Continual Learning for Natural Language Generations with Transformer Calibration

Pang Yang, Dingcheng Li, Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA
{pengyang5612, dingchengl, pingli98}@gmail.com

Abstract

Conventional natural language process (NLP) generation models are trained offline with a given dataset for a particular task, which is referred to as isolated learning. Research on sequence-to-sequence language generation aims to study continual learning model to constantly learning from sequentially encountered tasks. However, continual learning studies often suffer from catastrophic forgetting, a persistent challenge for lifelong learning. In this paper, we present a novel NLP transformer model that attempts to mitigate catastrophic forgetting in online continual learning from a new perspective, i.e., attention calibration. We model the attention in the transformer as a calibrated unit in a general formulation, where the attention calibration could give benefits to balance the stability and plasticity of continual learning algorithms through influencing both their forward inference path and backward optimization path. Our empirical experiments, paraphrase generation and dialog response generation, demonstrate that this work outperforms state-of-the-art models by a considerable margin and effectively mitigate the forgetting.

1 Introduction

Sequence-to-sequence (Seq2Seq) generation has been widely applied in artificial learning (AI) system to deal with various challenging tasks, e.g., paraphrase, dialogue system (Bordes et al., 2016), machine translation, etc. In addition, powerful representation learning (e.g., Transformer) have been used in Seq2Seq models, which have taken the state-of-the-art of generation models to a new level. Generally, nature language generation (NLG) models leverage an encoder to create a vector representation for source inputs, and then pass this representation into a decoder so as to output a target sequence word by word. For example, Bart (Lewis et al., 2019) is such a transformer-based NLG architecture that is equipped with the BERT-type network structure (Devlin et al., 2019) as its encoder and with the GPT-type structure as the decoder.

Despite the remarkable ability on sequence generation, the conventional paradigm aims to learn a Seq2Seq model on the whole available dataset, which limits its ability in accumulating knowledge in continual learning scenario. When switching to a new task from some previously learned ones, the fine-tuned model on the new task sometimes faces a significant performance drop on previous learned data, where such a phenomenon is also referred to as *catastrophic forgetting* (Parisi et al., 2019; Mai et al., 2021; Yin et al., 2021; Li et al., 2022a,b). In contrast, humans and animals exhibit remarkable ability to deal with new tasks by effectively adapting their acquired knowledge without forgetting the previously learned skills. If one desires to build a human-like NLG model, continual learning ability is a necessary skill for achieving this goal.

The existing replay-based continual learning approaches have taken into account of different perspectives of the model training process to remedy the *catastrophic forgetting* dilemma, such as regularizing the parameter change during training (Chaudhry et al., 2018; Parisi et al., 2019), selective memory storage or replay (Aljundi et al., 2019), Bayesian and variational Bayesian training (Kirkpatrick et al., 2017; Nguyen et al., 2018), and task-specific parameterization of the model (Pham et al., 2021; Singh et al., 2020). In this paper,we tackle the problem from a novel angle that is distinct to all the aforementioned attempts, i.e., seeking a better balance between stability and plasticity with neuron calibration. Specifically, we refer to neuron calibration as a process of mathematically adjusting the transformation functions in various layers of transformer-based architecture. In this way, the neuron calibration is able to prioritize both model parameter and feature map that are suitable to new tasks. In detail, our proposed neuron calibration approach regularizes the param-

Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), pages 40 - 49

eter update against catastrophic forgetting via posing a trainable soft mask on the attention and feature maps, which then influences both the model inference process and the model training process through the forward inference path and the backward optimization path.

The contributions of our work are three-fold: (i) we introduce a general and light-weight feature calibration approach to tackle task-incremental continual learning problems where the models are formulated as feed-forward transformer-based function approximations; (ii) we formulate a novel task-incremental learning paradigm to train the calibrated model with an interleaved optimization scheme to mitigate the forgetting issue; (iii) we indicate through extensive empirical experiments that the proposed method could outperform the recent continual learning algorithms on Seq2Seq language generation applications.

2 Related Work

Continual Learning. Existing continual learning methods can be classified into three categories. The *regularization approaches* (Li and Hoiem, 2017; Zenke et al., 2017; Schwarz et al., 2018) impose a regularization constraint to the objective function to mitigate the catastrophic forgetting. The *rehearsal approaches* (Rolnick et al., 2019; Aljundi et al., 2019; Buzzega et al., 2020; Wang et al., 2022) allocate a small memory buffer to store and replay the exemplar from the previous task to consolidate the historical knowledge. The *architectural* approaches (Rusu et al., 2016; Serra et al., 2018; Singh et al., 2020; von Oswald et al., 2020) avoid catastrophic forgetting through approximating the training of the task-specific network and allowing the expansion of the parameters during continual learning. Nonetheless, all these methods are confined to supervised classification problem, which limits their application in real-life problems. Lifelong GAN (Zhai et al., 2019) tackles the generation problem of continual learning and learn task-specific representation on shared parameters. Their method is restricted to image generation tasks and not directly applicable to NLP benchmark datasets.

Continual Language Generation. Few work has been done in continual learning for Seq2seq language generation. The most relevant work is from Mi et al. (2020), which propose a continual learning framework that builds a human-like dialogue system in an incremental learning manner. Specifically, this method combines the memory replay with the regularization technique to address the catastrophic forgetting, and empirically achieves a promising result on the MultiWoZ-2.0 dataset. Nonetheless, their system is specifically designed for the dialogue task and lacks generalization to Seq2Seq tasks. Our method differs from Mi et al. (2020) in terms of the following three points: (i) our method is built upon a neuron calibration approach, where such contribution is orthogonal to that from all the previous works; (ii) our proposed method does not engage any task-specific part; (iii) we do not store the historical exemplar from the episodic memories during training. In addition, our proposed method could be adapted to various seq2seq language generation applications, such as summarization, translation, paraphrases, dialog response generation.

3 Method

3.1 Preliminary

We introduce the setting of online continual learning. Formally, we denote the sequence of training tasks in continual learning as $\{\mathcal{T}_1, \cdots, \mathcal{T}_T\}$. The tasks come and go in an online fashion, and the training data for each task is available only at that time slot. When the new task arrives, the previous task's data is deleted and cannot be used any more. For the t-th task, we denote its training dataset as $\mathcal{D}_t$. The objective of the task is to learn a transformer-based generation model. Our work tackles the natural language generation (NLG)-based continual learning problems and thus the model is typically modeled as a feed-forward transformer with L-blocks (i.e, $\{l_i\}_{i=1}^{L}$), with its corresponding parameters denoted as $\{\theta_i\}_{i=1}^{L}$.

3.2 Transformer Calibration

We introduce a general calibration mechanism to tackle the continue learning problems on Seq2Seq generation, where the models are parameterized by the transformer-based NLG models. By applying neuron calibration, we aim to adapt the transformation function in the deep transformer layers. Our proposed learning paradigm with neuron calibration could perform both model selection and feature selection to effectively avoid catastrophic change on the model parameters while accomplishing a stable consolidation of knowledge among tasks. In this framework, the calibration module is independent from the pre-trained base model in order to

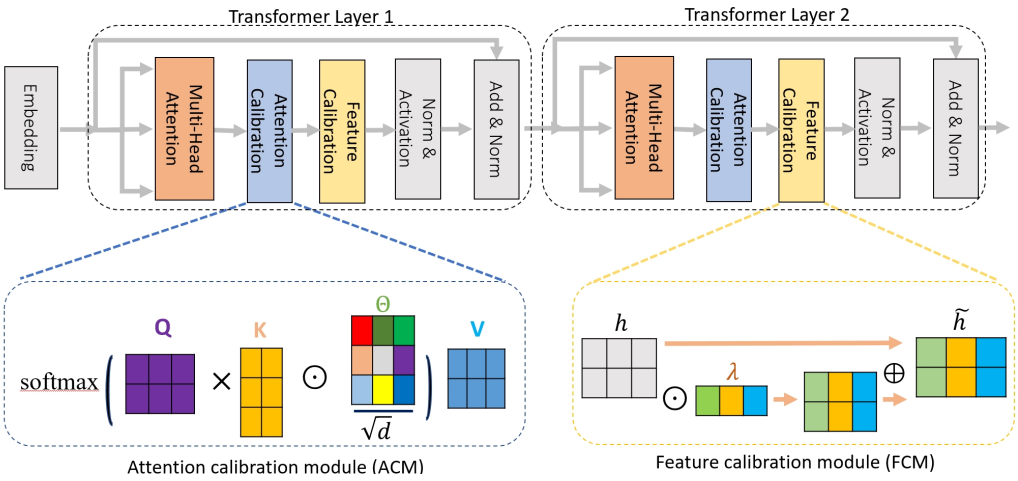

Figure 1: Overview of our proposed transformer calibration for continual learning framework. This method consists of two types of calibration modules: attention calibration module (ACM) and feature calibration module (FCM), which are sequentially applied to the layers in the multi-head attention model (as shown in the figure) to calibrate the attention signals and feature maps, respectively.

preserve the learned knowledge and avoid catastrophic forgetting. Figure 1 provides an illustration of our neuron calibration process.

Formally, we introduce two types of general calibration modules to be applied on the transformer-based NLG models: (i) attention calibration module (ACM) and (ii) feature calibration module (FCM). The attention calibration module learns to scale the attentions of the *transformer function* whereas the feature calibration module learns to scale the *feature map* output from the transformer block. When calibrating the i-th layer of the transformer block, we use A_i to denote its scaled attention function after applying attention calibration (ACM). Meanwhile, we use $\mathbf{h}_i$ and $\widetilde{\mathbf{h}}_i$ to denote the output feature maps before and after applying feature calibration (FCM), respectively.

We first introduce the formulation for ACM. To calibrate the attention, we first define a learnable matrix $\Phi_i \in \mathbb{R}^{N\times N}$, which presents the importance of each pair of words, where N is the maximal number of words in the sentence and a subset of parameters is used according to sentence length. The scale dot-product attention is formulated as:

$$\text{Atten} = \text{Softmax}\left(Q_i K_i^\top \odot \left(\frac{\Phi_i}{\sqrt{d}}\right)\right) V_i \quad (1)$$

where $\odot$ is the element-wise product. As Φ_i is learned across the sequential tasks, the task-aware attention can serve as a task representation instead of traditional task embedding. The overall calibrated attention can be decoupled into two parts: the $QK^\top$ term presents the content-based attention, and $\Phi_i/\sqrt{d}$ term acts as the soft mask for attention calibration. This united design offers more task adaptation by suppressing the unrelated attention values and highlighting the important ones. With the ACM, the calibrator module plays a crucial role during the model training process: at the forward inference path, it scales the value of the attention in the attention block to make prediction; at the backward learning path, it serves as a prioritized weight to regularize the update on parameters.

By applying attention calibration on transformer blocks, the attention function at the i-th layer $\text{Atten}(Q_i, K_i, V_i, \Phi_i)$ is parameterized by Φ_i and produces the output as follows,

$$h_i = \mathcal{F}_{A_i}(h_{i-1}), \text{s.t. } A_i = \text{Atten}(Q_i, K_i, V_i, \Phi_i) \quad (2)$$

The output h_i of the attention function is then processed by a feature calibration module (FCM) to generate the calibrated feature map for that layer. We use $\Omega_{\lambda_i}(\cdot)$ to denote the feature transformation function at the i-th layer, parameterized by λ_i. With FCM, the calibration parameters also interact with the feature map h_i with a multiplicative operation. Specifically, the calibrated feature is computed as:

$$\Omega_{\lambda_i}(h_i) = \text{tile}(\lambda_i) \odot h_i, \quad \lambda_i \in \mathbb{R}^d, h_i \in \mathbb{R}^{N\times d} \quad (3)$$

given the dimension of feature map d.

In the end, the outputs from (2) and (3) get added up in an element-wise manner by a residual connection. This is followed by normalization and activation operations to produce a final output for that layer. In summary, the overall calibration process

for the i-th layer could be formulated as follows,

$$\tilde{h}_i = \sigma\left(\mathcal{LN}\left(\Omega_{\lambda_i}\left(\mathcal{F}_{A_i}\left(h_{i-1}\right)\right) \oplus \mathcal{F}_{A_i}\left(h_{i-1}\right)\right)\right), \tag{4}$$

where $\mathcal{LN}(\cdot)$ denotes the layer normalization, $\oplus$ denotes an element-wise addition operator, and $\sigma(\cdot)$ is an activation function. Then $\tilde{h}_i$ is sent as input to the $i+1$-th layer in the feed-forward network. All the aforementioned calibrator parameters are initialized with a value of 1 at the start of training. We illustrate an example case of applying the calibration on a transformer-based model in Figure 1.

3.3 Learning Calibration Parameters

We propose an interleaved learning paradigm to train the calibrated transformer model. In the training procedure, we aim to exploit the training of the calibrator parameters to mitigate the catastrophic forgetting on the continual learning. Since the 'forgetting' in the training is often attributed to dramatic changes in parameter values, we design the learning objective for the calibrator learning as to regularize the parameter change after accessing the new knowledge not to be biased too much from the model values learned from previous ones.

To formulate the objective function for the calibrated model training, we inherit the *elastic weight consolidation* (EWC) approach proposed in Kirkpatrick et al. (2017) . Specifically, EWC approximates the true posterior distribution for the continual learning parameters by a Gaussian distribution given by the mean from the previous tasks and a diagonal precision from the Fisher information matrix. In this work, we formulate a weight calibration process to prevent the catastrophic change on model parameters. Then we train the calibrator parameters with the following loss function,

$$\mathcal{L}_c = \underbrace{\text{vec}\left(\theta - \theta^t\right)^\top \Lambda_t \text{vec}\left(\theta - \theta^t\right)}_{\text{term (a)}} + \underbrace{\beta \mathcal{L}_t(\Psi, \lambda, \theta)}_{\text{term (b)}} \tag{5}$$

where β is a trade-off parameter, and the operator *vec* $(\cdot)$ stacks the tensor into a vector.

The matrix Λ_t in term (a) are the Fisher information matrix, which is obtained from the data training loss for previous observed tasks, while the $\mathcal{L}_t(\Psi, \lambda, \theta)$ in term (b) is the loss for the current task. The two terms perform the consolidation process to retain the essential parameters towards past knowledge when the base model parameters are trained to absorb new tasks. To consolidate the knowledge on the calibrated model, the Fisher information matrix is computed upon the gradients on calibrated parameters.

3.4 Optimization

We formulate the optimization process to train the calibrated model under an iterative optimization schema, with the parameters from the base model and those from the calibration module being optimized by the loss function (5). During the interleaved optimization process, we first fix θ_t and take gradient steps with regard to $\{\Psi, \lambda\}$ as follows:

$$\Psi_{t+1} \leftarrow \Psi_t - \alpha \bigtriangledown_\Psi \mathcal{L}_c\left((\Psi, \lambda), \theta_t, \mathcal{D}_t\right), \tag{6}$$

$$\lambda_{t+1} \leftarrow \lambda_t - \alpha \bigtriangledown_\lambda \mathcal{L}_c\left((\Psi, \lambda), \theta_t, \mathcal{D}_t\right), \tag{7}$$

Then, we go on to optimize the base model parameter when the inference takes place with the updated base model,

$$\theta_{t+1} \leftarrow \theta_t - \alpha \bigtriangledown_\theta \mathcal{L}_c\left(\theta, (\psi_{t+1}, \lambda_{t+1}), \mathcal{D}_t\right) \tag{8}$$

where α is the learning rate. By employing the calibrated parameterization of the transformer-based network, and optimizing it with the iterative learning scheme, our method achieves the trade-off between new data adaptation and past knowledge consolidation. We present the details in Algorithm 1.

Algorithm 1: Transformer Calibration for Continual Learning Algorithm (TCCL)

Input: Base model θ, calibrator (Φ, λ) learning rate α, trade-off parameter β, training data $\{\mathcal{D}_1^{tr}, ..., \mathcal{D}_T^{tr}\}$, test data $\{\mathcal{D}_1^{te}, ..., \mathcal{D}_T^{te}\}$

Output: Base model $\mathcal{F}_\theta$, calibrator $\mathcal{F}_{(\Phi,\lambda)}$.

function *train_and_eval*

- Randomly initialize θ, Ψ and λ.
- **for** $t \leftarrow 1$ *to* T **do**
  - **for** $b \leftarrow 1$ *to* n_{batch} **do**
    - Observe a batch of data $\mathcal{B}^t = \{x_i, y_i\}_{i=1}^{bs}$ from $\mathcal{D}_t^{tr}$.
    - $\Phi' \leftarrow \Phi - \alpha \nabla_\Phi \mathcal{L}_c(\mathcal{B}^t; \theta, \Phi, \lambda)$
    - $\lambda' \leftarrow \lambda - \alpha \nabla_\psi \mathcal{L}_c(\mathcal{B}^t; \theta, \Phi, \lambda)$
    - $\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_c(\mathcal{B}^t; \theta, \Phi', \lambda')$
    - Compute Λ_t according to $\nabla_\theta \mathcal{L}_c$
  - **for** $te \leftarrow 1$ *to* t **do**
    - Evaluate testing accuracy for the current model on $\mathcal{D}_{1,...,t}^{te}$: $\hat{y}_{1,...,t} \leftarrow \mathcal{F}(\mathcal{D}_{1,...,t}^{te}; \theta_t, \Phi_t, \lambda_t)$

4 Empirical Experiments

We evaluated the proposed algorithm on seq2seq generation tasks. We applied the algorithms on two datasets for seq2seq generation tasks in the continual learning. We also conducted the ablation study with respect to attention calibration and feature calibration to evaluate the robustness and effectiveness of the proposed calibration techniques.

4.1 Application: Paraphrase Generation

Dataset. For paraphrase generation, we train the model over three existing paraphrase datasets, Quora[1], Twitter[2] and Wiki_data (linked-wikitext2)[3], in a sequential manner, where the model observes the three sequential tasks (i.e., datasets) one by one. See Table 1 for Statistics of the datasets.

	train	valid	test
Quora	111,947	8,000	37,316
Twitter	85,970	1,000	3,000
Wiki_data	78,392	8,154	9,324
total	276,309	17,154	49,640

Table 1: Statistics of Dataset on Paraphrase Generation

Experimental Setting. We exploit the SOTA generation model, BART, as the generation model backbone in the continual learning framework. We compare our approach with the following baselines:

- Finetune: for each new task, the model is initialized with the parameters learned from previous observed tasks, and then fine-tuned with data of the current new task.
- Full: the model is trained with all the available instances from three datasets together, which regarded as the up-bounded performance for the continual learning techniques.
- EWC: the EWC (Kirkpatrick et al., 2017) is introduced in the objective function to train the model over the sequential tasks.

For evaluation metrics, we use Bleu4, RougeL and Meteor for the Seq2Seq generation tasks. To measure the forgetting rates of different methods, we basically exploit the model learned on t-th task to evaluate its performance on previous tasks, i.e., $1, \cdots, t-1$ task. We tune the learning rate α from $\{10^{-3}, 10^{-2}, ..., 10^{0}\}$ for both model parameter and calibrator parameter, and trade-off parameter β from $\{0.1, 0.5, 1, 5, 10\}$. Meanwhile, the batch size is set to be $\{128, 256, 512\}$ on all datasets. All training and evaluation experiments are performed using Tesla V100S GPUs. The whole learning process takes around 0.5 GPU day.

4.1.1 Experimental Results

Accuracy Measurement: Table 2 presents the accuracy results in the continual learning setting, where the model is evaluated after the model has been trained on sequential tasks one after another. In the table, the first three models are *independent* baselines trained on either one of three datasets. As expected, model trained on new dataset may suffer the significant performance drop on previous instances, due to the data distribution gap between old and new datasets. For example, twitter includes the short casual text while Wiki_data contains formal academic text.

For the fine-tune, the model is trained in a Quora-Tweeter-Wiki (QTW) order, in which the model is initialized with the model parameters learned on the previous task and then fine tuned over the following task. We observe that finetune results on Quora and Wiki_data are comparable with those when building the model from scratch. In addition, EWC can achieve a better performance than Finetune and independent training over any evaluation metrics on Quora and most metrics on Twitter and Wiki, demonstrating the effectiveness of EWC in continual learning. Nonetheless, our calibration model consistently achieves the best performance across all sequential tasks, demonstrating that the calibration model yields a promising domain adaptation in continual learning.

Forgetting Measurement. Table 3 presents the results when the current models are evaluated on testing data from the previous tasks. The purpose of this experimental setting is to measure the forgetting rate of the models in the sequential training. In the order of QTW, the results are evaluated on Quora after the model is trained on Twitter, as well as on Quora and Twitter after the model is trained on Wiki. Our method is compared with independent baseline, finetune and EWC. Table 3 indicates that our method obtains a less performance drop than Finetune and EWC, with a low forgetting rate. Moreover, after the model is trained on

[1]https://huggingface.co/datasets/quora

[2]https://metatext.io/datasets/paraphrase-and-semantic-similarity-in-twitter-(pit)

[3]https://paperswithcode.com/dataset/wikitext-2

	Quora Test			Twitter Test			Wiki Test		
Models	bleu4*	rougeL	meteor	bleu4*	rougeL	meteor	bleu4*	rougeL	meteor
Quora-trained	30.11	55.85	57.17	2.12	6.13	5.49	4.51	11.21	12.13
Twitter-trained	3.18	11.46	9.01	35.47	57.49	54.57	4.60	9.76	7.50
Wiki_data-trained	22.38	43.44	46.23	9.32	17.93	21.03	42.12	73.86	73.10
Finetune	30.11	55.85	57.17	35.79	56.32	54.93	42.12	73.86	73.10
EWC	30.25	56.16	57.98	33.52	54.41	54.21	42.15	73.53	73.59
Ours	**32.14**	**58.12**	**59.13**	**36.81**	**58.46**	**55.32**	**44.47**	**74.49**	**73.66**
Full	33.99	59.56	61.67	38.56	58.76	56.01	46.86	76.59	75.91

Table 2: Results of model evaluations on QTW setting
(bleu4* denotes a more strict scoring version for the baseline evaluation)

Train: Twitter → Test: Quora

Models	bleu4*	rougeL	meteor
Quora-trained	30.11	55.85	57.17
Finetune	15.80	46.59	**47.31**
EWC	15.63	41.53	46.03
Ours	**15.93**	**46.65**	45.81

Train: Wiki_data → Test: Quora

Models	bleu4*	rougeL	meteor
Quora-trained	30.11	55.85	57.17
Finetune	19.07	51.76	55.95
EWC	19.63	49.35	53.02
Ours	**21.39**	**53.62**	**56.44**

Train: Wiki_data → Test: Twitter

Models	bleu4*	rougeL	meteor
Twitter-based	35.79	56.32	54.93
Finetune	14.09	37.97	45.89
EWC	14.84	38.65	46.33
Ours	**16.62**	**40.25**	**48.44**

Table 3: Results of all the methods when testing new models on previous domains (from 2nd row to the last).

Wiki, the performance on Quora is even improved from the one after trained on Twitter. Moreover, this work outperforms EWC on all the evaluation domains with a noticeable margin, which demonstrates that our calibration module is effective to boost the performance for continual learning via properly regularizing the parameter update against catastrophic forgetting. Overall, the empirical result demonstrates that the calibration mechanism can mitigate the forgetting issue greatly.

Ablation Study. We conduct the ablation study where several simplified versions of the calibration framework are evaluated in order to understand the effects of different components. Specifically, we evaluate the model variants without attention calibration module (i.e., w/o ACM), or feature calibration module (i.e., w/o FCM), or EWC regularization term (i.e., w/o R), and present the comparison result in Table 4. From the table, we can observe that (i) equipped with ACM or FCM, the performance is apparently better than the original backbone since dropping the calibration module ("w/o ACM" and "w/o FCM") would degrade the performance; (ii) EWC regularization is also effective, indicated by the better result than the one without EWC regularization term ("w/o R"). Overall, the results demonstrate that calibrating on latent feature and attention value is a promising direction.

	Quora Test		Wiki_data Test	
Models	bleu4*	meteor	bleu4*	meteor
Finetune	30.11	57.17	42.12	73.10
w/o FCM	33.32	59.32	43.33	73.10
w/o ACM	32.25	58.91	42.15	72.59
w/o R	33.77	59.57	43.51	72.93
Ours	**35.44**	**61.45**	**44.47**	**73.66**

Table 4: Ablation studies on the proposed calibration components and regularizion terms.

Next we aim to investigate the effect of the attention calibration that is performed on three different attentions in the transformer model. Specifically, we equipped the calibration component on either one of the self-attention of encoder, the self-attention of decoder and the encoder-decoder (ED) attention. The comparison results in Table 5 indicate that (i) the self-attention calibration on encoder is more effective to boost the performance; (ii) the calibration on encoder-decoder attention yields

	Quora Test		
Model Variants	bleu4*	rougeL	meteor
Self-Attention (E)	33.31	59.94	59.56
Self-Attention (D)	32.65	58.76	58.34
ED-Attention (D)	34.81	60.55	60.33
Ours (All)	**35.44**	**61.37**	**61.45**

Table 5: Ablation studies of the calibration different attention blocks in language model.

SOURCE	BART	Ours	TARGET
What is the best home workout to reduce waist fat?	How can I reduce my waist fat through a diet?	What is best home remedy for reducing belly fats?	What is best home remedy for reducing belly fats?
What's it like to be in a relationship with a married man?	What is it like for a married man to be in a relationship?	What's it like to be in a relationship with a married man?	What's it like to be in a relationship with a married man?
which provides a conventional sonic underscore to the onscreen action	which provides a sonic underscore to the onscreen action	which provides a conventional sonic underscoring to the onscreen action	which provides a conventional underscore to the onscreen action
Example gymnasium scene's first encounter with Angela	Example gymnasium scene, Angela 's first encounter with Angela	For example, the gymnasium scene, Pfaster 's first encounter with Angela	One example is the gymnasium scene, Lester 's first encounter with Angela.

Table 6: Examples of the generated paraphrases by BART and Ours on QTW data setting.

much better results than other two self-attentions. Overall, the results demonstrate that the attention calibration plays an important role for boosting the performance of the transformer-based generation model.

Case Study. In Table 6, we perform the case studies on paraphrase generation tasks. All examples are results generated by the final model, e.g., the model trained on Wiki_data is used to generate samples on Quora, Twitter, Wiki_data. Among the four examples, the first two is from Quora, and the others from Wiki_data. We compare our generated sentence with ones from BART backbone. From the table, we observe that our method has a better generation on all four cases. In those generation samples, the colored parts are key words. Yet, BART model either fails to generate those key words or creates the examples of false causality. In contrast, our method is able to generate key words in all cases with correct word relations.

4.2 Application: Dialog Response Generation

Dataset. The proposed model is evaluated on the dialog response generation task using the MultiWoZ-2.0 dataset (Budzianowski et al., 2018), which contains 6 domains (Attraction, Hotel, Restaurant, Booking, Taxi and Train) and 7 DA intents ("Inform, Request, Select, Recommend, Book, Offer-Booked, No-Offer"). We follow the setting (Mi et al., 2020) to generate the train/validation/test splits of MultiWoz. The details of the dataset is present in Table 7.

Experimental Setting. To evaluate the method performance, we exploit the slot error rate (SER) and BLEU4 score as the evaluation metrics. The lower value of SER indicates a better performance. To estimate the forgetting rate, the above metrics are reported in two continual learning settings (Kemker et al., 2018): $\Omega_{all} = \frac{1}{T}\sum_{i=1}^{T}\Omega_{all,i}$ and $\Omega_{first} = \frac{1}{T}\sum_{i=1}^{T}\Omega_{first,i}$, where T is total number of tasks in the sequential order. $\Omega_{all,i}$ is the test performance on all the tasks evaluated by the model learned with the i-th task, while $\Omega_{first,i}$ is the test result on the first task after the i-th task has been learned.

Domain and Intents of MultiWoZ-2.0 Data			
Domains	#. Total	Intents	#. Total
Attraction	8,823	Inform	28,700
Hotel	10,918	Request	7,621
Restaurant	10,997	Select	865
Booking	8,154	Book	4,525
Taxi	3,535	Recommend	3,678
Train	13,326	Offer-Booked	2,099
		No-Offer	1,703

Table 7: Statistics on the Dialog Response dataset

Our work exploits the well-known seq2seq generation model, conditional variational encoder (CVAE) as the backbone model, and the proposed model is compared with the following baselines:

a) **Finetune**: the model trained from previous observed tasks is used to be fine-tuned with data of the current new task.
b) **Full**: this model is trained with the data from current tasks and all historical tasks together.
c) **ARPER** (Mi et al., 2020): the model introduces memory replay and adaptive regularization together to mitigate the catastrophic forgetting issue.
d) **ER**: the model with the chosen exemplars that best approximate the mean DA vector (Rebuffi et al., 2017).

For CVAE, we equipped the feature calibration module on the backbone, due to no attention on the CVAE. In the following experiment, we follow

the setting (Mi et al., 2020) and utilize the selected exemplars to compute the Fisher information as in the function (5).

4.2.1 Comparison Result

We conduct comparison experiments with baselines with various number of exemplars. The first one is that all methods do not use any exemplars. The reason for this comparison is that our proposed method is memory-free, i.e., no memory buffer required to store and replay the exemplar for data rehearsal. In such setting, ARPER reduces to the general regularization technique. Table 8 gives the evidence that without any exemplars, our method achieves a better performance than ARPER in both Ω_{all} and Ω_{first}, with a noticeable margin. We observe that the ARPER severely relies on the exemplars. Without the exemplars, the ARPER suffer a significant performance drop in terms of the accuracy, even poorer than Finetune.

With the increased number of exemplars, our method can obtain a better performance since the fisher matrix in our objective can cumulative the informative data throughout the training process. In addition, ER and APRE are memory-based techniques and are obviously beneficial from the exemplars. Nonetheless, our method can consistently outperform APRER and ER in both settings of 250 exemplars and 500 exemplars. That indicates that our memory-free calibration technique can effectively exploit the exemplar knowledge without the need of data storage for the exemplars.

4.2.2 Dynamic Results in Continual Learning

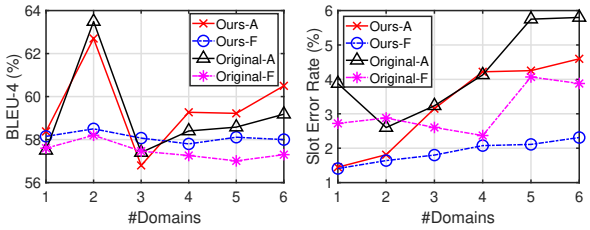

Figure 2: BLEU-4 and SER on all observed domains (solid) and on the first domain (dashed) over the six continually observed domains using 250 exemplars.

Figure 2 presents the comparison results along the six continually observed domains of dialog response. We compare the performance of the calibrated model with the original CVAE backbone. With more tasks continually learned, our method gradually performs better performance than the original backbone. On the first task (dashed lines), the calibrated model outperforms the original one on both metrics. These results illustrate the advantage of our calibration components throughout the entire continual learning process.

Zero exemplars in total

	Ω_{all}		Ω_{first}	
Models	SER	BLEU4	SER	BLEU4
Finetune	64.46	0.361	107.27	0.253
ER	67.23	0.360	105.33	0.181
ARPER	63.54	0.360	102.87	0.192
Ours	**56.90**	**0.395**	**68.60**	**0.258**
ALL	4.26	0.599	3.60	0.616

250 exemplars in total

	Ω_{all}		Ω_{first}	
Models	SER	BLEU4	SER	BLEU4
Finetune	64.46	0.361	107.27	0.253
ER	16.89	0.535	9.89	0.532
ARPER	5.22	0.590	2.99	0.624
Ours	**4.41**	**0.603**	**2.33**	**0.635**
ALL	4.26	0.599	3.60	0.616

500 exemplars in total

	Ω_{all}		Ω_{first}	
Models	SER	BLEU4	SER	BLEU4
Finetune	64.46	0.361	107.27	0.253
ER	12.25	0.555	4.53	0.568
ARPER	5.12	0.598	2.81	0.627
Ours	**4.33**	**0.606**	**2.21**	**0.638**
ALL	4.26	0.599	3.60	0.616

Table 8: Average Results of all the methods when learning six domains using 0/250/500 exemplars. (BLEU4 follows the setting in Mi et al. (2020))

5 Conclusions

We propose an efficient seq2seq generation model with the calibration on the transformer, where a fixed architecture network after calibration can dynamically adjust the function with respect to each individual task. To optimize our method, we further propose a reproductive learning equipped with an iterative optimization objective that trade-off between plasticity and stability. Moreover, our calibration module is very light-weight without introducing any task-specific parameters. Extensive empirical experiments indicate that our approach outperforms the baselines and achieves a promising result. We also indicate that the calibration module and interleaved optimization play a vital role to boost the performance. Finally, extending the calibration module to multi-lingual pre-trained model is a promising future research direction.

References

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. 2019. Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11849–11860, Vancouver, Canada.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. pages 5016–5026.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. 2020. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual.

Arslan Chaudhry, Puneet Kumar Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the 15th European Conference on Computer Vision (ECCV), Part XI*, pages 556–572, Munich, Germany.

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 (NAACL-HLT)*, pages 4171–4186, Minneapolis, MN.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L. Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 3390–3398, New Orleans, LN.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, and Agnieszka Grabska-Barwinska. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Dingcheng Li, Zheng Chen, Eunah Cho, Jie Hao, Xiaohu Liu, Fan Xing, Chenlei Guo, and Yang Liu. 2022a. Overcoming catastrophic forgetting during domain adaptation of seq2seq language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5441–5454.

Dingcheng Li, Peng Yang, Zhuoyi Wang, and Ping Li. 2022b. Power norm based lifelong learning for paraphrase generations. *preprint*.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. 2021. Online continual learning in image classification: An empirical survey. *arXiv preprint arXiv:2101.10423*.

Fei Mi, Liangwei Chen, Mengjie Zhao, Minlie Huang, and Boi Faltings. 2020. Continual learning for natural language generation in task-oriented dialog systems. In *Findings of the Association for Computational Linguistics (EMNLP Findings)*, volume EMNLP 2020, pages 3461–3474, Online Event.

Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. 2018. Variational continual learning. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada.

German Ignacio Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

Quang Pham, Chenghao Liu, Doyen Sahoo, and Steven C. H. Hoi. 2021. Contextual transformation networks for online continual learning. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. iCaRL: Incremental classifier and representation learning. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542, Honolulu, HI.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 348–358, Vancouver, Canada.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018.

Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4535–4544, Stockholm, Sweden.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR.

Pravendra Singh, Vinay Kumar Verma, Pratik Mazumder, Lawrence Carin, and Piyush Rai. 2020. Calibrating cnns for lifelong learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual.

Johannes von Oswald, Christian Henning, João Sacramento, and Benjamin F. Grewe. 2020. Continual learning with hypernetworks. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia.

Zhuoyi Wang, Dingcheng Li, and Ping Li. 2022. Latent coreset sampling based data-free continual learning. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM)*, Atlanta, GA.

Haiyan Yin, Peng Yang, and Ping Li. 2021. Mitigating forgetting in online continual learning with neuron calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10260–10272, virtual.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3987–3995, Sydney, Australia.

Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. 2019. Lifelong GAN: continual learning for conditional image generation. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2759–2768, Seoul, Korea.