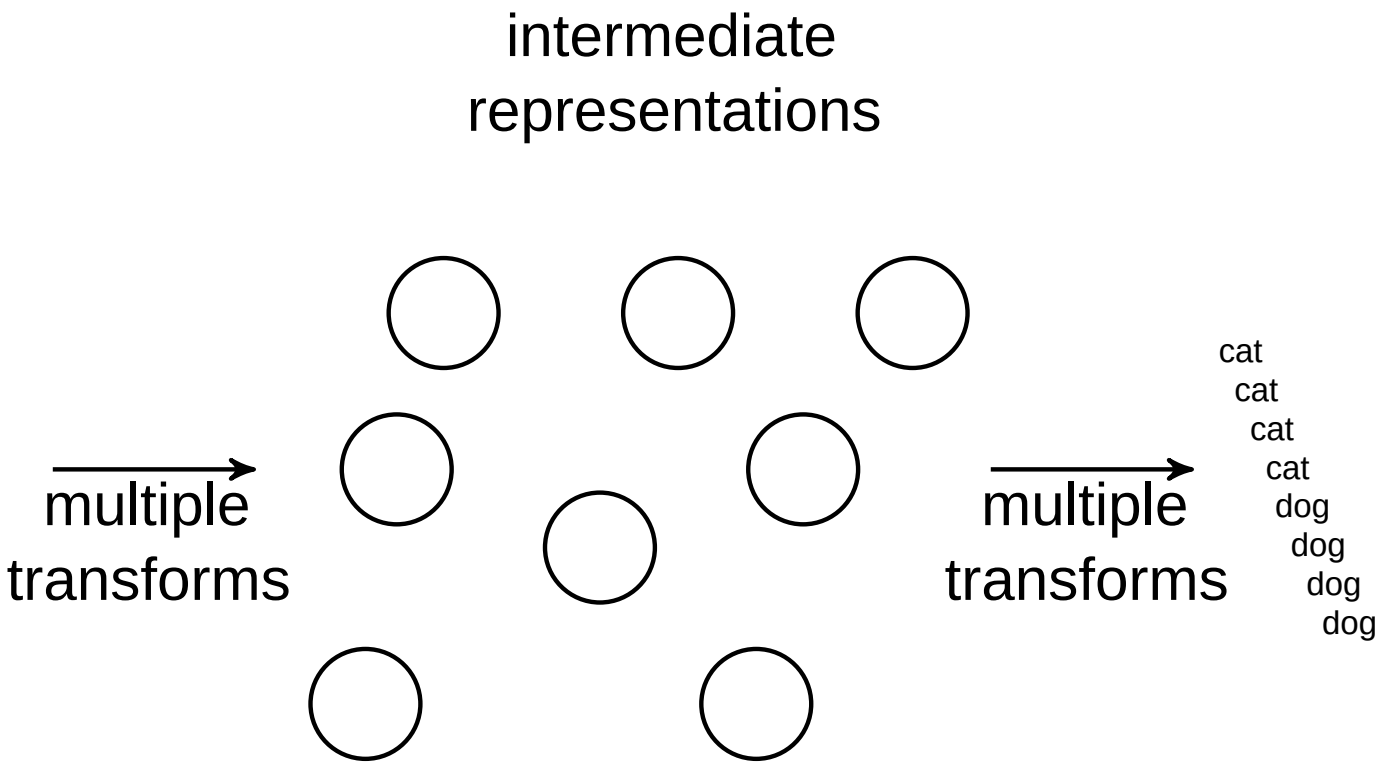
intermediate
representations

multiple
transforms

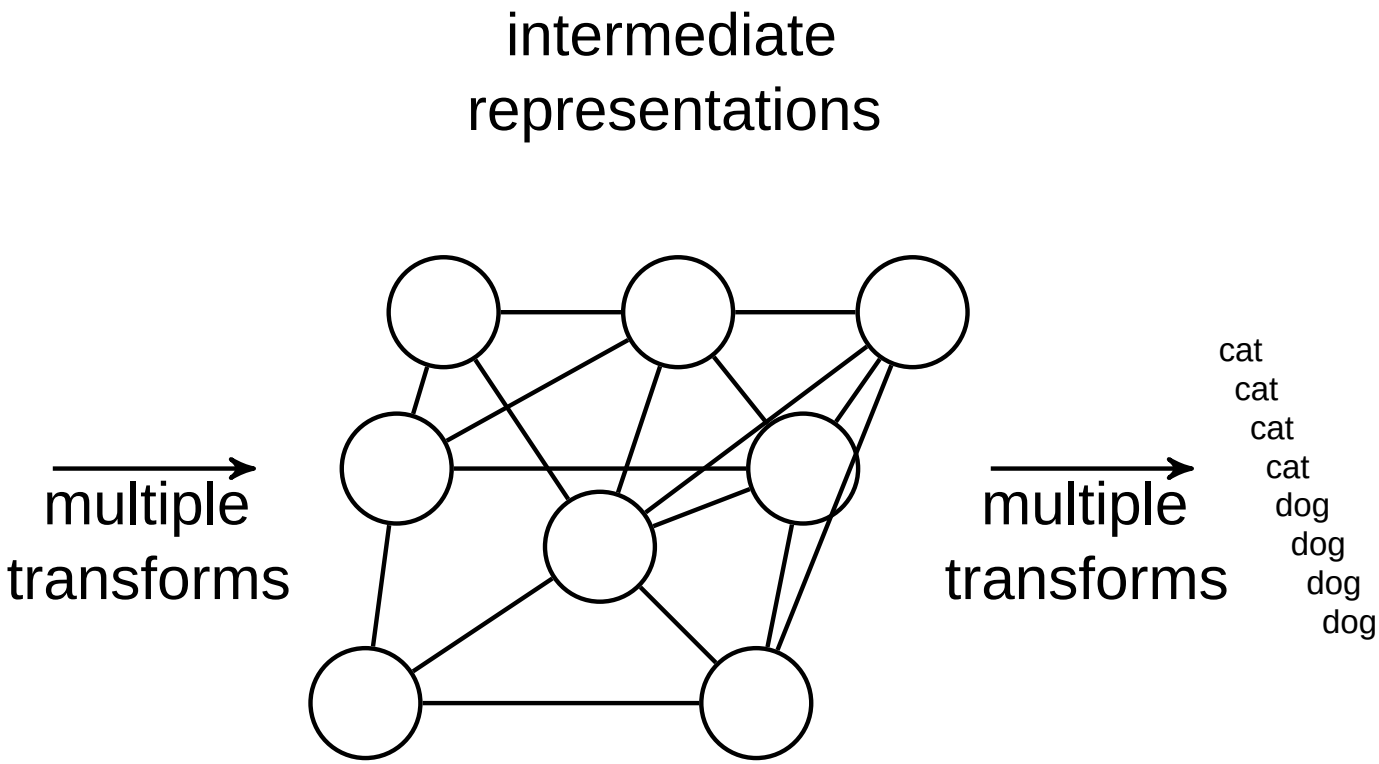
multiple
transforms



Graph representation of latent spaces



Graph representation of latent spaces



Relational Knowledge Distillation (RKD) - distance between pairs of examples

$$L_{RKD\text{-}D} = \sum_{\ell^2} \sum_{(x;x') \in \mathcal{X}^2} L_d \left(\frac{k x_{S,\ell} - x'_{S,\ell} k_2}{\sigma_{S,\ell}}; \frac{k x_{T,\ell} - x'_{T,\ell} k_2}{\sigma_{T,\ell}} \right) \quad (4)$$

Graph Knowledge Distillation (GKD)

$$L_{GKD} = \sum_{\ell^2} L_d(G_{S,\ell}(X); G_{T,\ell}(X)) : \quad (5)$$

$$L_{GKD} = \sum_{\ell^2} \left\| D_{S,\ell}^{\frac{1}{2}} A_{S,\ell} D_{S,\ell}^{\frac{1}{2}} - D_{T,\ell}^{\frac{1}{2}} A_{T,\ell} D_{T,\ell}^{\frac{1}{2}} k_2^2 \right\| : \quad (6)$$

Empirical experiments and analysis

Outline

Error rate comparison against RKD-D in CIFAR-10;

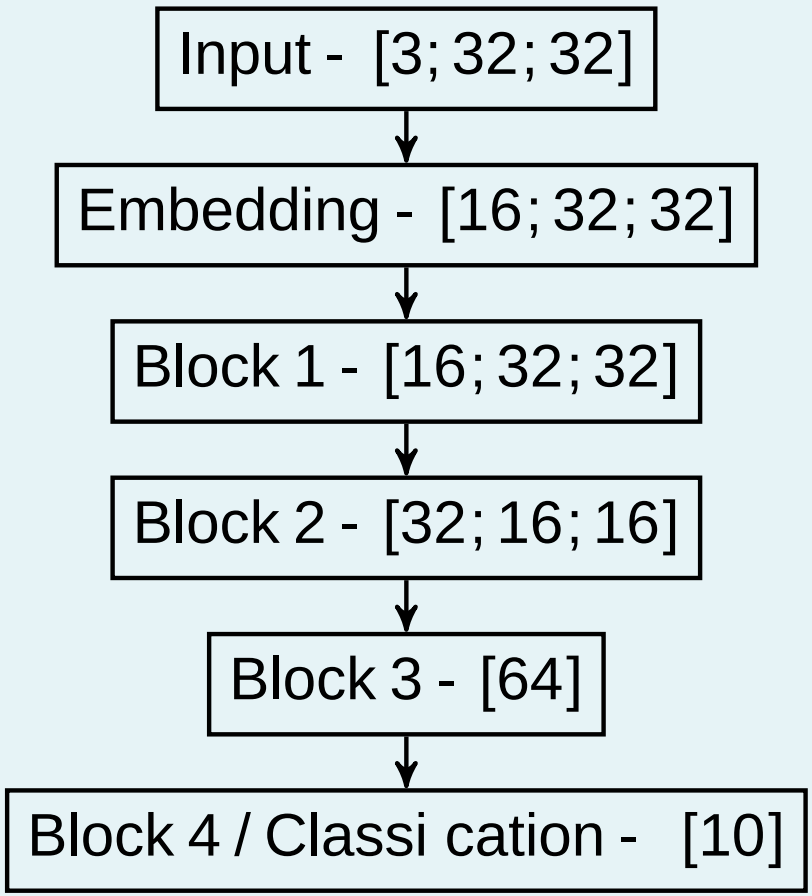
Classi cation consistency;

Graph signal smoothness analysis;

E ect of using task speci c graphs.

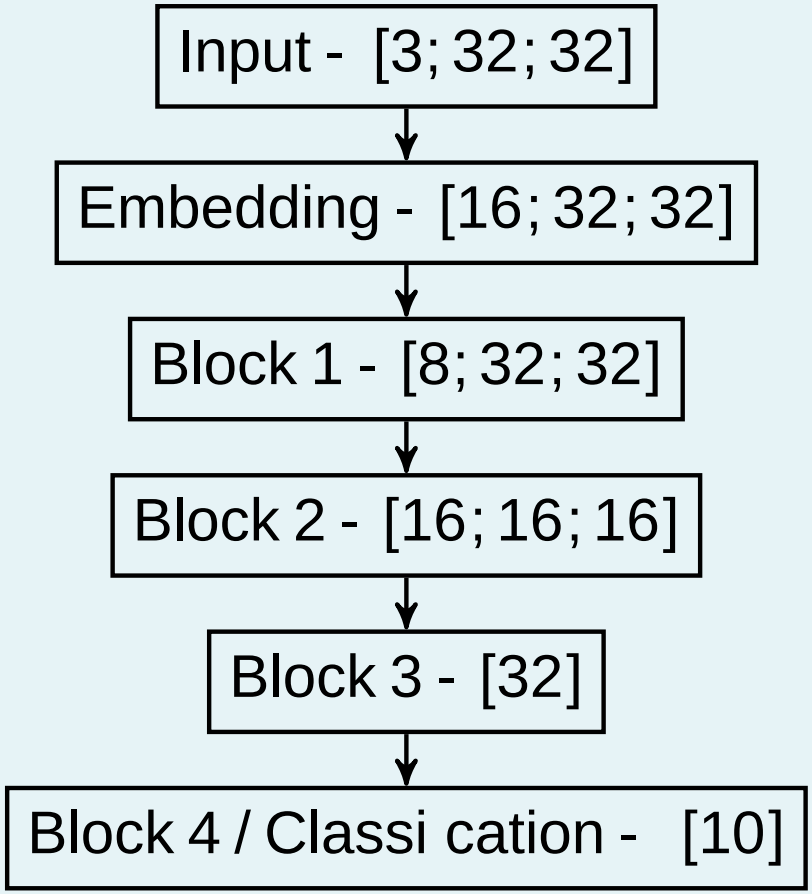
Neural net architectures

Teacher - WideResnet-28-1



Student - WideResnet-28-0.5

4 times smaller (parameters and FLOPS) than the teacher



Empirical experiments and analysis

CIFAR-10 error rate

Table: Median error rate and standard deviation on the CIFAR-10 dataset.

Method	CIFAR-10	Relative size
Teacher	7.27% (0.26)	100%
Student without KD (Baseline)	10.34% (0.27)	27%

Empirical experiments and analysis

CIFAR-10

Table: Median error rate and standard deviation comparison on the CIFAR-10.

Method	CIFAR-10	Relative size
Teacher	7.27% (0.26)	100%
Student without KD (baseline)	10.34% (0.27)	27%
RKD-D	10.05% (0.28)	27%
GKD	9.71% (0.27)	27%
GKD (inter-class graph)	9.31% (0.25)	27%

Empirical experiments and analysis

Classi cation consistency with teacher

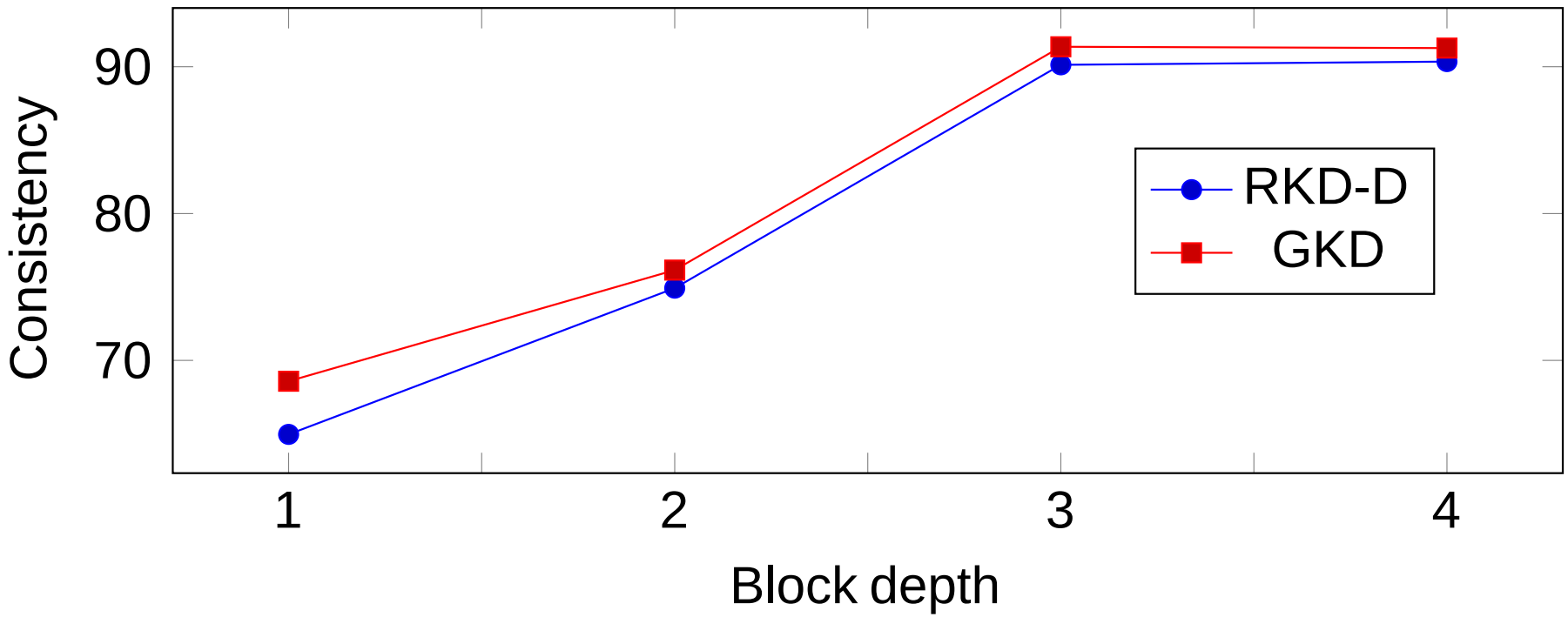


Figure: Analysis of the consistency of classi cation compared to the teacher, across blocks of RKD-D and GKD students.

Empirical experiments and analysis

Graph signal smoothness analysis

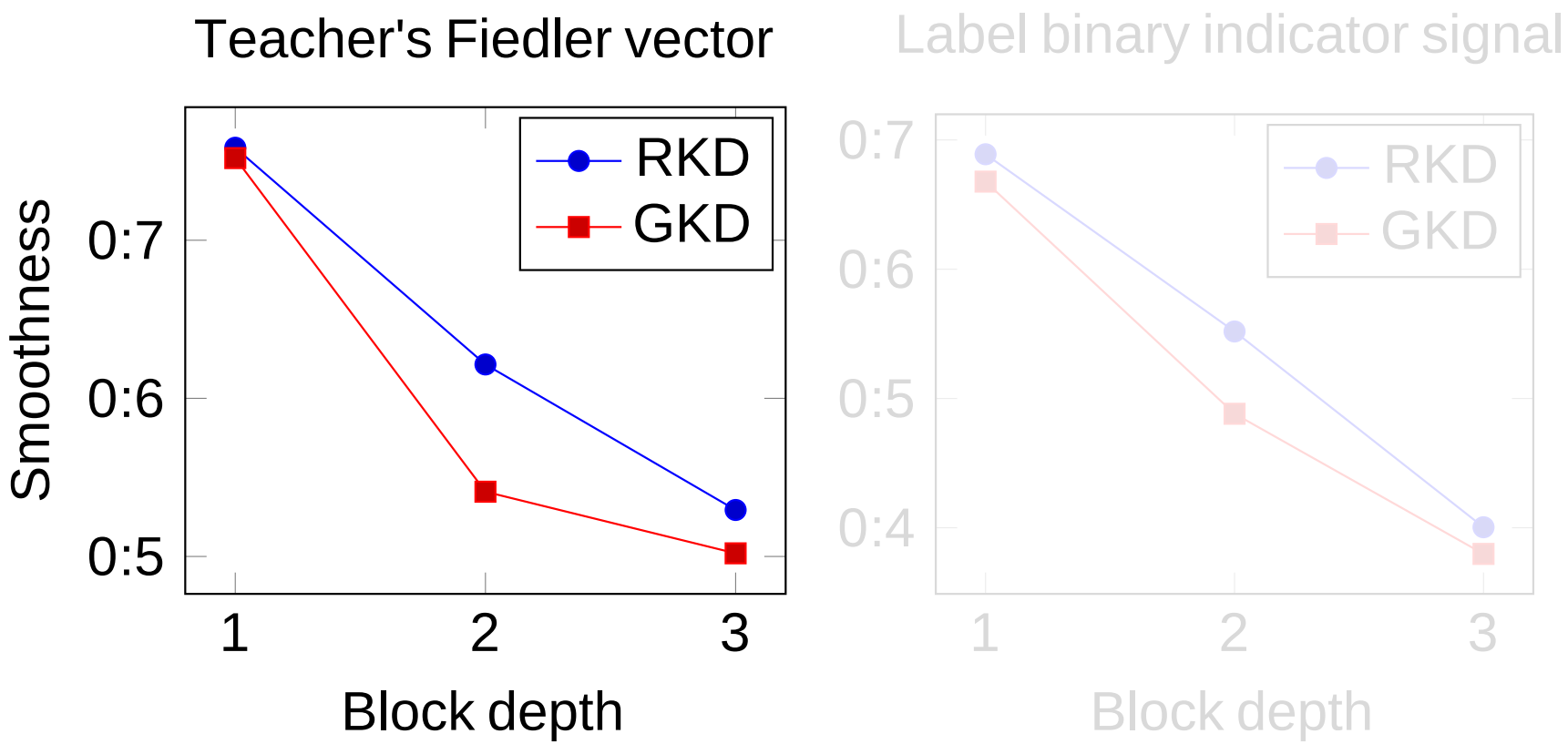


Figure: Analysis of the smoothness evolution across layers of the RKD and GKD students

Empirical experiments and analysis

Graph signal smoothness analysis

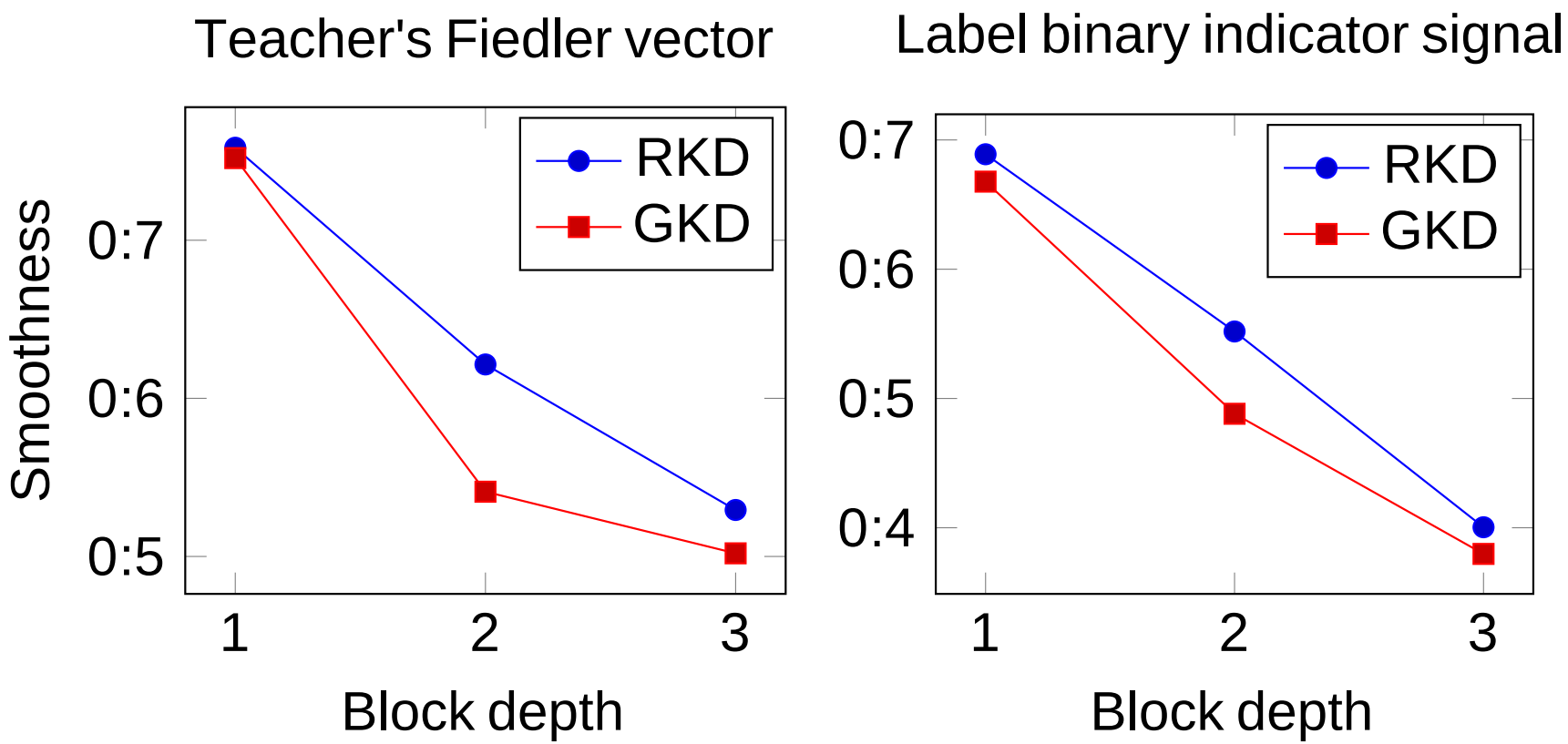


Figure: Analysis of the smoothness evolution across layers of the RKD and GKD students

Empirical experiments and analysis

Task specific graphs

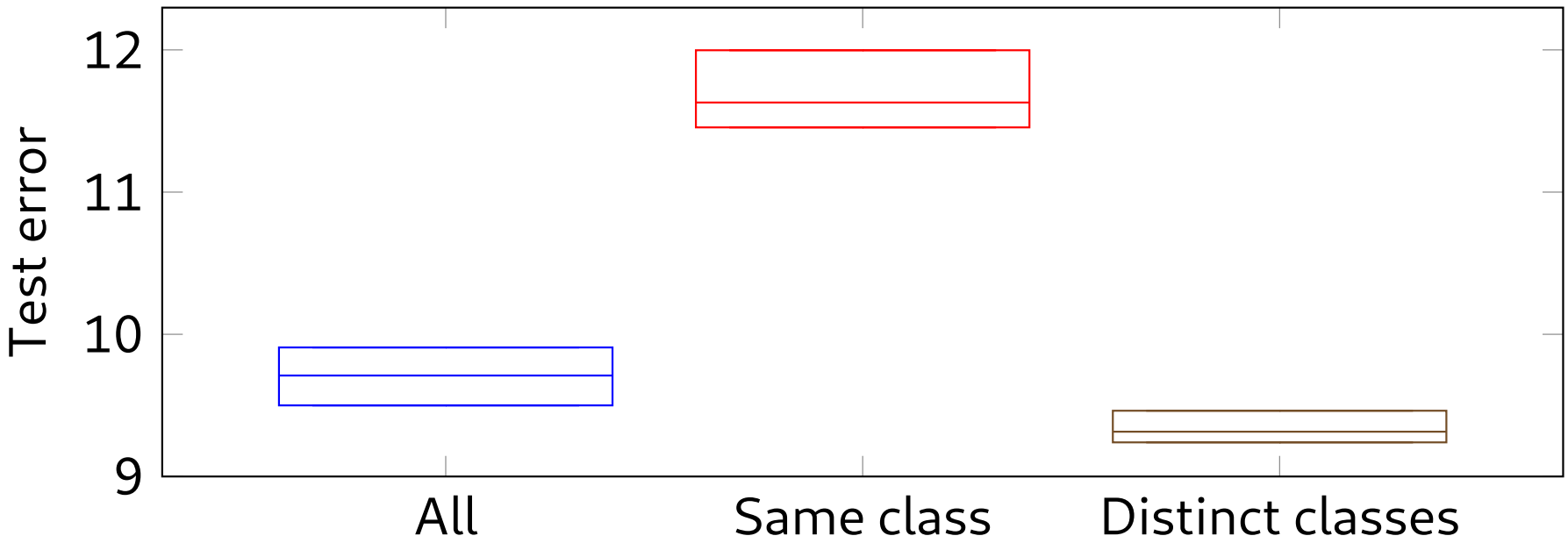


Figure: Analysis of the effect of task specific graphs. A graph of distinct classes has edges only between nodes of different classes, while same class graphs has edges only between nodes of the same class.

Conclusion

Wrap up

Graphs can be used as a proxy to the **geometry** of latent representations in deep neural networks;

Using graphs for knowledge distillation allows us to improve the performance of compressed student networks;

We were able to analyze the intermediate representations of our student networks.

Future work

Small gains, could be combined with other approaches;

More relevant graph distances, such as spectral distance;

Train the network block-wise instead of end-to-end.

Conclusion

Wrap up

Graphs can be used as a proxy to the **geometry** of latent representations in deep neural networks;

Using graphs for knowledge distillation allows us to improve the performance of compressed student networks;

We were able to analyze the intermediate representations of our student networks.

Future work

Small gains, could be combined with other approaches;

More relevant graph distances, such as spectral distance;

Train the network block-wise instead of end-to-end.

Thank you for watching this presentation.

I will be happy to answer any questions you have via e-mail: carlos.rosarkoslassance@imt-atlantique.fr.
Code available at `github.com/cadurosar/graph_kd`

References

Hinton et al., 2014, “Distilling the Knowledge in a Neural Network.”, NIPS Workshop;

Romero et al., 2015, “Fitnets:Hints for thin deep nets.”, ICLR;

Koratana, et al., 2019, "LIT: Learned intermediate representation training for model compression.", ICML;

Park et al., 2019, “Relational knowledge distillation.”, CVPR;

Liu et al., 2019, “Knowledge Distillation via Instance Relationship Graph.”, CVPR;

Lee et al., 2019, “Graph-based Knowledge Distillation by Multi-head Attention Network.”, BMVC.