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

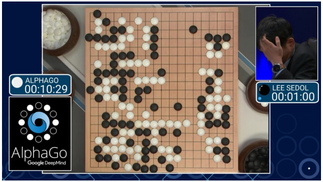
ICASSP 2020

Outline

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

Motivation


Phone
Mail
Messages
What can I help you with?

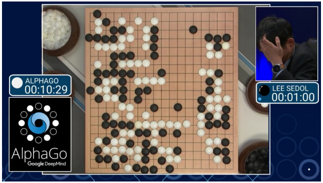

ALPHAGO
00:10:29
AlphaGo
Google DeepMind
LEE SEDOL
00:01:00

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 $\approx$ 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 $\approx$ 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 influencing all layers;
- **Solution**:
  - Enforce [student latent space = teacher latent space];
- **Drawback**: intermediate representation dimensions may not match.

- Modern neural networks tend to be very deep;
- Distilling only the output does not guarantee influencing 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 influencing 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.

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.

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

Training NN with KD

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{KD}} \cdot \mathcal{L}_{\text{KD}} \tag{1}$$

Individual Knowledge Distillation (IKD)

$$\mathcal{L}_{\text{IKD}} = \sum_{\ell \in \Lambda} \sum_{\mathbf{x} \in \mathcal{X}} \mathcal{L}_d(\mathbf{x}_{S_\ell}, \mathbf{x}_{T_\ell}) \tag{2}$$

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

$$\mathcal{L}_{\text{RKD-D}} = \sum_{\ell \in \Lambda} \sum_{(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2} \mathcal{L}_d\left(\frac{\|\mathbf{x}_{S_\ell} - \mathbf{x}'_{S_\ell}\|_2}{\Delta_{S_\ell}}, \frac{\|\mathbf{x}_{T_\ell} - \mathbf{x}'_{T_\ell}\|_2}{\Delta_{T_\ell}}\right) \tag{3}$$

Training NN with KD

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{KD}} \cdot \mathcal{L}_{\text{KD}} \tag{1}$$

Individual Knowledge Distillation (IKD)

$$\mathcal{L}_{\text{IKD}} = \sum_{\ell \in \Lambda} \sum_{\mathbf{x} \in X} \mathcal{L}_d(\mathbf{x}_{S_\ell}, \mathbf{x}_{T_\ell}) \tag{2}$$

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

$$\mathcal{L}_{\text{RKD-D}} = \sum_{\ell \in \Lambda} \sum_{(\mathbf{x},\mathbf{x}') \in X^2} \mathcal{L}_d\left(\frac{\|\mathbf{x}_{S_\ell} - \mathbf{x}'_{S_\ell}\|_2}{\Delta_{S_\ell}}, \frac{\|\mathbf{x}_{T_\ell} - \mathbf{x}'_{T_\ell}\|_2}{\Delta_{T_\ell}}\right) \tag{3}$$

Training NN with KD

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{KD}} \cdot \mathcal{L}_{\text{KD}} \tag{1}$$

Individual Knowledge Distillation (IKD)

$$\mathcal{L}_{\text{IKD}} = \sum_{\ell \in \Lambda} \sum_{\mathbf{x} \in X} \mathcal{L}_d(\mathbf{x}_{S_\ell}, \mathbf{x}_{T_\ell}) \tag{2}$$

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

$$\mathcal{L}_{\text{RKD-D}} = \sum_{\ell \in \Lambda} \sum_{(\mathbf{x},\mathbf{x}') \in X^2} \mathcal{L}_d\left(\frac{\|\mathbf{x}_{S_\ell} - \mathbf{x}'_{S_\ell}\|_2}{\Delta_{S_\ell}}, \frac{\|\mathbf{x}_{T_\ell} - \mathbf{x}'_{T_\ell}\|_2}{\Delta_{T_\ell}}\right) \tag{3}$$

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 $\rightarrow$ easier to compare;
  - More meaningful relational distances;
  - Graph variations:
    1. Task specific graphs (inter/intra-class graphs);
    2. Localized graphs (k-neighbors graphs);
    3. 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:

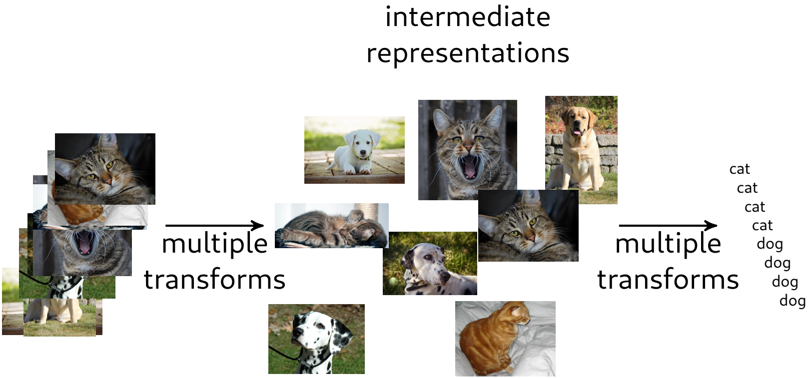
- Use graphs to represent latent spaces;
- Student should mimick the teacher's graphs;
- Introducing a graph formalism opens research directions:
  - Graph Signal Processing (GSP) analysis of the results;
  - Better normalization $\rightarrow$ easier to compare;
  - More meaningful relational distances;
  - Graph variations:
    1. Task specific graphs (inter/intra-class graphs);
    2. Localized graphs (k-neighbors graphs);
    3. 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

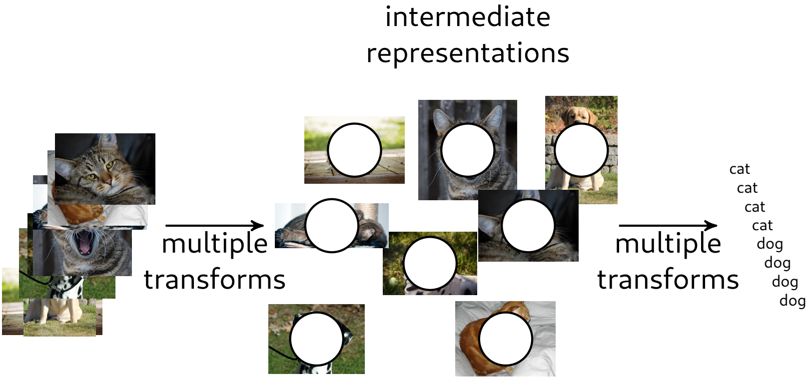
We propose to use graphs to distillate knowledge:

- 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 $\rightarrow$ easier to compare;
    - More meaningful relational distances;
    - Graph variations:
        1. Task specific graphs (inter/intra-class graphs);
        2. Localized graphs (k-neighbors graphs);
        3. Smoothed graphs (adjacency matrix to power p).
- Form of RKD.
- Concurrently proposed by Liu et al., 2019; Lee et al., 2019; and this work.

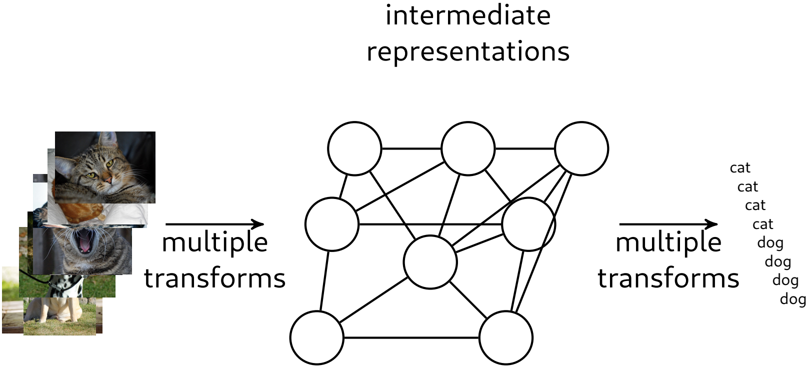
Graph representation of latent spaces



Graph representation of latent spaces



Graph representation of latent spaces



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

$$\mathcal{L}_{\text{RKD-D}} = \sum_{\ell \in \Lambda} \sum_{(\mathbf{x},\mathbf{x}') \in X^2} \mathcal{L}_d \left(\frac{\|\mathbf{x}_{S_\ell} - \mathbf{x}'_{S_\ell}\|_2}{\Delta_{S_\ell}}, \frac{\|\mathbf{x}_{T_\ell} - \mathbf{x}'_{T_\ell}\|_2}{\Delta_{T_\ell}} \right) \quad (4)$$

Graph Knowledge Distillation (GKD)

$$\mathcal{L}_{\text{GKD}} = \sum_{\ell \in \Lambda} \mathcal{L}_d(\mathcal{G}_{S_\ell}(X), \mathcal{G}_{T_\ell}(X)) . \quad (5)$$

$$\mathcal{L}_{\text{GKD}} = \sum_{\ell \in \Lambda} \|\mathbf{D}_{S_\ell}^{-\frac{1}{2}} \mathbf{A}_{S_\ell} \mathbf{D}_{S_\ell}^{-\frac{1}{2}} - \mathbf{D}_{T_\ell}^{-\frac{1}{2}} \mathbf{A}_{T_\ell} \mathbf{D}_{T_\ell}^{-\frac{1}{2}}\|_2^2 . \quad (6)$$

1. Error rate comparison against RKD-D in CIFAR-10;
2. Classification consistency;
3. Graph signal smoothness analysis;
4. Effect of using task specific graphs.

Neural net architectures

Teacher - WideResnet-28-1

Input - [3, 32, 32]
↓
Embedding - [16, 32, 32]
↓
Block 1 - [16, 32, 32]
↓
Block 2 - [32, 16, 16]
↓
Block 3 - [64]
↓
Block 4 / Classification - [10]

Student - WideResnet-28-0.5

$\approx$ 4 times smaller (parameters and FLOPS) than the teacher

Input - [3, 32, 32]
↓
Embedding - [16, 32, 32]
↓
Block 1 - [8, 32, 32]
↓
Block 2 - [16, 16, 16]
↓
Block 3 - [32]
↓
Block 4 / Classification - [10]

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

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

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

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

Empirical experiments and analysis

Classification consistency with teacher

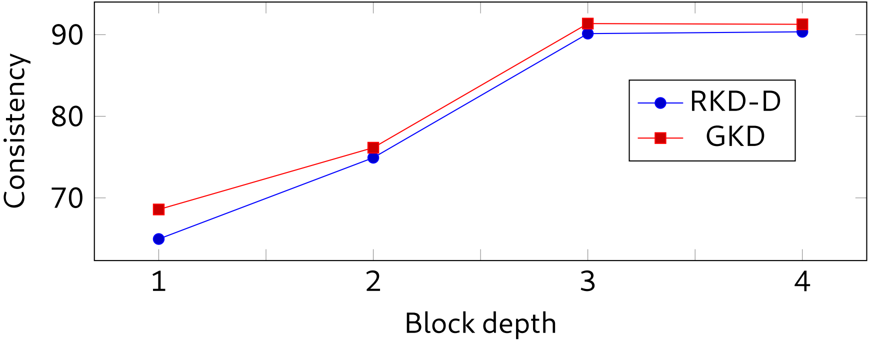


Figure: Analysis of the consistency of classification compared to the teacher, across blocks of RKD-D and GKD students.

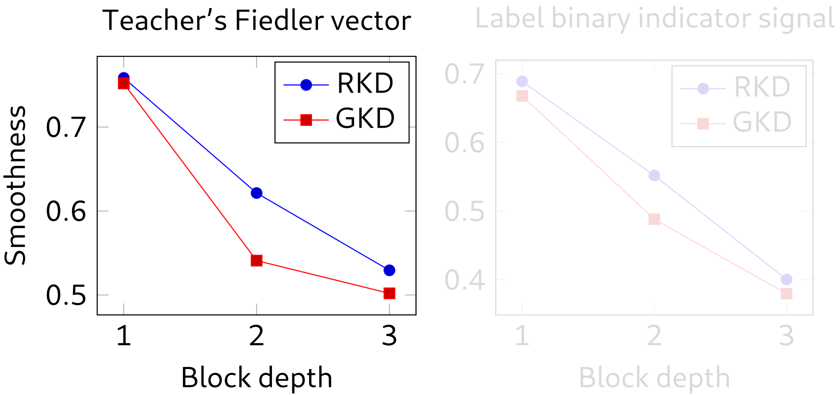


Figure: Analysis of the smoothness evolution across layers of the RKD and GKD students

Empirical experiments and analysis

Graph signal smoothness analysis

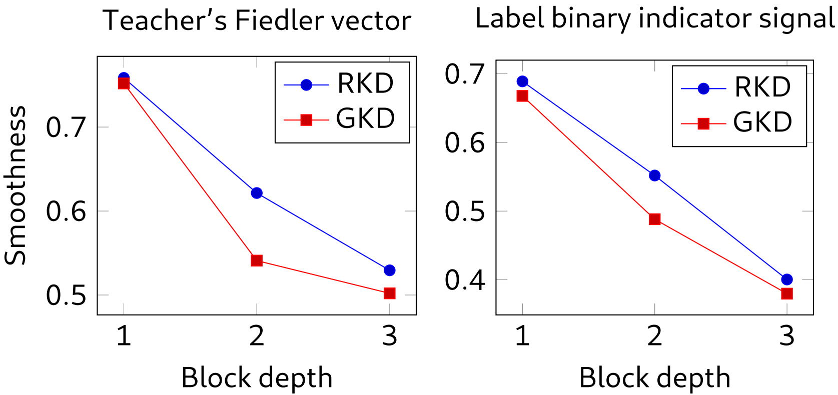


Figure: Analysis of the smoothness evolution across layers of the RKD and GKD students

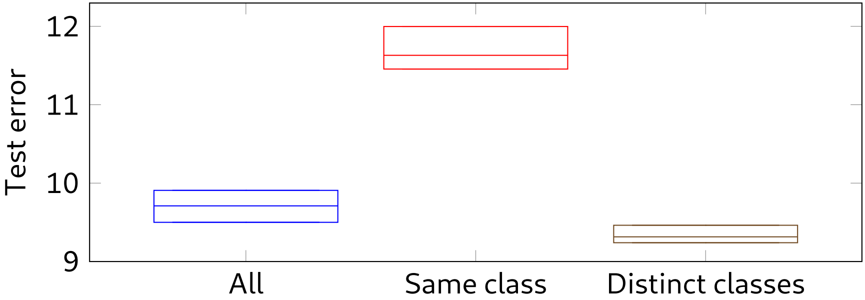


Figure: Analysis of the effect of task specific graphs. A graph of distinct classes has edges only between nodes of different classes, while same class graphs has edges only between nodes of the same class.

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.