Congratulations! You passed!

Grade received 85.71% To pass 80% or higher

Go to next item

Knowledge Distillation

Total points 7

1. True Or False: The goal of knowledge distillation is optimizing the network implementation: **1 / 1 point**

- (•) False
- ( ) True

Correct
Exactly! Rather than optimizing, distillation seeks to create a more efficient model.

2. In knowledge distillation, the teacher will be trained using a _____. **0 / 1 point**

- ( ) A Standard objective function
- ( ) A Soft Target
- ( ) GoogLeNet
- (•) K-L divergence

Incorrect
You might need to recheck that, K-L divergence is a metric for comparing predictions.

3. True Or False: DistilBERT is a bigger version of BERT with a modified architecture, but the same number of layers. **1 / 1 point**

- (•) False
- ( ) True

Correct
You're right! it's a smaller version of BERT: they reduced the numbers of layers and kept the rest of the architecture identical

4. True Or False: In knowledge distillation, the "teacher" network is deployed in production as it is able to mimic the complex feature relationships of the "student" network. **1 / 1 point**

- (•) False
- ( ) True

Correct
Exactly! It's actually the "student" network the one deployed to mimic the "teacher" network.

5. For a multi-class classification problem, which ones of the following statements are true regarding the training cost functions of the "student" and the "teacher" networks? (Select all that apply) **1 / 1 point**

- [x] Soft targets encode more information about the knowledge learned by the teacher than its output class prediction per example.

Correct
That's right! Soft targets provide more information that the output class predicted per example as they include information about all the classes per training example through the probability distribution.

- [ ] They both share the same cost functions,
- [ ] The teacher network is trained to maximize its accuracy and the the student network uses a cost function to output the same classes as the teacher network.
- [x] The teacher network is trained to maximize its accuracy and the the student network uses a cost function to approximate the probability distributions of the predictions of the teacher network.

Correct
That's right!

6. When the softmax temperature ___, the soft targets defined by the teacher network become less informative **1 / 1 point**

- ○ increases
- ◉ decreases
- ○ is equal to 1

Correct
That's right! The softness of the teacher's distribution is worse, thus less informative.

7. Generally, knowledge distillation is done by blending two loss functions and involves several hyperparameters. Here, L_h is the cross-entropy loss from the hard labels and LKL is the Kullback–Leibler divergence loss from the teacher labels. Which of the following statements are correct about the hyperparameters of knowledge distillation? (Select all that apply) **1 / 1 point**

- ☑ When computing the the "standard" loss between the student's predicted class probabilities and the ground-truth "hard" labels, we use a value of the softmax temperature T equal to 1

Correct
That's right! This way, the student loss function would be a classical softmax function

- ☑ In case of heavy data augmentation after training the teacher network, the alpha hyperparameter should be high in the student network loss function

Correct
That's correct! This high alpha parameter would reduce the influence of the hard labels that went through aggressive perturbations due to data augmentation

- ☐ In case of heavy data augmentation after training the teacher network, the alpha hyperparameter should be low in the student network loss function
- ☐ When computing the the "standard" loss between the student's predicted class probabilities and the ground-truth "hard" labels, we use the same value of the softmax temperature T to compute the softmax on the teacher's logits