Bayes-enhanced Lifelong Attention Networks for Sentiment Classification

Hao Wang [†‡*], **Shuai Wang** [♯*], **Sahisnu Mazumder** [§], **Bing Liu** [§], **Yan Yang** [†], **Tianrui Li** [†]
[†]School of Information Science and Technology, Southwest Jiaotong University
[§]Department of Computer Science, University of Illinois at Chicago
[‡]Zhejiang Lab; [♯]Amazon AI
{cshaowang, shuaiwanghk, sahisnumazumder}@gmail.com
liub@uic.edu; {yyang, trli}@swjtu.edu.cn

Abstract

The classic deep learning paradigm learns a model from the training data of a single task and the learned model is also tested on the same task. This paper studies the problem of learning a sequence of tasks (sentiment classification tasks in our case). After each sentiment classification task is learned, its knowledge is retained to help future task learning. Following this setting, we explore attention neural networks and propose a Bayes-enhanced Lifelong Attention Network (BLAN). The key idea is to exploit the generative parameters of naïve Bayes to learn attention knowledge. The learned knowledge from each task is stored in a knowledge base and later used to build lifelong attentions. The constructed lifelong attentions are then used to enhance the attention of the network to help new task learning. Experimental results on product reviews from Amazon.com show the effectiveness of the proposed model.

1 Introduction

In recent years, deep learning methods have brought sentiment classification to a new height with big data, see a survey paper (Zhang et al., 2018). However, these deep learning methods primarily focus on learning a model from the training data of one task and the learned model is tested on the same task. In this paper, we study the problem of learning a sequence of sentiment classification tasks. This learning setting is called *lifelong learning*. Lifelong learning is a continual learning process where the learner learns a sequence of tasks; after each task is learned, its knowledge is retained and later used to help new/future task learning (Thrun, 1998; Silver et al., 2013; Ruvolo and Eaton, 2013; Mitchell et al., 2018; Parisi et al., 2019). Lifelong learning has been introduced extensively in a book, see (Chen and Liu, 2018). The first lifelong learning method for sentiment classification (called lifelong sentiment classification) was introduced in (Chen et al., 2015). Following Chen et al. (2015), several other lifelong sentiment classification approaches were proposed respectively in (Xia et al., 2017; Wang et al., 2019; Lv et al., 2019; Xu et al., 2020). We will discuss these and other related work in the next section. Among these, (Lv et al., 2019) is a pioneering lifelong sentiment classification method based on deep learning. Following these existing works, we treat the classification in each domain (e.g., a type of product) as a learning task. Thus, we use the terms *domain* and *task* interchangeably throughout the paper. Formally, we study the following problem.

Problem Definition: At any point in time, a learner has learned n sentiment classification tasks. Each task $T_i|_{i=1}^{n}$ has its own data D_i. The learned knowledge is retained in a knowledge base. When faced with a new sentiment classification task T_{n+1}, the knowledge in the knowledge base is leveraged to help learn the new task T_{n+1}. After T_{n+1} is learned, its knowledge is also incorporated into the knowledge base for future task learning.

The above-defined problem is a reformulation from the work of Chen et al. (2015). As can be seen, the main challenges of this problem are 2-fold: (1) *what forms of knowledge should be retained from each task* and (2) *how does the learner use the knowledge to help future task learning*? These are also the

*Work was done while Hao Wang and Shuai Wang were at the University of Illinois at Chicago.

Proceedings of the 28th International Conference on Computational Linguistics, pages 580–591
Barcelona, Spain (Online), December 8-13, 2020

challenges faced by lifelong learning. This paper studies the problem of lifelong learning for sentiment classification. Below let us consider an example in sentiment classification.

"*I bought this camera last month. It works good although a little heavy. Would recommend this.*"

We (humans) can observe that (1) the first sentence shows a neutral sentiment (or no sentiment), (2) the second sentence expresses both positive (i.e., good) and negative (i.e., heavy) sentiments, and (3) the third sentence shows a strong positive (i.e., recommend) sentiment. The overall sentiment polarity of this example is positive. One reason why we can understand these clearly is that we give different focuses/attentions to each word and each sentence. We give strong attentions to the words (**good**, **heavy**, and **recommend**) and **the last two sentences**. Another crucial reason is that we have accumulated a large amount of knowledge in the past, and use it now to help understanding. For example, we know that the word "good" is positive for all products, and the word "heavy" is negative for camera but may be positive for some other products. Such a problem is called *domain generality* and *domain specificity* as introduced in (Li and Zong, 2008). The problem is particularly acute in lifelong learning as lifelong learning needs to tackle a large number of (possibly never ending) domains/tasks. Thus, we motivate and propose to use attention from previous tasks as knowledge to help future task learning.

Attention has become enormously popular as an essential component of neural networks especially for a wide range of natural language processing tasks (Chaudhari et al., 2019). Attention neural networks was first introduced for machine translation (Bahdanau et al., 2014). Since then, a large body of sentiment classification models using attention mechanisms have been arrived — to name a few (Zhou et al., 2016; Yang et al., 2016; Gu et al., 2018; He et al., 2018; Peng et al., 2019; Zhang et al., 2019; Yu et al., 2019). However, none of them is a lifelong sentiment classification method. Most of them learn in isolation as the classic deep learning paradigm learned, although some methods work for multi-domain sentiment classification which focus on leveraging data in source domain to help target domain. So, they are far from sufficient for our goal as we aim at building a model to learn a sequence of sentiment classification tasks and learn them successively. In addition, *whether attention is explainable (or interpretable)* was recently discussed in (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019). Letarte et al. (2018) and Clark et al. (2019) show the importance of attention for neural networks.

In this paper, we introduce a knowledge-enhanced attention model which we call BLAN (as shorthand for Bayes-enhanced Lifelong Attention Network) for learning a sequence of sentiment classification tasks. The idea is that the model parameters of naïve Bayes are first generated from each task. These parameters are then transformed to knowledge, retained in the knowledge base and later used to build lifelong attentions. Frequent Pattern Mining (Agrawal et al., 1994) is used in extracting lifelong attentions. The extracted lifelong attentions are used to enhance attention neural networks. This enhancement is especially useful when the training data of a task is small and thus, attention neural networks do not work well. The proposed lifelong attention mechanism using naïve Bayes and Frequent Pattern Mining is helpful and promising. These will be clear later. The experimental results demonstrate the superiority of the proposed model against multiple baseline methods. In summary, the paper makes the following contributions:

- The paper studies a problem of learning a sequence of sentiment classification tasks, where the tasks are learned incrementally.
- It proposes a Bayes-enhanced Lifelong Attention Network to address the aforementioned problem. The proposed method exploits the generative parameters of naïve Bayes to learn attention knowledge and the knowledge learned from each previous task is retained and later used to build lifelong attention to enhance neural networks for new task learning.
- Experimental results using two Amazon review datasets (involving 20 and 24 types of product domains) show that the proposed method achieves remarkable results.

2 Related Work

Lifelong learning, an advanced machine learning paradigm has been recently researched significantly in a relatively short time, see the book (Chen and Liu, 2018). In brief, lifelong learning aims to imitate

"human learning" process and capability, i.e., *accumulating the knowledge learned in the past and using it to help future learning and problem solving*. Lifelong learning for sentiment classification (called lifelong sentiment classification) was first introduced in (Chen et al., 2015), which is based on optimization considering the knowledge from previous tasks to help future task learning. Following this work, Xia et al. (2017) investigated lifelong sentiment classification based on voting of each individual task classifier. Wang et al. (2019) proposed a heuristic naïve Bayes which achieves forward knowledge transfer to improve any future task learning and also achieves backward knowledge transfer to improve the model of any past task without retraining using the past task training data. Xu et al. (2020) proposed a continuous naïve Bayes learning framework for e-commerce product review sentiment classification by extending the parameter estimation mechanism in naïve Bayes. Lv et al. (2019) introduced a deep learning method for lifelong sentiment classification by jointly training two networks, a feature learning network and a knowledge retention network. Such a method is the first lifelong sentiment classification method based on deep learning. We have compared this deep learning method in our experiments and show that it is inferior to our proposed method. (Shu et al., 2017; Wang et al., 2018a; Wang et al., 2018b) also proposed lifelong learning methods for sentiment classification, but they focus on aspect-based analysis, which is different from document-level sentiment classification (our focus in this paper).

There are currently several research topics that are closely related to lifelong learning — most notably, *multi-task learning* and *transfer learning*.

Multi-task learning vs. lifelong learning: Multi-task learning (Zhang and Yang, 2017) jointly optimizes the learning of multiple tasks. Although it is possible to make it continual, multi-task learning does not retain any knowledge. Lifelong learning is a continual learning process where the learner learns task incrementally, accumulates knowledge from each task and later uses it to help future task learning.

Transfer learning vs. lifelong learning: Transfer learning (Pan and Yang, 2010) uses the source domain labeled data to help target domain learning which has no or little labeled data. Unlike lifelong learning, transfer learning is not continual and has no knowledge retention. The source should be similar to the target (which are normally selected by the user). Domain adaptation is a type of transfer learning by establishing knowledge transfer from a labeled source domain to an unlabeled target domain. Domain adaptation with multiple sources (or called Multi-source Domain Adaption) aims to use the labeled data collected from multiple sources to help target domain learning (Sun et al., 2015; Wu and Huang, 2016). However, these methods are only one-directional, i.e., sources helps target because the target has no or little labeled data.

Our work is also related to knowledge-enhanced attention neural networks. Knowledge-enhanced attention models have been introduced in (Chen et al., 2017; Zhou et al., 2018; Yang et al., 2019; Peters et al., 2019; Chen et al., 2019; Huang et al., 2020). However, they do not concern sentiment classification task. Although there are a large body of attention models for sentiment classification as we mentioned in the previous section, none of them uses enhanced attention mechanisms. They mainly use self-attention learned from the networks. Our proposed attention mechanism is a combination of self-attention and enhanced-attention. We detail our attention mechanism and the proposed model next.

3 Proposed Model

The proposed BLAN model is a hierarchical architecture as shown in Figure 1. The key characteristic of BLAN is the Bayes-enhanced lifelong attention (to build knowledge-enhanced attention). This is done by using the generative model parameters $P_{w_i|c_k}$ of the naïve Bayes, formulated as

$$P_{w_i|c_k} = \frac{\lambda + N_{c_k,w_i}}{\lambda|V| + \sum_{v=1}^{|V|} N_{c_k,w_v}} \tag{1}$$

where w_i is a word, c_k is the positive (+) or negative (-) class in our case, N_{c_k,w_i} is the number of times that word w_i occurs in the training documents of class c_k, $|V|$ is the vocabulary V size, and λ is the smoothing parameter. Below we provide an "*ideal*" case to illustrate how our model uses these parameters to perform Bayes-enhanced lifelong attention in a lifelong learning setting.

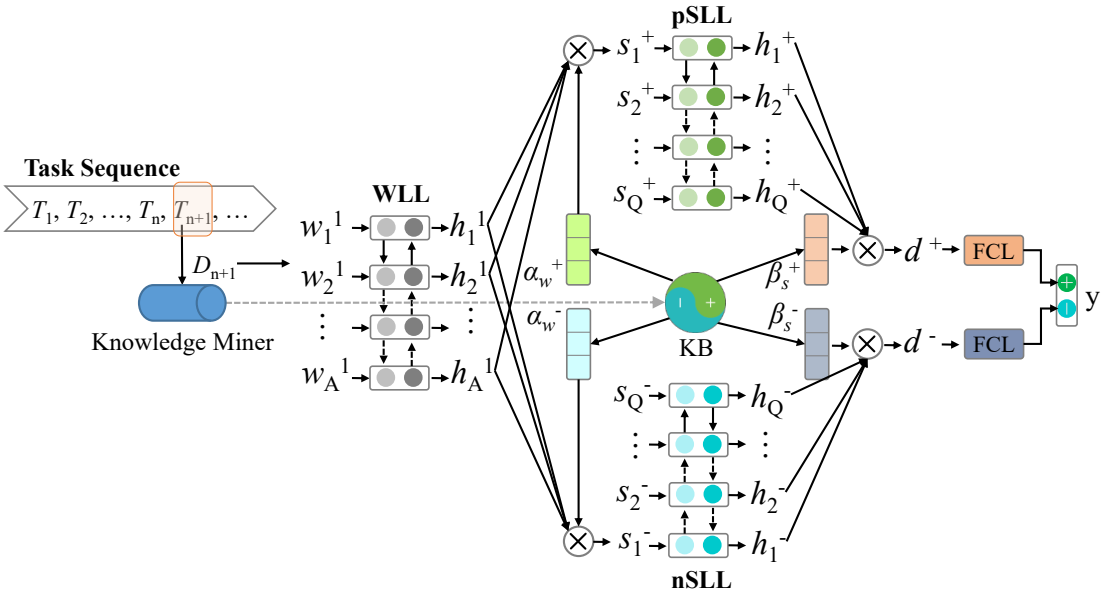


Figure 1: Architecture of the proposed Bayes-enhanced Lifelong Attention Networks (BLAN).

From Parameters to Knowledge (Attention): Recall the example mentioned in the introduction. Suppose the three words "*good, heavy, recommend*" appear in the camera domain. We first transform the parameters to ratios as shown in Table 1, where $R_{w|+} = P_{w|+}/P_{w|-}$ and $R_{w|-} = P_{w|-}/P_{w|+}$. Then, the ratios are treated as attentions (i.e., knowledge) and retained in the Knowledge Base (KB). We use the ratios (not the original parameters) as we aim to reinforce the sentiment intensity for a sentiment word. Thus the KB stores two types of attention knowledge (i.e., positive ratio $R_{w|+}$ and negative ratio $R_{w|-}$) [1] for each task. From Table 1, we can see that the attention knowledge for each word here is useful, also meaningful and interpretable in such a way.

Word	**Parameters** $\rightarrow$	**Ratios**
	$P_{w\|+}$, $P_{w\|-}$	$R_{w\|+}$, $R_{w\|-}$
Good	0.8, 0.2	4.0, 1/4
Heavy	0.3, 0.7	3/7, 7/3
Recommend	0.9, 0.1	9.0, 1/9

Table 1: An "*ideal*" case from parameters to ratios. (Note, this is an ideal case. In practice, it is hard to achieve such a result for each word but each word should have a relative sentiment polarity.)

Lifelong Attention: Suppose we are given a task sequence $(T_1, ..., T_n, T_{n+1}, ...)$, our model has learned the first n tasks. The above two types of knowledge (i.e., $R_{w|+}$ and $R_{w|-}$) of each task has also been retained in the KB. The knowledge mining process is marked using a Knowledge Miner in Figure 1. Again, we use the terms *domain* and *task* interchangeably throughout the paper because we treat the classification in each domain as a learning task following the existing lifelong sentiment classification works. We now focus on using the retained knowledge and the newly learned knowledge from the new/target task T_{n+1} to generate lifelong attentions (namely *lifelong word-level attention* and *lifelong sentence-level attention* as shown later) for the target task. As introduced early, we need to tackle the problems of domain-generality and domain-specificity. For each word w in the target task, we rely on the knowledge from all domains to solve the former. We filter/remove the domain-specific knowledge from previous domains to address the latter when helping new task/domain learning. Here we follow Wang

[1] In practice, we can store one of the $R_{w|+}$ and $R_{w|-}$ into the KB because $R_{w|+} = 1/R_{w|-}$.

et al. (2018a) and propose to filter the domain-specific knowledge using Frequent Pattern Mining (FPM) (Agrawal et al., 1994). A frequent pattern is a set of items that appear frequently (above a minimum frequency threshold, e.g., 10, called *minimum support*) in a database of transactions. Taking positive attention knowledge as an example, we treat the positive attention knowledge of all words in each domain as items. A set of items in one attention distribution form one transaction. Given all transactions, FPM can find frequent items. Then, we treat those infrequent items as domain-specific knowledge and filter out the domain-specific knowledge of each previous domain. Let $\hat{R}^t_{w|+}$ and $\hat{R}^t_{w|-}$ denote the results of the word w after performing FPM in the t-th previous domain/task. The lifelong word-level attentions [2] of this word for the target task T_{n+1} are formulated as

$$\alpha^+_{l,w} = \sum_{t=1}^{n} \hat{R}^t_{w|+} + R^{n+1}_{w|+}, \quad \alpha^-_{l,w} = \sum_{t=1}^{n} \hat{R}^t_{w|-} + R^{n+1}_{w|-}. \tag{2}$$

Assume that all words in a sentence s are conditionally independent given the class (which follows the naïve Bayesian assumption). Then, the lifelong sentence-level attentions of s are formulated as

$$\beta^+_{l,s} = \prod_{i=1}^{A} \alpha^+_{l,w_i}, \quad \beta^-_{l,s} = \prod_{i=1}^{A} \alpha^-_{l,w_i} \tag{3}$$

where A is the number of words in s.

Bayes-enhanced Lifelong Attention Network: When the target task T_{n+1} comes, our model first extracts the above two types of lifelong attentions. The extracted lifelong attentions are used to enhance the hierarchical neural networks. Given a document d with Q sentences $\{s_1, ..., s_Q\}$ and the sentence s_j with words $\{w^j_1, ..., w^j_A\}$, where j denotes the j-th sentence s_j, the proposed BLAN works as follows:

Word-level learner (WLL): For each word in the sentence s_j, WLL first looks up the embedding vector from pre-trained word embedding vectors. The embedding vectors of all words in this sentence are fed into a bidirectional GRU (Bi-GRU) (Bahdanau et al., 2014). Then WLL can learn a representation $h^j_i = [\overrightarrow{h}^j_i, \overleftarrow{h}^j_i]$ for each word (e.g., w^j_i) by concatenating the forward hidden state $\overrightarrow{h}^j_i$ and the backward hidden state $\overleftarrow{h}^j_i$. The representations of these words in sentence s_j with positive and negative attentions are fed into a positive sentence-level learner (pSLL in Figure 1) and a negative sentence-level learner (nSLL in Figure 1), respectively.

Let $\mathbf{H}^j_w = [h^j_1, ..., h^j_A]$ and $\mathbf{w}^j = [w^j_1, ..., w^j_A]$, then the network attentions $\ddot{\alpha}_{\mathbf{w}^j}$ in a vector form are formulated as

$$\begin{aligned} \mathbf{u}_{\mathbf{w}^j} &= tanh(\mathbb{W}_w \mathbf{H}^j_w + \mathbf{b}_w) \\ \ddot{\alpha}_{\mathbf{w}^j} &= softmax(\mathbf{v}_w \mathbf{u}_{\mathbf{w}^j}) \end{aligned} \tag{4}$$

where $tanh()$ is a $tanh$ function, $\mathbb{W}_w$, $\mathbf{b}_w$ and $\mathbf{v}_w$ are trainable parameters.

From Eq. (4), we can see that the values of any attention of the network is a quantity bounded in magnitude $\mathcal{O}(1)$ as we perform $softmax$. To keep lifelong attentions in the same scale, we also perform $softmax$ for lifelong word-level attentions:

$$\widetilde{\alpha}^+_{l,\mathbf{w}^j} = softmax(\alpha^+_{l,\mathbf{w}^j}), \quad \widetilde{\alpha}^-_{l,\mathbf{w}^j} = softmax(\alpha^-_{l,\mathbf{w}^j}) \tag{5}$$

where $\alpha^+_{l,\mathbf{w}^j}$ and $\alpha^-_{l,\mathbf{w}^j}$ are lifelong word-level attentions for words $\mathbf{w}^j$ using Eq. (2).

[2]Here let us describe some notational conventions. Throughout the paper, word-level attention is denoted by symbol α and sentence-level attention is denoted by symbol β.

For each word w_i^j ($i = 1, ..., A$) in $\mathbf{w}^j$, we propose to use the lifelong word-level attention to enhance the network's attention, formulated as -

$$\hat{\alpha}^+_{w_i^j} = \ddot{\alpha}_{w_i^j} + \widetilde{\alpha}^+_{l,w_i^j}, \hat{\alpha}^-_{w_i^j} = \ddot{\alpha}_{w_i^j} + \widetilde{\alpha}^-_{l,w_i^j} \tag{6}$$

$$\alpha^+_{w_i^j} = \frac{\hat{\alpha}^+_{w_i^j}}{\sum_{i=1}^{A} \hat{\alpha}^+_{w_i^j}}, \alpha^-_{w_i^j} = \frac{\hat{\alpha}^-_{w_i^j}}{\sum_{i=1}^{A} \hat{\alpha}^-_{w_i^j}} \tag{7}$$

where $\alpha^+_{w_i^j}$ and $\alpha^-_{w_i^j}$ are the final attentions on the word w_i^j. Then, positive and negative unified representations of the words in sentence s_j are respectively formulated as

$$s_j^+ = \mathbf{H}_w^j(\alpha^+_{\mathbf{w}^j})^T, \quad s_j^- = \mathbf{H}_w^j(\alpha^-_{\mathbf{w}^j})^T. \tag{8}$$

Sentence-level learner (SLL): Two SLLs (positive and negative, i.e., pSLL and nSLL in Figure 1) receive the output (i.e., Eq. (8)) of WLL. Similar to WLL, pSLL and nSLL use Bi-GRU to learn representations ($h_j^+ = [\overrightarrow{h}_j^+, \overleftarrow{h}_j^+]$ and $h_j^- = [\overrightarrow{h}_j^-, \overleftarrow{h}_j^-]$) for each sentence.

Let $\mathbf{s} = [s_1, ..., s_Q]$ denote the sentences of document d, then we have $\mathbf{H}_s^+ = [h_1^+, ..., h_Q^+]$ and $\mathbf{H}_s^- = [h_1^-, ..., h_Q^-]$. Then, the network's sentence-level attentions $\ddot{\beta}^+_{\mathbf{s}}$ and $\ddot{\beta}^-_{\mathbf{s}}$ in a vector form are formulated as

$$\begin{aligned} \mathbf{u}^+ &= tanh(\mathbb{W}_s^+\mathbf{H}_s^+ + \mathbf{b}_s^+) \\ \ddot{\beta}^+_{\mathbf{s}} &= softmax(\mathbf{v}_s^+\mathbf{u}^+) \end{aligned} \tag{9}$$

$$\begin{aligned} \mathbf{u}^- &= tanh(\mathbb{W}_s^-\mathbf{H}_s^- + \mathbf{b}_s^-) \\ \ddot{\beta}^-_{\mathbf{s}} &= softmax(\mathbf{v}_s^-\mathbf{u}^-) \end{aligned} \tag{10}$$

where $\mathbb{W}_s^+$, $\mathbb{W}_s^-$, $\mathbf{b}_s^+$, $\mathbf{b}_s^-$, $\mathbf{v}^+$ and $\mathbf{v}^-$ are trainable parameters.

As introduced above, we use Eq. (3) to compute the lifelong sentence-level attentions $\beta^+_{l,\mathbf{s}}$ and $\beta^-_{l,\mathbf{s}}$ for the sentences $\mathbf{s}$ in document d. These lifelong sentence-level attentions are further normalized as

$$\widetilde{\beta}^+_{l,\mathbf{s}} = softmax(\beta^+_{l,\mathbf{s}}), \quad \widetilde{\beta}^-_{l,\mathbf{s}} = softmax(\beta^-_{l,\mathbf{s}}). \tag{11}$$

Similar to WLL, each SLL also learns a unified representation of the sentences $\{s_1, ..., s_Q\}$ in the document d by using knowledge-enhanced attentions. For each sentence s_j ($j = 1, ..., Q$), we propose to use Bayes-enhanced lifelong sentence-level attention to enhance the network's attention:

$$\hat{\beta}^+_{s_j} = \ddot{\beta}^+_{s_j} + \widetilde{\beta}^+_{l,s_j}, \hat{\beta}^-_{s_j} = \ddot{\beta}^-_{s_j} + \widetilde{\beta}^-_{l,s_j} \tag{12}$$

$$\beta^+_{s_j} = \frac{\hat{\beta}^+_{s_j}}{\sum_{j=1}^{Q} \hat{\beta}^+_{s_j}}, \beta^-_{s_j} = \frac{\hat{\beta}^-_{s_j}}{\sum_{j=1}^{Q} \hat{\beta}^-_{s_j}} \tag{13}$$

where $\beta^+_{s_j}$ and $\beta^-_{s_j}$ are the final attentions on the sentence s_j.

Then, the unified representation of all sentences in this document d is formulated as follows

$$d^+ = \mathbf{H}_s^+(\beta^+_{\mathbf{s}})^T, \quad d^- = \mathbf{H}_s^-(\beta^-_{\mathbf{s}})^T. \tag{14}$$

Document-level learner (DLL): The learned d^+ and d^- are fed into a full connection layer (FCL) to produce the positive sentiment score $P_{+|d}$ and the negative sentiment score $P_{-|d}$ for this document. DLL learns the class label by performing $\arg\max\{P_{+|d}, P_{-|d}\}$. For the objective function, the motivation is that the proposed model should assign a large positive sentiment score $P_{+|d}$ (meanwhile, a small or even zero negative sentiment score $P_{-|d}$) to a positive document, and the reverse results to a negative document. Thus, we propose to maximize the objection function below

$$\max \prod_{d \in D} \left\{ \theta(d)^{\delta(d \in +)} \hat{\theta}(d)^{\delta(d \in -)} \right\} \tag{15}$$

Algorithm 1: BLAN in lifelong learning for sentiment classification

Input : Target domain training documents D, knowledge $R^t_{w|+}$ and $R^t_{w|-}$ $(t = 1, ..., n)$ of the previous n tasks (stored in the KB)

Output: BLAN in the target domain

Learn knowledge $R^{n+1}_{w|+}$ and $R^{n+1}_{w|-}$ of the target domain and store the newly learned knowledge in the KB ;
for *each document* d *in a training batch* **do**
  for *each sentence* s_j *in* d **do**
    for *each word* w^j_i *in the sentence* s_j **do**
      Learn word-level attention $\alpha^+_{w^j_i}$ and $\alpha^-_{w^j_i}$ using Eq. (7) ;
    end
    Learn unified representation s^+_j and s^-_j for the sentence s_j using Eq. (8) ;
    Learn sentence-level attention $\beta^+_{s_j}$ and $\beta^-_{s_j}$ using Eq. (13) ;
  end
  Learn unified representation d^+ and d^- for the document d using Eq. (14) ;
end
Update parameters of the BLAN networks using loss function (i.e., Eq. (17)) with Adam optimizer.

where D are the training documents, $\theta(d) = sigmoid(P_{+|d} - P_{-|d})$, $\hat{\theta}(d) = sigmoid(P_{-|d} - P_{+|d})$, and the activation function $\delta(z)$ is defined as

$$\delta(z) = \begin{cases} 1, \text{ if } z \text{ is true,} \\ 0, \text{ otherwise.} \end{cases} \tag{16}$$

In practice, we solve a logarithmic minimization problem with a regularization term. Formally, the objective function to train our model is formulated as follows

$$\mathcal{L} = \sum_{d \in D} \left\{ \begin{array}{c} -\log(\theta(d)^{\delta(d \in +)}) - \log(\hat{\theta}(d)^{\delta(d \in -)}) \\ +\lambda(P_{+|d} + P_{-|d}) \end{array} \right\} \tag{17}$$

where λ is a regularization parameter (set to 0.01 in training). For the new/target domain T_{n+1}, the training process of the proposed BLAN model is presented in Algorithm 1.

4 Experiments

4.1 Datasets

Since the most related work to ours is lifelong sentiment classification (Chen et al., 2015; Xia et al., 2017; Wang et al., 2019; Lv et al., 2019), we evaluate sentiment classification using the same dataset as in (Chen et al., 2015). Wang et al. (2019) and Lv et al. (2019) also used this dataset. The dataset contains Amazon reviews of 20 types of product domains, denoted by **Amazon(20)** [3]. Each domain consists of 1,000 reviews. Each review has been assigned a sentiment label, i.e., positive (+) or negative (-). We did not use the dataset in (Xia et al., 2017) as their data for each task/domain are from the same domain. In addition, we sampled a large Amazon review dataset from the SNAP corpus [4], which contains product reviews from 24 types of product domains. Following the existing works (Pang et al., 2002; Blitzer et al., 2007; Chen et al., 2015), we treat reviews with rating > 3 as positive and reviews with rating < 3 as negative. We randomly sampled 10,000 reviews for each domain from the SNAP corpus. The sampled dataset is denoted by **SNAP(24)**.

[3] http://anthology.aclweb.org/attachments/P/P15/P15-2123.Datasets.zip
[4] http://jmcauley.ucsd.edu/data/amazon/

The Dataset Amazon(20)			
Alarm Clock	30.51	Home Theater System	28.84
Baby	16.45	Jewelry	12.21
Bag	11.97	Keyboard	22.66
Cable Modem	12.53	Magazine Subscriptions	26.88
Dumbbell	16.04	Movies TV	10.86
Flashlight	11.69	Projector	20.24
Jewelry	19.50	Rice Cooker	18.64
Gloves	13.76	Sandal	12.11
Graphics Card	14.58	Vacuum	22.07
Headphone	20.99	Video Games	20.93
The Dataset SNAP(24)			
Amazon Instant Video	10.96	Health and Personal Care	10.64
Apps for Android	18.39	Home and Kitchen	10.22
Automotive	5.99	Kindle Store	6.42
Baby	11.85	Movies and TV	13.74
Beauty	12.47	Musical Instruments	4.94
Books	9.28	Office Products	5.94
CDs and Vinyl	9.32	Patio Lawn and Garden	10.26
Cell Phones and Accessories	14.07	Pet Supplies	12.45
Clothing Shoes and Jewelry	10.84	Sports and Outdoors	7.12
Digital Music	10.02	Tools and Home Improvement	8.17
Electronics	12.40	Toys and Games	7.27
Grocery and Gourmet Food	10.24	Video Games	14.02

Table 2: Name of each domain and the proportion of negative reviews in each domain.

The name of each domain from the datasets Amazon(20) and SNAP(24), and the proportion of negative reviews in each domain are shown in Table 2. We can see that the proportion of negative reviews in each domain is in the range $[4\%, 31\%]$. *It is worth stressing that the negative reviews in each domain in both datasets are minorities and thus very hard to classify*.

4.2 Baselines

In this work, we use naïve Bayes (NB) parameters to enhance attention neural networks and propose a lifelong sentiment classification model. So, we compare with three types of methods, namely naïve Bayes, neural networks, and lifelong sentiment classification.

- *Naïve Bayes*: we compare our model with the classic **NB** method. We also compare with a strong variant of NB, called NBSVM (Wang and Manning, 2012). There are two models of NBSVM with uni-grams and bi-grams, denoted by **NBSVM-uni** and **NBSVM-bi**, respectively.
- *Neural Networks*: we compare with **Conv-GRNN** (Tang et al., 2015), **LSTM-GRNN** (Tang et al., 2015) and **HAN** (an attention based model) (Yang et al., 2016).
- *Lifelong Sentiment Classification*: we compare with the the existing works **LLV** (Xia et al., 2017), **LSC** (Chen et al., 2015), **LNB** (Wang et al., 2019) and **SRK** (Lv et al., 2019).

For our BLAN, we also created a variant without the Bayes-enhanced attentions, denoted by **nBLAN**. nBLAN is to evaluate our Bayes-enhanced attentions are helpful.

4.3 Settings

Our experiment settings and model parameters are set as follows.

Training set, validation set and test set. For each dataset, the review documents in each domain are partitioned into training set, validation set and test set with 80%, 10% and 10% respectively. For each review document, we use NLTK [5] to split it into sentences, then tokenize and lemmatize each sentence into a sequence of words. We filter out words that appear less than 3 (and 5) times in the training data of each domain of the dataset Amazon(20) (and SNAP(24)) in building the vocabulary of each domain.

[5]NLTK: Natural Language Toolkit, see `http://www.nltk.org/`

Method	Non-lifelong SC Evaluation				Lifelong SC Evaluation			
	Amazon(20)		SNAP(24)		Amazon(20)		SNAP(24)	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
NB	85.29	47.68	81.99	38.49	82.51	53.01	79.26	37.17
NBSVM-uni	88.04	49.31	92.31	44.46	91.07	67.25	93.07	57.30
NBSVM-bi	85.54	28.80	91.27	31.15	90.52	64.59	93.89	61.53
Conv-GRNN	87.68	54.97	92.23	50.63	91.39	72.77	93.39	64.75
LSTM-GRNN	85.43	35.83	92.43	47.27	91.64	70.39	93.88	61.53
HAN	86.57	47.83	93.01	56.06	91.17	68.90	93.94	66.05
LLV	87.12	40.85	90.98	30.38	86.98	48.02	90.76	31.10
LSC	87.71	46.95	91.01	33.59	88.75	58.62	90.94	41.81
LNB	86.92	47.18	91.97	39.86	88.69	61.25	91.74	43.25
SRK	85.43	45.23	90.44	40.69	87.25	60.46	90.57	57.63
nBLAN	86.67	50.00	87.12	56.17	91.50	71.07	93.80	66.76
BLAN	**88.66**	**60.78**	**93.13**	**57.88**	**92.38**	**74.78**	**94.65**	**69.66**

Table 3: (Evaluation results) Average ACC and F1 on the datasets Amazon(20) and SNAP(24). Note, the F1 of negative class should be more reliable than the ACC of both classes as negative reviews are minorities in each dataset.

Model parameters. For the NB, the smoothing parameter is set to 1 (Laplacian smoothing). For the other non-neural networks models (i.e., NBSVM-uni, NBSVM-bi, LLV, LSC and LNB), we use their default parameter settings. In addition, we followed Pang et al. (2002) to deal with negation words as performed in (Chen et al., 2015) for NB, LSC and LNB. For the neural networks models (i.e., Conv-GRNN, LSTM-GRNN, HAN, SRK, nBLAN and BLAN), we use pre-trained GloVe.840B [6] (Pennington et al., 2014) to initialize the word embeddings and the embedding dimension is 300. The number of convolutional filters is 3 with kernel sizes 1, 2 and 3 respectively. The size of LSTM and GRU hidden states is 300. The network parameters are updated using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001. The learning rate is clipped gradually using a norm of 0.9 in performing the Adam optimization. The dropout rate is set to 0.5 in the input layer. We use Accuracy (ACC) of both classes in evaluation. We also use F1 of the negative class because the number of reviews in the negative class is small (see Table 2) and thus not easy to classify.

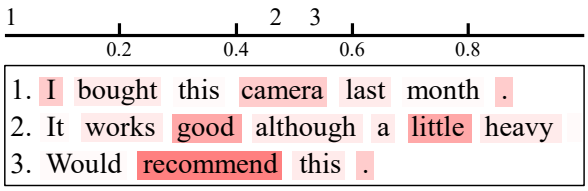
4.4 Result Analysis

We evaluate sentiment classification performance using both *non-lifelong setting* and *lifelong setting*. For the non-lifelong setting, we train and test all systems on each single domain (no knowledge from other domains). For the lifelong setting, we follow (Chen et al., 2015) to treat each domain as the future/target domain with the remaining domains as the past domains. That is, *we run each system 20 (and 24) times on the dataset Amazon(20) (and SNAP(24))*. Note that those non-lifelong baselines are tailored for a single domain. To have a fair comparison, we combine the training data of all domains (past and target domains) to train each non-lifelong model and test the model on each target domain test data. The average accuracy (ACC) of both classes and the F1 of negative class over the 20 domains of Amazon(20) and 24 domains of SNAP(24) are shown in Table 3.

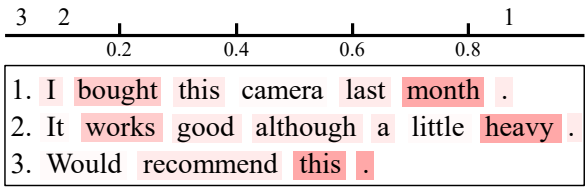
From the results, we make the following observations:

- The results of each method in lifelong setting are better than in non-lifelong setting on each dataset. This show that lifelong learning is a promising research direction.
- Our BLAN achieves the best ACC and F1 on both datasets respectively. The results clearly show the superiority of BLAN.

[6] `http://nlp.stanford.edu/data/glove.840B.300d.zip`



(a) Positive attention



(b) Negative attention

Figure 2: *Left*: Positive lifelong word-level and sentence-level attentions, and *Right*: Negative lifelong word-level and sentence-level attentions. (Color online)

- BLAN outperforms the existing lifelong sentiment classification baselines (i.e., LLV, LSC, LNB and SRK) markedly. This shows that our work presents a new margin to beat.
- Attention based baseline (i.e., HAN) is inferior to our BLAN, especially on the smaller dataset Amazon(20). nBLAN (a variant of BLAN without Bayes-enhanced attention) is also inferior to BLAN. This shows that Bayes-enhanced attention (i.e., lifelong attention) works and is useful.
- Most of the neural network baselines (i.e., Conv-GRNN, LSTM-GRNN, HAN and SRK) are slightly superior to the non-neural network baselines (i.e., NB, NBSVM-uni, NBSVM-bi, LLV, LSC and LNB), especially on the large dataset SNAP(24). This shows that neural network based methods are superior in dealing with sentiment classification using large-scale data.

4.5 Case Study

Recall the example given in the introduction:

"*I bought this camera last month. It works good although a little heavy. Would recommend this.*"

The final learned lifelong attentions from the dataset Amazon(20) are shown in Figure 2, where each number above the axis corresponds to a sentence ID and its location on the axis represents the attention value (i.e., the value below the axis) of that sentence. The color depth in the textbox represents the attention value of each word. We can see that our model works well for attention learning. The learned lifelong attentions here are also interpretable in our human cognition.

5 Conclusions

In this work, we studied deep learning for a sequence of sentiment classification tasks. After each task learned, its knowledge was retained to help subsequent task learning. This problem is called lifelong sentiment classification. This paper proposed a novel knowledge-enhanced attention network model for lifelong sentiment classification. The proposed model learns to build two types of lifelong attentions by exploiting the model parameters of naïve Bayes. The built lifelong attentions are used to enhance the overall attention networks. Experimental results showed that the proposed model outperforms a wide range of baselines markedly. The proposed model works for two-class (positive and negative) problem. In our future work, we plan to extend it to multi-class problem. We also plan to address catastrophic forgetting problem (Kirkpatrick et al., 2017) in neural networks and bring our method to a new height.

Acknowledgements

Hao Wang and Yan Yang's work was supported in part by a grant from the National Natural Science Foundation of China (grant no. 61976247), and a grant from Miaozi Project in Science and Technology Innovation Program of Sichuan Province (grant no. 2020132).

References

Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *VLDB*, volume 1215, pages 487–499.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473v7*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, pages 440–447.

Sneha Chaudhari, Gungor Polatkan, Rohan Ramanath, and Varun Mithal. 2019. An attentive survey of attention models. *arXiv preprint arXiv:1904.02874*.

Zhiyuan Chen and Bing Liu. 2018. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–207.

Zhiyuan Chen, Nianzu Ma, and Bing Liu. 2015. Lifelong learning for sentiment classification. In *ACL*, pages 750–756.

Qin Chen, Qinmin Hu, Jimmy Xiangji Huang, Liang He, and Weijie An. 2017. Enhancing recurrent neural networks with positional attention for question answering. In *SIGIR*, pages 993–996.

Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. 2019. Deep short text classification with knowledge powered attention. In *AAAI*, pages 6252–6259.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Shuqin Gu, Lipeng Zhang, Yuexian Hou, and Yin Song. 2018. A position-aware bidirectional attention network for aspect-level sentiment analysis. In *COLING*, pages 774–784.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Effective attention modeling for aspect-level sentiment classification. In *COLING*, pages 1121–1131.

Weiyi Huang, Qiang Qu, and Min Yang. 2020. Interactive knowledge-enhanced attention network for answer selection. *Neural Computing and Applications*, pages 1–17.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *NAACL*, pages 3543–3556.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*, pages 1–15.

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

Gaël Letarte, Frédérik Paradis, Philippe Giguère, and François Laviolette. 2018. Importance of self-attention for sentiment analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 267–275.

Shoushan Li and Chengqing Zong. 2008. Multi-domain sentiment classification. In *ACL*, pages 257–260.

Guangyi Lv, Shuai Wang, Bing Liu, Enhong Chen, and Kun Zhang. 2019. Sentiment classification by leveraging the shared knowledge from a sequence of domains. In *DASFAA*, pages 795–811.

Tom Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Bo Yang, Justin Betteridge, Andrew Carlson, B Dalvi, Matt Gardner, Bryan Kisiel, et al. 2018. Never-ending learning. *Communications of the ACM*, 61(5):103–115.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *ACL*, pages 79–86.

German I 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.

Hao Peng, Jianxin Li, Senzhang Wang, Lihong Wang, Qiran Gong, Renyu Yang, Bo Li, Philip Yu, and Lifang He. 2019. Hierarchical taxonomy-aware and attentional graph capsule RCNNs for large-scale multi-label text classification. *IEEE Transactions on Knowledge and Data Engineering*, Early Access:1–14.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. In *EMNLP-IJCNLP*, pages 43–54.

Paul Ruvolo and Eric Eaton. 2013. ELLA: An efficient lifelong learning algorithm. In *ICML*, pages 507–515.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *ACL*, pages 2931–2951.

Lei Shu, Hu Xu, and Bing Liu. 2017. Lifelong learning CRF for supervised aspect extraction. In *ACL*, pages 148–157.

Daniel L Silver, Qiang Yang, and Lianghao Li. 2013. Lifelong machine learning systems: Beyond learning algorithms. In *AAAI Spring Symposium: Lifelong Machine Learning*, volume 13, pages 48–55.

Shiliang Sun, Honglei Shi, and Yuanbin Wu. 2015. A survey of multi-source domain adaptation. *Information Fusion*, 24:84–92.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*, pages 1422–1432.

Sebastian Thrun. 1998. Lifelong learning algorithms. In *Learning to Learn*, pages 181–209.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *ACL*, pages 90–94.

Shuai Wang, Guangyi Lv, Sahisnu Mazumder, Geli Fei, and Bing Liu. 2018a. Lifelong learning memory networks for aspect sentiment classification. In *IEEE BigData*, pages 861–870.

Shuai Wang, Sahisnu Mazumder, Bing Liu, Mianwei Zhou, and Yi Chang. 2018b. Target-sensitive memory networks for aspect sentiment classification. In *ACL*, volume 1, pages 957–967.

Hao Wang, Bing Liu, Shuai Wang, Nianzu Ma, and Yan Yang. 2019. Forward and backward knowledge transfer for sentiment classification. In *ACML*, pages 457–472.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *EMNLP-IJCNLP*, pages 11–20.

Fangzhao Wu and Yongfeng Huang. 2016. Sentiment domain adaptation with multiple sources. In *ACL*, pages 301–310.

Rui Xia, Jie Jiang, and Huihui He. 2017. Distantly supervised lifelong learning for large-scale social media sentiment analysis. *IEEE Transactions on Affective Computing*, 8(4):480–491.

Feng Xu, Zhenchun Pan, and Rui Xia. 2020. E-commerce product review sentiment classification based on a naïve bayes continuous learning framework. *Information Processing & Management*, page 102221.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*, pages 1480–1489.

Min Yang, Lei Chen, Xiaojun Chen, Qingyao Wu, Wei Zhou, and Ying Shen. 2019. Knowledge-enhanced hierarchical attention for community question answering with multi-task and adaptive learning. In *IJCAI*, pages 5349–5355.

Jianfei Yu, Jing Jiang, and Rui Xia. 2019. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Kai Zhang, Hefu Zhang, Qi Liu, Hongke Zhao, Hengshu Zhu, and Enhong Chen. 2019. Interactive attention transfer network for cross-domain sentiment classification. In *AAAI*, pages 5773–5780.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Attention-based LSTM network for cross-lingual sentiment classification. In *EMNLP*, pages 247–256.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.