Hard Gate Knowledge Distillation - Leverage Calibration for a Robust and Reliable Language Model

Dongkyu Lee[1,3*] **Zhiliang Tian**[2†] **Yingxiu Zhao**[1]
Ka Chun Cheung[3] **Nevin L. Zhang**[1]
[1]Department of Computer Science and Engineering, HKUST
[2]College of Computer, National University of Defense Technology
[3]NVIDIA AI Technology Center, NVIDIA
[1]{dleear, yzhaocx, lzhang}@cse.ust.hk
[2]tianzhilianghit@gmail.com [3]chcheung@nvidia.com

Abstract

In knowledge distillation, a student model is trained with supervisions from both knowledge from a teacher and observations drawn from a training data distribution. Knowledge of a teacher is considered a subject that holds inter-class relations which send a meaningful supervision to a student; hence, much effort has been put to find such knowledge to be distilled. In this paper, we explore a question that has been given little attention: "*when to distill such knowledge*." The question is answered in our work with the concept of model calibration; we view a teacher model not only as a source of knowledge but also as a gauge to detect miscalibration of a student. This simple and yet novel view leads to a hard gate knowledge distillation scheme that switches between learning from a teacher model and training data. We verify the gating mechanism in the context of natural language generation at both the token-level and the sentence-level. Empirical comparisons with strong baselines show that hard gate knowledge distillation not only improves model generalization, but also significantly lowers model calibration error.

1 Introduction

In recent years, the deep learning community has achieved marked performance gains across a variety of tasks (Brown et al., 2020; Devlin et al., 2018). In the meantime, some deep learning models have become excessively large, limiting their applicability in some scenarios. To cope with the issue, Hinton et al. (2015) proposed knowledge distillation (KD), in which knowledge of a large network, called a teacher network, is transferred to a relatively small model, called a student model.

The benefits of KD have been widely witnessed across multiple domains (Romero et al., 2015; Jiao et al., 2020). Recently, it has been observed that KD can be used in both reducing model size and improving model generalization (Tang et al., 2021; Furlanello et al., 2018). Hinton et al. (2015) argue that a distribution, defined by a teacher, holds inter-class relations, commonly referred to as the *dark knowledge*, and that such distribution brings a meaningful supervision to a student. Therefore, a large body of research in KD has viewed a teacher as a source of knowledge and has focused on *finding a meaningful knowledge* to be transferred (Romero et al., 2015; Bulò et al., 2016; Park et al., 2019; Yuan et al., 2020; Kim et al., 2021).

In this work, we focus on *when to distill knowledge of a teacher*. This is a central question to ask, as a model can benefit from the adaptive control of supervision between ground truth and a teacher; When a model is trained to increase the predictive score of a prediction, a one-hot encoded supervision, without incorporating teacher model, sends a direct signal in increasing the score (Müller et al., 2019). In another case, when a model is trained to learn knowledge of a teacher, a teacher's output without fusing a ground truth sends more direct signal in minimizing the knowledge gap between the student and the teacher. However, the question of "when" has not been answered. For this reason, previous works choose to learn from both of the supervisions.

We give an answer to the question from the perspective of model calibration. Model calibration refers to how well a predicted probability of a model reflects the true accuracy. Therefore, a well-calibrated predictive score represents the **likelihood of correctness of a prediction** (Guo et al., 2017). In this light, such score can be viewed as a gauge to detect a miscalibration of a student in training; when a student makes a prediction with a probability mass that is higher than the expected accuracy of the prediction (overconfidence), a student model is trained with only supervision from a

[*] This work was done while Dongkyu was an intern at NVIDIA

[†] Corresponding author

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9793–9803
December 7-11, 2022 ©2022 Association for Computational Linguistics

teacher. In the case of underconfidence, a student is trained with only supervision from ground-truth.

Switching supervision is supported by two widely accepted ideas: 1) the close link between miscalibration and overfitting, and 2) the regularization effect of KD. Guo et al. (2017) empirically find that a model overfits to negative log likelihood (NLL) training, leading to miscalibration, and Mukhoti et al. (2020) further support the claim. Therefore, we utilize the regularization effect held in KD training (Yuan et al., 2020). Aside from the inter-class relations held in knowledge, recent findings suggest that KD is a form of adaptive regularization (Tang et al., 2021; Yuan et al., 2020), where a teacher enforces a student to distribute probability mass on output space more evenly.

Taking all these factors into account, we present a simple, yet novel KD method, called Hard gate Knowledge Distillation (HKD). Given a calibrated teacher model, the teacher gates supervisions between knowledge and observation for each instance/time step, selecting which objective the student should be optimized to. We introduce two levels of hard gates: the token-level and the sentence-level which are instance-specific hard gates computed on the fly during forward propagation. Our work validates the proposed idea on a task in the Natural Language Generation (NLG) domain, as there is an inseparable relation between the quality of an output and model calibration (Kumar and Sarawagi, 2019).

The contributions of the proposed method are as follows:

- To the best of our knowledge, this work is the first attempt to leverage knowledge and observations in KD with a hard gate which is instance-specific.
- Our work introduces a novel view and role of a teacher model in student-teacher framework which improve model generalization and model calibration of a student by a significant margin across multiple datasets.

2 Preliminaries & Related Work

2.1 Knowledge Distillation

The conventional logit-based KD (Hinton et al., 2015) aims to minimize the distance between the probability distribution mapped by a teacher and that of a student, while another objective is to maximize the likelihood of predicting ground truth. Following is the loss of an instance $(\mathbf{x}^i, \mathbf{y}^i) \in \mathcal{X} \times \mathcal{Y}$ at time-step t, where i indicates the index of the sample.[1]

$$\begin{aligned}\mathcal{L}_{kd} = -\sum_{v}^{|V|}(1-\alpha)y^i_{t,v}\log P_\theta(y^i_{t,v}|\mathbf{c}^i_{<t}) \\ + \alpha P_\phi(y^i_{t,v}|\mathbf{c}^i_{<t};\tau)\log P_\theta(y^i_{t,v}|\mathbf{c}^i_{<t};\tau)\end{aligned} \quad (1)$$

V and τ denote a set of vocabularies and a temperature respectively. ϕ and θ denote parameters of a teacher and those of a student. α denotes a balancing parameter which in this work is termed a **gate**, and $\mathbf{c}_{<t}$ is a context at time step t, hence made of input $\mathbf{x}$ and preceding tokens $\mathbf{y}_{<t}$. The gate is set to a value between 0 and 1, which indicates a *soft gate, and it is shared among instances and remains fixed throughout training* (Park et al., 2019; Hinton et al., 2015; Yuan et al., 2020). Therefore, a student model is trained with a soft target $\tilde{y}^i_t$, a result of linear interpolation between a ground truth and a distribution mapped by a teacher.

Numerous studies have attempted to find meaningful knowledge to be distilled. Starting with interclass relations on logit space (Park et al., 2019; Hinton et al., 2015), the scope of knowledge expanded to feature-level (Romero et al., 2015) to encourage a student to maintain similar intermediate representations to those of a teacher. Recent studies find that even a model with an identical model structure to that of a student can suit the role as a teacher; thus it is commonly referred to as Self-Knowledge Distillation (Yuan et al., 2020; Kim et al., 2021; Liu et al., 2021). (Yuan et al., 2020; Tang et al., 2021) argue that the success is brought by KD's close link to label smoothing (Szegedy et al., 2016), with KD holding a regularization effect. In this regard, there have been attempts to explore the importance of the soft gate. PS-KD (Kim et al., 2021) linearly increases the value of the gate in the course of training. Similar to our work, Zhu and Wang (2021) propose a hard gate mechanism in KD; however the work utilizes an iteration-specific hard gate, and the gates only apply to distillation loss of KD.

2.2 Calibration

A model is said to be well-calibrated when the predictive confidence truly reflects true accuracy (Guo et al., 2017).

$$P(\hat{Y} = Y | P(\hat{Y}|X) = p) = p, \forall p \in [0, 1] \quad (2)$$

[1] Loss equations are illustrated in time-step level hereinafter as a natural language generation task can be viewed as a sequence of classification.

Therefore, when a model makes predictions with probability of p, the accuracy of the predictions is expected to be p. The quantity is commonly approximated with Expected Calibration Error and Maximum Calibration Error (Naeini et al., 2015).

There have been continuous efforts in lowering the calibration error of a model, and one of the simplest, yet effective methods is temperature scaling (Guo et al., 2017). Temperature scaling is a parametric post-hoc calibration method, where a single parameter is learned; with model parameters fixed, the single parameter is learned to lower the negative log likelihood on validation dataset. This simple calibration method has been widely appreciated for its ability to improve the reliability of a model (Müller et al., 2019).

3 Approach

In this section, we first discuss a new interpretation of a teacher under KD training and introduce methods that switch supervision between knowledge and observations with an instance-specific hard gate.

3.1 A View on Teacher Model

When a teacher model is well-calibrated, via calibration method such as temperature scaling (Guo et al., 2017), the predictive score of a teacher can be used to *estimate* the true likelihood of correctness. In this light, a teacher can be used to evaluate if a student model makes a miscalibrated prediction, either resulting in underconfidence or overconfidence. Furthermore, given a calibrated teacher, minimizing the knowledge gap provides a meaningful insight which is more than learning the inter-class relations, as the objective *extends to improving calibration* of a student. By minimizing the KL divergence between the two probability distributions, the prediction of a student is expected to reflect the calibrated predictive score.

3.2 Hard Gate

From the novel view of a teacher, our work presents two instance-specific hard gates: the token-level and the sentence-level hard gate.

3.2.1 Token-Level Gate

When a predictive score of a prediction by a student is high compared to an *approximated* likelihood of the correctness of the prediction, a student is supervised to distribute the probability mass to other remaining classes, hence learning to output a smooth distribution. In another case, in which the predictive score is less than the approximation, the student is learned with supervision that increases the probability, learning from a sample drawn from the data distribution.

In every time step, instance-specific hard gates are computed on the fly during forward propagation as follows:

$$a_t^i = \begin{cases} 1, & \text{if } P_\theta(y_{t,j}^i|\mathbf{c}_{<t}^i) > f(y_{t,j}^i, \mathbf{c}_{<t}^i) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$P_\theta(y_{t,j}^i|\mathbf{c}_{<t}^i)$ and $f(y_{t,j}^i, \mathbf{c}_{<t}^i)$ are conditional probability of a ground truth index j mapped by a student model and the true likelihood of $y_{t,j}^i$ occurring in the given context. Since the true likelihood of correctness cannot be obtained, we approximate the quantity with a teacher network with enhanced calibration ability $f(y_{t,j}^i, \mathbf{c}_{<t}^i) \approx P_\phi(y_{t,j}^i|\mathbf{c}_{<t}; \tau)$.

Supervision from Observations ($\alpha = 0$) When the hard gate is computed to be 0, it is an indication of *underfitting and underconfidence* by a student on the instance. The student needs further training so that the likelihood of predicting the target index is increased. Due to the normalizing activation layer, softmax, a direct way of escalating the probability mass on the ground truth index is to minimize the KL divergence with one-hot encoded ground truth (Müller et al., 2019), without incorporating knowledge. That being the case, when the hard gate is set to 0, supervision to a student solely comes from ground truth.

Supervision from Knowledge ($\alpha = 1$) In another case when the gate is set to 1, it is an indication of *overconfidence* evaluated by the approximated quantity mapped by a teacher. Therefore, a student is trained to distribute the probability mass on output space more evenly; the student learns to close the gap between its probability distribution and that of a teacher.

This gating mechanism can be viewed as smoothing of labels, hence presenting a regularization effect. The entropy of supervisions by the proposed method, conventional logit-based KD ($\tilde{\mathbf{y}}_t^i$), and one-hot encoded target (hard target) are as follows:

$$H(P_\phi(Y|\mathbf{c}_{<t}^i; \tau)) \geq H(\tilde{\mathbf{y}}_t^i) \geq H(\mathbf{y}_t^i) \tag{4}$$

where $\tilde{\mathbf{y}}_t^i$ is the soft target that is a linear interpolation of a ground truth and a probability distribution mapped by a teacher. The entropies illustrate how

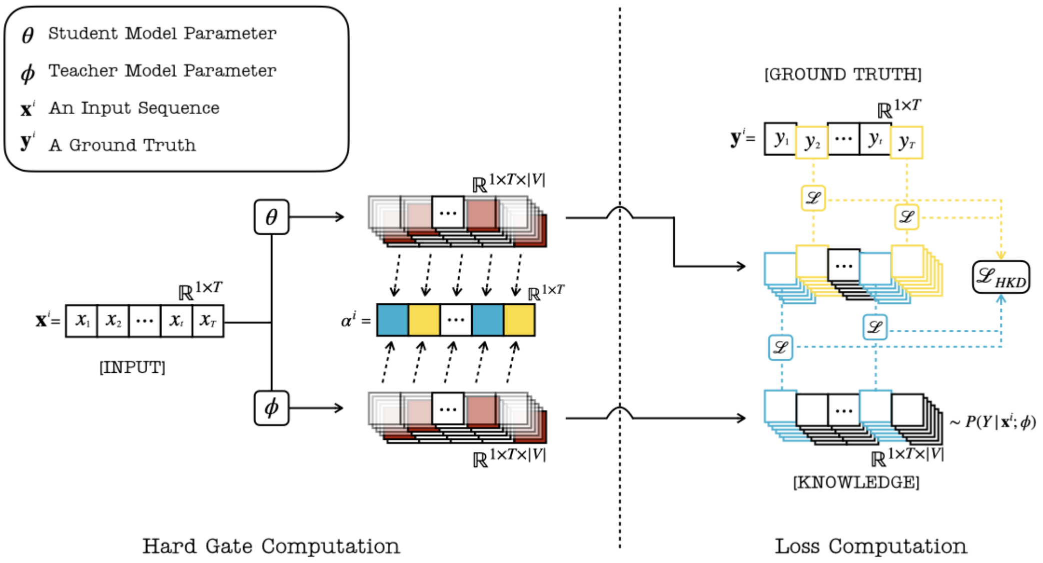


Figure 1: Overview of the proposed token-level hard gate KD. The conditional probability distributions mapped by a student and a teacher are in comparison to compute the instance and time step-specific hard gates. The loss of the instance is computed according to the hard gates.

the proposed method regularizes a student by presenting high entropy supervision. Specifically, the following proposition holds.

Proposition 1 (Opposite Gradient) *When $\alpha = 1$, the sign of the expectation of the gradient by the proposed KD method with respect to logit on "incorrect" classes is guaranteed to be opposite to that of the cross entropy with hard target.*

$$\mathbb{E}_{i \neq j} \frac{\partial \mathcal{L}_{hkd}}{\partial z_i} < 0 \leq \mathbb{E}_{i \neq j} \frac{\partial \mathcal{L}_{ce}}{\partial z_i} \tag{5}$$

The gradient of a sample with respect to a logit z by the cross entropy is as follows[2]:

$$\frac{\partial \mathcal{L}_{ce}}{\partial z_i} = P_\theta(y_i|\mathbf{c}_{<t}) - y_i \tag{6}$$

When $\alpha = 1$, the gradient of the proposed method is

$$\frac{\partial \mathcal{L}_{hkd}}{\partial z_i} = P_\theta(y_i|\mathbf{c}_{<t}) - P_\phi(y_i|\mathbf{c}_{<t}; \tau) \tag{7}$$

Then, it is straightforward to compute the sum of the quantities except the target index j.

$$\sum_{i, i \neq j} \frac{\partial \mathcal{L}_{ce}}{\partial z_i} = 1 - P_\theta(y_j|\mathbf{c}_{<t}) \tag{8}$$

$$\sum_{i, i \neq j} \frac{\partial \mathcal{L}_{hkd}}{\partial z_i} = P_\phi(y_j|\mathbf{c}_{<t}) - P_\theta(y_j|\mathbf{c}_{<t}) \tag{9}$$

As Equation 8 is guaranteed to be greater than or equal to 0, Equation 9 must be smaller than 0, since $P_\phi(y_j|\mathbf{c}_{<t}) < P_\theta(y_j|\mathbf{c}_{<t})$. The Proposition 1 is not guaranteed in conventional logit-based KD, while it holds true within the proposed approach.

The cross entropy with hard target forces a student to decrease the probability mass on the other classes, while the proposed method sends gradients that have opposite direction to that of the cross entropy. In return, the proposed method pushes a student to increase the probability mass on the other output space, regularizing the student.

In addition to the regularization effect, the inter-class relations are given more directly to the student than that of the conventional logit-based KD. The conventional KD shrinks the dark knowledge $P_\phi(y^i_{t,v}|\mathbf{c}^i_{<t}; \tau)$ by α as in Equation 1. This, however, is different in the proposed method as α is set to 1, and hence the amount of dark knowledge remains unchanged.

3.2.2 Sentence-Level Gate

A natural language generation task is a sequence of classification. In this regard, in addition to the token-level gate, we propose to compute hard gates on the sentence-level. In particular, the gates are determined by comparing the sentence probabilities mapped by the two models in KD.

$$\forall_t a^i_t = \begin{cases} 1, & \text{if } P_\theta(\mathbf{y}^i|\mathbf{x}^i) > f(\mathbf{y}^i, \mathbf{x}^i) \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

[2] For notational simplicity, time step is omitted.

where $P(\mathbf{y}^i|\mathbf{x}^i)$ is a sentence probability computed as the product of the conditional probabilities of time steps $\prod P_\theta(y^i_{t,j}|\mathbf{c}^i_{<t})$. As in the token-level gate, the true likelihood of the sentence appearing is approximated with a teacher $f(\mathbf{y}^i, \mathbf{x}^i) \approx \prod_t P_\phi(y^i_{t,j}|\mathbf{c}^i_{<t})$.

A probability of a sentence defined by a language model is a reflection of how likely a model predicts the sentence. If a student model defines a sentence probability that is higher than that of a calibrated teacher, this is a possible sign of overconfidence in the sentence. Therefore, in such case, the student only receives supervision from knowledge. In the opposite case, as in the token-level gate, the student is solely trained with ground truth.

Sentence-level computes hard gates in a more cautious manner than token-level does. A sentence probability is a product of probabilities of words within the sentence; hence a miscalibrated probability of a word can cause much change in the final probability. This aspect is depicted in Figure 3 in which sentence-level gates and token-level gates differ in the ratio of α in the course of training.

3.3 Final Loss

The final loss function is as follows:

$$\begin{aligned}\mathcal{L}_{hkd} = -\sum_{v}^{|V|}(1-\alpha^i_t)y^i_{t,v}\log P(y^i_{t,v}|\mathbf{x}^i;\theta)\\ + \alpha^i_t P_\tau(y^i_{t,v}|\mathbf{x}^i;\phi)\log P(y^i_{t,v}|\mathbf{x}^i;\theta)\end{aligned} \tag{11}$$

The α in both token and sentence-level is computed during the forward propagation. Therefore, the following propositions can be made.

Proposition 2 *When expected α approaches 0, $\mathcal{L}_{hkd}$ reduces to MLE with hard targets. In other case, when the expected value approaches 1, $\mathcal{L}_{hkd}$ reduces to minimizing KL divergence between probability distribution of a student and that of a teacher.*

$$\begin{aligned}\lim_{\mathbb{E}[\alpha]\to 0}\mathcal{L}_{hkd} &= \mathcal{L}_{ce}(P(Y|X;\theta), Y)\\ \lim_{\mathbb{E}[\alpha]\to 1}\mathcal{L}_{hkd} &= \mathcal{L}_{ce}(P(Y|X;\theta), P_\tau(Y|X;\phi))\end{aligned} \tag{12}$$

where $\mathcal{L}_{ce}$ indicates the cross-entropy loss. We empirically observe that the former case is seen in the early stages of training which is depicted in Figure 3. In the other case in which the expected value converges to 1, the loss reduces to solely minimizing the distance of the distributions mapped by the models without any observation from empirical training distribution.

One difference to notice is the temperature scaling in Equation 11. The proposed KD solely applies temperature scaling on the logit of a teacher for the purpose of calibrating the teacher's output. This is a marked difference from the existing logit-based KD, where both a student and a teacher logits are scaled with a pre-defined temperature as in Equation 1. This distinction encourages a student model to mimic a probability distribution of a teacher which contains inter-class relations as well as calibrated predictive scores.

4 Experiment

4.1 Dataset & Experiment Setup

We validate the proposed gating mechanisms on three popular translation tasks: IWSLT14 German to English (DE-EN), IWSLT15 English to Vietnamese (EN-VI), and Multi30K DE-EN[3]. There are two core reasons for conducting experiments on a NLG task. First, our method suits NLG tasks by nature (token and sentence-level). Second, calibration has inseparable relation with NLG, as popular generation schemes, such as top-k, top-p, and beam search, are affected by the calibration ability of a language model. The generation schemes start by assuming that a predictive score represents likelihood of an event (Müller et al., 2019).

All of the experiments are conducted on a single Telsa V100, and both a student and a teacher model follow transformer architecture (Vaswani et al., 2017). The proposed method is tested on self-knowledge distillation environment, considering the efficiency of computation and general applicability[4]. The hyperparameters are identical to the specified configuration in `fairseq` (Ott et al., 2019)[5]. The teacher network is trained with uniform label smoothing (Szegedy et al., 2016) in our environment; nonetheless a teacher trained with the regular cross-entropy training with hard targets is a valid option, which we show in the ablation study. For evaluation, we comprehensively validate previous methods and the proposed methods with popular translation evaluation metrics: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie,

[3] The details of the datasets are reported in Appendix A.

[4] Self-knowledge distillation can be used even when strong teacher does not exist or cannot be obtained (Yuan et al., 2020).

[5] https://github.com/facebookresearch/fairseq/blob/main/examples/translation/README.md

		Calibration		Generalization				
Dataset	**Method**	**ECE (↓)**	**MCE (↓)**	**BLEU**	**METEOR**	**WER (↓)**	**ROUGE-L**	**NIST**
`Multi30K`	Base	14.95	26.01	40.64	73.31	38.76	69.21	7.96
	LS-Uniform	9.17	17.22	42.55	74.78	38.00	70.17	8.08
	LS-Unigram	9.12	17.78	42.52	74.61	37.54	70.30	8.13
	ConfPen	48.21	52.53	43.14	75.03	37.95	70.54	8.12
	Loras	20.27	40.86	41.78	74.31	38.03	69.83	8.12
	TF-KD	21.18	42.87	41.77	74.33	37.55	69.94	8.12
	PS-KD	14.75	26.69	41.95	74.26	38.45	69.75	8.05
	SD	6.87	12.38	43.41	75.41	37.50	70.76	8.15
	Beta	11.71	21.90	41.9	74.24	38.56	69.66	8.03
	HKD-T (Ours)	2.86	6.92	**43.96**	**75.64**	**36.38**	**71.32**	8.26
	HKD-S (Ours)	**2.25**	**3.78**	43.78	75.48	36.58	71.20	**8.27**
`IWSLT15`	Base	14.05	20.38	30.17	59.09	54.02	63.91	7.18
	LS-Uniform	8.53	12.13	30.74	59.7	53.62	64.33	7.24
	LS-Unigram	7.89	11.71	30.8	59.57	53.23	64.34	7.27
	ConfPen	43.94	59.28	31.10	59.65	53.00	64.50	7.31
	Loras	12.41	19.15	30.13	58.98	53.79	63.72	7.21
	TF-KD	13.30	19.29	30.17	58.94	54.01	63.83	7.16
	PS-KD	9.34	14.18	31.21	60.04	52.71	64.71	7.34
	SD	5.01	9.64	30.86	59.54	53.39	64.36	7.27
	Beta	9.57	14.97	30.48	59.29	53.39	64.07	7.26
	HKD-T (Ours)	2.17	5.12	**31.96**	**60.54**	**52.21**	**65.01**	**7.40**
	HKD-S (Ours)	**1.38**	**4.70**	31.85	60.35	52.24	64.94	7.39
`IWSLT14`	Base	12.98	19.29	35.96	64.70	48.17	61.82	8.47
	LS-Uniform	6.43	9.98	36.82	65.30	47.50	62.41	8.60
	LS-Unigram	6.12	9.46	36.97	65.38	47.30	62.57	8.62
	ConfPen	48.19	57.58	37.11	65.55	47.09	62.67	8.66
	Loras	10.54	15.29	36.32	64.93	48.78	61.98	8.43
	TF-KD	12.20	17.60	36.35	64.88	47.84	62.11	8.53
	PS-KD	5.63	8.88	37.49	65.81	46.48	63.07	8.71
	SD	7.82	13.71	37.35	65.76	47.33	62.70	8.63
	Beta	8.63	13.21	36.91	65.39	47.87	62.40	8.55
	HKD-T (Ours)	1.43	3.52	**38.27**	**66.54**	45.74	**63.73**	**8.85**
	HKD-S (Ours)	**1.27**	**3.22**	38.08	66.44	**45.72**	63.65	8.84

Table 1: Scores are reported in percentage by averaging three runs with different random seeds. A bold number indicates the best performance within each corpus tested, and the underlined numbers are the second best performing scores. HKD-T and HKD-S denote the proposed method with the token-level and the sentence-level hard gates respectively.

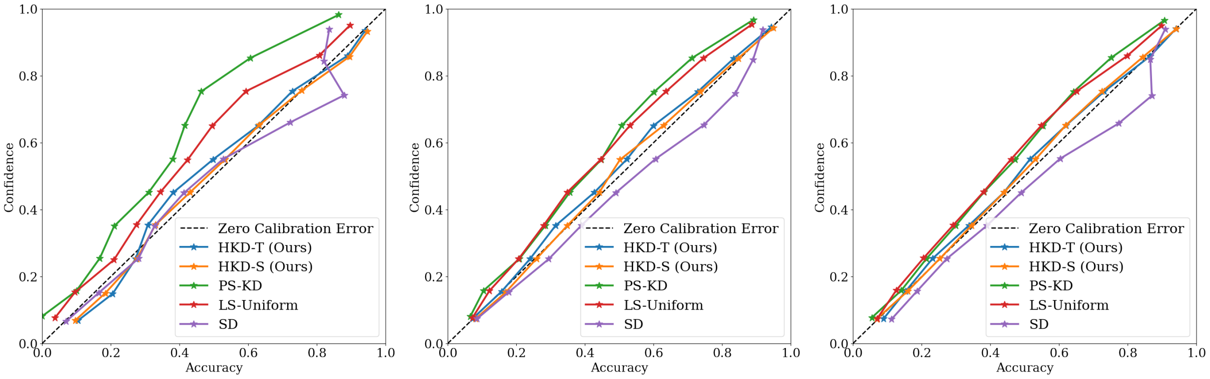
2005), Word Error Rate (WER), ROUGE-L (Lin, 2004), and NIST (Doddington, 2002). For quantifying the level of model calibration, we report Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) (Naeini et al., 2015).

4.2 Baselines

As this paper lies in a branch of KD, though is closely linked to regularization methods, we compare the proposed methods with baselines from both of the domains.

Base Method The base method in this work is the cross-entropy with hard targets (Base).

Regularizers Although Label Smoothing (LS) was first introduced to enhance model performance in (Szegedy et al., 2016), it has been found to help in model calibration as well. The prior label distribution is commonly set with a uniform distribution (LS-Uniform) (Vaswani et al., 2017; Lewis et al., 2020), yet unigram distribution (LS-Unigram) is another valid choice. Similar to label smoothing, ConfPen (Pereyra et al., 2017) prevents a model from outputting a peak distribution by penalizing high confident predictions. Loras (Ghoshal et al., 2021) theoretically find that the generalization error largely depends on prior label distribution; thus, it jointly learns model parameters and prior label distribution.



(a) Multi30K DE-EN

(b) IWSLT15 EN-VI

(c) IWSLT14 DE-EN

Figure 2: Reliability diagrams of the five methods on the three corpora tested. The dashed line indicates zero calibration error.

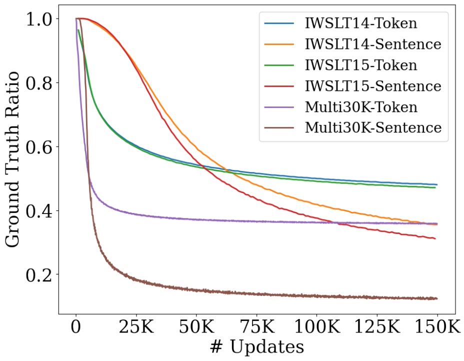


Figure 3: This figure plots the change of ground truth supervision ratio in the course of training.

KD-Methods Yuan et al. (2020) empirically find that even a weak teacher can improve a student. Accordingly, the authors present TF-KD where a pre-trained teacher with identical model structure to that of a student is utilized in KD. PS-KD (Kim et al., 2021) is similar, but the core difference is that a teacher is the previous checkpoint of a student in training. Zhang and Sabuncu (2020) introduce instance-specific label smoothing methods, SD and Beta, which make use of self-knowledge and encourage diversity in predictions.

4.3 Experimental Result

The automatic evaluation results are reported in Table 1. Both of the proposed methods achieve noticeable gains compared to the Base method and the strong baseline methods. The improvements are seen across every metric and corpus tested. Our methods excel not just in n-gram matching (BLEU) and harmonic mean of unigram precision and recall (METEOR), but also in having the longest common subsequence with references (ROUGE-L) and outputting informative n-grams (NIST). Moreover, as clearly depicted with low WER, the outputs of our systems require the least amount of modification to be converted into reference sequences. Without adding any additional learnable parameter compared to the base method, our token-level HKD illustrates superior performance to that of the base method, absolute gain of 3.32 BLEU score and relative improvement of 8.16% on Multi30K. Sentence-level hard gate method also illustrates competitive results to those of the token-level hard gate. On every corpus and metric tested, both token and sentence-level gates outperform the strong baselines by a large margin.

Figure 3 depicts the change of ground truth supervision ratio in the course of training. In each corpus, the ratio of ground truth decays throughout the training. In the early stage of training, most of the supervisions come from ground truth as a student is underfitted, leading to the low expected α. The ratio decreases as the student model learns to map the task distribution, increase in the ratio of supervision from knowledge. The ratios converge at a certain point in each dataset, illustrating how the proposed methods hold self-regulating property in switching the supervisions. A noteworthy point is the *correlation of the ratio and the corpus size*. Under Multi30K training, which is the smallest in terms of training dataset size, the student receives majority of supervision from the teacher which the ratio is higher than 0.6 with token-level and 0.8 with sentence-level hard gate. This empirically shows that the proposed systems, in an environment with the risk of overconfidence and overfitting demonstrate strong regularization effect in training. The methods enforce a student model

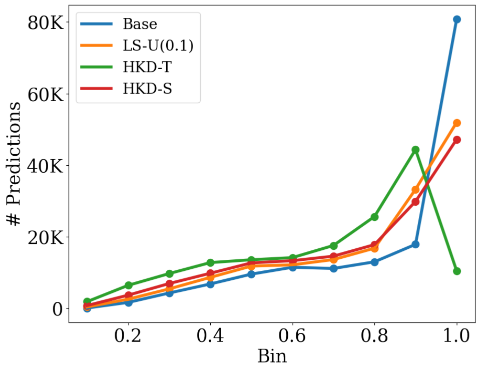


Figure 4: This figure shows the confidence histogram of the three approaches (Base, LS-Uniform, and HKD) on IWSLT14. The predictions are binned to 10 bins based on their confidence scores.

to learn from knowledge to avoid possible degradation in model generalization and calibration.

4.3.1 Model Calibration

Following (Müller et al., 2019), our work evaluates model calibration by formulating the generation process as the next token prediction task. Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) of the models on the corpora tested are reported in Table 1.

It is clearly demonstrated that the proposed methods, especially with sentence-level hard gate, lead to a calibrated student. The ECE and MCE scores of the base method are high as modern neural networks are found to be overconfident (Guo et al., 2017). The amount of error is mitigated to some extent with the introduction of regularizers and KD methods. For instance, LS with uniform distribution lowers the ECE and MCE score to 6.43 and 9.98 on IWSLT14 corpus. The errors are reduced to around half of those of the base method. Nevertheless, the most noticeable gain in calibration is seen across the proposed methods. HKD-S achieves 1.27 in ECE and 3.22 in MCE, illustrating remarkable improvement in model calibration. The absolute improvement compared to the base method is approximately 11.7 score in ECE and 16 in MCE on IWSLT14 dataset. The proposed methods enhance the model calibration of a student by a large margin, with such gain observed across the corpora. Reliability diagrams in Figure 2 further support the claim. Despite the baselines improving model calibration, the methods tend to show either underconfident or overconfident predictions. LS-Uniform and PS-KD consistently make overconfident predictions, while SD suffers from underconfidence. On the other hand, the reliability diagrams of our work display calibrated results that mainly conform with the low ECE and MCE scores in Table 1.

Method	BLEU	ECE
Base	35.96	12.98
ϕ trained with LS	38.27	1.43
ϕ trained with CE (τ=1.0)	37.79	5.87
ϕ trained with CE (τ=1.5)	38.57	1.87

Table 2: Ablation study on the proposed approach with different temperatures and knowledge. CE denotes training a model with the cross-entropy with hard targets.

Furthermore, our methods make predictions more evenly distributed as illustrated in Figure 4. HKD-T and HKD-S do not make predictions solely with low confidence; the number of predictions with confidence scores between 0.0 to 0.9 by our methods outnumbers those of the other methods, demonstrating an ability to make decision with *diverse confidence*.

4.4 Ablation Study

In order to further validate the proposed method, we conduct an ablation study with different teachers and varying temperature values, and the results are shown in Table 2. In a case where a teacher is trained with the cross-entropy with hard targets, both BLEU score and ECE score are enhanced compared to those of Base method. However, with a proper temperature control (enhanced calibration), both BLEU and ECE improve significantly. The BLUE score is higher than that of a model trained with LS-Uniform, and the ECE score is also competitive. This empirically validates that with a proper control of temperature, the proposed KD systems are compatible with a *wide choice of teacher*.

5 Conclusion

In this study, we present hard gate knowledge distillation, a mechanism that switches supervision between knowledge and ground truth at either the sequence-level or the token-level. This originates from the novel view and role of a teacher in KD. The proposed method is simple yet effective in improving model generalization and calibration, achieving superior performances compared to those of the strong baselines.

Limitations

As in previous KD methods, the proposed approaches utilize a teacher model, hence inevitably causing the computation cost to increase. In addition, as two forward passes are needed, one by a teacher and the other by a student, the training time is longer than non-KD training methods. Lastly, the proposed idea does not suit a natural language understanding task due to the introduction of the token-level and sentence-level gate.

Ethical Consideration

The proposed idea is a student-teacher framework; hence, a teacher model can greatly affect a student model. If a teacher model is trained with a dataset with biased information or misinformation, the student is likely to learn such features while minimizing the knowledge gap. One can mitigate the concern to some extent if fact checking system or biased detection system is employed. This is not the fundamental solution to the problem that KD training faces, yet the level of danger is expected to be mitigated to some extent.

Acknowledgement

Research on this paper was supported by Hong Kong Research Grants Council (Grant No. 16204920).

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. 2016. Dropout distillation. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 99–107, New York, New York, USA. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1607–1616. PMLR.

Asish Ghoshal, Xilun Chen, Sonal Gupta, Luke Zettlemoyer, and Yashar Mehdad. 2021. Learning better structured representations using low-rank adaptive label smoothing. In *International Conference on Learning Representations*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. 2021. Self-knowledge distillation with progressive refinement of targets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6567–6576.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *CoRR*, abs/1903.00802.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Sheng Shen, and Mirella Lapata. 2021. Noisy self-knowledge distillation for text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–703, Online. Association for Computational Linguistics.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. 2020. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, volume 33, pages 15288–15299. Curran Associates, Inc.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Association for the Advancement of Artificial Intelligence*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3976.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Jiaxi Tang, Rakesh Shivanna, Zhe Zhao, Dong Lin, Anima Singh, Ed H. Chi, and Sagar Jain. 2021. Understanding and improving knowledge distillation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. 2020. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911.

Zhilu Zhang and Mert R. Sabuncu. 2020. Self-distillation as instance-specific label smoothing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yichen Zhu and Yi Wang. 2021. Student customized knowledge distillation: Bridging the gap between student and teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5057–5066.

A Dataset & Implementation Details

Multi30K dataset is the smallest in size among the corpora tested, 28K sentence pairs for training, 1K for validation, and 1K for testing. IWSLT15 EN-VI comprises 133K pairs in the training set, 1.5K in the validation, and 1.3K in the testing set. Lastly, IWSLT14 DE-EN corpus contains around 170K, 7K, and 7K pairs of sentences for training, validation, and testing dataset respectively. Words are processed into sequence of subword units with `subword-nmt` (Sennrich et al., 2016)[6].

The structure of models, both the student and the teacher, in all of the experiments follow transformer architecture (Vaswani et al., 2017). Specifically, both the encoder and the decoder are composed of 6 transformer layers with 4 attention heads, and the hidden dimension size is set to 512. The dropout

[6] `https://github.com/rsennrich/subword-nmt`

probability is set to 0.3, and the maximum number of tokens within a batch is 4096. For the temperature in our experiments, we have tested {0.8, 1.0, 1.5, 2.0, 2.5}, and find that 1.0 works the best with a self-teacher trained with label smoothing. For a self-teacher trained with hard targets, 1.5 for temperature value illustrates the best performance, as the temperature smoothed the output of the teacher. We report random seeds used in our work for reproducibility, the seeds being `{0000, 3333, 5555}`.