November 2020

Use of Temperature in Dynamic Knowledge Distillation for Joint Optimization Model

Anonymous

Use of Temperature in Dynamic Knowledge Distillation for Joint Optimization Model

ABSTRACT

In a joint optimization model, information from a large, complex teacher model is transported to small, light student models using knowledge distillation. Dynamic knowledge distillation allows the student models to learn from the teacher model on the fly. However, the performance of a joint optimization model that uses dynamic knowledge distillation suffers if the teacher model contains too much noise from the negative labels, or does not have enough information from the negative labels. This disclosure describes techniques to implement dynamic knowledge distillation by using temperature to control the amount of information transmission about negative labels from a teacher model to a student model in a joint optimization model. Greater amount of information about the negative labels can be transmitted by setting the temperature high, while noise from the negative labels in the teacher model can be suppressed by setting the temperature low.

KEYWORDS

- Joint model optimization
- Ranking model
- Ad ranking
- Deep learning
- Teacher model
- Student model
- Prediction distribution
- Negative label
- Dynamic structure neural network
- Knowledge distillation

BACKGROUND

One of the most important challenges for content providers, social media platforms, and other online entities is to match content to the audience, e.g., to connect advertisers to customers by delivering the most relevant advertisements (ads) to the target audience. To achieve this,

content providers need to compare and rank hundreds of potentially relevant ads for each advertising opportunity. Ranking systems that are utilized to rank ads use a joint optimization model to simultaneously train multiple models that have some shared uniform low level arches. Using a joint optimization model saves the computing infrastructure cost and improves the consistency across models. Thus, the models trained by one joint model can serve multiple stages in an ads ranking system and improve system performance.

In the joint optimization model, information from a large, complex teacher model is transported to small, light student models using knowledge distillation [1]. A static knowledge distillation process uses a fixed teacher label, while a dynamic knowledge distillation process uses a concurrent teacher label from the joint model to teach the student model. Applying dynamic knowledge distillation on joint optimization models can improve the consistency between models and the performance of the student model.

The advantage of dynamic knowledge distillation is that it allows the student models to learn from both positive and negative labels in the teacher model. Information is passed about labels in the teacher model in accordance with their prediction probability. The prediction probability distribution of the teacher model is a soft target, and the sample input labels are a hard target. Compared with a hard target (labels that take only 0 or 1 value), a soft target (possibility distribution) contains more information. Two models can produce an identical hard target prediction while containing different soft target possibility distributions. Transporting the whole soft target to the student model can give the student model more possible values about the positive and negative labels due to the soft target.

However, the performance of the joint optimization model that uses dynamic knowledge distillation suffers if the teacher model contains too much noise from negative labels or does not have enough information from the negative labels.

DESCRIPTION

This disclosure describes techniques to improve dynamic knowledge distillation by using temperature to adjust the probability distribution output of the teacher model in a joint optimization model. Temperature is implemented by adding a filter layer before the output SoftMax/Sigmoid layer in the teacher model. The filter layer divides every value from the input by a constant value T (distillation temperature), and passes on the result to the Sigmoid/SoftMax layer. The output of the SoftMax/Sigmoid layer are the teacher labels that go into the loss function of the student model for dynamic knowledge distillation.

For T=1, the original input teacher labels are retained. For T < 1 the probability distribution is sharper than the original input, while for T > 1 the probability distribution is flatter than the original input. Figure 1 illustrates the extreme use cases (T=0, T= infinity) for the transformation of teacher labels using temperature.

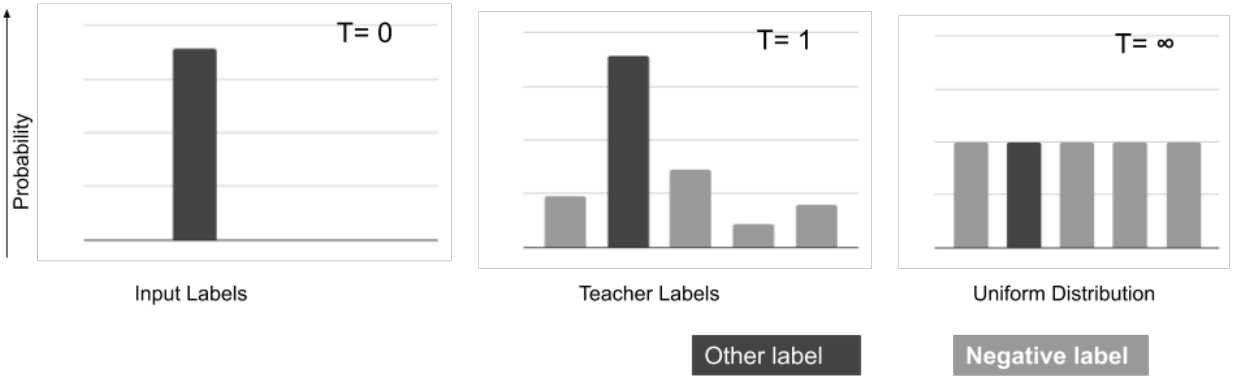

Figure 1: Transformation of input labels using temperature

The center chart represents T=1, where the original probability distribution of the teacher labels is retained. The chart on the right represents T= ∞, which is a uniform distribution. As shown in the chart on the left, for T = 0 the probability distribution is reduced to binary labels.

The techniques described in this disclosure can be illustrated by considering a joint optimization model consisting of a Deep & Wide (D&W) neural network and a Dynamic structure neural network (DSNN) as shown in Figure 2 below.

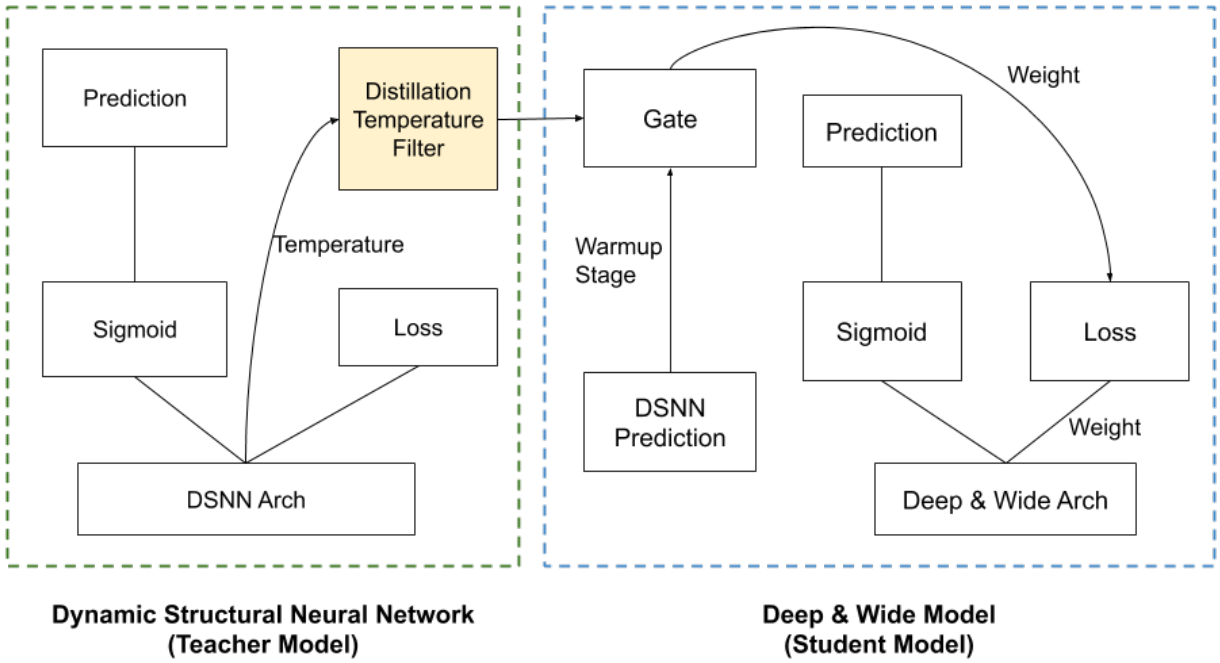

Figure 2: Joint model optimization using static/dynamic knowledge distillation

In the joint optimization model illustrated in Figure 2, the DSNN model functions as the teacher model while the D&W model is the student model. A distillation temperature filter is applied to the output of the teacher model and the resultant probability distribution is passed on to the input gate in the student model to implement dynamic knowledge distillation. Additionally, the input gate also receives a fixed teacher label, which is the model prediction result from the teacher model to implement static knowledge distillation. Static knowledge

distillation is used as a warmup method and dynamic knowledge distillation is used as the optimization method. A weight value *W* is used in the loss function to balance the contribution of teacher label and input labels.

Equations 1-3 demonstrate the process for joint estimation using knowledge distillation with temperature.

$$E(x|T) = (W)\sum_i \widehat{H}(x|T) + (1-W)\sum_i \overline{H}(x) \quad (1)$$

$$\widehat{H}(x|T) = -\hat{y}_i(x|T)\log y_i(x) - (1-\hat{y}_i(x|T))\log(1-y_i(x)) \quad (2)$$

$$\overline{H}(x) = -\bar{y}_i(x)\log y_i(x) - (1-\bar{y}_i(x))\log(1-y_i(x)) \quad (3)$$

In the above equations, *E* is the final estimation, *W* is the weight for teacher labels, $\widehat{H}$ is the cross entropy for teacher labels, $\overline{H}$ is the cross entropy for input labels, *y* is the distilled model prediction, $\hat{y}$ is the teacher model prediction, and $\bar{y}$ is the input sample label.

Temperature is implemented as:

$$\hat{y}_i(x|T) = 1/(1+e^{-x/T})$$

The following principles can be used for setting the temperature in the joint optimization model:

- The temperature should be set high, if the teacher model has a small positive/negative label ratio or if more information is to be transmitted from the negative labels in the teacher model.
- The temperature should be set low, if the teacher model has a large positive/negative label ratio or if the noise from the negative labels in the teacher model is to be suppressed.

CONCLUSION

This disclosure describes techniques to implement dynamic knowledge distillation by using temperature to control the amount of information transmission about negative labels from a teacher model to a student model in a joint optimization model. Greater amount of information about the negative labels can be transmitted by setting the temperature high, while noise from the negative labels in the teacher model can be suppressed by setting the temperature low.

REFERENCES

1. Buciluǎ, Cristian, Rich Caruana, and Alexandru Niculescu-Mizil. "Model compression." In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 535-541. 2006.
2. Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).