1512 Project

(a) What is the purpose of using KD in this paper? [0.5 Marks]

The purpose is to transfer the knowledge from a well-trained cumbersome model to a small model that is suitable for deployment to many users with the more stringent requirement on latency and computational resources.

(b) In the paper, what knowledge is transferred from the teacher model to the student model? [0.5 Marks]

(c) What is the temperature hyperparameter *T*? Why do we use it when transferring knowledge from one model to another? What effect does the temperature hyperparameter have in KD? [0.5 Marks]

(d) Explain in detail the loss functions on which the teacher and student models are trained in this paper. How does the task balance parameter affect student learning? [0.5 Marks]

(e) Can we look at the KD as a regularization technique, here? Explain your rationale. [0.5

Marks]

Resources

<https://www.cs.toronto.edu/~hinton/absps/subclassdistillation.pdf>

<https://medium.com/analytics-vidhya/knowledge-distillation-in-a-deep-neural-network-c9dd59aff89b>

<https://intellabs.github.io/distiller/knowledge_distillation.html>

<https://keras.io/examples/vision/knowledge_distillation/>