Knowledge Distillation : knowledge transfer between two networks during training

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

Course organisation

Sessions

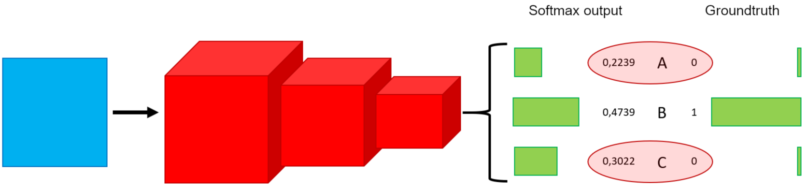
1. Introduction/Refresher on Deep Learning
2. Quantization,
3. Pruning,
4. Data Augmentation,
5. Factorization,
6. Distillation,
7. Embedded Software and Hardware for DL,
8. Final session.

Course organisation

Sessions

1. Introduction/Refresher on Deep Learning
2. Quantization,
3. Pruning,
4. Data Augmentation,
5. Factorization,
6. Distillation,
7. Embedded Software and Hardware for DL,
8. Final session.

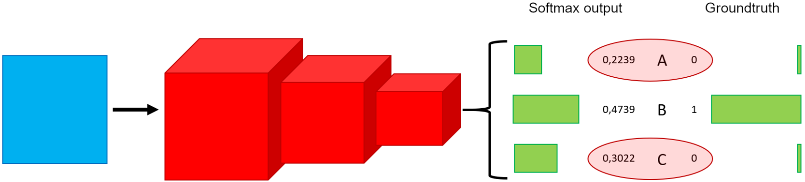
Original observation



Distilling the Knowledge in a Neural Network, Hinton & al. 2015

- The values of wrong classes give hints on the network's ability to generalize
- These soft labels can serve as a more relevant groundtruth to train another network

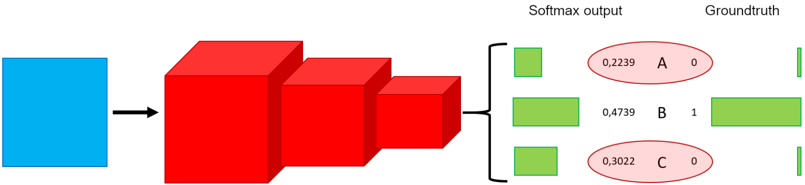
Original observation



Distilling the Knowledge in a Neural Network, Hinton & al. 2015

- The values of wrong classes give hints on the network's ability to generalize
- These soft labels can serve as a more relevant groundtruth to train another network

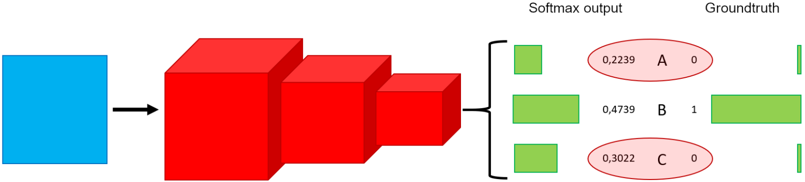
Original observation



Distilling the Knowledge in a Neural Network, Hinton & al. 2015

- The values of wrong classes give hints on the network's ability to generalize
- These soft labels can serve as a more relevant groundtruth to train another network

Original observation



Distilling the Knowledge in a Neural Network, Hinton & al. 2015

- The values of wrong classes give hints on the network's ability to generalize
- These soft labels can serve as a more relevant groundtruth to train another network

Basic principle

Hinton's distillation

$$\mathcal{L}_{KD} = \underbrace{H(y_{true}, P_s)}_{\text{supervised term}} + \underbrace{\lambda D_{KL}(P_T, P_S)}_{\text{distillation's term}} \text{ with } D_{KL}(P_T, P_S) = \sum_i P_T(i) log(\frac{P_T(i)}{P_S(i)})$$

with P_S the student's output, P_T the teacher's output, H the cross-entropy loss and D_{KL} the Kullback-Leibler divergence (or relative entropy).

Let's consider the *softmax* outputs: $q_i = \frac{e^{z_i/T}}{\sum_i e^{z_j/T}}$

- T is 1 during inference but is superior to 1 for the distillation term (therefore, the outputs are softer)
- Since the amplitude of the outputs is $1/T^2$, the result must be multiplied by T^2

Basic principle

Hinton's distillation

$$\mathcal{L}_{KD} = \underbrace{H(y_{true}, P_s)}_{\text{supervised term}} + \underbrace{\lambda D_{KL}(P_T, P_S)}_{\text{distillation's term}} \text{ with } D_{KL}(P_T, P_S) = \sum_i P_T(i) log(\frac{P_T(i)}{P_S(i)})$$

with P_S the student's output, P_T the teacher's output, H the cross-entropy loss and D_{KL} the Kullback-Leibler divergence (or relative entropy).

Let's consider the *softmax* outputs: $q_i = \frac{e^{z_i/T}}{\sum_i e^{z_j/T}}$

- T is 1 during inference but is superior to 1 for the distillation term (therefore, the outputs are softer)
- Since the amplitude of the outputs is $1/T^2$, the result must be multiplied by T^2

Basic principle

Hinton's distillation

$\mathcal{L}_{KD} = \underbrace{H(y_{true}, P_s)}_{\text{supervised term}} + \underbrace{\lambda D_{KL}(P_T, P_S)}_{\text{distillation's term}}$ with $D_{KL}(P_T, P_S) = \sum_i P_T(i) log(\frac{P_T(i)}{P_S(i)})$

with P_S the student's output, P_T the teacher's output, H the cross-entropy loss and D_{KL} the Kullback-Leibler divergence (or relative entropy).

Let's consider the *softmax* outputs: $q_i = \frac{e^{z_i/T}}{\sum_i e^{z_j/T}}$

- T is 1 during inference but is superior to 1 for the distillation term (therefore, the outputs are softer)
- Since the amplitude of the outputs is $1/T^2$, the result must be multiplied by T^2

Basic principle

Hinton's distillation

$$\mathcal{L}_{KD} = \underbrace{H(y_{true}, P_s)}_{\text{supervised term}} + \underbrace{\lambda D_{KL}(P_T, P_S)}_{\text{distillation's term}} \text{ with } D_{KL}(P_T, P_S) = \sum_i P_T(i) log(\frac{P_T(i)}{P_S(i)})$$

with P_S the student's output, P_T the teacher's output, H the cross-entropy loss and D_{KL} the Kullback-Leibler divergence (or relative entropy).

Let's consider the *softmax* outputs: $q_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}$

- T is 1 during inference but is superior to 1 for the distillation term (therefore, the outputs are softer)
- Since the amplitude of the outputs is $1/T^2$, the result must be multiplied by T^2

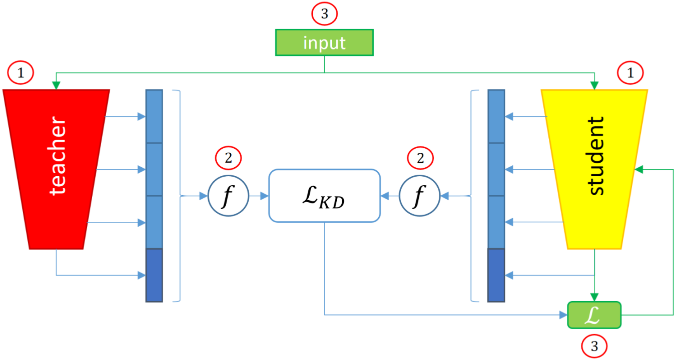
Many methods...

Table 5 Performance comparison of different knowledge distillation methods on CIFAR10. Note that ↑ indicates the performance improvement of the student network learned by each method comparing with the corresponding baseline model.

Offline Distillation				
Methods	Knowledge	Teacher (baseline)	Student (baseline)	Accuracies
FSP (Yim et al., 2017)	RelK	ResNet26 (91.91)	ResNet8 (87.91)	88.70 (0.79 ↑)
FT (Kim et al., 2018)	FeaK	ResNet56 (93.61)	ResNet20 (92.22)	93.15 (0.93 ↑)
IRG (Liu et al., 2019g)	RelK	ResNet20 (91.45)	ResNet20-x0.5 (88.36)	90.69 (2.33 ↑)
SP (Tung and Mori, 2019)	RelK	WRN-40-1 (93.49)	WRN-16-1 (91.26)	91.87 (0.61 ↑)
SP (Tung and Mori, 2019)	RelK	WRN-40-2 (95.76)	WRN-16-8 (94.82)	95.45 (0.63 ↑)
FN (Xu et al., 2020b)	FeaK	ResNet110 (94.29)	ResNet56 (93.63)	94.14 (0.51 ↑)
FN (Xu et al., 2020b)	FeaK	ResNet56 (93.63)	ResNet20 (92.11)	92.67 (0.56 ↑)
AdaIN (Yang et al., 2020a)	FeaK	ResNet26 (93.58)	ResNet8 (87.78)	89.02 (1.24 ↑)
AdaIN (Yang et al., 2020a)	FeaK	WRN-40-2 (95.07)	WRN-16-2 (93.98)	94.67 (0.69 ↑)
AE-KD (Du et al., 2020)	FeaK	ResNet56 (—)	MobileNetV2 (75.97)	77.07 (1.10 ↑)
JointRD (Li et al., 2020b)	FeaK	ResNet34 (95.39)	plain-CNN 34 (93.73)	94.78 (1.05 ↑)
TOFD (Zhang et al., 2020a)	FeaK	ResNet152 (—)	ResNeXt50-4 (94.49)	97.09 (2.60 ↑)
TOFD (Zhang et al., 2020a)	FeaK	ResNet152 (—)	MobileNetV2 (90.43)	93.34 (2.91 ↑)
CTKD (Zhao et al., 2020a)	RelK, FeaK	WRN-40-1 (93.43)	WRN-16-1 (91.28)	92.50 (1.22 ↑)
CTKD (Zhao et al., 2020a)	RelK, FeaK	WRN-40-2 (94.70)	WRN-16-2 (93.68)	94.42 (0.74 ↑)

Knowledge Distillation: A Survey, Gou et al. 2020

Evolution of the literature



1. Which teacher and student to choose?
2. What knowledge to extract?
3. What type of learning?

Which teacher and student to choose?

Teacher

- **A big network**
- Multiple networks

Student

- **A smaller network**
- A quantized network
- The same network

Two main philosophies :

- **Compression** : using a bigger network to improve a smaller, less expensive one
- **Optimisation** : using distillation to improve a network's performance, e.g.: *Born-Again Neural Networks*, Furlanello & al., 2018

Which teacher and student to choose?

Teacher

- **A big network**
- Multiple networks

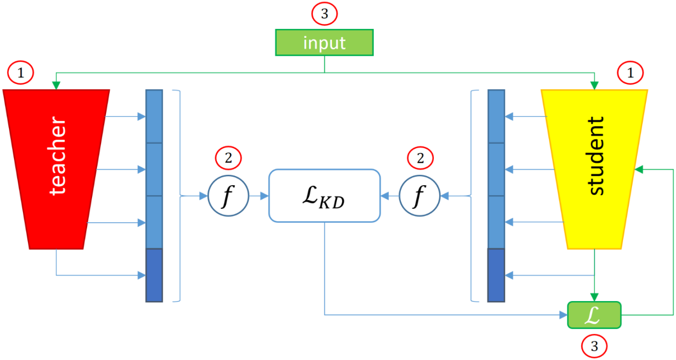
Student

- **A smaller network**
- A quantized network
- The same network

Two main philosophies :

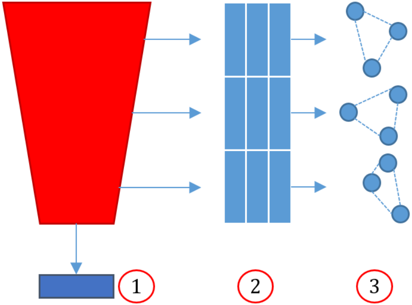
- **Compression** : using a bigger network to improve a smaller, less expensive one
- **Optimisation** : using distillation to improve a network's performance, e.g.: *Born-Again Neural Networks*, Furlanello & al., 2018

Evolution of the literature



1. Which teacher and student to choose?
2. What knowledge to extract?
3. What type of learning?

What knowledge to extract? 1/3

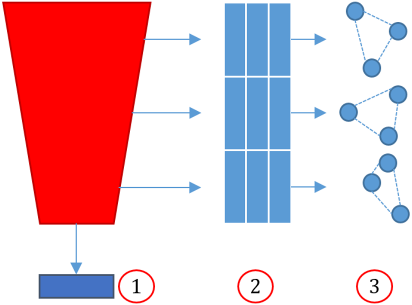


1. Output logits
2. Intermediate representations
3. Relations

Three representative articles:

1. *Distilling the Knowledge in a Neural Network*, Hinton & al. 2015
2. *FitNets : hints for thin deep nets*, Romero & al., 2014)
3. *Relational Knowledge Distillation*, Park & al., 2019

What knowledge to extract? 1/3

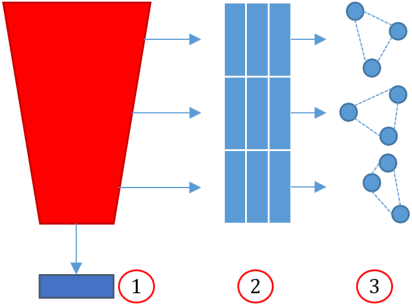


1. Output logits
2. Intermediate representations
3. Relations

Three representative articles:

1. *Distilling the Knowledge in a Neural Network*, Hinton & al. 2015
2. *FitNets : hints for thin deep nets*, Romero & al., 2014)
3. *Relational Knowledge Distillation*, Park & al., 2019

What knowledge to extract? 1/3



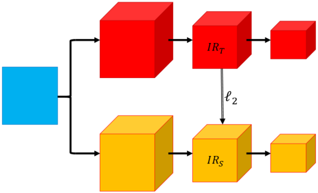
1. Output logits
2. Intermediate representations
3. Relations

Three representative articles:

1. *Distilling the Knowledge in a Neural Network*, Hinton & al. 2015
2. *FitNets : hints for thin deep nets*, Romero & al., 2014)
3. *Relational Knowledge Distillation*, Park & al., 2019

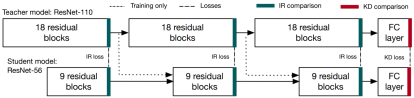
What knowledge to extract? 2/3

FitNets : hints for thin deep nets, Romero et al., 2014



- Distillation using intermediate représentations
- $\mathcal{L}_{IR} = \|IR_T - IR_S\|_2$

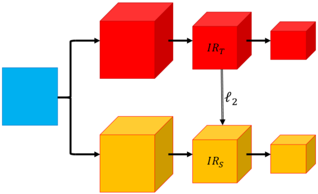
LIT: Block-wise intermediate representation training for model compression, Koratana & al., 2018



- The network is sliced into independently trained blocks
- If dimensions don't match: a linear (or 1×1 convolution) layer is inserted

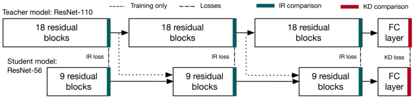
What knowledge to extract? 2/3

FitNets : hints for thin deep nets, Romero et al., 2014



- Distillation using intermediate représentations
- $\mathcal{L}_{IR} = \|IR_T - IR_S\|_2$

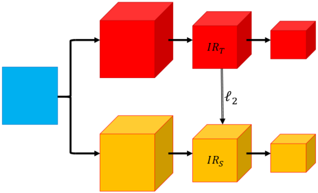
LIT: Block-wise intermediate representation training for model compression, Koratana & al., 2018



- The network is sliced into independently trained blocks
- If dimensions don't match: a linear (or 1×1 convolution) layer is inserted

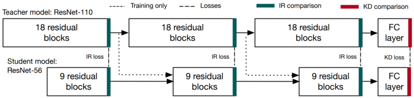
What knowledge to extract? 2/3

FitNets : hints for thin deep nets, Romero et al., 2014



- Distillation using intermediate représentations
- $\mathcal{L}_{IR} = \|IR_T - IR_S\|_2$

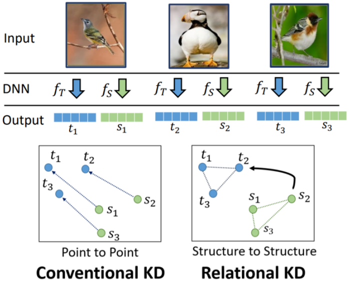
LIT: Block-wise intermediate representation training for model compression, Koratana & al., 2018



- The network is sliced into independently trained blocks
- If dimensions don't match: a linear (or 1×1 convolution) layer is inserted

RKD: learning how to discriminate data 2/2

Relational Knowledge Distillation, Park & al., 2019



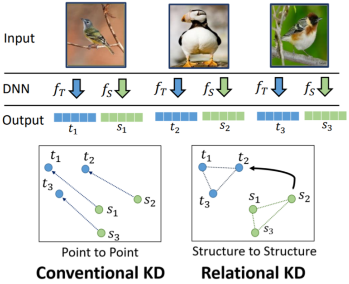
- Abstraction of IR
- For each batch, ℓ_2 norm between pairs of IR
- We compare these distances for the student and the teacher and add $\mathcal{L}_{RKD}$ to the loss

Relational Knowledge Distillation

$\mathcal{L}_{RKD} = \sum_{i,j \in \mathcal{X}^N} \ell(\phi(t_i, t_j), \phi(s_i, s_j))$ avec $\phi(t_i, t_j) = \frac{1}{\mu}\|t_i - t_j\|_2$,
is ℓ the Huber norm, a.k.a. "smooth ℓ_1 norm", μ is a normalization term and $\mathcal{X}^N$ is the training batch

RKD: learning how to discriminate data 2/2

Relational Knowledge Distillation, Park & al., 2019



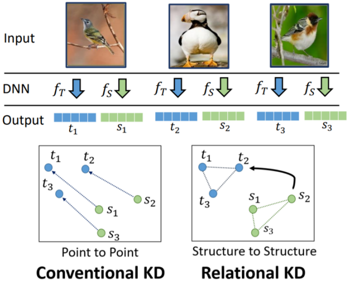
- Abstraction of IR
- For each batch, ℓ_2 norm between pairs of IR
- We compare these distances for the student and the teacher and add $\mathcal{L}_{RKD}$ to the loss

Relational Knowledge Distillation

$\mathcal{L}_{RKD} = \sum_{i,j \in \mathcal{X}^N} \ell(\phi(t_i, t_j), \phi(s_i, s_j))$ avec $\phi(t_i, t_j) = \frac{1}{\mu} \|t_i - t_j\|_2$,
is ℓ the Huber norm, a.k.a. "smooth ℓ_1 norm", μ is a normalization term and $\mathcal{X}^N$ is the training batch

RKD: learning how to discriminate data 2/2

Relational Knowledge Distillation, Park & al., 2019

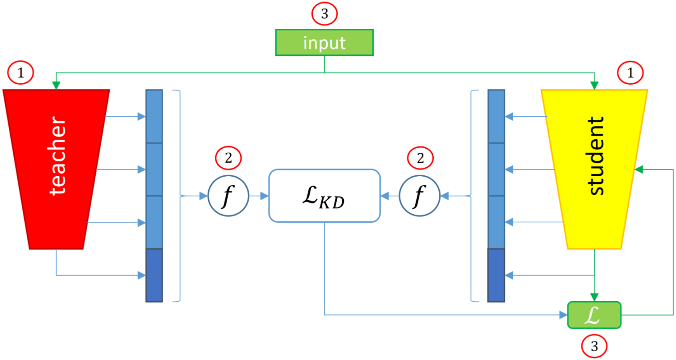


- Abstraction of IR
- For each batch, ℓ_2 norm between pairs of IR
- We compare these distances for the student and the teacher and add $\mathcal{L}_{RKD}$ to the loss

Relational Knowledge Distillation

$\mathcal{L}_{RKD} = \sum_{i,j \in \mathcal{X}^N} \ell(\phi(t_i, t_j), \phi(s_i, s_j))$ avec $\phi(t_i, t_j) = \frac{1}{\mu}\|t_i - t_j\|_2$,
is ℓ the Huber norm, a.k.a. "smooth ℓ_1 norm", μ is a normalization term and $\mathcal{X}^N$ is the training batch

Evolution of the literature



1. Which teacher and student to choose?
2. What knowledge to extract?
3. What type of learning?

What type of learning?

A very rich domain...

- The teacher is...
  - ... **pre-trained** (offline)
  - ... trained at the same time (online)
  - ... also the student (self-distillation)
- The inputs data are...
  - ... **the same for both teacher and student**
  - ... different (cross-modal distillation, ex: *SoundNet: Learning Sound Representations from Unlabeled Video*, Aytar et al. 2016)
  - ... synthesized (data-free distillation or adversarial distillation)
- The training is...
  - ... **single**
  - ... iterative

What type of learning?

A very rich domain...

- The teacher is...
  - ... **pre-trained** (offline)
  - ... trained at the same time (online)
  - ... also the student (self-distillation)
- The inputs data are...
  - ... **the same for both teacher and student**
  - ... different (cross-modal distillation, ex: *SoundNet: Learning Sound Representations from Unlabeled Video*, Aytar et al. 2016)
  - ... synthesized (data-free distillation or adversarial distillation)
- The training is...
  - ... **single**
  - ... iterative

What type of learning?

A very rich domain...

- The teacher is...
  - ... **pre-trained** (offline)
  - ... trained at the same time (online)
  - ... also the student (self-distillation)
- The inputs data are...
  - ... **the same for both teacher and student**
  - ... different (cross-modal distillation, ex: *SoundNet: Learning Sound Representations from Unlabeled Video*, Aytar et al. 2016)
  - ... synthesized (data-free distillation or adversarial distillation)
- The training is...
  - ... **single**
  - ... iterative

What type of learning?

A very rich domain...

- The teacher is...
  - ... **pre-trained** (offline)
  - ... trained at the same time (online)
  - ... also the student (self-distillation)
- The inputs data are...
  - ... **the same for both teacher and student**
  - ... different (cross-modal distillation, ex: *SoundNet: Learning Sound Representations from Unlabeled Video*, Aytar et al. 2016)
  - ... synthesized (data-free distillation or adversarial distillation)
- The training is...
  - ... **single**
  - ... iterative