GOVERN: Gradient Orientation Vote Ensemble for Multi-Teacher Reinforced Distillation

Wenjie Zhou[1*] **Zhenxin Ding**[1] **Xiaodong Zhang**[1]
Haibo Shi[1] **Junfeng Wang**[1] **Dawei Yin**[1]
[1]Baidu Inc., Beijing, China
wjzhou013@pku.edu.cn zhenxinding@gmail.com zxdcs@pku.edu.cn
haiboshi@outlook.com wangjunfeng@baidu.com yindawei@acm.org

Abstract

Pre-trained language models have become an integral component of question-answering systems, achieving remarkable performance. However, for practical deployment, it is crucial to perform knowledge distillation to maintain high performance while operating under computational constraints. In this paper, we address a key question: given the importance of unsupervised distillation for student model performance, how can knowledge from multiple teacher models be effectively ensemble during this stage without the guidance of labels? We propose a novel algorithm, GOVERN, to tackle this issue. GOVERN has demonstrated significant improvements in both offline and online experiments, enabling the student model to achieve results comparable to that of teacher ensembles. Our experiments show that GOVERN remarkably requires a mere 1% of the ensemble method's inference budget to achieve 99.5% of performance. The proposed algorithm has been successfully deployed in a real-world commercial question-answering system, demonstrating its real-world applicability.

1 Introduction

Traditional search engine aims at deliver relevant web pages to satisfy users' question, while sometimes the single paragraph that answer the question might buried deep in a web page, it asks for a web-based Open domain Question Answering (OpenQA) system to find that needle-in-a-haystack info (e.g. Qu et al., 2021; Zhang et al., 2023).

BERT-liked pre-trained language models have achieved state-of-the-art performance in OpenQA (e.g. Zhang et al., 2021). However, due to computational costs, the direct application of these models in real-time search engines like Google is currently unfeasible. For instance, the top-performing models on the Natural Question dataset, R2-D2 (Fajcik et al., 2021) and UnitedQA (Cheng et al., 2021) come with 1.29B and 2.09B model parameters. Further complicating matters is the fact that ensemble methods, which can enhance performance, entail even greater computational overheads.

The distillation of knowledge from multiple teachers has emerged as a powerful technique for improving the performance and generalization of DNN while reducing the computational cost. This two-stage training paradigm, which training large model with limited labeled data as teacher and then using it to generate soft label on large amount unlabeled data for the purpose of student training, was first proposed by Hinton et al. (2015). Since the knowledge from single teacher may be biased and inaccurate, ensemble distillation from multiple teachers was considered by previous works to achieve more promising performance (e.g. You et al., 2017; Fukuda et al., 2017a).

Several dynamic distillation methods were proposed to solve the problem that different teacher is good at different sample and low-quality teachers may mislead the student. e.g. Yuan et al. (2021) proposed a novel RL-based approach to dynamically assigns weights among teachers, Cai et al. (2022) ensembles multi-teacher logits supervised by human-annotated labels in an iterative way. But these dynamic teacher selection methods need supervision signal as guidance, that means they can not apply to unsupervised distillation which is the most important stage in distillation (Su et al., 2021).

In this paper, we propose **Gradient Orientation Vote Ensemble Reinforced distillatioN** (**GOVERN**) to do sample-wise dynamic teacher selection without the need of label guidance.

Our main contributions are summarized as follows:

- We propose GOVERN to do sample-wise dynamic teacher selection without the need of label guidance. We also give a proof that GOVERN can perform better than mean ensemble. To the best of our knowledge, GOVERN is

*Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1650–1658
November 12-16, 2024

the first method which can find sample-wise high-quality teachers without label guidance.

- We propose a novel distillation framework for industrial applications that integrates the GOVERN method into both unsupervised and supervised distillation stages. This framework enhances the performance of student neural networks, enabling them to achieve results comparable to those of ensemble methods. The potential benefits of this approach make it a valuable contribution to industrial OpenQA systems.

- Extensive experiments show that GOVERN is benefit in both distillation stage and can boost the real-world question answering system.

2 Answer Selection Task

In a web-based Open domain Question Answering (OpenQA) system, the primary objective is to select the relevant paragraphs $A_q = a_{i\,i=1}^{N} \subset P_q$ which can solve the custom's question $q \in Q$, where P_q is a collection of paragraphs obtained in web pages retrieved by search engine. A classic framework of this system is made up of two-stage modules including retriever and ranker, where both modules can be distilled down to a task of classifying the relevance between a question and an answer. Our work focus on improving the performance of classification model with the limit of model size.

The classification model assesses the relevance of a paragraph, denoted as p, to a specific question, denoted as q, by calculating the relevance score, $f(q, p; \theta)$. This scoring function, f, which is parameterized by θ, symbolizes the degree of relevance between the question q and the paragraph p. In practical application, a score threshold is established for the purpose of classification.

During training, the classification model is optimized by minimizing the loss over training data:

$$\min_{\theta} \sum_{q \in Q} \sum_{p \in P_q} l(y_p^q, f(q, p; \theta)) \tag{1}$$

where l is the loss function such as cross-entropy loss, margin loss or MSE loss, and y_p^q is the relevance label of q-p pair.

3 Methodology

We use multiple teachers ensemble distillation as the method to improving the performance of online model with the constriction of computational cost. Within a frequently employed Knowledge Distillation (KD) framework, a large teacher model, denoted as T, is meticulously pretrained or fine-tuned well ahead of time. The knowledge contained within the teacher model is subsequently transferred to a smaller student model, denoted as S, by minimizing the disparity between the two. This process can be mathematically formulated:

$$\min_{\theta} \sum_{x} l(f^S(x; \theta), f^T(x; \Theta)) \tag{2}$$

where x embodies the input sample, while $f^S(\cdot)$ and $f^T(\cdot)$ denote the scoring function of the teacher and student models respectively. Additionally, $L(\cdot)$ serves as a loss function that calculates the variation between the behaviors of the two models.

Specifically, we first utilize unsupervised distillation on a vast amount of task-specific, unlabeled data, followed by supervised distillation on the labeled data. The procedures of the distillation can be viewed in Figure 1.

3.1 Unsupervised Distillation

Unsupervised distillation, performed on a substantial amount of task-specific and unlabeled data, is vital for enhancing the performance of the student network. However, due to the absence of supervised signals, the prevalent unsupervised ensemble distillation method resorts to mean-ensemble to amalgamate the abilities of multiple models (You et al., 2017). Other studies have employed a weighted approach whereby individual teacher models are assigned varying weights to accentuate the contribution of higher performing models to knowledge transfer (e.g. Fukuda et al., 2017b; Kwon et al., 2020; Du et al., 2020; Liu et al., 2020). Methods to determine these weighting coefficients encompass weighting based on experience, calculating the weights based on logistic regression model, latent factor or multi-objective optimization in the gradient space.

While these weighting methods do account for the performance differences among various teachers, they employ a uniform weighting coefficient for all samples during the distillation process. This approach neglects the varying emphasis on each teacher's abilities and their respective confidence levels regarding different samples.

Here, we propose a novel unsupervised voting method called **Gradient Orientation Vote Ensemble Reinforced distillatioN** (**GOVERN**), which does not rely on any human-annotated signals and

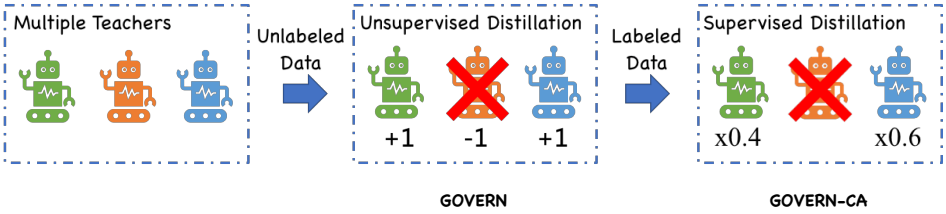

Figure 1: Procedures of Gradient Orientation Vote Ensemble Reinforced Distillation

dynamically assigns different teachers to different samples. In the following, we will introduce the implementation of this unsupervised distillation method and then mathematically prove its superiority over the mean-ensemble method.

It is noted that previous works like UniKD(Wu et al. (2022)) and wVID(Iliopoulos et al. (2022)) have explored the dynamic assignment of weights. But these methods are used to to evaluate the significance of unlabeled examples, rather than assessing the importance of teachers. These methods could be synergistically integrated with the GOVERN framework, as they enhance unsupervised distillation from distinct perspectives.

3.1.1 GOVERN

In unsupervised distillation using mean-ensemble, for a sample, the distilled-model calculates $logit_0$, and N teacher-models calculate $logit_i$ respectively $(1 <= i <= N)$. The distillation loss is:

$$Dist(logit_0, Mean(logit_1, ..., logit_N)) \quad (3)$$

where $Dist$ is a distance metric function that can be selected from MSE, cross-entropy, etc.

We take each teacher's gradient descent orientation into consideration while doing ensemble. Specifically, when $logit_i > logit_0$, the gradient of $Dist(logit_0, logit_i)$ calculated is greater than 0, otherwise it is less than 0, so the gradient descent orientation is noted as:

$$\begin{aligned} Grad_i &= SIGN(gradient(logit_0, logit_i)) \\ &= \begin{cases} 1 & logit_i > logit_0 \\ 0 & logit_i = logit_0 \\ -1 & logit_i < logit_0 \end{cases} \end{aligned} \quad (4)$$

The voted result is calculated as:

$$\chi(sample) = \begin{cases} 1 & \Sigma_{i=1}^{N} Grad_i > 0 \\ 0 & \Sigma_{i=1}^{N} Grad_i = 0 \\ -1 & \Sigma_{i=1}^{N} Grad_i < 0 \end{cases} \quad (5)$$

Each teacher is considerate as a voter in this way, then the loss for unsupervised distillation is represented as below:

$$W_i = \begin{cases} 1 & \chi * Grad_i \geq 0 \\ 0 & \chi * Grad_i < 0 \end{cases} \quad (6)$$

$$\mathcal{L}_{UD} = MSE(logit_0, \frac{\sum_{i=1}^{N} W_i logit_i}{\sum_{i=1}^{N} W_i}) \quad (7)$$

that means, we restrict our approach to guiding the student model's training under the current sample solely by utilizing the majority of teacher models with consistent gradient orientations.

In Appendix A, we give a prove that the sample-wise dynamic weighting ensemble algorithm GOVERN is better than mean-ensemble.

3.2 Supervised Distillation: GOVERN-CA

Inspired by Confidence-Aware Multi-teacher Knowledge Distillation (CA-MKD) proposed by Zhang et al. (2022), we further develop GOVERN algorithm with the help of human label. On each training sample, we select the teachers which share the same gradient descent orientation with the human label. Furthermore, we assign weights among these selected teachers to reflect their sample-wise confidence by calculating the cross entropy loss between the prediction of teachers and human label:

$$y(sample) = \begin{cases} 1, & if \quad positive \\ -1, & if \quad negative \end{cases} \quad (8)$$

$$W_i = \begin{cases} 1 & y * Grad_i > 0 \\ 0 & y * Grad_i \leq 0 \end{cases} \quad (9)$$

$$\omega_i = \frac{W_i}{\Sigma_j W_j}(1 - \frac{exp(L_{CE}^i)}{\Sigma_j W_j exp(L_{CE}^j)}) \quad (10)$$

where L_{CE}^i denotes the cross entropy loss between the prediction of i-th teacher and human label, $Grad_i$ is defined in (5). The loss for supervised distillation is aggregated with calculated weights:

$$\mathcal{L}_{SD} = MSE(logit_0, \Sigma_{i=1} \omega_i logit_i) \quad (11)$$

Thereby, we only select teacher with the correct gradient descent orientation. Besides, the teacher

whose prediction closely align with the ground-truth labels is assigned a greater weight ω_i. This weighting is attributed to the model's substantial confidence in making accurate judgments, thereby providing correct guidance.

Dataset	#Question	#Question-Paragraph Pair
unlabeled data	3,126,132	100M
train data	190,211	2,472,749
test data	3,301	93,446

Table 1: Dataset Statistic

4 Experiments and Results

4.1 Dataset

The questions and relevant web-pages we use are collected from a commercial search engine, the objective is to select a paragraph which can answer the question from the web-pages. We set question-paragraph pairs as samples need to be classified. Hundred millions of unlabeled pairs are collected for unsupervised distillation, and we obtained millions of labeled pairs which are used for teacher's fine-tune through crowd-sourcing annotators. The statistic of dataset is summarized in Table 1.

4.2 Experiment Details

Teacher Architecture In order to obtain multiple models with different structure and ability, we use the series of pretrained models ERNIE-2.0 (Sun et al., 2020) with different layer and fine-tune them on different samplings of the total labeled data. The specific structural parameters for each teacher model can be found in Table 2. Each model has been trained using a sample of 90% of the total data for training purposes.

Student Architecture Considering the computing resources and time consuming, we use the 12-layer transformer structure for online deployment.

In the training procedure, we use the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. For all teacher models, we set the learning rate as 2e-5, the batch size as 64, and the warm-up step as 1000. The maximum length of input text is set as 384 and cross-entropy is used as loss function. In the distillation stage, we set the warm-up step as 1000, the learning rate as 2e-5 and the batch size as 64. The maximum length of input text is set as 384 and MSE is used as loss function. The best checkpoint is picked according to the performance on dev-set.

4.3 Evaluation Metrics

The metrics we used for experimental evaluation are introduced as below.

Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. $P = T_p/(T_p + F_p), R = T_p/(T_p + F_n)$,where T_p, F_p and F_n represent for the number of true positives, false positives and false negatives.

Different threshold of a classifier leads to different Precision-Recall, follow the need of online system, we take recall value where precision equals to 90% as evaluation metrics.

q R@P=90% This metric only takes the paragraph with highest predicted score among all candidates under given question into consideration. A question is noted as T_p if the score of selected answer is higher than threshold and the label is positive, while F_p means the score of selected answer is higher than threshold but the label is negative. If the score of selected answer is lower than threshold but it does exist a positive answer for this question, we note it as F_n. This question granularity metric follows the behavior of web-based OpenQA system since system only displays the best answer was found, so it can best imitate model's performance in online system.

qp R@P=90% This metric takes every qp-pair sample into consideration so it can reflect model's general ability to find answers.

We also conduct a comparison called Good or Same or Bad (GSB) evaluation between two systems by inviting professional annotators to estimate which system produced a greater answer for each given question (Zhao et al., 2011). The gain of a new system can be formulated as:

$$\Delta_{GSB} = \frac{\#Good - \#Bad}{\#Good + \#Same + \#Bad} \quad (12)$$

where #Good (or #Bad) denotes the number of questions that the new (or base) system provides better answer and #Same denotes the number of questions that answer are equal in quality.

r(query_change) The query change ratio, defined as the proportion of sessions where users initiate a subsequent search following their initial query, serves as an online user behavior metric. This study reports only the difference in the query change ratio between the experimental and baseline methods, withholding absolute values.

Lower query change ratio reflects better performance as users are satisfy with the initial response, obviating the necessity for further queries.

Model	Architecture				Results	
	n_{params}	n_{layers}	d_{model}	n_{heads}	q R@P=90%	qp R@P=90%
Teacher1-125M	125M	12	768	12	79.51%	70.52%
Teacher2-350M	350M	24	1024	16	81.79%	73.92%
Teacher3-1.5B	1.5B	48	1536	24	82.55%	73.09%
Teacher4-10B	10B	48	4096	64	83.06%	73.31%
Ensemble Model						
Mean Ensemble	-	-	-	-	84.16%	76.71%
Logistic Regression Weighted Ensemble	-	-	-	-	83.44%	76.91%
Distilled Model						
Mean Ensemble Distillation on unlabeled data	125M	12	768	12	82.04% (0.07)	74.63% (0.12)
LR Ensemble Distillation on unlabeled data	125M	12	768	12	81.98% (0.11)	75.24% (0.12)
GOVERN on unlabeled data	125M	12	768	12	**83.65%** (0.08)	**76.02%** (0.14)
+ CA-MKD on labeled data	125M	12	768	12	82.68% (0.03)	75.67% (0.05)
+ GOVERN-CA on labeled data	125M	12	768	12	**83.69%** (0.06)	**76.43%** (0.09)

Table 2: Results of offline experiments. Metrics denoted in **bold** represent the best results in the unsupervised distillation phase, while **underscored and bolded** denote the best results in the supervised distillation phase. All distilled results are average taken over 5 random seeds with standard deviation in parenthesis.

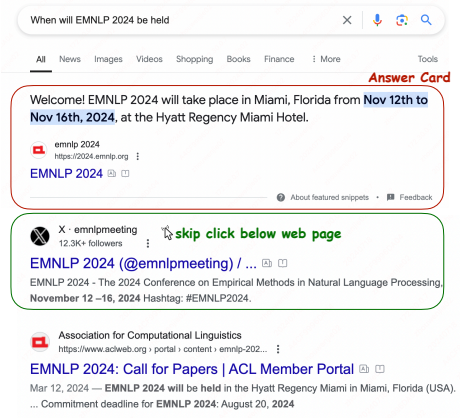

Figure 2: The Answer Card is retrieve by the question answering system. Web pages below are not display in answer card format.

r(skip_click) The skip click ratio, quantified as the proportion of instances where users click on web pages below the answer card (figure 2), indicates potential dissatisfaction with the answer provided. Due to confidentiality constraints, we report only the differential in skip click ratios between the experimental and baseline methods.

4.4 Main Results

The main results of distillation methods comparison are shown in Table 2, we also display the results of teachers and ensemble methods. The methods used in the offline comparison experiments include:

Mean Ensemble We simply average the output of all teachers as the final predict score.

Logistic Regression Weighted Ensemble We trained a logistic regression model based on a dev-set to determine the weighting coefficients, and use these to obtain the weighted-sum of scores.

MED(Mean Ensemble Distillation) The predict score produced by Mean Ensemble Teachers is used as the optimizing object of student.

LRED(LR Ensemble Distillation) This variant uses Logistic Regression Weighted Ensemble Teachers for distillation instead of Mean Ensemble.

CA-MKD This an algorithm proposed by Zhang et al. (2022) which adaptive assigns sample-wise reliability for each teacher prediction with the help of ground-truth labels, with those teacher predictions close to one-hot labels assigned large weights.

It is noted that the 125M distilled model outperforms the 10B teacher model. This could be attributed to the limited size of the training set, which comprises only 19K distinct questions and 2.5M labeled question-pair examples. Such a dataset may not be sufficiently large to leverage the full potential of the larger model. Additionally, an increase in the performance of the teacher models was noted throughout a finetuning epoch, suggesting that these models are underfitted.

4.5 Online Experiment

To investigate the effectiveness of our proposed method in the real production environment, we deploy the proposed model in a commercial search engine, and conduct online experiments for com-

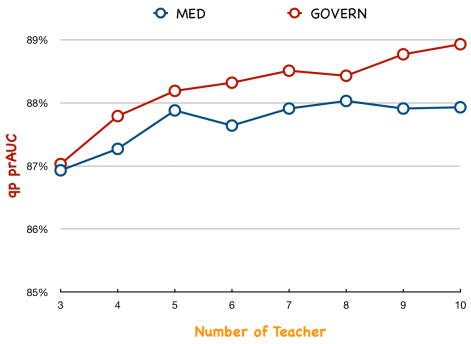

Figure 3: The effect of the number of teachers.

parison of MED and GOVERN.

	Random	Tail
Δ_{GSB}	+4.5%	+7.75%
$G:S:B$	27: 364: 9	39: 353: 8
Δ_{query_change}	-0.68%	-1.03%
Δ_{skip_click}	-3.46%	-4.76%

Table 3: Results of online experiments.

In contrast to random questions, tail questions are defined as those with a search frequency of less than 10 times per week. Given that heterogeneous search questions adhere to long-tail distributions, these tail questions constitute a significant portion of the questions processed by the search engine. It is evident that the proposed method consistently enhances the performance of the online QA system.

4.6 Ablation Study

Due to computational resource limitations, our ablation study utilized a 12-layer transformer as the teacher model and a 4-layer transformer as the student model. We divided the training data into ten folds, training each of the ten distinct teacher models on nine folds. The distillation process involved fifty million unlabeled samples, with the training epoch set to one.

The metric we report in this section is **qp prAUC**. This metric computes the area under the precision-recall curve where precision-recall is computed based on every qp-pair. It gives an overall measurement of classification ability.

Number of Teachers The impact of varying the number of teachers is illustrated in Figure 3. Experimental results indicate that the GOVERN algorithm consistently improves as the number of teachers increases. In contrast, mean-ensemble methods reach a performance plateau relatively quickly.

Effect of Single Teacher We further investigate the impact of varying the performance of a single teacher, with results presented in Table 4. The findings suggest that the GOVERN algorithm has the capacity to effectively select high-performing teachers, while simultaneously disregarding the noise generated by less effective ones.

	qp prAUC
GOVERN with 5-teachers	88.19%
replace one teacher with 10B model	89.03%
replace one teacher with 4-layer model	88.11%

Table 4: Effect of Single Teacher.

5 Related Work

Following the seminal work of Hinton et al. (2015), several studies have sought to develop advanced ensemble algorithms for distillation. We categorize these works into two groups based on their dependency on ground-truth labels.

Unsupervised Ensemble Distillation There are a few works focused on the ensemble method on unsupervised data (Li and Wang, 2019; Sui et al., 2020), these works simply use the average output of multiple teachers as the distillation signal. Recently, Wu et al. (2022) and Iliopoulos et al. (2022) made efforts on distillation with unlabeled examples, but these studies primarily concentrate on dynamically assigning weight to unlabeled data. These approaches do not address the issue of teachers specializing in varying sample distributions.

Supervised Ensemble Distillation The idea of dynamic knowledge distillation with the help of ground-truth label was first explored by Du et al. (2020) and Li et al. (2021). Yuan et al. (2021) proposed a novel RL-based approach, which dynamically assigns weights to teacher models at instance level. Cai et al. (2022) proposed algorithm ensembles multi-teacher logits supervised by human-annotated labels in an iterative way. Zhang et al. (2022) introduced confidence-aware mechanism on both predictions and intermediate features for multi-teacher knowledge distillation.

6 Conclusion

In this paper, we present a novel algorithm, GOVERN, which dynamically selects teachers based on their gradient descent orientation. It does not require ground-truth labels, making it suitable for unsupervised distillation stages. Additionally, it

can be integrated with existing supervised ensemble methods. The effectiveness of our method is affirmed through extensive experimentation.

Limitations

The GOVERN algorithm does not currently account for the varying performance levels of teachers. This could be a shortcoming as it may be beneficial to assign a higher weight to more competent teachers, even if they share the same gradient descent orientation as other selected teachers.

As mentioned in section 3.1, existing dynamic methods are typically used to assign significance to samples, allowing GOVERN to integrate with them. We leave such integration as future work.

Theoretically, GOVERN is a general method that can be applied to other classification tasks. We conducted experiments specifically on the QA task because our team is responsible for the question-answering function in a search engine. We encourage readers to explore its application in different use cases.

References

Lianshang Cai, Linhao Zhang, Dehong Ma, Jun Fan, Daiting Shi, Yi Wu, Zhicong Cheng, Simiu Gu, and Dawei Yin. 2022. PILE: Pairwise iterative logits ensemble for multi-teacher labeled distillation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 587–595, Abu Dhabi, UAE. Association for Computational Linguistics.

Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2021. UnitedQA: A hybrid approach for open domain question answering. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3080–3090, Online. Association for Computational Linguistics.

Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang. 2020. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. In Neural Information Processing Systems.

Martin Fajcik, Martin Docekal, Karel Ondrej, and Pavel Smrz. 2021. R2-D2: A modular baseline for open-domain question answering. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 854–870, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017a. Efficient knowledge distillation from an ensemble of teachers. In Interspeech.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017b. Efficient knowledge distillation from an ensemble of teachers. In Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017, pages 3697–3701. ISCA.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Fotis Iliopoulos, Vasilis Kontonis, Cenk Baykal, Gaurav Menghani, Khoa Trinh, and Erik Vee. 2022. Weighted distillation with unlabeled examples. In NeurIPS.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Kisoo Kwon, Hwidong Na, Hoshik Lee, and Nam Soo Kim. 2020. Adaptive knowledge distillation based on entropy. In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020, pages 7409–7413. IEEE.

Daliang Li and Junpu Wang. 2019. Fedmd: Heterogenous federated learning via model distillation. CoRR, abs/1910.03581.

Lei Li, Yankai Lin, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. 2021. Dynamic knowledge distillation for pre-trained language models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 379–389, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuang Liu, Wei Zhang, and Jun Wang. 2020. Adaptive multi-teacher multi-level knowledge distillation. Neurocomputing, 415:106–113.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5835–5847, Online. Association for Computational Linguistics.

Álvaro Romaniega Sancho. 2022. On the probability of the condorcet jury theorem or the miracle of aggregation. Math. Soc. Sci., 119:41–55.

Weiyue Su, Xuyi Chen, Shikun Feng, Jiaxiang Liu, Weixin Liu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-tiny : A progressive distillation framework for pretrained transformer compression. CoRR, abs/2106.02241.

Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuantao Xie, and Weijian Sun. 2020. FedED: Federated learning via ensemble distillation for medical relation extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2118–2128, Online. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 8968–8975.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. Unified and effective ensemble knowledge distillation. CoRR, abs/2204.00548.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. 2021. Reinforced multi-teacher selection for knowledge distillation. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 14284–14291. AAAI Press.

Hailin Zhang, Defang Chen, and Can Wang. 2022. Confidence-aware multi-teacher knowledge distillation. In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022, pages 4498–4502. IEEE.

Hang Zhang, Yeyun Gong, Yelong Shen, Weisheng Li, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Poolingformer: Long document modeling with pooling attention. In International Conference on Machine Learning, pages 12437–12446. PMLR.

Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023. A survey for efficient open domain question answering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14447–14465, Toronto, Canada. Association for Computational Linguistics.

Shiqi Zhao, Haifeng Wang, Chao Li, Ting Liu, and Yi Guan. 2011. Automatically generating questions from queries for community-based question answering. In Proceedings of 5th International Joint Conference on Natural Language Processing, pages 929–937, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

A Appendix

In this section, we mathematically prove that the sample-wise dynamic weighting ensemble algorithm GOVERN is better than mean-ensemble. We only make the proof on positive samples, as for the negative samples, the proof process is the same due to the symmetry.

A.1 Discrete Situation

First, we consider the discrete case where each $teacher_i$ can be viewed as a classifier. For a binary classification model with precision of p, the probability of correct classification after each sampling follows a Bernoulli distribution. Thus, the expected classification precision of a single teacher is p, and the variance is p(1-p).

To simplify computation, we assume the performance of the N teachers is consistent, i.e., $p = p_1 = ... = p_N$, where p_i is the precision of T_i. The mean ensemble of N teachers is formulated as:

$$X_{ME} = \frac{\Sigma_i X_i}{N} \tag{13}$$

given that X_i which follows Bernoulli distribution are independent and identically distribute, we obtain the conclusion that $E(X_{ME}) = p$, $D(X_{ME}) = p(1-p)/N$.

Due to the fact $E(X_{ME}) = E(X_{M_i})$ and $D(X_{ME}) < D(X_{M_i})$, we conclude the following lemma:

Lemma 1. *Compared to the prediction from single model, although the mean ensemble result demonstrates better robustness, it keeping the expected precision the same.*

Next, we consider the case where N teachers form a vote-ensemble classifier based on the principle of maximum voting. Then the expectation of the classifier is as follows:

$$p_0 = \sum_{m=\frac{N+1}{2}}^{N} C_N^m p^m (1-p)^{N-m} \tag{14}$$

Utilizing mathematical induction, it is trivial to prove when $p > 1/2$, $p_0 > p$. This is called Condorcet's jury theorem and details of proof can be found in (Sancho, 2022). Now we can state the following lemma:

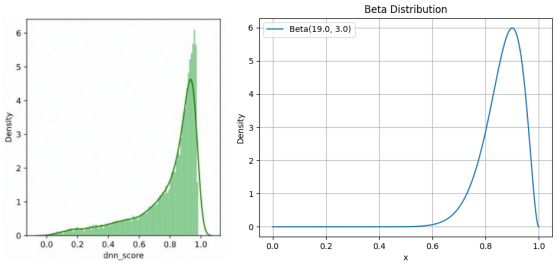

Figure 4: Left part shows the distribution of our model's output on test set, and right part shows the distribution of $Beta(19.0, 3.0)$. We can see that the model's output keep similar distribution with Beta function.

Lemma 2. *In discrete situation, vote-ensemble shows higher expected precision compared with mean-ensemble.*

A.2 Consecutive Situation

It is noted that in the setting of distillation, we take model as scorer rather than a simple classifier, and the output of the scorer is a float in $[0, 1]$. The distribution of the output is subject to Beta distribution, which is the conjugate distribution of Bernoulli distribution. This assumption can also be empirically verified as Figure 4 shows.

To simplify computation, we assume all teachers is subject to the same distribution, i.e., $X_i \sim B(b_1, b_2), \forall i \in \{1, ..., N\}$. Then we have:

$$E(X_{ME}) = \frac{\sum E(X_i)}{n} = \frac{b_1}{b_1 + b_2} \tag{15}$$

$$D(X_{ME}) = \frac{\sum D(X_i)}{n^2} \tag{16}$$

$$= \frac{(b_1 * b_2)}{n * (b_1 + b_2)^2 * (b_1 + b_2 + 1)} \tag{17}$$

So *Lemma 1* still holds in consecutive situation.

Next, we consider the case where N teachers calculate the ensemble scores by utilizing GOVERN method. We conduct numerical simulation using Monte-Carlo sampling to verify the superiority of GOVERN.

We set 10 teachers with same distribution as $X_i \sim B(20.0, 2.0), \forall i \in \{1, ..., N\}$, and set student as $X_0 \sim B(19.0, 3.0)$. The number of simulation is set to 1M.

The simulation result is shown in figure 5. We can see that the expectation of mean-ensemble is same with single teacher's output, while the variance is lower. This result is consist with *Lemma 1*. Under the setting of GOVERN, it shows higher expectation compare with mean-ensemble, and keeps comparable variance. This verifies that GOVERN can obtain a better score with high expectation for distillation, and keep comparable robustness like mean-ensemble.

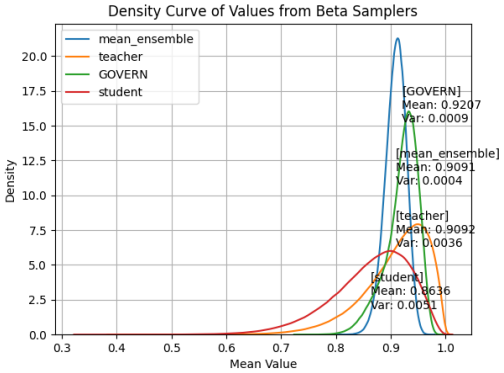

Figure 5