Achieving Forgetting Prevention and Knowledge Transfer in Continual Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Continual learning learns a sequence of tasks incrementally with the goal of achieving two main objectives: *overcoming catastrophic forgetting* (CF) and *encouraging knowledge transfer* (KT) *across tasks*. However, most existing techniques focus only on overcoming CF and have no mechanism to encourage KT, and thus do not do well in KT. Although several papers claimed that they could deal with both CF and KT, our experiments show that they suffer from serious CF when the tasks do not have much shared knowledge. In this paper, we study the continual learning of a sequence of natural language processing (NLP) tasks. In NLP, fine-tuning a BERT-like language model using in-domain data is regarded as one of the most effective approaches. However, this approach suffers from serious CF for continual learning. In this paper, we present a novel model called AFK to solve these problems. Experimental results demonstrate the effectiveness of AFK.[1]

1 Introduction

This paper studies continual learning (CL) of a sequence of natural language processing (NLP) tasks in the *task continual learning* (Task-CL) setting. It has two objectives: (i) preventing catastrophic forgetting (CF), and (ii) encouraging knowledge transfer across tasks. (ii) is particularly important because many tasks in NLP share similar knowledge (e.g., sentiment analysis) that can be leveraged to achieve better accuracy. CF means that in learning a new task, the existing network parameters learned for the previous tasks may be modified, which degrades the performance of previous tasks [39]. In the Task-CL setting, the task id is provided for each test case during testing so that the specific model for the task in the network can be applied to classify the test case. Another popular CL setting is *class continual learning*, which does not provide the task id during testing.

Most existing CL papers focus on dealing with CF [41]. There are also some papers that perform knowledge transfer. To achieve both objectives is highly challenging. To overcome CF in the Task-CL setting, we don't want the training of the new task to update the model parameters learned for previous tasks to achieve model separation. But to transfer knowledge across tasks, we want the new task to leverage the knowledge learned from previous tasks for learning a better model (forward transfer) and also want the new task to enhance the performance of similar previous tasks (backward transfer). This means it is necessary to update previous model parameters. This is a dilemma. Although several papers claimed that they could deal with both [23, 36], they were only tested using tasks with strong shared knowledge. When we tested them with tasks that don't have much shared knowledge, they suffer from severe CF (see Sec. 5.4). Those existing papers that focus on dealing with CF do not do well with knowledge transfer as they have no explicit mechanism to facilitate the transfer.

[1]The code has been uploaded as part of Supplementary.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

In NLP, fine-tuning a BERT [10] like language model has been seen as one of the most effective techniques in applications [61, 54]. However, fine-tuning works poorly for continual learning. This is because the fine-tuned BERT for a task captures highly task-specific information [40], which is difficult to be used by other tasks. When fine-tuning for a new task, it has to update the already fine-tuned parameters for previous tasks, which causes serious forgetting (see Sec. 5.4).

This paper proposes a novel neural architecture to achieve both the aforementioned objectives and solve the problem with BERT fine-tuning in the Task-CL setting. The proposed system is called **AFK** (*A*chieving *F*orgetting avoidance and *K*nowledge transfer in continual learning). The key novelty of AFK is its central network, called CBA (*C*ontinual learning *B*ERT *A*dapter). First, CBA can adapt BERT to achieve the power of BERT fine-tuning without causing forgetting of BERT fine-tuning. This idea is inspired by the Adapter-BERT in [18], which was originally proposed for parameter efficient transfer learning. However, CBA is entirely different from the adapter in [18], which is a 2-layer fully-connected network. Second, CBA uses CapsuleNet [17] to model each task and a proposed *transfer routing* algorithm to identify and transfer knowledge across tasks to achieve improved accuracy. Third, CBA learns and uses task masks to protect task-specific knowledge to avoid forgetting. Experimental results show that AFK outperforms existing state-of-the-art baselines.

2 Related Work

Catastrophic Forgetting: Existing work in CL mainly focused on overcoming CF using the following approaches. (1) *Regularization-based approaches,* such as those in [26, 29, 48, 65], add a regularization in the loss to consolidate weights for previous tasks when learning a new task. (2) *Replay-based approaches*, such as those in [43, 35, 5], retain an exemplar set of training data for old tasks and use them to avoid forgetting when learning a new task. The methods in [51, 21, 44, 16] encode training data for previous tasks into data generators and re-generate later when learning a new task to avoid forgetting. (3) *Parameter isolation-based approaches,* such as those in [49, 38, 12], allocate model parameters dedicated to different tasks and mask them out when learning a new task. (4) *Gradient projection-based approaches* [64] ensure the gradient updates occur only in the orthogonal direction to the input of old tasks and thus will not affect old tasks. Tackling CF avoids only model deterioration. These methods typically perform worse than learning each task separately.

Some NLP applications have used CL and dealt with CF. For example, [50] proposed a CL model for slot filling in dialogue systems, and [37] reported a benchmark dataset for dialogue CL. An existing CL method GEM [35] was employed to deal with degrading accuracy in cross-lingual modeling [34]. Other related work includes language modeling [55, 9], language learning [30], sentence embedding [32], machine translation [24], and question answering [14]. More details can be found in the survey [2]. We focus on Task-CL of supervised NLP classification tasks.

Knowledge Transfer: Ideally, learning from a sequence of tasks should *also* allow multiple tasks to support each other via knowledge transfer. In computer vision, Progressive Network [45] does forward transfer but it is for class continual learning (Class-CL). CAT [22] (a Task-CL system) works on a mixed sequence of similar and dissimilar tasks and can transfer knowledge among similar tasks.

Knowledge transfer in this paper is closely related to *Lifelong learning* (LL), which aims to improve the new/last task learning without handling CF [53, 46, 6]. In the NLP area, NELL [4] performs LL information extraction, and several other papers worked on lifelong document sentiment classification (DSC) and aspect sentiment classification (ASC). [7] and [58] proposed two Naive Bayesian methods to help improve the new task learning. [60] proposed a LL approach based on voting. [52] used LL for aspect extraction. [42] and [59] used neural networks for DSC and ASC, respectively. Several papers also studied lifelong topic modeling [6, 15]. However, all these works do not deal with CF.

SRK [36] and KAN [23] claim that they can deal with both CF and knowledge transfer in continual sentiment classification. However, they have two critical weaknesses: (i) Their RNN architectures cannot use adapters to fine-tune a BERT-like language model, which significantly limits their power. (ii) As they were mainly designed for knowledge transfer, they suffer from serious CF (see Sec. 5.4).

Capsule Networks. CapsuleNet has been used in several NLP applications [8, 66]. This work adopt capsules [17] for continual learning. The capsules in our work also have a new routing algorithm, which is instrumental for knowledge transfer.

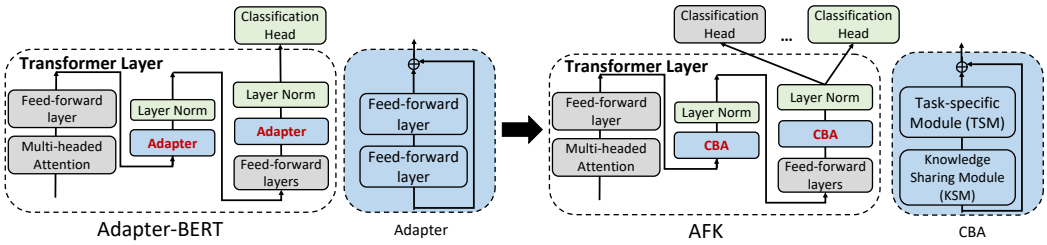

Figure 1: **(A).** Adapter-BERT [18] (left) and its adapter (right), which is added twice to each Transformer layer. **(B).** The proposed system AFK, which replaces the adapter with CBA (continual learning BERT adapter). CBA has two sub-modules: knowledge sharing module (KSM) and task-specific module (TSM). Each of these modules has a skip-connection.

3 AFK Architecture

This section describes the general architecture of AFK. The details about its key component CBA is presented in the next section. Due to its good performance, BERT [10] and its transformer [57] architecture are used as the base in our model AFK. As discussed in Sec. 1, BERT fine-tuning is prone to CF in CL. To solve the problem, we propose to use the Adapter-BERT idea [18].

Adapting BERT for CL. Adapter-BERT was originally designed for parameter efficient transfer learning. It inserts a 2-layer fully connected network (adapter) in each transformer layer of BERT (see Figure 1(A)). During fine-tuning, only the weights from adapters (yellow boxes) and normalization layers (green boxes) are trainable. All other BERT parameters (grey boxes) are frozen. Adapter-BERT has no mechanism for CL. We propose a CL Adapter, called CBA (Figure 1(B)).

Continual learning (CL) BERT Adapter (CBA). CBA employs the capsule network (CapsNet) [17, 47] for the CL purpose. In the classic neural network, a neuron outputs a scalar, real-valued activation as a feature detector. CapsNets replaces that with a vector-output capsule to preserve additional information. A simple CapsNet consists of two capsule layers. The first layer stores low-level feature maps, and the second layer generates the classification probability with each capsule corresponding to one class. CapsNet uses a *dynamic routing* algorithm to make each lower-level capsule to send its output to a similar (or "agreed", computed by dot product) higher-level capsule. This property can already be used to group similar tasks and their shareable features to produce a CL system (see the ablation study in Section 5.4). One of the key ideas of CBA (see Figure 2(A)) is a *transfer capsule layer* with a new *transfer routing* algorithm to explicitly identify transferable features/knowledge from previous tasks to transfer to the new task. Additionally, transfer routing avoids the need for hyper-parameter tuning on the number of iterations of dynamic routing [47] to update the agreements.

4 Continual Learning BERT Adapter (CBA)

The architecture of our *continual learning BERT adapter* (CBA) is shown in Figure 2(A). CBA takes two inputs: (1) hidden states $h^{(t)}$ from the feed-forward layer inside a transformer layer and (2) task ID t, which is required by task continual learning (Task-CL). The outputs are hidden states with features suitable for the t-th task for classification. Inside CBA, there are two modules: (1) *knowledge sharing module* (KSM) for identifying and transferring the shareable knowledge from similar previous tasks to the new task t, and (2) *task specific module* (TSM) for learning task specific neurons and their masks (which can protect the neurons from being updated by future tasks to deal with CF). Since TSM is differentiable, the whole system AFK can be trained end-to-end.

4.1 Knowledge Sharing Module (KSM)

KSM achieves knowledge transfer among similar tasks via a *task capsule layer* (TK-Layer), a *transfer capsule layer* (TR-Layer), and a *transfer routing* mechanism.

4.1.1 Task Capsule Layer (TK-Layer)

Each capsule in the TK-Layer represents a task, and it prepares the low-level features derived from each task (Figure 2(A)). As such, a capsule is added to the TK-Layer for each new task.

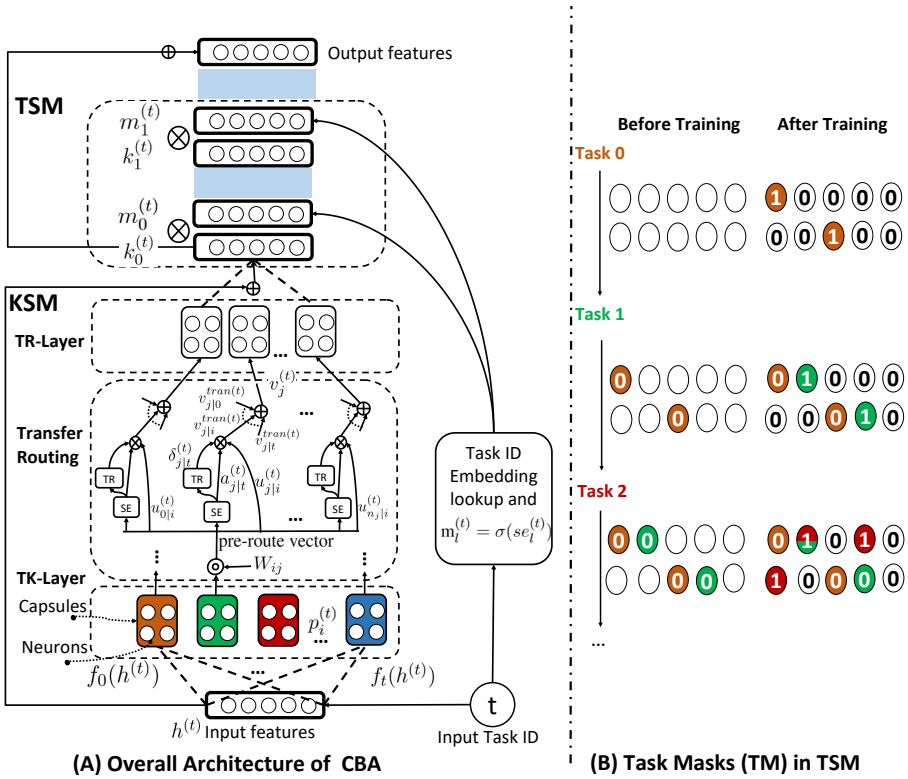

Figure 2: **(A)** Architecture of CBA. **(B)** Illustration of task masking. Cells/neurons in brown are used by task 0, in green by task 1 and in red by task 2. The neuron with two colors is used by two tasks

This incremental growth is efficient and easy because these capsules are discrete and do not share parameters. Also, each capsule is simply a 2-layer fully connected network with a small number of parameters. Let $h^{(t)} \in \mathbb{R}^{d_t \times d_e}$ be the input of CBA, where d_t is the number of tokens, d_e the number of dimensions, and t is the current task. In the TK-Layer, we have one capsule for each task. Assume we have learned t tasks so far. The capsule for the i-th ($i \leq t$) task is

$$p_i^{(t)} = f_i(h^{(t)}), \tag{1}$$

where $f_i(\cdot) = \text{MLP}_i(\cdot)$ denotes a 2-layer fully-connected network.

4.1.2 Transfer Routing and Transfer Capsule Layer

Each capsule in the *transfer capsule layer* (TR-Layer) represents the transferable representation extracted from TK-Layer. As shown in Figure 2(A), *transfer routing* between the lower-level capsules in TK-Layer and high-level capsules in TR-Layer has three components: *pre-route vector generator* (*PVG*), *similarity estimator (SE)* and *task router (TR)*. Given the task capsules in the TK-Layer, we first transform the feature through a trainable weight matrix. We call the output of this transformation the *pre-route vector*. Each SE estimates the similarity between a previous task and the current task using the pre-route vector, resulting in a similarity score for each higher-level capsule. Additionally, each SE is augmented by a TR module, a differentiable task router acting as a gate. This router estimates a binary signal that decides whether to connect or disconnect the current route between the two consecutive capsule layers (i.e. TK-Layer and TR-Layer in CBA). The binary signal estimated by TR can be seen as a differentiable binary attention. Conceptually, each SE and TR pair together learns the connectivity between capsules in a stochastic and differentiable manner, which can be seen as a task similarity-based connectivity search mechanism. This transfer routing identifies the shared features/knowledge from multiple task capsules and helps knowledge transfer across similar tasks. Next, we discuss the *pre-route vector generator*, *similarity estimator* and *task router*.

Pre-route Vector Generator (PVG). We first turns each transfer capsule $p_i^{(t)}$ into a pre-route vector,

$$u_{j|i}^{(t)} = W_{ij} p_i^{(t)}, \tag{2}$$

where $W_{ij} \in \mathbb{R}^{d_s \times d_k}$ is the weight matrix, d_s and d_k are the dimensions of task capsule i (also representing a task) and transfer capsule j, and t is the current task. The number of transfer capsules n_j is a hyperparameter detailed in Section 5.

Similarity Estimator (SE). Since tasks i and t are different, it is crucial to determine what in task i's representation is transferable. Inspired by [63], we use a convolution layer and activation units to compare task i with task t to determine the transferable proportion from the previous task i. In SE, we compute the task similarity as follows:

$$q_{j|t}^{(t)} = \text{MaxPool}(\text{Relu}(u_{j|t}^{(t)} * W_q + b_q)), \tag{3}$$

$$a_{j|i}^{(t)} = \text{MaxPool}(\text{Relu}(u_{j|i}^{(t)} * W_a + f_a(q_{j|t}^{(t)}) + b_a)), \tag{4}$$

where $b_a, b_q \in \mathbb{R}$ are the bias, $W_a, W_q \in \mathbb{R}^{d_e \times d_w}$ are convolutions filters and d_w is the windows size. We extract important features from the current task representation $u_{j|t}^{(t)}$ via the convolution network in Eq. (3). The MaxPool helps remove the insignificant features to generate a fixed-size vector with the size equal to the number of filters n_w. Similarly, we extract important features from the previous task i's representation $u_{j|i}^{(t)}$. Using the important features for the current and previous tasks, we compute a similarity score between them in Eq. (4) with ReLU activation. Note f_a is a 1-layer fully-connected network to match the dimensions. As a result, $a_{j|i}^{(t)}$ indicates how similar the representation of the i-th task is to the current task t. For those tasks with a very low $a_{j|i}^{(t)}$, their representations are less similar to the current task and thus has little transferable knowledge.

Task Router (TR). TR controls which previous task representation should flow to the next layer with the goal of letting only the transferable information to flow. Given the similarity $a_{j|i}^{(t)}$, TR estimates a binary decision signal $\delta_{j|t}^{(t)} \in \{$`0:disconnect, 1:connect`$\}$. We first apply a convolution layer with 2 output channels and 1×1 kernel size to generate un-normalized decision value. To estimate the binary decision, we need to generate a decision chosen from the set of two mutually exclusive and exhaustive events (disconnect and connect). In our work, we adopt the Gumbel-Softmax [20] to help make the TR gate differentiable.

$$\delta_{j|i}^{(t)} = \text{Gumbel_softmax}(a_{j|i}^{(t)} * W_\delta + b_\delta). \tag{5}$$

Given the similarity $a_{j|t}^{(t)}$, binary decision $\delta_{j|t}^{(t)}$ and the pre-route vector $u_{j|t}^{(t)}$, we compute the transferable representation $v_{j|i}^{\text{tran}(t)}$ and final output $v_j^{(t)}$ as follows:

$$v_{j|i}^{\text{tran}(t)} = a_{j|i}^{(t)} \otimes v_{j|i}^{(t)}, \qquad v_j^{(t)} = \sum_{\substack{i=1 \\ \delta_{ij}^{(t)}=1}}^{n+1} v_{j|i}^{\text{tran}(t)}. \tag{6}$$

This makes sure only task capsules for tasks that are salient or similar to the new task are used, and the others task capsules are ignored (and thus protected) to learn more general shareable knowledge. As many NLP applications have similar tasks, such learning of task-sharing features can be very important. Note that in backpropagation, only the similar tasks with connected gate ($\delta_{ij}^{(t)}$=1) are updated, encouraging backward knowledge transfer of similar tasks.

4.2 Task Specific Module (TSM)

Although knowledge sharing is important, it is equally essential to preserve task-specific knowledge of previous tasks to prevent forgetting (CF). To achieve this, we use task masks (Figure 2(B)). Specifically, we first detect the neurons used by each old task and then block off or mask out all the *used* neurons when learning a new task.

The task-specific module consists of differentiable layers (CBA uses a 2-layer fully-connected network). Each layer's output is further applied with a task mask to indicate which neurons should be protected for that task to overcome CF and forbids gradient updates for those neurons during backpropagation for a new task. Those tasks with overlapping masks indicate some parameter sharing. Due to KSM, the features flowing into those overlapping neurons enable the related old tasks to also

improve in learning the new task. Here we borrow the hard attention idea in [49] and leverage the task ID embedding to train the task mask. Further details can be found in Supplementary.

Illustration. The task masking process is illustrated in Figure 2(B), which shows the learning process of three tasks. Before training, those solid cells with a 0 are the neurons that have been used by some previous tasks and should be protected (masked). Those empty cells are free neurons (not used). After training, those solid cells with a 1 are neurons that are important for the current task, which will be used as masks in the future. Those solid cells with a 0 are masked as they are important for previous tasks. Those non-solid cells with a 0 are neurons that are not used so far.

Let us walk through the learning process of the three tasks. After training task 0, we obtain its useful neurons indicated by the 1 entries. Before training task 1, those useful neurons for task 0 are first masked (those previous 1's entries are turned to 0's). After training task 1, two neurons with 1 are used by the task. When task 2 arrives, all used neurons by tasks 0 and 1 are masked before training, i.e., their entries are set to 0. After training task 2, we see that tasks 2 and 1 have a shared neuron (the cell with two colors, red and green), which is used by both of tasks.

5 Experiments

We evaluate AFK using three applications. We follow the continual learning (CL) evaluation method given in [28]. AFK first learns a sequence of tasks. After a task is trained, the training data of the task is no longer accessible. After all tasks are learned, their task models are tested using their respective test sets. In training each task, we use its validation set to decide when to stop training.

5.1 Three Applications and Their Datasets

The first two applications (and datasets) are used to show the knowledge transferability of AFK because their tasks are similar and have shared knowledge. Catastrophic forgetting (CF) is not a major concern for them. The third application (and dataset) is mainly used to test AFK's ability to overcome CF as its tasks are very different and have little shared knowledge to transfer.

1. Document Sentiment Classification (DSC). This application is to classify each full product review into one of the two opinion classes (*positive* and *negative*). The training data of each task consists of reviews of a particular type of product. We adopt the text classification formulation in [10], where a `[CLS]` token is used to predict the opinion polarity.

We employ a set of 10 DSC datasets (reviews of 10 products) to produce sequences of tasks. The products are Sports, Toys, Tools, Video, Pet, Musical, Movies, Garden, Offices, and Kindle [23]. Two experiments are conducted: (1) using small data in each task: 100 positive and 100 negative training reviews per task; (2) using the full data in each task: 2500 positive and 2500 negative training reviews per task [23]. (1) is more useful in practice because labeling a large number of examples is very costly. The same validation reviews (250 positive and 250 negative) and the same test reviews (250 positive and 250 negative) are used in both experiments.

2. Aspect Sentiment Classification (ASC). It classifies a review sentence on the aspect-level sentiment (one of *positive*, *negative*, and *neutral*). For example, the sentence "*The picture is great but the sound is lousy*" about a TV expresses a *positive* opinion about the aspect "picture" and a *negative* opinion about the aspect "sound." We adopt the ASC formulation in [61], where the aspect term and sentence are concatenated via `[SEP]` in BERT. The opinion is predicted on top of the `[CLS]` token.

We employ a set of 19 ASC datasets (review sentences of 19 products) to produce sequences of tasks. Each dataset represents a task. The datasets are from 4 sources: (1) **HL5Domains** [19] with reviews of 5 products; (2) **Liu3Domains** [31] with reviews of 3 products; (3) **Ding9Domains** [11] with reviews of 9 products; and (4) **SemEval14** with reviews of 2 products - SemEval 2014 Task 4 for laptop and restaurant. For (1), (2) and (3), we use about 10% of the original data as the validate data, another about 10% of the original data as the testing data. For (4), we use 150 examples from the training set for validation. To be consistent with existing research [56], sentences with conflict opinions about a single aspect are not used. Statistics of the 19 datasets are given in Table 1. **3. Text classification using 20News data**. This dataset [27] has 20 classes and each class has about 1000 documents. The data split for train/validation/test is 1600/200/200. We created 10 tasks, 2 classes per task. Since this is topic-based text classification data, the classes are very different and have little

Data source	Liu3domain			HL5domain					Ding9domain									SemEval14	
Task/ domain	Speaker	Router	Computer	Nokia6610	Nikon4300	Creative	CanonG3	ApexAD	CanonD500	Canon100	Diaper	Hitachi	Ipod	Linksys	MicroMP3	Nokia6600	Norton	Restaurant	Laptop
Train	352	245	283	271	162	677	228	343	118	175	191	212	153	176	484	362	194	3452	2163
Val.	44	31	35	34	20	85	29	43	15	22	24	26	19	22	61	45	24	150	150
Test	44	31	36	34	21	85	29	43	15	22	24	27	20	23	61	46	25	1120	638

Table 1: Statistics of datasets for ASC. The datasets statistics for DSC and 20News have been described in the text. More detailed data statistics are given in *Supplementary*.

shared knowledge. As mentioned above, this application (and dataset) is mainly used to show AFK's ability to overcome forgetting.

5.2 Baselines

We setup 35 baselines, including both *standalone* and *continual learning* methods.

Standalone (SDL) Baselines: The SDL setting builds a model for each task independently using a separate network. It clearly has no knowledge transfer or forgetting. We have 4 baselines under SDL, **BERT**, **BERT (Frozen)**, **Adapter-BERT** and **W2V** (word2vec embeddings). For **BERT**, we use trainable BERT, i.e., fine-tuning. **BERT (Frozen)** uses frozen BERT with a trainable CNN text classification network [25] on top of it.; **Adapter-BERT** adapts the BERT as in [18], where only the adapter blocks are trainable; **W2V** uses embeddings trained on the Amazon review data in [62] using FastText [13]. For ASC, we adopt the ASC classification network in [63], which takes aspect term and review sentence as input. For DSC and 20News, we adopt the classification network in [10].

Continual Learning (CL) Baselines. CL setting includes 4 baselines with *no forgetting handling* (**NFH**) (corresponding to the 4 standalone baselines), and 25 baselines from 9 state-of-the art *task continual learning* (Task-CL) methods that deal with forgetting (CF). NFH baselines learn the tasks one by one with no awareness of forgetting/transfer.

The 10 state-of-the-art CL systems are: **KAN** [23] and **SRK** [36] are Task-CL methods for document sentiment classification. The rest were designed initially for image classification. Therefore, we replace their original MLP or CNN image classification network with CNN for text classification [25]. **HAT** [49] is one of the best Task-CL methods with almost no forgetting. **CAT** [22] is similar to HAT but can work with a mixed sequence. **UCL** [1] is a recent Task-CL method. **EWC** [26] is a popular regularization-based class continual learning (Class-CL) method, which was adapted for Task-CL by only training on the corresponding head of the specific task ID during training and only considering the corresponding head's prediction during testing. **OWM** [64] is also a Class-CL method, which we also adapt to Task-CL. **L2** [26] is a classic regularization based Class-CL method, which we adapt to Task-CL. **A-GEM** [5] is an efficient version of the replay Task-CL method GEM [35], which penalizes the previous task loss from being increased. **DER++** [3] is a recent replay method using distillation to regularize the new task training and it can function as a Task-CL method.

From the 10 systems, we created 10 baselines using **W2V** embeddings with the aspect term added before the sentence so that the Task-CL methods can take both aspect and the review sentence for ASC; 7 baselines using **Adapter-BERT** (SRK, KAN and CAT's architectures cannot work with adapters); and 10 baselines using **BERT (Frozen)** (which replaces W2V embeddings). The BERT formulation in Section 3 naturally takes both aspect and review sentence in the ASC case.

5.3 Hyper-parameters

Unless otherwise stated, the same hyper-parameters are used in experiments for ASC, DSC and 20News datasets. For the knowledge sharing module (KSM), we employ 2 layers of fully connected network with dimensions 768 in the TK-Layer. We employ 3 transfer capsules. We experimented with 2 to 15 capsules and selected 3 based on the validation data accuracy. For the task specific module (TSM), we use 2000 dimensions as the final and the hidden layers of the TSM. The task ID embeddings have 2000 dimensions. A fully connected layer with softmax output is used as the classification heads in the last layer of the BERT, together with the categorical cross-entropy loss. dropout of 0.5 between fully connected layers. The training of BERT, Adapter-BERT and AFK

Scenarios	Category	Model	ASC		DSC (small)		DSC (full)		20News		20News (FR)	
			Acc.	MF1	Acc.	MF1	Acc.	MF1	Acc.	MF1	Acc.	MF1
Non-continual Learning (SDL)	BERT	SDL	85.84	76.35	78.04	74.17	87.84	86.80	96.49	96.48	—	—
	BERT (Frozen)	SDL	78.14	58.13	73.88	67.97	85.34	80.17	96.49	96.48	—	—
	Adapter-BERT	SDL	85.96	78.07	76.31	71.04	88.30	87.31	96.20	96.19	—	—
	W2V	SDL	77.01	51.89	62.06	53.80	69.57	65.51	94.72	94.72	—	—
Continual Learning (CL)	BERT	NFH	49.60	43.08	73.87	69.44	73.08	71.81	52.50	39.22	24.29	30.52
	BERT (Frozen)	NFH	85.51	76.64	83.12	79.23	61.88	45.79	83.28	81.81	8.76	9.73
	Adapter-BERT	NFH	54.03	44.81	63.76	53.95	64.94	63.40	68.29	61.70	30.59	3.79
	W2V	NFH	82.69	73.56	65.16	57.48	70.40	68.03	90.74	90.59	4.30	4.47
	BERT (frozen)	L2	56.04	38.40	59.17	48.39	69.80	62.63	72.14	65.39	24.57	32.05
		A-GEM	86.06	78.44	59.33	45.94	70.67	61.77	93.31	92.95	4.09	4.48
		DER++	84.27	75.08	72.29	66.28	86.70	85.46	60.44	49.67	10.54	12.16
		KAN	85.49	77.38	77.27	72.34	82.32	81.23	73.07	69.97	15.52	18.87
		SRK	84.76	78.52	78.58	76.03	83.99	82.66	79.64	77.89	12.06	13.97
		EWC	86.37	74.52	82.38	78.41	72.77	65.76	80.26	78.60	3.50	3.03
		UCL	83.89	74.82	80.12	74.13	74.76	69.48	94.65	94.63	0.48	0.48
		OWM	87.02	79.31	58.07	42.63	86.30	85.36	84.54	82.73	13.80	15.81
		HAT	86.74	78.16	79.48	72.78	87.29	86.14	93.51	92.93	2.26	2.89
		CAT	83.68	68.64	67.41	56.22	87.34	86.51	95.17	95.16	0.80	0.81
	Adapter-BERT	L2	63.97	52.43	67.26	62.76	73.03	71.50	69.56	65.50	23.12	27.39
		A-GEM	45.88	28.21	62.89	55.96	71.22	69.94	60.29	50.40	40.22	51.20
		DER++	47.63	35.54	70.52	63.56	59.67	57.82	58.95	49.58	36.39	45.30
		EWC	56.30	49.58	58.23	51.03	62.69	61.51	61.86	53.94	37.79	46.58
		UCL	64.46	36.64	48.30	32.07	57.06	55.86	51.75	36.06	4.70	6.60
		OWM	72.99	66.51	73.97	71.96	85.46	84.57	71.10	66.25	27.38	32.76
		HAT	86.14	78.52	80.83	78.41	88.00	87.26	95.22	95.21	0.33	0.34
	W2V	L2	60.36	39.13	54.34	43.19	57.71	48.00	59.54	54.40	7.83	11.89
		A-GEM	81.33	63.35	69.80	60.07	77.67	70.75	90.72	90.60	3.94	4.08
		DER++	83.27	69.93	77.51	73.13	74.79	66.68	89.28	89.19	4.32	4.42
		KAN	72.06	40.01	57.13	43.75	69.35	64.78	57.92	51.65	20.98	27.02
		SRK	71.01	39.63	64.47	55.93	69.65	65.25	61.07	58.47	7.26	8.81
		EWC	84.16	72.29	64.82	57.20	70.00	65.11	91.86	91.80	2.64	2.71
		UCL	84.41	75.99	56.23	41.34	70.56	67.01	90.61	90.46	4.53	4.70
		OWM	82.70	71.18	53.40	38.44	67.15	65.42	71.97	68.75	24.00	27.56
		HAT	80.83	63.63	62.57	50.83	69.75	65.44	67.73	64.43	26.04	29.70
		CAT	76.28	54.65	55.19	35.28	79.58	75.99	70.38	68.04	24.37	26.95
	AFK (forward)		87.89	80.25	83.75	82.55	89.86	89.16	95.63	95.62	—	—
	AFK		89.47	83.62	84.34	83.29	89.31	88.75	95.25	95.23	0.42	0.43

Table 2: Accuracy (Acc.) and Macro-F1 (MF1), averaged over 5 random sequences. The architectures of SRK, KAN and CAT cannot work with Adapter-BERT. "—" means not applicable. Due to limited space, ***standard deviation***, *running time* and *network size* are given in Supplementary.

follows that of [61]. We adopt $\text{BERT}_{\textbf{BASE}}$ (uncased). The maximum input length is set to 128 which is long enough for both ASC and DSC. We use Adam optimizer and set the learning rate to 3e-5. We use 10 epochs for SemEval datasets and 30 epochs for the other datasets in the ASC application. For DSC, we use 20 epochs. For 20News, we use 10 epochs. These are selected based on the validation results. The batch size is set to 32 for all cases. For all the other CL baselines, we use the code provided by their authors and adapt them for text classification. We also adopt their original parameters, including learning rate, early stopping, and batch size.

5.4 Evaluation Results and Analysis

Since the order of the tasks in a sequence may impact the final results, we randomly sampled 5 task sequences and averaged their results. We compute both accuracy and Macro-F1, where Macro-F1 is the primary metric as the imbalanced classes (in ASC) introduce biases on accuracy. Table 2 gives the average results of all the 19 tasks (or datasets) for ASC and 10 tasks (or datasets) for DSC and 20News over the 5 random task sequences.

Overall Performance. Table 2 shows that AFK clearly outperforms all baselines.

(1). For the standalone (SDL) baselines, BERT (with fine-tuning) and Adapter-BERT perform similarly. W2V and BERT (Frozen) are poorer, which is understandable.

(2). Comparing SDL (standalone learning) and NFH (continual learning with no forgetting handling), we see NFH is much better than SDL for W2V and BERT (Frozen). This indicates ASC and DSC tasks have similarities and thus shared knowledge. Catastrophic forgetting (CF) is not a major issue for W2V and BERT (Frozen). However, for 20News, NFH variants have serious CF. NFH with BERT (fine-tuning) is much worse than SDL and Adapter-BERT, which we explained in Introduction.

(3). Unlike BERT and Adapter-BERT, AFK does very well in both forgetting avoidance and knowledge transfer (outperforming all baselines). For baselines, EWC, UCL, OWM and HAT, although

Model	ASC		DSC (small)		DSC (full)		20News	
	Acc.	MF1	Acc.	MF1	Acc.	MF1	Acc.	MF1
AFK (-KSM;-TSM)	0.5403	0.4481	0.6376	0.5395	0.6494	0.6340	0.6829	0.6170
AFK (-TSM)	0.8312	0.7107	0.7085	0.6759	0.8545	0.8380	0.8275	0.8064
AFK (-KSM)	0.8614	0.7852	0.8083	0.7841	0.8800	0.8726	0.9522	0.9521
AFK (-TR/KSM)	0.8819	0.8155	0.8244	0.8119	0.8831	0.8762	0.9476	0.9469
AFK (dynamic routing)	0.8829	0.8140	0.8201	0.8063	0.7976	0.7651	0.9507	0.9504
AFK	**0.8947**	**0.8362**	**0.8434**	**0.8329**	**0.8931**	**0.8875**	**0.9525**	**0.9523**

Table 3: Ablation experiment results.

they perform better than NFH, they are significantly poorer than AFK as they don't have methods to encourage knowledge transfer for ASC and DSC. KAN and SRK do knowledge transfer but they are weaker than many other CL methods. They perform very poorly for 20News as they have limited ability to overcome CF. CAT works well with large datasets, but is weak for small datasets.

(4). AFK's improvements over SDL variants for DSC (large) is less than for DCS (small). This is understandable because when the training data is large, learning a separate model is already good, and knowledge transfer is less important.

(5). Compared with the SDL results, we can see that AFK has the least forgetting on 20News.

Effectiveness of Knowledge Transfer. For forward transfer (AFK(forward) in Table 2), we use the test accuracy and MF1 of each task when it was first learned. For backward transfer (AFK in Table 2), we use the final result after all tasks are learned. The average results of AFK forward (**AFK(forward)**) and backward (**AFK**) are given in Table 2. We observe that forward transfer of AFK is highly effective for the three datasets with similar tasks. For DSC, the less the data, the more effective is the transfer, which is reasonable Backward transfer improves the accuracy and MF1 of ASC and DSC (small). For DSC (full), it is slightly weaker and for 20News, it is also slightly weaker due to a very small amount of forgetting, but the less than 0.0055 CF is negligible.

Overcoming Forgetting. To validate AFK's effectiveness in dealing forgetting with a sequence of dissimilar tasks, we compute the *Forgetting Rate* [33] as follows FR $= \frac{1}{t-1}\sum_{i=1}^{t-1} A_{i,i} - A_{t,i}$ where $A_{i,i}$ is the forward test accuracy and $A_{t,i}$ is the test accuracy after training the last task t. We average over all tasks except the last one because the last one obviously has no forgetting. We report the forgetting rate FR (averaged over 5 random task sequences) for the 20News data on the two evaluation metrics in the last two columns of Table 2 (the other two datasets are mainly for knowledge transfer). AKF has very low FR values which indicate very little forgetting just like HAT with AFK having the highest accuracy and macro-F1 scores..

Ablation Study. The results of ablation experiments are in Table 3. "-KSM;-TSM" means that the knowledge sharing module (KSM) and the task specific module (TSM) are not used, and we simply deploy the Adapter-BERT. "-KSM" means that the knowledge sharing module (KSM) is not used. "-TSM" means that the task specific module (TSM) is not used. "-TR/KSM" means that the task router (TR) in KSM is not used. We directly send the transferable representation $v_{j|i}^{\text{tran}(t)}$ to the next layer. "dynamic routing" means that we replace our transfer routing (-(SE&TR)/KSM) with dynamic routing [47], which is one of the most popular routing algorithms in capsule networks. Table 3 shows that the full AFK system gives the best results, indicating every component contributes to the model.

6 Conclusion

This paper investigated task continual learning (Task-CL) using BERT fine-tuning. It proposed a novel technique called AFK to enable pre-trained BERT to be used for CL with fine-tuning. The key component of AFK is a Continual learning BERT Adapter (CBA) that adapts BERT for Task-CL. CBA is a capsule network with a new transfer routing mechanism to encourage knowledge transfer among tasks and also to isolate task-specific knowledge to avoid forgetting. Experimental results using three NLP applications showed that AFK markedly improves the performance of both the new task and the old tasks via knowledge transfer and is also effective at overcoming catastrophic forgetting. One limitation of our work is efficiency due to the use of capsules. Capsules try to represent a group of neurons in a vector reflecting properties of an entity. In NLP, an entity is a sentence/document which contains many tokens (e.g. 128) and features (e.g. 768 in $\text{BERT}_{\text{BASE}}$). Grouping them makes the capsule very large (e.g. 128×768), which slows down training.

References

[1] H. Ahn, S. Cha, D. Lee, and T. Moon. Uncertainty-based continual learning with adaptive regularization. In *NIPS*, 2019.

[2] M. Biesialska, K. Biesialska, and M. R. Costa-jussà. Continual lifelong learning in natural language processing: A survey. *arXiv preprint arXiv:2012.09823*, 2020.

[3] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara. Dark experience for general continual learning: a strong, simple baseline. *arXiv preprint arXiv:2004.07211*, 2020.

[4] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka, and T. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, volume 24, 2010.

[5] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny. Efficient lifelong learning with A-GEM. In *ICLR*, 2019.

[6] Z. Chen and B. Liu. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–207, 2018.

[7] Z. Chen, N. Ma, and B. Liu. Lifelong learning for sentiment classification. In *ACL*, 2015.

[8] Z. Chen and T. Qian. Transfer capsule network for aspect level sentiment classification. In A. Korhonen, D. R. Traum, and L. Màrquez, editors, *ACL*, 2019.

[9] Y.-S. Chuang, S.-Y. Su, and Y.-N. Chen. Lifelong language knowledge distillation. *arXiv preprint arXiv:2010.02123*, 2020.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *NAACL-HLT*, 2019.

[11] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, 2008.

[12] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *CoRR*, 2017.

[13] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. In *LREC*, 2018.

[14] C. Greco, B. Plank, R. Fernández, and R. Bernardi. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. *arXiv preprint arXiv:1906.04229*, 2019.

[15] P. Gupta, Y. Chaudhary, T. Runkler, and H. Schuetze. Neural topic modeling with continual lifelong learning. In *International Conference on Machine Learning*, pages 3907–3917. PMLR, 2020.

[16] X. He and H. Jaeger. Overcoming catastrophic interference using conceptor-aided backpropagation. In *ICLR*, 2018.

[17] G. E. Hinton, A. Krizhevsky, and S. D. Wang. Transforming auto-encoders. In *International conference on artificial neural networks*, 2011.

[18] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for NLP. In K. Chaudhuri and R. Salakhutdinov, editors, *ICML*, 2019.

[19] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD*, 2004.

[20] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

[21] N. Kamra, U. Gupta, and Y. Liu. Deep generative dual memory network for continual learning. *CoRR*, 2017.

[22] Z. Ke, B. Liu, and X. Huang. Continual learning of a mixed sequence of similar and dissimilar tasks. In *NeurIPS*, 2020.

[23] Z. Ke, B. Liu, H. Wang, and L. Shu. Continual learning with knowledge transfer for sentiment classification. In *ECML-PKDD*, 2020.

[24] H. Khayrallah, B. Thompson, K. Duh, and P. Koehn. Regularized training objective for continued training for domain adaptation in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 36–44, 2018.

[25] Y. Kim. Convolutional neural networks for sentence classification. In A. Moschitti, B. Pang, and W. Daelemans, editors, *EMNLP*, 2014.

[26] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, 2016.

[27] K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, 1995.

[28] M. D. Lange, R. Aljundi, M. Masana, and T. Tuytelaars. Continual learning: A comparative study on how to defy forgetting in classification tasks. *CoRR*, 2019.

[29] S. Lee, J. Kim, J. Jun, J. Ha, and B. Zhang. Overcoming catastrophic forgetting by incremental moment matching. In *NIPS*.

[30] Y. Li, L. Zhao, K. Church, and M. Elhoseiny. Compositional language continual learning. In *International Conference on Learning Representations*, 2019.

[31] Q. Liu, Z. Gao, B. Liu, and Y. Zhang. Automated rule selection for aspect extraction in opinion mining. In *IJCAI*, 2015.

[32] T. Liu, L. Ungar, and J. Sedoc. Continual learning for sentence representations using conceptors. *arXiv preprint arXiv:1904.09187*, 2019.

[33] Y. Liu, Y. Su, A. Liu, B. Schiele, and Q. Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *CVPR*, 2020.

[34] Z. Liu, G. I. Winata, A. Madotto, and P. Fung. Exploring fine-tuning techniques for pre-trained cross-lingual models via continual learning. *arXiv preprint arXiv:2004.14218*, 2020.

[35] D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. In *NIPS*, 2017.

[36] G. Lv, S. Wang, B. Liu, E. Chen, and K. Zhang. Sentiment classification by leveraging the shared knowledge from a sequence of domains. In *DASFAA*, 2019.

[37] A. Madotto, Z. Lin, Z. Zhou, S. Moon, P. Crook, B. Liu, Z. Yu, E. Cho, and Z. Wang. Continual learning in task-oriented dialogue systems. *arXiv preprint arXiv:2012.15504*, 2020.

[38] A. Mallya and S. Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *CVPR*, 2018.

[39] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. 1989.

[40] A. Merchant, E. Rahimtoroghi, E. Pavlick, and I. Tenney. What happens to BERT embeddings during fine-tuning? *CoRR*, 2020.

[41] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019.

[42] Q. Qin, W. Hu, and B. Liu. Using the past knowledge to improve sentiment classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1124–1133, 2020.

[43] S. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017.

[44] M. Rostami, S. Kolouri, and P. K. Pilly. Complementary learning for overcoming catastrophic forgetting using experience replay. In *IJCAI*, 2019.

[45] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive neural networks. *CoRR*, 2016.

[46] P. Ruvolo and E. Eaton. ELLA: an efficient lifelong learning algorithm. In *ICML*, pages 507–515, 2013.

[47] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. In *NIPS*, 2017.

[48] A. Seff, A. Beatson, D. Suo, and H. Liu. Continual learning in generative adversarial nets. *CoRR*, abs/1705.08395, 2017.

[49] J. Serrà, D. Suris, M. Miron, and A. Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *ICML*, 2018.

[50] Y. Shen, X. Zeng, and H. Jin. A progressive model to enable continual learning for semantic slot filling. In *EMNLP-IJCNLP*, Nov. 2019.

[51] H. Shin, J. K. Lee, J. Kim, and J. Kim. Continual learning with deep generative replay. In *NIPS*, 2017.

[52] L. Shu, H. Xu, and B. Liu. Lifelong learning CRF for supervised aspect extraction. In *ACL*, 2017.

[53] D. L. Silver, Q. Yang, and L. Li. Lifelong machine learning systems: Beyond learning algorithms. In *Lifelong Machine Learning, Papers from the 2013 AAAI Spring Symposium, Palo Alto, California, USA, March 25-27, 2013*, 2013.

[54] C. Sun, L. Huang, and X. Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *NAACL*, 2019.

[55] F.-K. Sun, C.-H. Ho, and H.-Y. Lee. Lamol: Language modeling for lifelong language learning. *arXiv preprint arXiv:1909.03329*, 2019.

[56] D. Tang, B. Qin, and T. Liu. Aspect level sentiment classification with deep memory network. In *EMNLP*, Nov. 2016.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[58] H. Wang, B. Liu, S. Wang, N. Ma, and Y. Yang. Forward and backward knowledge transfer for sentiment classification. In *ACML*, 2019.

[59] S. Wang, G. Lv, S. Mazumder, G. Fei, and B. Liu. Lifelong learning memory networks for aspect sentiment classification. In *IEEE International Conference on Big Data*, 2018.

[60] R. Xia, J. Jiang, and H. He. Distantly supervised lifelong learning for large-scale social media sentiment analysis. *IEEE Trans. Affective Computing*, 8(4):480–491, 2017.

[61] H. Xu, B. Liu, L. Shu, and P. S. Yu. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In J. Burstein, C. Doran, and T. Solorio, editors, *NAACL-HLT*, 2019.

[62] H. Xu, S. Xie, L. Shu, and P. S. Yu. Dual attention network for product compatibility and function satisfiability analysis. In *AAAI*, 2018.

[63] W. Xue and T. Li. Aspect based sentiment analysis with gated convolutional networks. In I. Gurevych and Y. Miyao, editors, *ACL*, 2018.

[64] G. Zeng, Y. Chen, B. Cui, and S. Yu. Continuous learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 2019.

[65] F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017.

[66] W. Zhao, H. Peng, S. Eger, E. Cambria, and M. Yang. Towards scalable and reliable capsule networks for challenging NLP applications. In A. Korhonen, D. R. Traum, and L. Màrquez, editors, *ACL*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We believe that our work does not have any negative societal impacts as we use only public domain datasets.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In supplemental material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In supplemental material
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In supplemental material
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]