Learning ULMFiT and Self-Distillation with Calibration for Medical Dialogue System

Shuang Ao
Doti Health Ltd, UK
The Open University, UK
ao.shuang@u.nus.edu

Xeno Acharya
Doti Health Ltd, UK
xeno.acharya@gmail.com

Abstract

A medical dialogue system is essential for healthcare service as providing primary clinical advice and diagnoses. It has been gradually adopted and practiced in medical organizations, largely due to the advancement of NLP. The introduction of state-of-the-art deep learning models and transfer learning techniques like Universal Language Model Fine Tuning (ULMFiT) and Knowledge Distillation (KD) largely contributes to the performance of NLP tasks. However, some deep neural networks are poorly calibrated and wrongly estimate the uncertainty. Hence the model is not trustworthy, especially in sensitive medical decision-making systems and safety tasks. In this paper, we investigate the well-calibrated model for ULMFiT and self-distillation (SD) in a medical dialogue system. The calibrated ULMFiT (CULMFiT) is obtained by incorporating label smoothing (LS) to achieve a well-calibrated model. Moreover, we apply the technique to recalibrate the confidence score called temperature scaling (TS) with KD to observe its correlation with network calibration. Furthermore, we use both fixed and optimal temperatures to fine-tune the whole model. All experiments are conducted on the consultation backpain dataset collected by experts then further validated using a large publicly medial dialogue corpus. We empirically show that our proposed methodologies outperform conventional methods in terms of accuracy and robustness.

1 Introduction

The medical dialogue system is becoming a necessary tool for the doctor-patient interaction as it provides the primary clinical advice and long-distance diagnoses, shortening the checking duration and reducing the manpower cost. It is gradually applied and accepted especially during the pandemic time.

In order to provide an integrated conversational system for back pain management, the system needs to be equipped with evidence on the aforementioned determinants of health. This could best be facilitated by incorporating this evidence through a medical dialogue system. However, the insufficient medical corpus is one of the biggest restrictions for the training of the neural conversational model. We build the dataset particularly for back pain consultation, including the query and suggestions regarding possible causes, symptoms, and treatment of back pain, which to our best knowledge, is the first medical conversational dataset subjected in the backpain field.

To achieve a promising accuracy of the sentence generation model, we choose the state-of-the-art transformer model (Vaswani et al., 2017) as the benchmark. As transfer learning has shown great success in machine learning tasks such as classification and regression (Pan and Yang, 2009), in this paper, we choose two well-known and efficient techniques: Universal Language Model Fine Tuning (ULMFiT) (Howard and Ruder, 2018) and Knowledge Distillation (KD) (Hinton et al., 2015). ULMFiT provides additional help to improve the model accuracy by transferring information from language modeling to NLP downstream tasks, such as conversational model, sentiment analysis, and Machine Translation. Due to its obvious advantages, we implement the pretrained language model on top of the conversation model to get better feature extraction. Furthermore, KD transfers the knowledge from a cumbersome model to a lighter-weight model so that the small model can replicate the result. KD has been used in the recent NLP research, such as text classification and sequence labeling (Yang et al., 2020) and got the promising result. However, due to the limitation of data size and robust model, the application KD is not flexible to some extent. To resolve these issues, Self Distillation (SD) (Yuan et al., 2019) is proposed, where the student model is used as the teacher model as

well. Results show that SD can almost replicate the accuracy regardless of a well-trained large model or big dataset such as in the image classification task (Zhang et al., 2019). SD has also been applied to NLP tasks such as language model and neural machine translation (Hahn and Choi, 2019) and obtains promising results.

Despite obtaining higher accuracy and better performance, modern deep learning models face drawbacks of miscalibration and overconfidence (Müller et al., 2019; Naeini et al., 2015; Lakshminarayanan et al., 2016). Recent studies resolve this issue by using techniques like label smoothing (Müller et al., 2019) and temperature scaling (Naeini et al., 2015), and Dirichlet calibration (Kull et al., 2019). These works show that the well-calibrated model can improve the model performance as well as feature representation. As for the NLP downstream tasks, research has shown that calibration benefits both sentence quality and length in the sentence classification (Jung et al., 2020), and helps to improve the model fine-tuning in text generation (Kong et al., 2020).

As transfer learning techniques and calibration contributes to NLP tasks, we investigate the correlation of improving calibrated feature representation with ULMFiT and SD. Label smoothing is integrated with ULMFiT to extract significant features from language modeling. To improve KD by recalibrating predicted probability, we incorporate temperature scaling (TS) with knowledge distillation loss. We also observe the correlation of a well-calibrated trained network in whole model fine-tuning. We conduct extensive experiments to validate our observations with two datasets of (1) the consultation back pain and (2) medical dialogue. Results show that a well-calibrated model is highly correlated with ULMFiT and SD, as well as fine-tuning, in terms of both accuracy and calibration error.

Our contributions can be concluded as following:
(1) We introduce the calibrated ULMFiT (CULMFiT) by applying label smoothing on conventional ULMFiT. Results are showing that the CULMFiT outperforms the vanilla ULMFiT, proving the impact of calibration of language modeling.
(2) We measure optimal calibrated temperature and replace the fixed temperature value in KD loss and demonstrate that calibrated temperature outperforms the fixed value.
(3) We incorporate temperature scaling with the whole model fine-tuning and observe that calibration benefits model performance and uncertainty.
(4) We build the consultation backpain dataset, consisting of patients' queries and clinicians' responses into conversational pairs.

2 Proposed Method

2.1 Preliminaries

ULMFiT Natural Language Processing has picked up the pace in recent years and caught researchers' attention greatly, essentially attributed to the conquer of inductive transfer learning, which was seen as the major obstacle that NLP was lagging behind Computer Vision (CV). Universal Language Fine Tuning (ULMFiT) was proposed (Howard and Ruder, 2018) as obtaining the success of passing the acquired knowledge of pre-trained model to other similar tasks. ULMFiT is to pretrain the model on a large general domain corpus such as Wikipedia data, then fine-tune it on the target tasks. As a source task trained with a large corpus, the pre-trained language model can capture most facets and contexts of the data, which is ideal for NLP downstream tasks. Hence including Text Classification that ULFMiT was firstly introduced with, it gets great success and applied in almost all NLP fields. It is believed that with the language model trained on the large-scale data, the model with small or medium data will also replicate similar results to the vanilla model.

Label Smoothing (LS) Label smoothing has been widely applied in various fields of deep learning, such as image classification (Real et al., 2019) and speech recognition (Chorowski and Jaitly, 2016). It achieves promising results since Szegedy et al. (Szegedy et al., 2016) first introduced it, then gets further development after the extension explanation on its mechanism of how it improves the model calibration (Müller et al., 2019). As the regularization technique to tackle the overconfidence of a model, label smoothing softens the one-hot labels in the penultimate layer's logit vectors, to improve the calibration and further help the robustness and reliability of the model. Here is the mathematical illustration of label smoothing: suppose $\hat{p}_c$ is the probability and p_c is the ground truth of the c-th class, where p_c is 1 for the correct class and 0 for the rest classes, the cross-entropy loss of network trained with a hard target can be

demonstrated as: $CE = -\sum_{c=1}^{C} p_c \log(\hat{p}_c)$.For a network trained with a label smoothing hyper-parameter α, the one-hot true value will be clipped as: $p_c^{LS} = p_c(1-\alpha) + \alpha/C$. Hence the cross-entropy loss with label smoothing can be illustrated as:

$$CE^{LS} = -\sum_{c=1}^{C} p_c^{LS} \log(\hat{p}_c). \quad (1)$$

Temperature Scaling (TS) It has been observed that most of the modern neural networks are poorly calibrated even with a high confidence score. To solve this issue and make the model better calibrated, among all possible factors that may influence the calibration, temperature scaling (TS), as a straightforward extension of Platt Scaling, has been verified as the most efficient and least time-consuming and computationally expensive way (Guo et al., 2017). A single scalar T (T>1) called temperature is applied on the logit then it passes to the softmax function (denoted as σ), which will not change the maximum value in it, so the prediction remains intact. Here is the equation for TS given the logit vector:

$$\hat{p_c}^{TS} = \max_c \sigma_(logit_c/T)^{(c)}. \quad (2)$$

Self-Distillation (SD) Knowledge distillation (KD) targets compressing a cumbersome teacher model into a lighter-weight student model. The distilled model can still replicate similar or better accuracy due to the privileged information captured by the teacher model. Suppose the logits for teacher model and student model are $logit^T$ and $logit^S$, and fixed T value as T^{fix}, the loss function with of Kullback-Leibler divergence (KL divergence) L_{KD} can be formulated as:

$$L_{KD} = \sum \mathrm{KL}\Big(\sigma\big(\frac{logit^T}{T^{fix}}\big), \sigma\big(\frac{logit^S}{T^{fix}}\big)\Big) \quad (3)$$

It is generally believed that the teacher model should be well-trained with a large corpus and has a bigger capacity than the student model. However, the insufficiency of the dataset and the untrustworthiness of the model are substantial restrictions to KD. Yuan et al (Yuan et al., 2019) argue that the student model can achieve similar results with a poor-trained or smaller teacher model, even under the circumstance of no teacher model, which is called self-distillation (SD). By making the model be their own teacher, SD is to train the student model first to get a pre-trained model, then using it as the teacher to train itself. It has been further proved the positive effect that self-distillation has on calibration (Zhang and Sabuncu, 2020).

2.2 ULMFiT with Label Smoothing

ULMFiT has obtained great success in NLP tasks as it transfers information from the pre-trained model to the target application domain, and LS helps in calibration and better uncertainty. We apply LS to ULMFiT to gain a calibrated ULMFiT (CULMFiT) to further improve the feature representation and extract more distinctive information from language modeling. Given θ^{ULMFiT} is the pre-trained ULMFiT weight, x as the input of the conversational model, the loss function of ULMFiT with LS can be written as follows:

$$CE_U^{LS} = -\sum_{c=1}^{C} p_c^{LS} \log(\hat{p_c}|x, \theta_U). \quad (4)$$

2.3 Self-Distillation with TS

Self-distillation (SD) has been proved to replicate the similar accuracy as the knowledge distillation (KD) with the teacher model training on student model, and temperature scaling helps to prevent miscalibration. We integrate TS on SD to attain a well-calibrated distilled model. For this purpose, we adopt KD loss of KL divergence with calibration as in the paper (Hinton et al., 2015). However, temperature set as a scalar value is a similar technique as network calibration, and the optimal temperature is expected to be a better option. In our work, we measure optimal T and assign it to the KD, aiming at preventing inappropriate calibration and investigating the relation between calibration and SD. Suppose the logits for the teacher model and student model are $logit^T$ and $logit^S$, and the optimal temperature is T^{opt}. The loss function with KL divergence L_{KD} can be formulated as:

$$L_{SD} = \sum \mathrm{KL}\Big(\sigma\big(\frac{logit^T}{T^{opt}}\big), \sigma\big(\frac{logit^S}{T^{opt}}\big)\Big) \quad (5)$$

The final loss L can be demonstrated as:

$$L = L_{SD} + L_{CE} \quad (6)$$

2.4 Fine-tuning with TS

As an approach of transfer learning, fine-tuning can propagate the acquired knowledge from one domain to another and enhances the learning capacity.

Table 1: Samples of the backpain dataset.

ID		Medical Dialogue
0	Enquiry	What is musculoskeletal pain condition?
	Reply	A great change of lifestyle and behaviour, such as too much workload, adjustments in the workplace, work breaks and sudden exercise would improvement of musculoskeletal pain.
1	Enquiry	Why my foot pain cause back pain?
	Reply	The possible reason is your spine's alignment or overstressing lower back muscles
2	Enquiry	The back pain cause me unable to carry groceries, what should I do?
	Reply	Try the grocery delivery or ask help from your close family or friends. If it is severe, contact your clinician immediately.
3	Enquiry	Will back pain influence the enjoyment between couples?
	Reply	Yes, studies have shown that higher lever of back pain can impair the leisure activities with the spouse.
4	Enquiry	I feel pain in my joints after exercise, what is the problem?
	Reply	If your joint feels particularly painful afterwards for longer than two hours after an exercise session, reduce the intensity of your next exercise session.

On the other hand, TS produces a well-calibrated confidence score. To further improve the information transformation and feature representation, we apply TS to the logit for cross-entropy loss calculation while fine-tuning the entire model. Given p_c^{TS} is the temperature scaled logit (as shown in formula 2), the loss function with TS can be illustrated as:

$$CE^{TS} = -\sum_{c=1}^{C} p_c \log(\hat{p}_c^{TS}). \tag{7}$$

3 Experiments

3.1 Datasets

Backpain Dataset To develop an evidence-based skillful conversational model, we collect the backpain dataset with pairs of the query from a patient and the response from a clinician. Table 1 shows samples of conversational pairs. Sources of queries are various sites people would generally ask health-related questions, such as Google and Quora, and responses are collated from either peer-reviewed journal articles (Hayden et al., 2005) (Henschke et al., 2010) (Cagnie et al., 2007) (Scheermesser et al., 2012) (Choi et al., 2010) (Van Dam et al., 2018) or other sources recognized for providing valid health advice and suggestions like NHS website [1]. It covers five highly related factors that cause back pain, namely sleep, mental health, exercise, nutrition, and social and environmental factors. The dataset contains 1000 conversational pairs for the train set and 200 pairs for the validation set, and the minimum and maximum length of the reply are 16 and 40.

[1] https://www.nhs.uk/conditions/back-pain/

MedDialog Due to the disadvantages of the small volume of our backpain dataset, we also use the MedDialog Dataset (Zeng et al., 2020) to further testify our hypothesis of calibration. It consists of conversational pairs of symptoms description from patients and follow-up questions and diagnoses from doctors, which covers various medical fields such as pathology and family medicine. We randomly divide the dataset into train and validation set with the ratio of 0.8 and 0.2.

3.2 Implementation Details

We choose the well-known transformer model as the benchmark in our project. The language modeling architecture for ULMFiT is the encoder part of the Transformer with Fully-Connected (FC) Layers, and the loss function is cross-entropy loss with label smoothing. To fine-tune the proposed model, we first get the optimal TS value, then apply it to re-calibrate the logit for the trained model. The GPU of Nvidia Tesla T4 with the memory of 16GB is used to conduct all the experiments in this work. The dataset is split with 0.8 and 0.2 for training and validation. All experiments are conducted with the Adam optimizer, 0.01 as the learning rate and batch size of 4. The best BLEU-1 score metric is used to find the best epoch.

4 Experiments

4.1 Results

4.1.1 Evaluation Metrics

We use the uni-gram similarity metrics BLEU-1 as the major evaluation for our dialogue system. To measure the word overlapping between the ground truth and prediction, we also apply Metric for Evaluation of Translation with Explicit Ordering (ME-

Table 2: Results of Backpain Dataset. Annotations of experimental models are as following: the vanilla transformer model and ULMFiT are labeled as Baseline; ULMFiT with label smoothing as CULMFiT; model with ULMFiT and fine tune with TS as Fine-tune.

Method	Model	BLEU-1	Perplexity	METEOR	ECE
Baseline	Transformer	0.4292	7.9895	0.4079	0.3702
	ULMFiT	0.4321	8.0603	0.4218	0.3764
LS	CULMFiT	**0.4632**	5.6155	**0.4552**	0.3674
TS	Fine-tune	0.4415	**5.2797**	0.4268	**0.2884**

Table 3: Results of MedDialog Dataset: the vanilla transformer model and ULMFiT are annotated as Baseline; ULMFiT is the Transformer model trained with the Medical Dialogue Dataset; regularized ULMFiT is annotated as CULMFiT; the proposed model fine tuned with TS is Fine-tune.

Method	Model	BLEU-1	Perplexity	METEOR	ECE
Baseline	Transformer	0.3387	11.5422	0.2280	0.2611
	ULMFiT	0.3609	**7.9134**	0.2556	0.3519
LS	CULMFiT	**0.3765**	10.2346	0.2578	0.3734
TS	Fine-tune	0.3747	12.6997	**0.2618**	**0.0580**

TEOR) metric (Banerjee and Lavie, 2005) in our work. Perplexity, as the measurement of model uncertainty to the training data, is calculated based on the cross-entropy loss for each sample. We use the Expected Calibration Error (ECE) (Naeini et al., 2015) to check the efficiency of calibration techniques. ECE divides predictions into N equally-spaced bins and takes the weighted mean of each bin's confidence gap. We choose N=15 bins in our work.

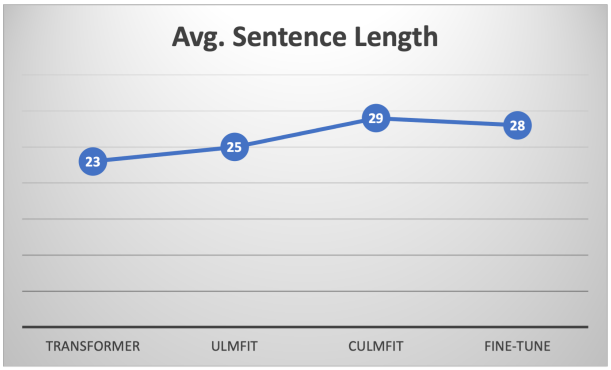

Figure 1: Average sentence length generated by each model in the backpain dataset.

4.1.2 Evaluation of the Backpain Dataset

The results of the dialogue system trained with the consultation backpain dataset are shown in table 2 and examples of generated responses are demonstrated in table 4. The calibrated ULMFiT with LS (CULMFiT) significantly outperforms the baseline transformer model by improving the BLEU-1 score by about 3.8%, and exceed the vanilla ULMFiT by approximately 1.5%. On the other hand, the fine-tuning TS improves both BLEU-1 score and ECE with 1% and 8%, respectively. Though the fine-tuning with TS does not provide the best BLEU-1 score, it provides the best calibrated confidence score with the lowest ECE. In terms of generated sentence length and quality, our proposed models of CULMFiT and fine-tuning outperform baseline models of transformer and ULMFiT. The diagram 1 illustrates that on average, the generated conversation length of CULMFiT and fine-tuning is longer than those from the benchmark models, where CULMFiT model produces the longest responses. Furthermore, the proposed models generate more logical and meaningful sentences. For example, in the first sample in table 4, the CULMFiT network predicts the verb "*stand*" and the fine-tuning model generates the phrase "*a short period every hour*" that exactly matches the ground truth, which makes the response more accurate for the symptom. Overall, CULMFiT demonstrates the best performance on most of the evaluation metrics. Evaluation results prove the effectiveness of proposed calibration techniques with the ULMFiT and fine-tuning on the probability and correctness adjustment.

4.1.3 Evaluation of MedDialog Dataset

To further verify the hypothesis that calibration benefits the model performance in both accuracy and robustness, we replicate the previous experi-

Table 4: Visualization of predicted responses. Query is the input and GT is the ground truth. Prediction is the response generated by the model.

	Sample		Prediction
Backpain	Query: What to do to solve long time sitting issue except exercise? GT: stand up and move about gently for a short period every hour would help relieve the muscle stiffness	Transformer	up and move for a short time would help the stiffness.
		ULMFiT	be up and move for a short time every hour would help muscle stiffness.
		CULMFiT	stand up and move around gently for a short period every hour would help to solve the muscle stiffness.
		Fine-tune	be up and move gently for a short period every hour would help the muscle stiffness.
MedDialog	Query: Hi doctor, I am 25 years old and I have a history of hair fall for almost 5 years. I am very concerned of it. Could you give me some advice and possible treatment? GT: Hi, as per you case history of hair fall, my treatment advice is to take good nutritious diet full of green leafy vegetables and milk, and to use a good herbal shampoo and coconut hair oil.	Transformer	Hello, per your case of hair fall, my advice be good diet with vegetables and milk, use shampoo and oil out of it.
		ULMFiT	Hello, per your case history of hair fall, my advice be take good herbal diet full of green leafy vegetables and milk, use a good shampoo and oil for it.
		CULMFiT	Hi, per you case history of hair fall, my treatment advice is to take nutritious diet of green vegetables and milk, and to use a good herbal shampoo and hair oil.
		Fine-tune	Hi, per you case history of hair fall, my treatment advice is to take herbal diet of green vegetables and milk, and to use a good herbal shampoo and green herbal oil.

Table 5: Results of self-distillation with Backpain dataset. Three methods are applied in this experiment: without self-distillation (standalone), self-distillation with a fixed value of TS (SD Fixed TS), and self-distillation with optimal TS (SD optimal TS). The fixed TS is 2. The optimal TS for the transformer model and CULMFiT is 3.025 and 4.789 respectively.

Method	Model	BLEU-1	Perplexity	METEOR	ECE
Standalone	Transformer	0.4292	7.9895	0.4079	0.3702
	CULMFiT	**0.4632**	**5.6155**	**0.4552**	0.3674
SD Fixed TS	Transformer	0.4331	7.8329	0.4221	0.3820
	CULMFiT	0.4236	6.3934	0.4135	0.1962
SD Optimal TS	Transformer	0.4334	7.8010	0.4187	0.3703
	CULMFiT	0.4473	5.8486	0.4402	**0.1788**

ments on the Medical Dialogue Dataset. The results of various evaluation metrics are illustrated in table 3, and the sample visualization is shown in 4. All results are mostly consistent with the previous experiments. For example, in the sample illustrated in the table 4, the length of predicted sentences from CULMFiT and fine-tuning model is longer than the baseline models. Besides, the adjective "*herbal*" for the noun "*shampoo*" from the proposed models can better explain the type of shampoo product, which makes the response more specific for the patient's inquiry. Overall, the proposed methodologies illustrate superior performance in most of the evaluation metrics. The calibrated ULMFiT (CULMFiT) with LS outperforms the benchmark and the vanilla ULMFiT by about 4% and 1.5% increment of BLEU-1 score correspondingly. The fine-tuning with the TS model significantly improves ECE by about 35%. Results from both experiments prove that calibration tech-

niques of LS and TS help to improve the robustness and uncertainty of the model.

4.1.4 Evaluation of Self-Distillation With TS

One of our observations is that the SD model with the optimal TS outperforms the one with fixed TS. All results are shown in table 5. We select the benchmark transformer model and the model with the calibrated ULMFiT in this experiment. It has been shown that SD with the optimal T value obtains better performance than with the fixed T (with $T = 4$) value for image classification (Hinton et al., 2015). Hence in our work, we also compare the SD with fixed T and optimal T applied in both benchmark and proposed model. To select the best fixed T value, we apply T values of 1.5, 2, 3, 4, and 5 and choose the one with the best BLEU-1 score. The diagram 2 indicates that $T = 2$ provides the best BLEU-1 score. Compared to the standalone, SD with fixed and optimal T of transformer and CULMFiT models in table 5, CLUMFiT without SD obtains the best BLEU-1 score, perplexity, and METEOR, while SD with optimal TS provides the best ECE. On the other hand, CULMFiT gets hampered with calibration, which has been evinced in the work (Müller et al., 2019). Overall, the performance of the model trained with optimal TS beats the one with fixed TS.

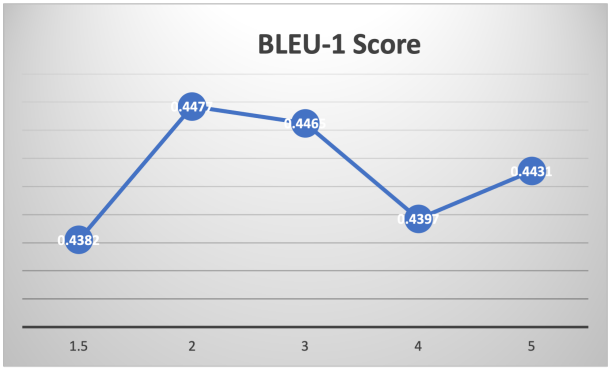

Figure 2: BLEU-1 score with different T values.

5 Discussion

In this paper, we apply calibration techniques of LS and TS to develop the medical dialogue system and get promising results. Table 2 and 3 are showing results that the well-calibrated model benefits ULMFiT, SD and fine-tuning. Table 5 demonstrates the observation of self-distillation on fixed and optimal temperature scaling. All our observations is presented with the sample visualization in table 4. Overall, the ULMFiT with LS provides the best BLEU-1 score and the fine-tuning TS improves the ECE mostly, which is consistent with experiments in both datasets. Despite the higher model performance in both accuracy and calibration, fine-tuning is a two-stage training, which can cause an additional computational burden. Even though LS and TS introduce additional computational parameters, the computational cost is negligible. On the other hand, ULMFiT with label smoothing hurts SD, which has been reported in (Müller et al., 2019).

6 Conclusion

In this paper, we propose the calibrated ULMFiT, self-distillation and fine-tuning to build a medical dialogue system. Label smoothing and temperature scaling are utilized to obtain calibrated network and improve the performance in terms of accuracy and robustness. We empirically demonstrate calibration is highly co-related with ULMFiT, SD and fine-tuning, which has been presented in table 2, 3,4 and 5. For future work, we will explore the calibration and knowledge-distillation impact on other NLP downstream tasks like Neural Machine Translation and Sentiment Analysis.

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.

Barbara Cagnie, Lieven Danneels, Damien Van Tiggelen, Veerle De Loose, and Dirk Cambier. 2007. Individual and work related risk factors for neck pain among office workers: a cross sectional study. *European Spine Journal*, 16(5):679–686.

Brian KL Choi, Jos H Verbeek, Wilson Wai-San Tam, and Johnny Y Jiang. 2010. Exercises for prevention of recurrences of low-back pain. *Cochrane Database of Systematic Reviews*, (1).

Jan Chorowski and Navdeep Jaitly. 2016. Towards better decoding and language model integration in sequence to sequence models. *arXiv preprint arXiv:1612.02695*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*.

Sangchul Hahn and Heeyoul Choi. 2019. Self-knowledge distillation in natural language processing. *arXiv preprint arXiv:1908.01851*.

Jill Hayden, Maurits W Van Tulder, Antti Malmivaara, and Bart W Koes. 2005. Exercise therapy for treatment of non-specific low back pain. *Cochrane database of systematic reviews*, (3).

Nicholas Henschke, Raymond WJG Ostelo, Maurits W van Tulder, Johan WS Vlaeyen, Stephen Morley, Willem JJ Assendelft, and Chris J Main. 2010. Behavioural treatment for chronic low-back pain. *Cochrane database of systematic reviews*, (7).

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Taehee Jung, Dongyeop Kang, Hua Cheng, Lucas Mentch, and Thomas Schaaf. 2020. Posterior calibrated training on sentence classification tasks. *arXiv preprint arXiv:2004.14500*.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in-and out-of-distribution data. *arXiv preprint arXiv:2010.11506*.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Hao Song, Peter Flach, et al. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *arXiv preprint arXiv:1910.12656*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2016. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4694–4703.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. 2019. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 4780–4789.

Mandy Scheermesser, Stefan Bachmann, Astrid Schämann, Peter Oesch, and Jan Kool. 2012. A qualitative study on the role of cultural background in patients' perspectives on rehabilitation. *BMC musculoskeletal disorders*, 13(1):1–13.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Nicholas T Van Dam, Marieke K van Vugt, David R Vago, Laura Schmalzl, Clifford D Saron, Andrew Olendzki, Ted Meissner, Sara W Lazar, Catherine E Kerr, Jolie Gorchov, et al. 2018. Mind the hype: A critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on psychological science*, 13(1):36–61.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ziqing Yang, Yiming Cui, Zhipeng Chen, Wanxiang Che, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Textbrewer: An open-source knowledge distillation toolkit for natural language processing. *arXiv preprint arXiv:2002.12620*.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. 2019. Revisit knowledge distillation: a teacher-free framework. *arXiv preprint arXiv:1909.11723*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: A large-scale medical dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722.

Zhilu Zhang and Mert R Sabuncu. 2020. Self-distillation as instance-specific label smoothing. *arXiv preprint arXiv:2006.05065*.