Attention Temperature Matters in Abstractive Summarization Distillation

Shengqiang Zhang[1] *† **Xingxing Zhang**[2*], **Hangbo Bao**[2†], **Furu Wei**[2]
[1] Peking University
[2] Microsoft Research Asia
sq.zhang@pku.edu.cn
{xizhang,t-habao,fuwei}@microsoft.com

Abstract

Recent progress of abstractive text summarization largely relies on large pre-trained sequence-to-sequence Transformer models, which are computationally expensive. This paper aims to distill these large models into smaller ones for faster inference and with minimal performance loss. Pseudo-labeling based methods are popular in sequence-to-sequence model distillation. In this paper, we find simply manipulating attention temperatures in Transformers can make pseudo labels easier to learn for student models. Our experiments on three summarization datasets show our proposed method consistently improves vanilla pseudo-labeling based methods. Further empirical analysis shows that both pseudo labels and summaries produced by our students are shorter and more abstractive. Our code is available at https://github.com/Shengqiang-Zhang/plate.

1 Introduction

Automatic document summarization is the task of rewriting a long document into its shorter form while still retaining its most important content. In the literature, there are mainly two kinds of methods for summarization: *extractive summarization* and *abstractive summarization* (Nenkova and McKeown, 2011). In this work, we focus on abstractive summarization, which is viewed as a sequence-to-sequence (Seq2Seq) learning problem, since recent abstractive models outperform their extractive counterparts and can produce more concise summaries (Raffel et al., 2020; Lewis et al., 2020; Zhang et al., 2020; Liu and Lapata, 2019). Recent progress of abstractive summarization largely relies on large pre-trained Transformer models (Raffel et al., 2020; Lewis et al., 2020; Zhang et al., 2020; Liu and Lapata, 2019; Bao et al., 2020). With these extremely large models, we can obtain state-of-the-art summarization results, but they are slow for online inference, which makes them difficult to be used in the production environment even with cutting-edge hardware. This paper aims to distill these large Transformer summarization models into smaller ones with minimal loss in performance.

Knowledge distillation is a class of methods that leverage the output of a (large) teacher model to guide the training of a (small) student model. In classification tasks, it is typically done by minimizing the distance between the teacher and student predictions (Hinton et al., 2015). As to Seq2Seq models, an effective distillation method is called pseudo-labeling (Kim and Rush, 2016), where the teacher model generates pseudo summaries for all documents in the training set and the resulting document–*pseudo*-summary pairs are used to train the student model.

In this paper, we argue that attention distributions of a Seq2Seq teacher model might be too sharp. As a result, pseudo labels generated from it are sub-optimal for student models. In the summarization task, we observe that 1) pseudo summaries generated from our teacher model copy more continuous text spans from original documents than reference summaries (56% 4-grams in pseudo summaries and 15% 4-grams in reference summaries are copied from their original documents on CNN/DailyMail dataset); 2) pseudo summaries tend to summarize the leading part of a document (measured on CNN/DailyMail, 74% of sentences in pseudo summaries and 64% of sentences in reference summaries are from the leading 40% sentences in original documents). We obtain the two numbers above by matching each sentence in a summary with the sentence in its original document that can produce maximum ROUGE (Lin, 2004) score between them. We call the two biases above the *copy bias* and the *leading bias*. In order to have an intuitive feeling, we select a rep-

* Equal contribution.
† Work done during the authors' internships at Microsoft Research Asia.

resentative example [1] and visualize its cross attention weights [2] (see the left graph in Figure 1). We observe that attention weights form three "lines", which indicates very time the decoder predicts the next word, its attention points to the next word in the input document. That may be the reason why multiple continuous spans of text are copied. Another phenomenon we observe is that all *high-value* attention weights (in deeper color) concentrate on the first 200 words in the input document, which reflects the leading bias. In either case, the attention distribution is too sharp (i.e., attention weights of the next word position or the leading part is much larger than other positions), which means our teacher model is over-confident.

Based on the observations above, we propose a simple method called PLATE (as shorthand for **P**seudo-labeling with **L**arger **A**ttention **TE**mperature) to smooth attention distributions of teacher models. Specifically, we re-scale attention weights in all attention modules with a higher temperature, which leads to *softer* attention distributions. Figure 1 intuitively shows the effect of using higher attention temperatures. Compared with the left graph, the right graph with higher attention temperature has shorter lines (less copy bias) with high attention weights, and positions of high attention weights extend to the first 450 words (less leading bias). Less *copy bias* in pseudo summaries encourages student models to be more abstractive, while less *leading bias* in pseudo summaries encourages student models to take advantage of longer context in documents.

Experiments on CNN/DailyMail, XSum, and New York Times datasets with student models of different sizes show PLATE consistently outperforms vanilla pseudo-labeling methods. Further empirical analysis shows that, with PLATE, both pseudo summaries generated by teacher models and summaries generated by student models are shorter and more abstractive, which matches the goal of abstractive summarization.

2 Related Work

Large pre-trained Seq2Seq Transformer models largely improve results of generation tasks including text summarization (Song et al., 2019; Lewis et al., 2020; Bao et al., 2020; Raffel et al., 2020; Zhang et al., 2020). These models are pre-trained using unsupervised text-to-text objectives. For example, T5 (Raffel et al., 2020) is pre-trained by predicting corrupted text spans. BART (Lewis et al., 2020) employs denoising auto-encoding objectives such as text infilling and sentence permutation during its pre-training. The pre-training objective of PEGASUS (Zhang et al., 2020) is tailored for the summarization task, which predicts the most "summary worthy" sentences in a document. Our method aims to make these large models faster.

In knowledge distillation, besides learning from gold labels in the training set, student models can learn from soft targets (Ba and Caruana, 2014; Hinton et al., 2015), intermediate hidden states (Romero et al., 2014), attentions (Zagoruyko and Komodakis, 2017; Wang et al., 2020), and target output derivatives (Czarnecki et al., 2017) of teacher models. Recent work for distillation of pre-trained Transformers (e.g., DistilBERT (Sanh et al., 2019), TinyBERT (Jiao et al., 2020), MobileBERT (Sun et al., 2020), BERT-of-Theseus (Xu et al., 2020a), MINILM (Wang et al., 2020)) focuses on natural language understanding tasks such as GLUE (Wang et al., 2018) or SQuAD (Rajpurkar et al., 2016) benchmarks. Most methods above are designed for classification models.

In Seq2Seq learning tasks such as summarization, we can apply distillation methods above to each step of sequence model predictions. However, the sequence-level knowledge of teacher mod-

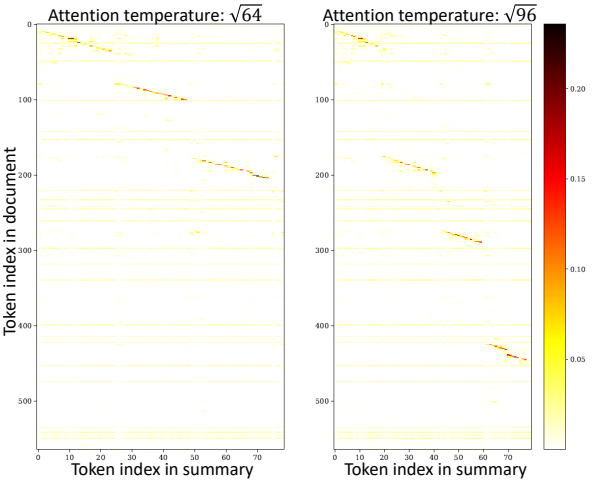

Figure 1: Visualization of teacher cross attention weights when generating pseudo labels with normal (left) and smoothed (right) attention weights. This example is generated by the BART teacher trained on CNNDM (see §4.4). Its training and inference hyperparameters are described in detail in §4.2.

[1] See the detailed example in Appendix E.

[2] We use cross attention because we can see how words in documents are selected during generation.

els is not well utilized. Therefore, Kim and Rush (2016) introduce a sequence-level knowledge distillation method (i.e., *pseudo-labeling*), where a student model is trained with pseudo labels generated by the teacher model using beam search decoding. Kim and Rush (2016) and later work (Kasai et al., 2020; Gu et al., 2017; Denkowski and Neubig, 2017) show *pseudo-labeling* achieves competitive performance for Seq2Seq tasks such as machine translation. Shleifer and Rush (2020) propose the *shrink and fine-tune* (SFT) approach for pre-trained summarization distillation, which re-finetunes a teacher model with some layers removed, and they show SFT outperforms *pseudo-labeling* and a modification of direct knowledge distillation (Jiao et al., 2020) on one of their datasets, but not others. Our method, which builds on top of *pseudo-labeling*, is conceptually simple and improves *pseudo-labeling* across different summarization datasets.

There is an interesting line of work called self-distillation or self-training (Furlanello et al., 2018; Xie et al., 2020; Deng et al., 2009; Liu et al., 2020; He et al., 2019), where the size of the student model is identical to the size of the teacher model. Our method can also be applied in self-distillation and can potentially be combined with the self-distillation methods above.

3 Summarization Distillation

3.1 Transformer based abstractive summarization

Abstractive summarization aims to rewrite a document into its shorter form (i.e., summary), which is a typical Seq2Seq learning problem. We adopt the Seq2Seq Transformer (Vaswani et al., 2017) model. Given a document $X = (x_1, x_2, \ldots, x_{|X|})$ and its gold summary $Y = (y_1, y_2, \ldots, y_{|Y|})$, we estimate the following conditional probability:

$$p(Y|X;\theta) = \prod_{t=1}^{|Y|} p(y_t|y_{<t}, X;\theta) \quad (1)$$

where θ is the model parameter and $y_{<t}$ stands for all tokens before position t (i.e., $(y_1, y_2, \ldots, y_{t-1})$).

The Seq2Seq Transformer model can be trained by minimizing the negative log-likelihood of gold document-summary pairs:

$$\mathcal{L}_{\text{G}}(\theta) = -\frac{1}{|Y|}\log p(Y|X;\theta) \quad (2)$$

where $|Y|$ is the number of tokens in summary Y.

3.2 Distillation with pseudo labels

Knowledge distillation refers to the task of transferring knowledge of a large teacher model (or a group of large teacher models) into a small student model. As to Seq2Seq learning tasks such as machine translation and summarization, pseudo-labeling based methods are usually used to imitate teacher predictions at the sequence level. Specifically, suppose we have a document X, and $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|\hat{Y}|})$ is a *pseudo* summary generated by a teacher model using beam search. The student can be trained by minimizing the negative log-likelihood of document-to-*pseudo*-summary pairs.

$$\mathcal{L}_{\text{PL}}(\theta) = -\frac{1}{|\hat{Y}|}\sum_{t=1}^{|\hat{Y}|}\log p(\hat{y}_t|\hat{y}_{<t}, X;\theta) \quad (3)$$

Strictly, all possible *pseudo* summaries from X should be taken into account. Unfortunately, the computational cost is prohibitive. We therefore use a single sample $\hat{Y}$ (which takes a large portion of probability mass from the teacher) instead as in Kim and Rush (2016).

3.3 Re-scaling attention temperatures

Both our teacher and student models are Seq2Seq Transformer models. The core part of a Transformer model is the attention module:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\tau})V \quad (4)$$

where Q, K, V are linear projections of hidden states of a layer and τ is the temperature of the attention module which is usually $\sqrt{d}$ (d is the hidden dimension size of that attention head).

Our distillation method PLATE works as follows. Assume we have a teacher model trained with $\tau = \sqrt{d}$. When the teacher generates pseudo labels with beam search, we use a higher attention temperature and set $\tau = \sqrt{\lambda\, d}$ where $\lambda > 1$ (λ is the attention temperature coefficient). Note that we only change the teacher's attention temperature during inference time. When we train our student model with pseudo labels, we still use a normal temperature (i.e., $\tau = \sqrt{d}$). We find that adjusting the student's attention temperature does not work. Probably because the student can easily adapt to the scaled attention temperature during training.

We find that $\lambda = 1.5$ or $\lambda = 2.0$ usually works well in practice. To encourage teacher models to

generate pseudo labels with more diversity, we further propose to use a random λ for each input document ($\lambda \sim U[a, b]$). Note that $U[a, b]$ is a uniform distribution and we typically set $a = 1.0$ and $b = 2.0$.

4 Experiments

4.1 Datasets

We conduct our experiments on three popular document summarization datasets: CNN/DailyMail (Hermann et al., 2015), XSum (Narayan et al., 2018), and New York Times (Sandhaus, 2008). All datasets are tokenized with the GPT-2 tokenizer (Radford et al., 2019), which is based on UTF-8 BPE (Sennrich et al., 2016).

CNNDM The CNN/DailyMail dataset (CNNDM; Hermann et al., 2015) contains online news articles from the CNN and DailyMail websites paired with their associated highlights as reference summaries. We follow the standard pre-processing steps described in See et al. (2017); Liu and Lapata (2019). [3] The resulting numbers of document-summary pairs for training, validation, and test are 287,227, 13,368, and 11,490, respectively.

XSum The XSum dataset is collected by harvesting online articles from the BBC with single sentence summaries, which is professionally written. The summaries are *extremely* abstractive. We use the official splits of Narayan et al. (2018). There are 204,045 articles for training; 11,332 articles for validation; and 11,334 articles for test.

NYT The New York Times dataset (NYT; Sandhaus, 2008) is composed of articles published by New York Times, and the summaries are written by library scientists. After applying the pre-processing procedures described in Durrett et al. (2016); Liu and Lapata (2019), we first obtain 110,540 articles with abstractive summaries. The test set is constructed by including the 9,076 articles published after January 1, 2007. The remaining 100,834 articles are further split into training and validation sets. After removing articles with summaries less than 50 words, we obtain the final dataset with 38,264 articles for training; 4,002 articles for validation; and 3,421 articles for test.

[3] Scripts are available at https://github.com/abisee/cnn-dailymail.

Model	# Param.	Latency (Millisecond)		
		CNNDM	XSum	NYT
BART	406M	1975	903	3272
BART 12-6	306M	1279	438	1692
BART 12-3	255M	924	289	1488
Transformer	70M	1028	406	1462

Table 1: Latency (in Milliseconds) on a V100 GPU and number of parameters (million) of our models.

4.2 Implementation details

Teacher/Student model settings We use BART Large (Lewis et al., 2020) as our teacher model, which has 12 layers in the encoder and decoder. The hidden size of each layer is 1024, and each layer contains 16 attention heads with a hidden size of 64. We have four kinds of student models. The first three student models are initialized from BART weights (therefore, their hidden sizes are the same as that of BART). All the three students have the 12 layers of BART encoder and differ in the number of decoder layers. They are denoted by BART 12-6, BART 12-3, and BART 12-12 with 6, 3, and 12 decoder layers, respectively. For BART 12-6 (or BART 12-3), the decoder is initialized from the first 6 (or 3) layers or the maximally spaced 6 (or 3) layers of BART decoder. The fourth student is the Transformer base model (Vaswani et al., 2017), which has 6 layers in each of the encoder and decoder. Each layer has a hidden size of 512 and 8 attention heads. This student is randomly initialized and denoted by Transformer. The latency statistics (Milliseconds) and numbers of parameters of above four models are in Table 1.

Training and inference Hyper-parameters for BART, BART 12-6, BART 12-3, and BART 12-12 are similar. Specifically, all models are optimized using Adam (Kingma and Ba, 2014) with $\beta_1 = 0.9, \beta_2 = 0.999$. Learning rates are tuned on validation sets (choose from 1e-5, 3e-5, 5e-5, 7e-5). We truncate all documents and summaries to 1024 sub-word tokens. We use a batch size of around 80 documents (we limit the max number of tokens on each GPU to 2048) and train our models for 20,000/15,000/6,000 steps with 500 warmup steps for CNNDM, XSum, and NYT, respectively. We also employ a weight decay of 0.01. For Transformer, the hyper-parameters of the Adam optimizer is a bit different, and we use $\beta_1 = 0.9, \beta_2 = 0.98$. Learning rates are picked from 1e-4, 3e-4, 5e-4, 7e-4 accord-

ing to validation sets. The weight decay is set to 0.0001. The warmup step we use is 4000. We train `Transformer` for 100 epochs and select the best model w.r.t. their ROUGE scores on validation sets. For all models above we apply a label smoothing of 0.1 to prevent overfitting (Pereyra et al., 2017).

During inference, as common wisdom, we apply beam search. The beam size, length penalty, and minimal length are 4, 2.0, and 55 on CNNDM; 6, 0.1, and 1 on XSum; and 4, 0.7, and 80 on NYT, respectively. All our models are trained on 8 NVIDIA V100 GPUs. The training is fairly fast. Training on CNNDM with the teacher model (i.e., `BART`) is most time-consuming. It takes about 45 minutes for one epoch, and we need 6 epochs in total.

4.3 Evaluations

We evaluate the quality of different summarization systems using ROUGE. On CNNDM and XSum datasets, we report full-length F1 based ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) scores. Following Durrett et al. (2016); Liu and Lapata (2019), we report limited-length recall based ROUGE-1, ROUGE-2, and ROUGE-L, where generated summaries are truncated to the lengths of gold summaries. All ROUGE scores are computed using the `ROUGE-1.5.5.pl` script [4].

Summaries generated by abstractive models may be ungrammatical or unfaithful to the original document. Additionally, we also measure the quality of generated summaries by eliciting human judgements. We randomly sample 50 documents from the test set of CNNDM. 12 annotators are invited (they are either native English speakers or graduate students with IELTS test score over 6.5). In the evaluation, participants are presented with a document and a list of outputs by different models. First, they are asked to evaluate the summaries on three dimensions: *fluency* (is the summary grammatically correct?), *faithfulness* (is the summary faithful to the original document?), and *coverage* (does the summary coverage important information of the document?). Then, they are asked to rank the summaries from best to worst as a way of determining the overall quality of summaries. Each document is ensured to be annotated by 3 different subjects.

4.4 Results

Our main results are shown in Table 2. The first block includes several recent abstractive summarization models based on large pre-trained Transformers. BERTSUM (Liu and Lapata, 2019) employs BERT (Devlin et al., 2019) as its encoder and uses randomly initialized decoder. T5 (Raffel et al., 2020), PEGASUS (Zhang et al., 2020) and BART (Lewis et al., 2020) are three popular large Seq2Seq Transformer models with different pre-training objectives. Our own fine-tuning version of BART (BART (ours)) is comparable or slightly better than the original reported BART results, and we use it as the teacher model on the three datasets.

The second block presents results of student models. Shleifer and Rush (2020) compare pseudo-labeling (BART-PL), knowledge distillation using both output and intermediate layers (BART-KD) as well as shrink and fine-tuning (BART-SFT) methods. They also use BART as teacher models. Note their settings of student models are `BART 12-6` on CNNDM and `BART 12-3` on XSum.

Results of our `BART 12-3` and `BART 12-6` student models are in the third and fourth block. We present results of students trained with gold labels (Gold) and regular pseudo labels (Regular) as well as pseudo labels with higher and random attention temperatures ($\text{PLATE}^{\text{B12-3}}_{\lambda=1.5}$, $\text{PLATE}^{\text{B12-3}}_{\lambda=2.0}$ and $\text{PLATE}^{\text{B12-3}}_{\text{rnd}}$). $\text{PLATE}^{\text{B12-3}}_{\lambda=1.5}$ means that the student uses attention temperature coefficient $\lambda = 1.5$ with architecture setting `BART 12-3`. $\text{PLATE}^{\text{B12-3}}_{\text{rnd}}$ means that we use random attention temperature of $\lambda \sim U[1.0, 2.0]$. We observe that using pseudo-labeling methods with higher attention temperatures consistently improves over its counterpart with normal attention temperatures (Regular) across all three datasets, and the differences between them are almost always significant measured with the ROUGE script [5] (see details in Table 2). Interestingly, our student models $\text{PLATE}^{\text{B12-3}}_{\lambda=2.0}$ and $\text{PLATE}^{\text{B12-6}}_{\lambda=2.0}$ outperform all models in comparison (including student models and even the teacher model) on CNNDM. Our best performing student model $\text{PLATE}^{\text{B12-3}}_{\lambda=1.5}$ outperforms BART-PL, BART-SFT, and BART-KD on XSum. Meanwhile, our method is conceptually simpler and can further be combined with their methods with additional train-

[4] with -c 95 -r 1000 -n 2 -a -m arguments.

[5] The script uses bootstrap re-sampling technology (Davison and Hinkley, 1997) to compute the 95% confidence interval following Lin (2004).

Model/Dataset		CNNDM			XSum			NYT		
		R1	R2	RL	R1	R2	RL	R1	R2	RL
Teacher										
BERTSUM (Liu and Lapata, 2019)		42.13	19.60	39.18	38.81	16.50	31.27	49.02	31.02	45.55
T5-11B (Raffel et al., 2020)		43.52	21.55	40.69	–	–	–	–	–	–
PEGASUS (Zhang et al., 2020)		44.17	21.47	41.11	47.21	24.56	39.25	–	–	–
BART (Lewis et al., 2020)		44.16	21.28	40.90	45.14	22.27	37.25	–	–	–
BART (ours)		44.71	21.52	41.44	45.50	22.26	36.98	55.41	36.59	51.11
Student										
BART-PL (Shleifer and Rush, 2020)		–	19.93	–	–	21.38	–	–	–	–
BART-KD (Shleifer and Rush, 2020)		–	20.95	–	–	21.63	–	–	–	–
BART-SFT (Shleifer and Rush, 2020)		–	21.21	–	–	21.08	–	–	–	–
BART 12-3	Gold	44.28	21.31	41.18	44.33	21.60	36.73	54.75	35.52	50.56
	Regular	43.65	21.10	40.40	44.40	21.63	36.44	53.82	35.12	49.45
	PLATE$_{\lambda=1.5}$	44.54*	21.70*	41.41*	**44.40**	**21.92**	**36.92***	54.47*	35.65	50.39*
	PLATE$_{\lambda=2.0}$	**44.65***	**21.78***	**41.71***	43.50	21.45	36.47	**54.96***	**35.72**	**51.05***
	PLATE$_{\text{rnd}}$	44.27*	21.50*	41.15*	44.21	21.70	36.81*	54.60*	35.70	50.53*
BART 12-6	Gold	44.00	21.08	40.76	44.88	21.75	36.72	55.07	35.91	50.69
	Regular	44.00	21.08	40.29	44.87	21.65	36.47	53.85	35.08	49.36
	PLATE$_{\lambda=1.5}$	44.29*	21.57*	41.13*	**45.13**	22.07*	**37.13***	54.41*	35.61*	50.29*
	PLATE$_{\lambda=2.0}$	**44.84***	**21.95***	**41.77***	44.51	21.79	36.92*	**55.07***	**35.92***	**51.05***
	PLATE$_{\text{rnd}}$	44.38*	21.65*	41.27*	45.00	**22.09***	37.09*	54.74*	35.88*	50.66*
BART 12-12	Regular	43.58	21.14	40.33	44.55	21.42	36.01	54.36	35.74	49.97
	PLATE$_{\lambda=1.5}$	44.72*	21.88*	41.55*	**45.22***	**22.30***	**37.22***	54.90	36.17	50.84*
	PLATE$_{\lambda=2.0}$	**45.08***	**21.98***	**42.07***	44.76	22.06*	37.09*	**55.70***	**36.28**	**51.70***
	PLATE$_{\text{rnd}}$	44.65*	21.80*	41.53*	44.60	21.86*	36.69*	55.15*	**36.28**	51.11*
Transformer	Gold	40.29	17.49	36.71	29.04	9.21	22.18	49.44	29.04	45.07
	Regular	41.00	18.35	37.65	**30.19**	9.79	22.88	49.97	31.00	45.88
	PLATE$_{\lambda=1.5}$	**41.19**	18.33	**38.01***	29.40	**10.11***	**22.95***	50.21	**31.14**	46.25
	PLATE$_{\lambda=2.0}$	41.15	**18.41**	38.00*	28.56	10.02*	22.83*	**50.35**	30.75	**46.39**

Table 2: Results of various models on CNNDM, XSum, and NYT datasets. ROUGE scores on CNNDM and XSum are F1 based and ROUGE scores on NYT are limited-length recall based. BART (ours) is our own implementation of BART fine-tuning. * indicates the model significantly outperforms the regular pseudo-labeling model (Regular).

ing objectives.

In Section 3.3, we also propose a variant of our method, which employs random attention temperatures (PLATE$_{\text{rnd}}$ in Table 2). We can see that though random temperature based method is not as good as our best fixed-temperature method, it in general produces decent results. Therefore, we recommend using this method when the computing budget is limited. Note that we also tried more extreme λ values as shown in Appendix B, and we find the value of 1.5 or 2.0 works better than others.

In the fifth block, we additionally conduct self-distillation experiments, which is not the focus of this work. Our method improves the teacher model on CNNDM; ROUGE-2/L scores are improved on XSum; while on NYT, there are improvements on ROUGE-1/L.

Results with the Transformer student (the sixth block) follow a similar trend, although the improvements are smaller. It may because the modeling power of Transformer without pre-training is not large enough to effectively model the differences in pseudo labels. It is also interesting to see that students distilled with pseudo-labeling do improve gold label based students using randomly initialized Transformer, but not with pre-trained models (i.e., BART 12-6 and BART 12-3), which may also be due to the strong modeling power of large pre-trained Transformers.

	Ref	Regular	PLATE$^{\text{B12-6}}_{\lambda=1.5}$	PLATE$^{\text{B12-6}}_{\lambda=2.0}$
rank	2.4	2.1	2.4	2.7*

Table 3: Human Evaluation on CNNDM dataset. * means significantly better than Regular.

Human evaluation We randomly sample 50 documents from the test set of CNNDM. We compare our best student model PLATE$^{\text{B12-6}}_{\lambda=2.0}$ against the

Attention Setting	R1	R2	RL
$\lambda_{\text{enc}} = \lambda_{\text{cross}} = \lambda_{\text{dec}} = 2.0$	45.65	22.59	42.60
– with $\lambda_{\text{enc}} = 1.0$	45.65	22.57	42.55
– with $\lambda_{\text{cross}} = 1.0$	44.45	21.52	41.22
– with $\lambda_{\text{dec}} = 1.0$	45.08	22.25	42.02

Table 4: Effects of re-scaling attention temperatures for encoder self-attention, decoder self-attention, and decoder cross-attention on the validation set of CNNDM.

Method	R1	R2	RL
Sampling (Edunov et al., 2018)	43.70	20.83	40.56
Nucleus Sampling	43.86	20.95	40.68
Output Layer $T = 0.5$	43.80	21.20	40.59
Regular	44.00	21.08	40.29
PLATE$_{\lambda=2.0}$ (Ours)	44.84	21.95	41.77

Table 5: Comparison with sampling and output layer temperature based distillation methods.

regular pseudo-labeling model (Regular), another model PLATE$^{\text{B12-6}}_{\lambda=1.5}$ and human reference (Ref). We ask human judges to rank the outputs of these models from best to worst. We convert the ranks to rank ratings (rank i to $5 - i$) and further conduct student t-test on these ratings. As shown in Table 3, PLATE$^{\text{B12-6}}_{\lambda=2.0}$ obtains the best ranking score and the difference between PLATE$^{\text{B12-6}}_{\lambda=2.0}$ and the regular pseudo-labeling based method Regular is significant ($p < 0.05$), which indicates our proposed method PLATE indeed produces better summaries.

Ablation study In a Transformer, there are three types of attention modules (i.e., encoder self-attention, decoder self-attention and decoder cross-attention), and we can scale attention temperatures for all of them or some of them. Let λ_{enc}, λ_{cross}, and λ_{dec} denote the attention temperature coefficient of the encoder self-attention module, the decoder cross-attention module, and the decoder self-attention module, respectively. As shown in Table 4, using large attention temperature coefficients (2.0) for all three types of attention modules leads to the best result. When setting the coefficient of the cross attention module to $\lambda_{\text{cross}} = 1.0$, the ROUGE scores drop most. Perhaps this is not surprising, since cross attentions are directly related to the selection of document contents for summarization. Besides, the attention temperature of the decoder self-attention is also crucial but not as important as the cross-attention (see the fourth row).

Comparison with sampling and tuning output layer temperature Sampling based methods can produce more diverse and richer outputs than its beam search based counterpart and has been proven useful in back translation (Edunov et al., 2018). We implement the sampling method in Edunov et al. (2018) and Nucleus Sampling (Holtzman et al., 2019), a more advanced sampling method, to generate pseudo labels for distillation. We use the `BART 12-6` as the student model, and the distillation results on CNNDM are in Table 5. As can be seen, both of the sampling based methods above perform worse than the regular beam search based pseudo-labeling method (Regular), let alone ours. Besides the attention temperatures, we can also tune the temperature T in the decoder output softmax layer. With a proper T (i.e., $T = 0.5$) during pseudo label generation, the resulting student model slightly outperforms the baseline student model with regular pseudo labeling method on ROUGE-2/L (see Table 5), but worse than PLATE$_{\lambda=2.0}$. More results with different Ts are in Appendix C.

4.5 Analysis

Why does our distillation method work? To answer this question, we first try to analyze the reasons from both the external characteristics of the summaries generated by the teacher model and the internal characteristics of the teacher's attention mechanism. Then, we will give an in-depth explanation.

Length and novel n-grams We first analyze the pseudo summaries generated by the teacher models. We calculate novel n-grams and lengths of generated summaries. Note that if an n-gram appears in the summary, but not in the original document, we call it a novel n-gram. Proportions of novel n-grams are used to measure the abstractiveness of summaries (See et al., 2017; Liu and Lapata, 2019). As shown in Table 6, when using a larger λ, pseudo summaries are shorter [6] and contain a larger portion of novel n-grams. It indicates that the teachers can produce more concise and abstractive summaries, which matches the goal of abstractive summarization. *Are these pseudo summaries of good quality?* The performance of the teacher with different attention temperatures on CNNDM test

[6] We also try changing the length penalty during teachers' inference to make pseudo summaries shorter, but we find this method does not help summarization distillation (see Appendix D for more details).

	λ Setting	CNNDM gold	CNNDM 1.0	CNNDM 1.5	CNNDM 2.0	XSum gold	XSum 1.0	XSum 1.5	XSum 2.0	NYT gold	NYT 1.0	NYT 1.5	NYT 2.0
Average Length													
Teacher	Avg. Len.	48.03	64.78	56.81	**52.16**	21.10	20.33	17.28	**15.66**	78.61	105.83	88.58	**79.05**
Student	Avg. Len.	67.51	82.31	73.10	**65.92**	21.01	22.46	18.69	**16.84**	92.61	109.78	98.16	**88.52**
Novel n-grams Ratio(%)													
Teacher	1-gram	25.24	7.89	9.15	**12.56**	46.78	38.68	39.05	**39.33**	12.96	4.04	4.34	**6.25**
	2-grams	61.08	23.60	27.38	**36.81**	87.83	80.50	81.91	**82.70**	45.90	22.54	23.14	**28.95**
	3-grams	77.49	35.43	40.54	**52.77**	97.17	93.09	94.27	**94.91**	65.12	39.20	39.88	**46.93**
	4-grams	85.13	44.10	49.66	**62.56**	99.08	96.78	97.64	**98.07**	75.21	51.09	51.63	**58.36**
Student	1-gram	23.55	4.58	5.07	**6.56**	46.80	37.33	38.01	**38.07**	10.36	3.46	3.37	**3.64**
	2-grams	58.52	15.16	16.64	**21.40**	87.89	78.74	80.56	**81.28**	41.16	21.21	20.50	**21.93**
	3-grams	75.50	24.36	26.58	**33.67**	97.21	91.99	93.55	**94.18**	60.65	37.60	36.67	**38.71**
	4-grams	83.49	31.70	34.36	**42.74**	99.12	96.10	97.25	**97.70**	71.48	49.56	48.47	**50.56**

Table 6: Statistics on outputs of teachers and students with different attention temperature coefficient λ. The student models are all with the `BART 12-6` setting. Inference hyper-parameters on the same dataset are the same.

λ	R1	R2	RL
1.0	44.71	21.52	41.44
1.5	44.92	21.72	41.84
2.0	44.38	21.02	41.50

Table 7: ROUGE of teacher models with different attention temperature coefficient λ on test set of CNNDM.

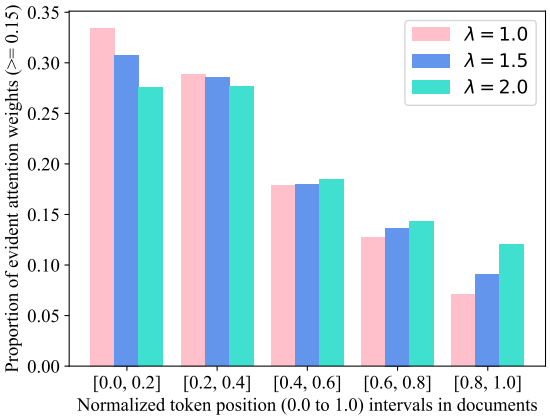

Figure 2: Distributions of *evident* cross attention weights (≥ 0.15) when teachers generate pseudo labels with different attn. temperatures w.r.t. token positions.

set is shown in Table 7. Their results are all decent and close to each other (at least for ROUGE-1 and ROUGE-L). Interestingly, compared with $\lambda = 1.0$, the performance of the teacher with $\lambda = 2.0$ is worse, but the resulting student is much better (see Table 2). Perhaps not surprisingly, the styles of summaries from students are similar with these from their teachers. Concise and abstractive teachers lead to concise and abstractive students (see Table 6). Conciseness and abstractiveness are good properties for summarization, which however may not be the case for other generation tasks such as machine translation. We apply PLATE to the WMT16 (Bojar et al., 2016) English-German translation task and use Transformer-big as the teacher and Transformer-base as the student. With $\lambda = 1.5$, we obtain a BLEU of 27.90, while the result of the regular pseudo-labeling is 27.79 (more details are in Appendix A).

Attention We have shown earlier in Figure 1 that with higher attention temperature, cross-attention modules of a teacher can attend to later parts in documents. We observe that students behave similarly, and we put more cross attention visualization of students in Appendix F. To obtain corpus-level statistics, we further calculate the *evident* cross-attention weight distributions of the teacher when generating pseudo labels on the training set of CNNDM. Note that an attention weight is evident if it is greater than 0.15, and these evident attention weights account for around 15% of all attention weights. Specifically, we normalize the token positions of each document to $(0.0, 1.0]$ and divide the normalized positions into five bins. The mean proportions of *evident* attentions for all bins are shown in Figure 2. Compared to the teacher with normal attention temperature (pink bar), teachers with higher attention temperatures (blue and green bars) attend less on the heading parts of documents while more on the tail parts of documents.

To sum up, teachers with higher attention temperatures can generate more concise and abstractive

pseudo summaries, which makes the teacher provide more *summary-like* pseudo labels to students. High-temperature teachers can alleviate the leading bias problems by providing pseudo labels with better coverage of source documents to students.

More explanation According to the study of Xu et al. (2020b), the prediction entropy correlates strongly with whether the model is copying or generating, as well as where in the sentence the token is (content selection). The decoder tends to copy when the model has a low prediction entropy and generate novel bigrams when the model has a high prediction entropy. They also find that high entropy of attention distribution strongly correlates with the model's high prediction entropy.

Our method with a higher attention temperature makes attention distributions of the teacher model smoother and leads to a higher entropy of attention distributions, which results in a higher prediction entropy. Therefore, the model with higher attention temperature tends to copy less and generate more novel tokens. The conclusion from Xu et al. (2020b) is in accordance with our observation in Table 6.

5 Conclusions

In this work, we propose a simple but effective extension of pseudo-labeling method PLATE for summarization distillation. Experiments on three datasets demonstrate that our method can consistently outperform the vanilla pseudo-labeling method. Further empirical analysis shows that by using our method, teacher models can generate more concise and abstractive summaries. As a result, summaries produced by student models also become more concise and abstractive. In the future, we would like to explore our method to other generation tasks as well as self-training with unlabeled data. We are also interested in combining our method with other distillation methods and extending our method for better teacher model training.

References

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *NIPS*.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. 2017. Sobolev training for neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4281–4290.

Anthony Christopher Davison and David Victor Hinkley. 1997. *Bootstrap methods and their application*. 1. Cambridge university press.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27, Vancouver. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018.

Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Karl Moritz Hermann, Tomás Kociskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

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.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A Smith. 2020. Deep encoder, shallow decoder: Reevaluating the speed-quality tradeoff in machine translation. *arXiv preprint arXiv:2006.10369*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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 and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yang Liu, Sheng Shen, and Mirella Lapata. 2020. Noisy self-knowledge distillation for text summarization. *arXiv preprint arXiv:2009.07032*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3):103–233.

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*. OpenReview.net.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

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.

Sam Shleifer and Alexander M Rush. 2020. Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936. PMLR.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. 2020a. BERT-of-theseus: Compressing BERT by progressive module replacing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7859–7869, Online. Association for Computational Linguistics.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020b. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281, Online. Association for Computational Linguistics.

Sergey Zagoruyko and Nikos Komodakis. 2017. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Model	λ	BLEU
`Transformer-Big` (teacher)	–	28.51
Student		
`Transformer-Base`	1.0	27.79
`Transformer-Base`	1.5	27.90
`Transformer-Base`	2.0	27.85

Table 8: Results of WMT En-De machine translation task on newstest2014. Student models are distilled from pseudo labels generated by the teacher with different attention temperatures (λ).

λ	R1	R2	RL
0.75	43.13	20.60	39.62
1.0	44.00	21.08	40.29
1.5	44.29	21.57	41.13
2.0	44.84	21.95	41.77
2.5	43.99	21.19	41.21
3.0	42.32	19.28	39.67

Table 9: ROUGE scores of `BART 12-6` student models with more values of λ on CNNDM dataset.

T	R1	R2	RL
0.5	43.80	21.20	40.59
1.0	44.00	21.08	40.29
1.5	42.81	20.43	39.56
2.0	42.76	20.34	39.53
Regular	44.00	21.08	40.29
PLATE$_{\lambda=2.0}$ (Ours)	44.84	21.95	41.77

Table 10: Distillation experiment results of changing the softmax temperature in the final decoder layer.

A Experiments of Applying PLATE to the Machine Translation Task

We apply our method on the WMT16 En-De translation task. We use `Transformer-Big` model as the teacher and `Transformer-Base` as the student. Our results on newstest2014 are shown in Table 8. The student models with our method ($\lambda = 1.5$ and $\lambda = 2.0$) slightly outperform the student with regular pseudo-labeling method ($\lambda = 1.0$). Note that the improvement is not as significant as in summarization tasks.

We speculate the reason may be that, unlike summarization, outputs of the machine translation task are relatively fixed. The strength of our method–conciseness and abstractiveness are good properties for summarization but seem not very beneficial to the translation task.

B Experiments of More λ Values

Besides the λ values of 1.5 and 2.0, we also try more values in a broader range. Table 9 shows the distillation performance of `BART 12-6` student models with more values of λ we try on CNNDM dataset (we also include the values of 1.0, 1.5, and 2.0 in table for convenient comparison). As can be seen, both lower and larger λ values are not helpful to the distillation. Though the suitable λ values may vary across datasets, we recommend considering the λ value 1.5 or 2.0 firstly in most cases.

C Experiments of Changing the Softmax Temperature in the Final Decoder Layer

It's a more direct idea to change the softmax temperature in the final decoder layer rather than attention temperatures, namely changing the T in equation 5 to some other values rather than the default value 1.0.

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{5}$$

However, our experiments demonstrate that this method does not help summarization distillation much. We use `BART` teacher models with different softmax temperatures in the final decoder layers to generate pseudo summaries and use the `BART 12-6` as student models. The experiment results are shown in table 10.

D Experiments of Shorter Pseudo Summaries with Smaller Length Penalty

Our method can make pseudo summaries shorter and more abstractive, so one natural idea is that whether just changing the inference hyperparameter length penalty to a smaller value, which can also make pseudo summaries shorter, can benefit abstractive summarization distillation. The experiment results are shown in Table 11, where the teacher is `BART`, and the student is `BART 12-6`. As can be seen from the table, teachers with smaller length penalty (i.e., 1.0 or 0.5) cannot teach better students than the Regular pseudo-labeling or our method.

Length Penalty	Avg. Len.	R1	R2	RL
1.0	64.39	42.96	20.67	39.76
0.5	60.26	43.49	20.89	40.14
Regular	64.78	44.00	21.08	40.29
PLATE$_{\lambda=2.0}$ (Ours)	52.16	44.84	21.95	41.77

Table 11: Distillation results of changing the teacher's inference hyper-parameter length penalty on CNNDM dataset. Avg. Len. represents the average length of the teacher generated pseudo summaries.

E The Example in Section 1

We present the detailed content of the example in Section 1 in table 12.

F Attention Visualization

We present more examples of student models' outputs and cross attention visualization here. The student models are with the `BART 12-6` setting and are trained on CNNDM and the following examples are from the validation set of CNNDM.

Example 1 Table 13 shows system outputs from different student models and Figure 3 illustrates the corresponding cross attention weights of these student models. Compared with the regular pseudo-labeling method (`[Regular]`), the summary generated by our method PLATE$^{\text{B12-6}}_{\lambda=1.5}$ omits the modifier "Nirvana frontman" and "Nirvana bassist" of the person "Kurt Cobain" and "Krist Novoselic", respectively and the resulting summary is shorter and more abstractive. The summary generated by our method PLATE$^{\text{B12-6}}_{\lambda=2.0}$ contains the text "will premiere on HBO on May 4", which is at the end of the source document and included in the reference (i.e., summary worthy), but is ignored by `[Regular]`. It indicates that our method can alleviate the leading bias problem. Figure 3 also shows that PLATE$^{\text{B12-6}}_{\lambda=2.0}$ can access the tail part of the document.

Example 2 The second example is shown in Table 14 (outputs) and Figure 4 (attention visualization). In this example, the source document is relatively long (over 700 words). As shown in Figure 4, the summary generated with the regular pseudo-labeling method `Regular` mainly focuses on the heading part of the source document (around the first 150 words), but our method PLATE$^{\text{B12-6}}_{\lambda=2.0}$ takes into account the tokens in the front, middle and tail of the source document. In Table 14, the summary from PLATE$^{\text{B12-6}}_{\lambda=2.0}$ contains the key sentence "Peter Bergen: Pilots are not different from other people, but they can be careless, lazy, inattentive and reckless", which is similar to the reference sentence "Peter Garrison: Pilots don't exist on different moral plane than the rest of us". The sentence "the human mind is the blackest of boxes" in the reference, which appears at the tail of the source document, is also included in summaries of PLATE$^{\text{B12-6}}_{\lambda=2.0}$. This example again demonstrates that our method can alleviate the leading bias problem and can make the generated summary have better coverage of source documents.

`[Reference]`: Mentally ill inmates in Miami **are housed on the "forgotten floor"** </s> Judge Steven Leifman says most are there as a result of **"avoidable felonies"** </s> While CNN tours facility, patient shouts: **"I am the son of the president"** </s> Leifman says the system is unjust and he's fighting for change.
`[PseudoLBL]`: Mentally ill inmates in Miami **are housed on the "forgotten floor"** of a **pretrial detention facility**. </s> Judge Steven **Leifman says about one-third of all people in Miami-Dade county jails are mentally ill**. </s> He says **they face drug charges or charges of assaulting an officer**, which are **"avoidable felonies"** </s> He says **the arrests often result from confrontations with police**, which **exacerbate their illness.**
`[Smoothed ]`: Mentally ill inmates in Miami **are housed on the "forgotten floor"** </s> Judge Steven Leifman says they are there because of **"avoidable felonies"** </s> He says many of them are in jail for drug or assault charges. </s> He says the system is unjust and he's trying **to change it.**

Table 12: Examples of reference summary (`[Reference]`), pseudo summary from the teacher model (`[PseudoLBL]`) and pseudo summary from the teacher with smoothed attention (`[Smoothed ]`). Text spans in **bold** are copied spans (with more than four words) from the original document.

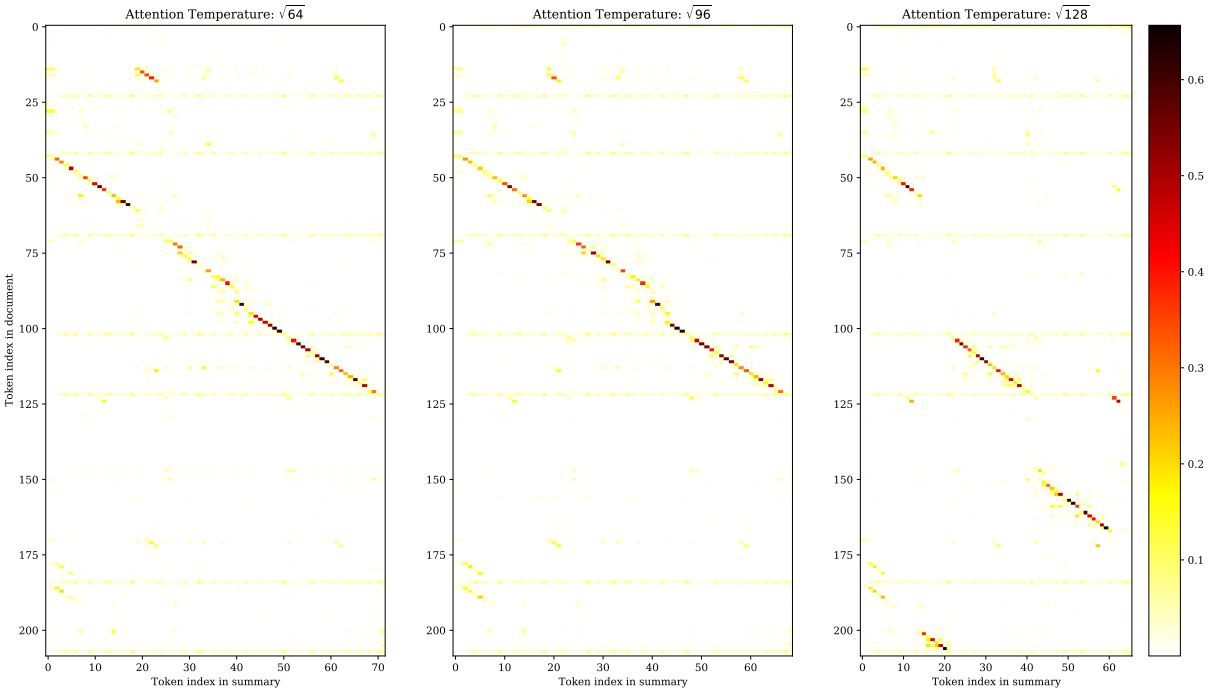

Figure 3: Example 1 of visualization of cross attention weight when the student generate summary with different attention temperatures.

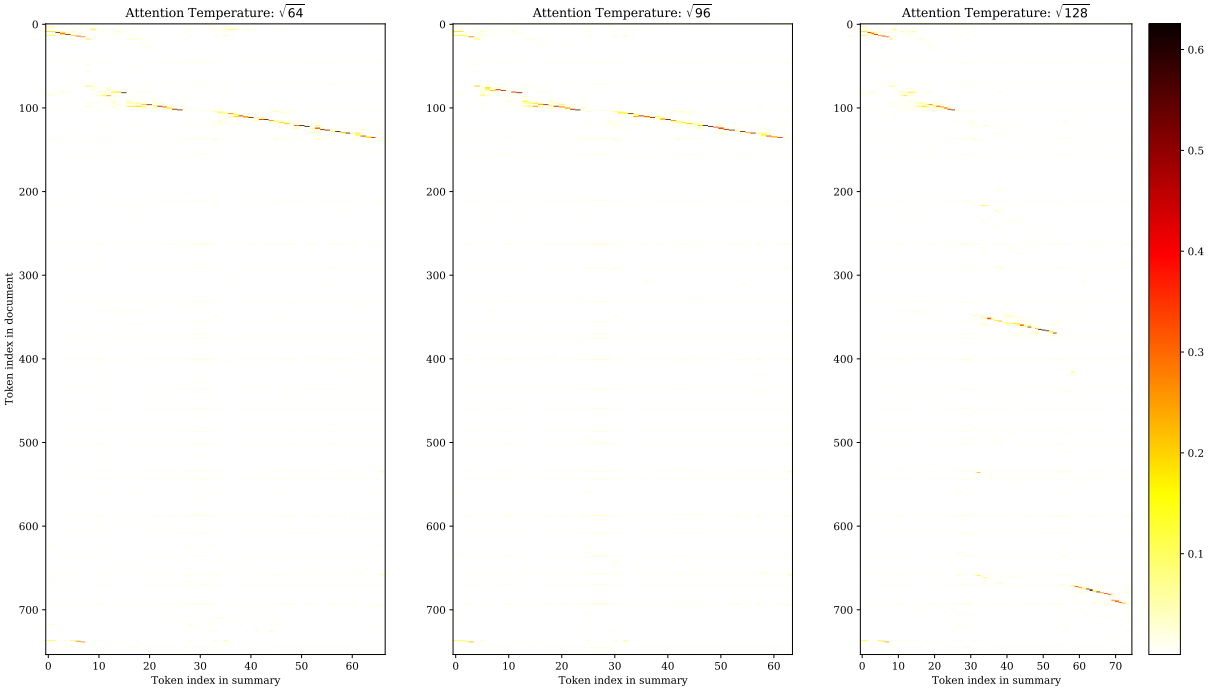

Figure 4: Example 2 of visualization of cross attention weight when the student generate summaries with different attention temperatures.

[Reference]: The trailer surfaced on Wednesday. The documentary is winning raves. It **premieres on HBO on May 4**.
[Regular]: "Montage of Heck" is directed by Brett Morgen and offers an intimate portrayal of **Nirvana frontman Kurt Cobain**. The trailer features interviews with those closest to the performer, including family members, Courtney Love and **Nirvana bassist Krist Novoselic**. A companion book containing art and archival documents from Cobain is being produced to accompany the film.
[PLATE$^{\text{B12-6}}_{\lambda=1.5}$]: "Montage of Heck" is directed by Brett Morgen and offers an intimate portrayal of **Kurt Cobain**. The trailer features bits of interviews with those closest to the performer, including family members, Courtney Love and **Krist Novoselic**. A companion book containing art and archival documents from Cobain is being produced to accompany the film.
[PLATE$^{\text{B12-6}}_{\lambda=2.0}$]: "Montage of Heck" is directed by Brett Morgen and **will premiere on HBO on May 4**. A companion book containing art and archival documents from Cobain is being produced to accompany the documentary. The soundtrack will include "a mind-blowing 12-minute acoustic Cobain unheard track," Morgen says.

Table 13: Example 1 of reference summary ([Reference]), summary generated from student with the regular pseudo-labeling method ([Regular]), and summaries generated from students with PLATE ([PLATE$^{\text{B12-6}}_{\lambda=1.5}$] and [PLATE$^{\text{B12-6}}_{\lambda=2.0}$].

[Reference]: Experts suspect first officer Andreas Lubitz locked pilot out of the cockpit of plane. **Peter Garrison: Pilots don't exist on different moral plane than the rest of us, and the human mind is the blackest of boxes.**
[Regular]: Germanwings first officer Andreas Lubitz is one of a handful of airline pilots who have used their airplanes to combine suicide with mass murder. Frida Ghitis: Why is this thought at once so fascinating and so horrifying? It is because of the incompatibility between what we want to believe about flying and what we now see.
[PLATE$^{\text{B12-6}}_{\lambda=1.5}$]: Andre Lubitz joins the short and infamous list of airline pilots who have used their airplanes to combine suicide with mass murder. Frida Ghitis: Why is this thought at once so fascinating and so horrifying? It is because of the incomp compatibility between what we want to believe about flying and what we now see.
[PLATE$^{\text{B12-6}}_{\lambda=2.0}$]: Germanwings first officer Andreas Lubitz is one of a handful of pilots who have used their airplanes to combine suicide with mass murder. **Peter Bergen: Pilots are not different from other people, but they can be careless, lazy, inattentive and reckless. He says the human mind is the blackest of boxes; no one can peer inside it.**

Table 14: Example 2 of reference summary ([Reference]), summary generated from student with the regular pseudo-labeling method ([Regular]), and summaries generated from students with PLATE ([PLATE$^{\text{B12-6}}_{\lambda=1.5}$] and [PLATE$^{\text{B12-6}}_{\lambda=2.0}$].