Diversity-Based Generalization for Unsupervised Text Classification under Domain Shift

Jitin Krishnan (✉), Hemant Purohit, and Huzefa Rangwala

George Mason University, Fairfax, VA, USA
{*jkrishn2, hpurohit, rangwala*}@gmu.edu

Abstract. Domain adaptation approaches seek to learn from a source domain and generalize it to an unseen target domain. At present, the state-of-the-art unsupervised domain adaptation approaches for subjective text classification problems leverage unlabeled target data along with labeled source data. In this paper, we propose a novel method for domain adaptation of *single-task* text classification problems based on a simple but effective idea of diversity-based generalization that does not require unlabeled target data but still matches the state-of-the-art in performance. Diversity plays the role of promoting the model to better generalize and be indiscriminate towards domain shift by forcing the model not to rely on same features for prediction. We apply this concept on the most explainable component of neural networks, the attention layer. To generate sufficient diversity, we create a multi-head attention model and infuse a diversity constraint between the attention heads such that each head will learn differently. We further expand upon our model by tri-training and designing a procedure with an additional diversity constraint between the attention heads of the tri-trained classifiers. Extensive evaluation using the standard benchmark dataset of Amazon reviews and a newly constructed dataset of Crisis events shows that our fully unsupervised method matches with the competing baselines that uses unlabeled target data. Our results demonstrate that machine learning architectures that ensure sufficient diversity can generalize better; encouraging future research to design ubiquitously usable learning models without using unlabeled target data.

Keywords: Text Classification · Unsupervised Domain Adaptation · Natural Language Processing · Neural Networks

1 Introduction

In natural language processing, domain adaptation of sequence classification problems has several applications ranging from sentiment analysis [4] to classifying social media posts during crisis events [1]. Knowledge learned from one domain, book reviews for instance, can be adapted to predict examples from a different domain such as reviews of electronics. Similarly, information about resource-need events learned from one natural disaster can be adapted to predict events from an ongoing crisis [20]. With the publication of Amazon reviews dataset [4] consisting of around 25 different domains, cross-domain sentiment analysis became

a common way to evaluate machine learning models for domain adaptation in text.

The top performing models in this line of research largely remain dependent on unlabeled target data. Although unlabeled data from the target domain tends to help, it is imperative to realize the extent of performance gain that source data alone can bring. We consider that the ideal criterion for no supervision in domain adaptation is having zero knowledge about the target domain beforehand; even if it is unlabeled. Our work can be viewed either as a strong baseline for future unsupervised cross-domain research that utilizes unlabeled target data or as a new direction in fully unsupervised domain adaptation without using any target data at all; which is necessary for tasks such as relevancy prediction for actionable information filtering in domains such as natural disasters that require timely and efficient methods. We scope our work to the following setting: **a)** *single-task transfer*, **b)** *single source and target*, and **c)** *without labeled and unlabeled target data available during training*. We compare and contrast our unsupervised methods to the existing counterparts. We do not consider any supervised or minimally supervised approaches in this work.

Contributions: a) We present a novel diversity-based generalization method using a multi-head attention model for domain adaption in unsupervised text classification tasks. **b)** To further improve the generalizability of our model and utilize additionally available unlabeled source data, we design a tri-training procedure with an additional diversity constraint between the attention heads of the tri-trained classifiers. **c)** Addressing the existing evaluation gap in component-level performance analysis, we show a systematic and incremental creation of our models by creating strong unsupervised baselines and improving upon existing work.

2   Related Work

Early works on domain adaptation such as Structural Correspondence Learning [5] make use of unlabeled target data to find a joint representation by automatically inducing correspondences among features from different domains. The importance of a good feature representation was later formally analyzed with a generalization bound by Ben-David et al. [3]. These studies realized the importance of finding commonality in features or pivots and minimizing the difference between the domains. Pan et al. [19] proposed a spectral feature alignment method to align domain-specific and domain-independent words into unified clusters via simultaneous co-clustering in a common latent space. Later, introduction of deep learning and neural networks helped remedy the problems of manual pivot selection and discrete feature representations. In order to learn better higher level representations, Stacked Denoising Autoencoders (SDA) [25] were introduced. Along with SDA, a more efficient version called marginalized SDA [6] with low computational cost and scalability has been utilized successfully in cross-domain tasks [10, 11]. Domain-Adversarial training of Neural Networks (DANN) [9] was proposed to effectively utilize unlabeled target data to create a classifier that is

indiscriminate toward different domains. In their work a negative gradient (gradient reversal) from a domain classifier branch is back-propagated to promote the features at the lower layers of the network incapable of discriminating domains. DANN became an essential component in many works that followed. Recent works such as Adversarial Memory Network (AMN) [15] bring interpretability by using attention to capture the pivots. Along with attention, they effectively use gradient reversal to learn domain indiscriminate features. Hierarchical Attention Network (HATN) [14] expands upon AMN by first extracting pivots and then jointly training pivot and non-pivot networks. Interactive Attention Transfer Network (IATN) [26], another closely related work to AMN and HATN, showed the importance of attending 'aspect' information. Another line of research, that approached domain adaption through innovation in training procedure, is tri-training [22, 28]. Tri-training utilizes three independently trained classifiers; of which one is trained only on unlabeled target data, pseudo-labeled by the other two. The final prediction is done by majority voting. Multi-task tri-training (MT-Tri) [21], on the other hand, introduced an orthogonality constraint between the two classifiers such that it can be trained jointly, reducing the compute time. This constraint is one of the inspirations for our work. Although diversity could be achieved through other means, we also focus on 'orthogonality'. All of these recent works used unlabeled target data for training classifiers. Our goal is to show that similar performance is achievable without using any target data at all.

Based on how the dataset is used, approaches to domain adaptation can vary from minimally-supervised to unsupervised. Minimally-supervised approaches such as Aligned Recurrent Transfer [7] utilize some *labeled* data from the target domain, while unsupervised approaches such as DANN, AMN, HATN, or IATN utilize only *unlabeled* target data making it a more realistic scenario in terms of usability where collecting labeled target data is expensive. However, many state-of-the-art unsupervised domain adaptation methods, strikingly, never compare with strong fully-unsupervised baselines where no target data is used. Newer methods have started using word vectors [18] for their input word representations. However, the baselines they compare with, utilize large 5000-dimension feature vector of the most frequent unigrams and bigrams as the input representation. In addition, many recent works present a complex system without conducting a component-wise analysis which makes it unclear as to how much each component (word vectors, gradient reversal, or attention) contributed to the performance boost as compared to a simple DANN architecture. To address these evaluation gaps, we perform a systematic and incremental construction of architectures such that individual performance gain is realized.

Advantages and Practical Utility: a) Our methods do not require any target data for training; making it out-of-the-box adaptable to any domain. **b)** We provide a method to utilize additionally available unlabeled source data. **c)** Our models are computationally cheaper (training converges quickly) when compared to the existing state-of-the-art models. **d)** Diversified attention can provide better quality of attended words which can be used for various downstream tasks such as knowledge graph construction.
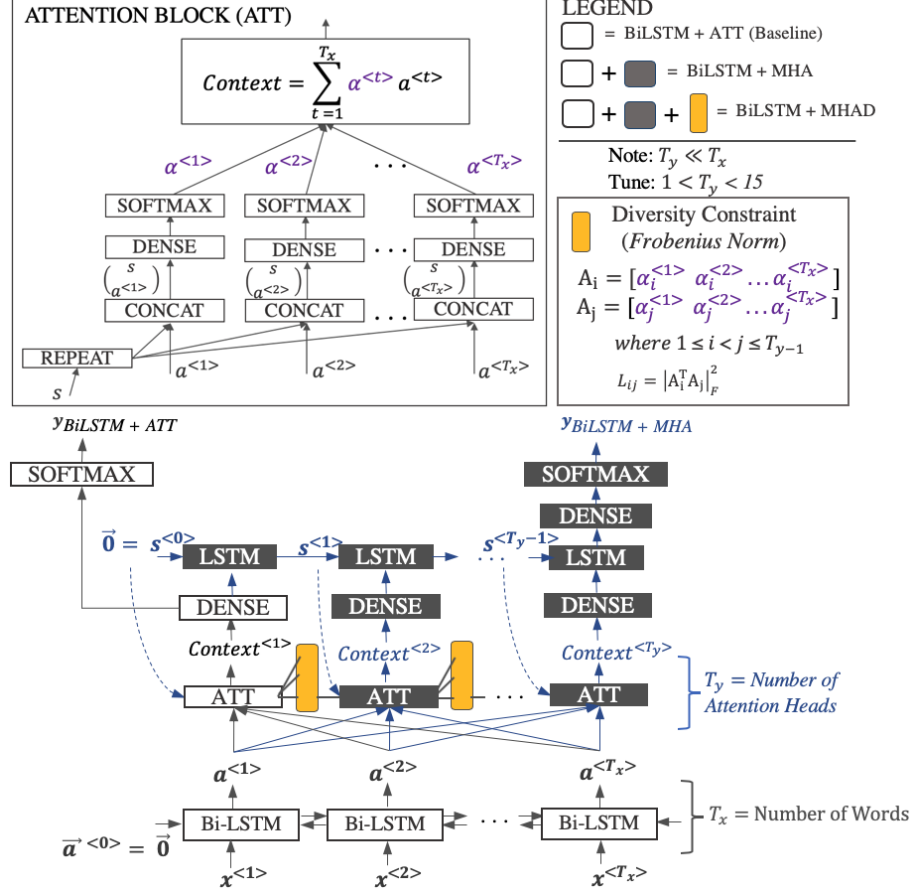
Fig. 1. Complete architecture of the multi-head attention model with diversity.

3   Methodology

3.1   Problem Definition and Notations

Given a source (D_s) and a target (D_t) domain, the goal is to train a classifier using data **only** from D_s and predict examples from the completely unseen D_t. X_s and X_t represent the set of labeled data from source and target domains respectively with their corresponding ground truth labels y_s and y_t. X_t and y_t are used for testing purposes only. X_s^u and X_t^u represent unlabeled data available from the source and target domain respectively. X_t^u (used in all of our competing models either for adversarial training or tri-training) is **never** used in our models. Finally, $[.]^{pl}$ represents data that is pseudo-labeled by the classifier. To summarize:
Input: X_s, y_s (and X_s^u for tri-training)
Output: $y_t^{pred} \leftarrow predict(X_t)$

3.2   Diversity-based Models

We introduce 4 models with one integral concept: *diversity*. Figure 1 provides an overview of the first two models and Figure 2 provides an overview of the last two. First is a multi-head attention baseline created to understand the naturally occurring diversity when multiple attention heads are connected. The second model enforces this diversity as a constraint such that all heads learn different features. The third model puts together three diversity-based classifiers and tri-trains them. Tri-training procedure in itself consists of an additional diversity constraint which forces two of the classifiers to learn differently. This is a one-step tri-training procedure intended for scenarios where no unlabeled source data is additionally available. When it is available, a full tri-training can be done until convergence, which is the fourth model.

Multi-Head Attention for Sequence Classification (BiLSTM+MHA): BiLSTM+ATT is a standard baseline attention architecture constructed using BiLSTM [12, 23] and attention mechanism [2, 16]. Bidirectional Long Short-Term Memory (BiLSTM) units have been successfully used in sequence modeling tasks because of their effectiveness in representing forward and backward dependencies in a sequence. For example, meanings of words like 'good' and 'bad' can be changed when they are prefixed with 'not' or suffixed with 'but'. Attention, on the other hand, provides task-specific benefits by attending the most relevant words such as 'excellent' or 'poor' in sentiment analysis. Attention and BiLSTM have been successfully combined previously for tasks such as relation extraction [27] to capture important semantic information in a sentence.

BiLSTM+MHA is an extension of the BiLSTM+ATT baseline by adding multiple attention heads as shown in Figure 1. This is similar to machine-translation-like architecture [16] where each attention head leads to an LSTM cell with memory carried from previous cells to predict the next word. To customize it to classification purpose, we simply use the output from the final LSTM cell. Setting the classification task this way gives more leniency for the model to learn, remember, and generalize. Multiple attention heads can learn differently and what is learned from the previous heads is transferred to the next. However, this does not guarantee diversity as we do not know if the attention heads will in fact learn differently. In order to enforce diversity, we introduce the following models.

Multi-Head Attention with Diversity (BiLSTM+MHAD): In order to guarantee that these attention heads learn differently and forcing the model not to rely on the same features, we create a *diversity constraint*, an additional loss term shown below.

$$L_d = \frac{1}{k} \sum_{i=1}^{T_y-2} \sum_{j=i+1}^{T_y-1} \|A_i^T A_j\|_F^2 \ ; \ \text{where } i \neq j \tag{1}$$

where $k = \frac{(T_y-2)(T_y-1)}{2}$, the total number of combinations. T_y is the total number of attention heads. A_i and A_j are i^{th} and j^{th} attention heads and $\|.\|_F^2$

is the squared Frobenius norm, similar to the orthogonality constraint used in [21]. We leave the last attention head from this loss term so that we have one layer that learns freely without any constraints. The complete architecture of this diversity-based model is shown in Figure 1. Resulting overall loss function, consisting of a binary cross entropy loss term and the diversity loss term, for N training examples is shown below.

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] + \gamma L_d \tag{2}$$

where γ is the hyperparameter to control how much diversity to be enforced within the model.

One-Step Diversity Tri-training (BiLSTM+MHAD-Tri-I): To further expand the concept of diversity, we tri-train the BiLSTM+MHAD models by adapting the multi-task tri-training procedure by [21]. In addition to applying the diversity constraint within each classifiers, an additional orthogonality loss is enforced between first two models m_1 and m_2. The third model m_3 is left out from the joint training. The loss term is shown below.

$$L_o = \frac{1}{k} \sum_{i=1}^{T_y} \sum_{j=1}^{T_y} \|A(m_1)_i^T A(m_2)_j\|_F^2 \tag{3}$$

where $k = \frac{(T_y-1)(T_y)}{2}$. $A(m_1)$ and $A(m_2)$ are the attention heads for models m_1 and m_2 respectively. T_y is the total number of attention heads of each model. The total tri-training diversity loss is given below.

$$L_{dtri} = \alpha L_o + \beta L_d \tag{4}$$

where α and β are the hyperparameters to control how much diversity to be enforced within and between the models.

For one-step diversity tri-training shown in Algorithm 1, we jointly train m_1 and m_2 with tri-training diversity loss L_{dtri}. m_3 is separately trained as a BiLSTM+MHAD model. For predictions, a majority voting rule is applied over the three classifiers. The overall loss function for N training examples is given below.

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] + L_{dtri} \tag{5}$$

Algorithm 1: One-Step Diversity Tri-training

Input: X_s
Output: m_1, m_2, m_3
$m_1, m_2 \leftarrow joint\_diversity\_train\_models(X_s)$
$m_3 \leftarrow diversity\_train\_model(X_s)$
apply majority vote over m_i

Tri-training until Convergence (BiLSTM + MHAD-Tri-II): Full tri-training, shown in Algorithm 2 and Figure 2, utilizes additionally available unlabeled source data. While the first two classifiers m_1 and m_2 are jointly trained on labeled source data, the third classifier m_3 is solely dedicated to the training over unlabeled data that is pseudo-labeled by m_1 and m_2. Similar to [21], we define a threshold value τ such that at least one out of the two models should predict with probability greater than τ to be considered successfully pseudo-labeled. We set τ to be 0.7. Starting with second iteration, m_1 is trained jointly with m_2 using a combination of labeled source data and unlabeled source data pseudo-labeled by m_2 and m_3. During joint-training, we give priority to the primary model by setting the loss weights accordingly. For example, while joint-training m_1 with m_2, losses for the models can be minimized in a $2 : 1$ ratio, giving priority to m_1. We continue this process until a convergence condition is met: $m_1 \approx m_2 \approx m_3$.

Algorithm 2: Tri-training [21] - Modified

Input: X_s, X_s^u
Output: m_1, m_2, m_3
while *convergence condition is not met* **do**
    for $i \in 1..3$ **do**
        $X^{pl} \leftarrow \emptyset$
        for $x \in X_s^u$ **do**
            if $p_i(x) = p_k(x)(j, k \neq i)$ **then**
                $X^{pl} \leftarrow X^{pl} \cup \{(x, p_j(x))\}$
            end
        end
        if $i = 3$ **then**
            $m_3 \leftarrow diversity\_train(X^{pl})$;   // Eq. 2
        else if $i = 1$ **then**
            $m_1 \leftarrow joint\_diversity\_train(X_s \cup X^{pl}, m_2)$;   // Eq. 5
        else
            $m_2 \leftarrow joint\_diversity\_train(X_s \cup X^{pl}, m_1)$;   // Eq. 5
        end
    end
end
apply majority vote over m_i

---

4 Experimental Evaluation

4.1 Benchmark Dataset: Amazon Reviews

We use the standard benchmark Amazon reviews dataset[1] [4] which is widely used for cross-domain sentiment analysis. We consider four domains: Books

---

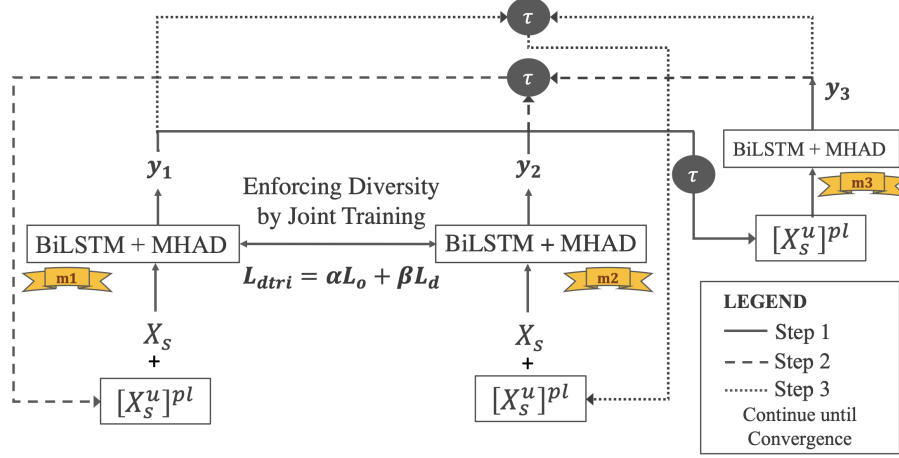[1] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

Fig. 2. Tri-training BiLSTM+MHAD models

(B), Kitchen (K), DVD (D), and Electronics (E). For a fair evaluation of the architectures, we use the exact same raw dataset[2] used by our top competitor model HATN [14], which is a part of Blitzer's original raw dataset. We also use the same 300-dimensional word vectors[3] [18]. Table 1 summarizes this dataset.

4.2   Crisis Dataset (Tweets)

Additionally, we construct a new dataset consisting of Twitter posts (tweets) collected during three hurricane crises by *CitizenHelper* [13] system: *Harvey* and *Irma* in 2017, and *Florence* in 2018. Similar to sentiment classification, our goal here is to classify whether a tweet text indicates an event or not. Using the crowd-sourcing platform Figure-Eight[4], three workers at minimum were assigned to give a binary label to each tweet. We define events to be actions that involve at least one noun/entity. Events could be past, present, or future actions. It could also be questions, news, or instructions about actions. Some examples are: '*A rescues B*', '*A is sending food to B*', '*A will move to location B*', and so on. Table 1 summarizes this dataset. Unfortunately, the labeled dataset for Florence and Irma consists of very low number of positive events. Consequently, we set up the experiments such that we train only on *Harvey* and test on *Florence* and *Irma*.

4.3   Experimental Setup

We follow the traditional cross domain sentiment classification set up where each experiment consists of a source domain (S) and a target domain (T). A model

---

will be trained on source data and tested on target data, represented as $S \to T$. We use all available labeled target data for testing. Crisis dataset is balanced before training and testing.

	Positive	Negative	Unlabeled	Average Number of Tokens	Vocabulary
Books	3000	3000	9750	182.0	105920
DVD	3000	3000	11843	197.5	117619
Kitchen	3000	3000	13856	102.0	52972
Elec.	3000	3000	17009	119.3	72458
Harvey	1122	960	10001	17.2	23562
Florence	201	1475	10001	17.1	26380
Irma	313	596	10001	15.3	20764

Table 1. Dataset Statistics

4.4   Implementation Details

We use Keras deep learning library with Adam optimizer ($lr = 0.005$, $beta_1 = 0.9$, $beta_2 = 0.999$, $decay = 0.01$) for our implmentations. Maximum epoch is 40 with an early stopping patience of 3. Batch size is 32 and validation split is 0.15.

We set the number of attention heads, $T_y = 5$ and number of words from each review, $T_x = 200$. To keep the model simple, we do not change this further. Dropouts are kept at 0.4. τ is kept at 0.7 and tri-training is stopped at 85% agreement. We set $\gamma = 0.01$, $\alpha = 0.05$ and $\beta = 0.01$. These values are obtained by performing a basic hyperparameter tuning using grid search.

4.5   Baselines and Modifications

Adversarial Learning Based Methods: DANN [9] introduced adversarial training by making use of unlabeled target domain data. Earlier layers of the deep neural network architecture are made domain invariant through back-propagating a negative gradient using a jointly trained domain classifier. It uses a 5000-dimensional feature vector of the most frequent unigrams and bigrams as the input representation. DAmSDA [10], on the other hand, uses mSDA [6] representation instead. We report the scores for DAmSDA and DANN from HATN [14]. For DANN, additionally, we create a customized implementation (**DANN**[+]) using BiLSTM and word vectors. This modified architecture simply consists of a shared BiLSTM layer followed by a dense layer for sentiment classification and the same BiLSTM layer followed by a gradient reversal layer and a dense layer for domain classification. Note that the accuracy for **DANN**[+] (our improved DANN) is **+3.2%** higher than what is reported in HATN.

Tri-training Based Methods: Multi-task tri-training (MT-Tri) [21] conducts tri-training on a multilayer perceptron model with an orthogonality loss between the final layers to enforce diversity between the jointly trained models. Unlabeled target data pseudo-labeled by the first two classifiers are fed to the third classifier. Three classifiers are optimized until none of the models' predictions change. We improve upon this model (**MT-Tri**$^+$) by using word vectors and BiLSTM.

Attention Based Methods: Recent works such as AMN [15], HATN [14], and IATN [26] use attention to identify sentiment pivots. Utilizing unlabeled target data, gradient reversal is an essential component in their models for domain classification. AMN expands DANN to an attention-based model. HATN improves AMN further by building pivot and non-pivot networks. The pivot network (P-Net) performs the same task as AMN by extracting pivots. The non-pivot network (NP-Net) takes a transformed input that hides previously extracted pivots, which is then jointly trained with P-Net. IATN incorporates 'aspect' information in addition to the sentence attentions. At the time of writing of this paper, the open source code[5] for IATN is still being prepared by its authors. We include the reported scores for reference purpose. IATN reports a 0.8% increase in performance as compared to HATN (85.9% versus 85.1%). IATN uses the same input settings and the dataset as HATN with one difference: 200-dimensional word vectors instead of 300. Meanwhile, we use the exact same dataset and GoogleNews word vectors used by HATN for all our experiments for both reproducibility as well as blind comparison.

Strong Unsupervised Baselines: To study component-wise performance, we construct two strong unsupervised baselines from standard neural network architectures: BiLSTM and BiLSTM+ATT. BiLSTM consists of traditional BiLSTM units with the final unit making the prediction. BiLSTM+ATT, as shown in Figure 1, adds a single attention layer on top of BiLSTM and the prediction is based on the output from the attention layer. Note that these two baselines still produce strong results and provide a reference for how much improvement following models make.

5   Results & Discussion

Tables 2, 3, and 4 show the competitive nature of our fully unsupervised methods when compared with the existing unsupervised counterparts that use unlabeled target data. Our experiments showed incrementally improving results when each component is added to the baselines. Attention with diversity improved the single-attention baseline and tri-training with diversity improved it even further. Using additionally available unlabeled source data proved to be fruitful for most of the domains. Note that for crisis dataset we only use *Harvey* for training because labeled data for *Florence* and *Irma* was just too low.

An implication of the diversity-based attention heads is shown in Figure 3. Diversity pushes the model not to rely on the same features. First exam-

---

[5] https://github.com/1146976048qq/IATN

Table 2. Classification accuracy scores showing that unlabeled target data is not necessary to achieve strong performance. **+**: improved implementations, ◇: reproduced implementations, ♠: strong unsupervised baselines constructed from standard neural network architectures, *****: reported scores from [14, 26] (see description). Our scores are averaged over 5 independent runs. **Note that only the models in the bottom table do not use any unlabeled target data.**

S → T	DANN*	DAmSDA*	IATN*	DANN+	MT-Tri+	AMN◇ /P-Net	HATN◇
B → D	83.42	86.12	86.80	82.85	84.67	87.07	87.70
B → E	76.27	79.02	86.50	81.03	84.62	82.98	86.20
B → K	77.90	81.05	85.90	82.01	84.78	84.85	87.08
K → B	74.17	80.55	84.70	79.38	80.98	83.50	84.83
K → D	75.32	82.18	84.40	79.04	78.89	82.83	84.73
K → E	85.53	88.00	87.60	86.00	85.87	86.72	89.08
E → B	73.53	79.92	81.80	78.92	80.64	83.28	83.62
E → K	84.53	85.80	88.70	86.43	89.62	89.80	90.12
E → D	76.27	82.63	84.10	77.83	79.97	83.37	83.87
D → B	80.77	85.17	87.00	84.32	85.67	87.85	88.02
D → E	76.35	76.17	86.90	81.74	84.48	84.65	86.78
D → K	78.15	82.60	85.80	83.29	85.05	84.28	87.00
AVG	78.52	82.43	85.90	81.78	83.77	85.10	**86.59**

S → T	BiLSTM♠	BiLSTM +ATT♠	BiLSTM +MHA	BiLSTM +MHAD	BiLSTM +MHAD-Tri-I	BiLSTM +MHAD-Tri-II
B → D	84.19	87.44	87.29	87.54	87.76	87.46
B → E	83.61	83.90	85.36	85.63	85.75	86.08
B → K	83.87	85.21	86.04	87.06	87.34	87.68
K → B	80.52	82.15	83.11	83.70	84.19	84.23
K → D	78.28	80.17	81.50	82.27	82.11	83.34
K → E	86.33	87.30	88.60	88.81	88.98	89.22
E → B	80.58	82.10	83.55	83.67	83.96	84.33
E → K	88.07	88.19	89.61	89.96	90.07	91.05
E → D	78.08	81.93	82.77	82.93	82.87	82.81
D → B	83.93	87.72	87.77	88.22	88.51	88.74
D → E	82.98	84.57	84.75	85.93	85.79	86.21
D → K	84.38	85.45	86.50	86.73	86.74	87.37
AVG	82.90	84.68	85.57	85.98	86.17	**86.54**

Table 3. Classification accuracy scores for crisis dataset.

S → T	HATN	BiLSTM +ATT	BiLSTM +MHA	BiLSTM +MHAD	BiLSTM +MHAD-Tri-I	BiLSTM +MHAD-Tri-II
H → F	80.01	74.88	74.32	75.69	76.00	78.11
H → I	58.53	63.84	64.32	65.10	65.02	64.38

Table 4. Training time in d-hh:mm:ss for H→F on a Dual Intel(R) Xeon(R) Gold 5120 CPU@2.2GHz with 28 cores and 1.5TB RAM.
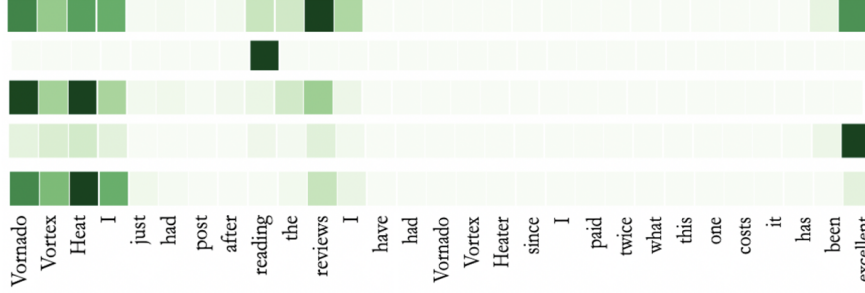
HATN	BiLSTM+MHA	BiLSTM+MHAD	BiLSTM+MHAD-Tri-I	BiLSTM+MHAD-Tri-II
1-08:31:09	00:50:31	01:29:28	2:07:23	6:21:18

Review Text: Vornado Vortex Heat. I just had to post after reading the reviews. I have had a Vornado Vortex Heater since 1994, I paid twice what this one costs and it has been excellent. $y_{true} = +ve$
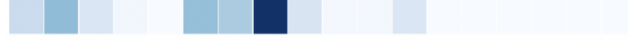
BiLSTM+ATT: $y_{pred} = -ve$

BiLSTM + MHAD with 5 diverse attention layers: $y_{pred} = +ve$

Review Text: Butter dish not tall enough. It looks nice but the lid touches the top of the butter and sticks to it. $y_{true} = -ve$

BiLSTM+ATT: $y_{pred} = +ve$

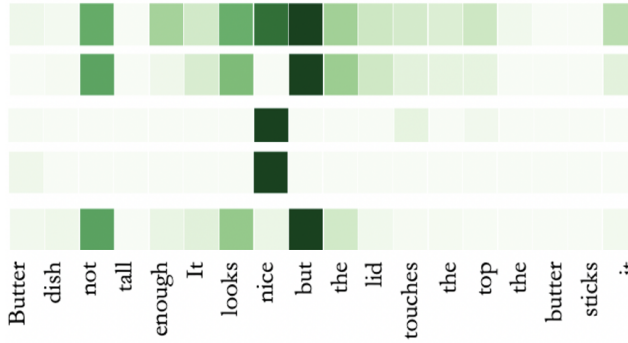BiLSTM + MHAD with 5 diverse attention layers: $y_{pred} = -ve$

Fig. 3. Two examples of *kitchen* review predictions by BiLSTM + ATT and BiLSTM + MHAD models trained on *book* reviews. When a single attention head fails to attend key words like 'excellent' or 'but', at least one of the diverse heads tends to make up for it.

ple shows misclassification by a single attention model that attends incorrect sentiment words like 'Vortex' and 'Heat'. However, with diversity, the model is lenient and look for alternate features. At least one of the T_y diverse heads

tends to find important words like 'excellent'. These examples also show that placing diversity on attention layers, rather than on any other hidden layers, provides an explainable understanding of which words the model deems to be important and can be used for subsequent pivot extraction like in AMN or HATN.

Computational Performance: To show that our work is practically useful for all communities alike, experiments are run on a CPU. A sample training time comparison is shown in Table 4. HATN needs gradient reversal to utilize unlabeled target data for the domain classifier branch and pivot extraction for joint training; subsequently making it much slower.

Gradient Reversal: To study the impact of gradient reversal procedure with BiLSTM, we conducted experiments with unlabeled target data. The performance of BiLSTM versus DANN$^+$ (improved DANN) models in Table 2 showed that, with a good dropout value for the BiLSTM units, gradient reversal did not help much. On a similar note, domain adversarial loss was found not to be helpful in tri-training experiments [21]. In our context, we speculate that this might be because the dropout in the BiLSTM layer drops individual words that can lead to a better generalization which is essentially the purpose of gradient reversal. This will be studied in our future work.

Table 5. Classification accuracy scores for three distinct combinations.

S → T	P-Net	BiLSTM +MHAD-Tri-II
Electronics → Yelp	88.45	89.15
Kitchen → IMDb	76.38	78.33
Yelp → IMDb	78.75	77.28

Additional Analysis: Generalizability of our models is further tested with three randomly selected experiments using very divergent domains such as Yelp[6] restaurant reviews and IMDb [17] movie reviews in addition to Amazon reviews; *Electronics (Amazon) → Yelp, Kitchen (Amazon) → IMDb*, and *Yelp → IMDb*. We randomly selected 2000 positive and negative reviews from Yelp and IMDb. Their accuracy scores on our final model when compared to PNet[7] of HATN is shown in Table 5. Once again, this shows that unlabeled target data is not always necessary; thus providing us with a fully unsupervised and computationally efficient alternative for domain adaptation in text classification tasks.

---

[6] https://www.yelp.com/dataset/challenge
[7] PNet is the first component of HATN which is computationally faster and within $\sim 1.5\%$ accuracy of HATN

6   Future Work

Experiments shown in the additional analysis section can be expanded to a lot more datasets and divergent domains; particularly applying to domains where unlabeled target data is not readily available such as during an onset of a natural or man-made disaster event. A careful empirical study of several existing complex architectures that employ adversarial training using the gradient reversal strategy is another direction. It is crucial to understand how much of the gain claimed by the adversarial training approach can actually be brought through by good generalization of the model without using any unlabeled data from the target domain. Recent progress in deep learning and natural language processing has seen impressive performance gain across various domains by using the transformer [8, 24] based models. Leveraging such models and studying the effect of encouraging diversity among the architectures appear to be a promising future direction.

7   Conclusion

Our study shows that machine learning architectures designed for achieving sufficient diversity in learning can generalize better for domain adaptation. Further, unlabeled target data, used often by state-of-the-art models, is not always necessary to produce strong performance for domain adaptation in subjective text classification problems. We introduced a novel diversity-based generalization approach for the domain shift problem using a multi-head attention model where attention heads are constrained to learn differently such that the classifier can leverage on alternative features. Experiments on the standard benchmark dataset of Amazon reviews and a newly constructed dataset of Crisis events showed that our fully unsupervised methods that completely avoid target data can indeed match the competing unsupervised baselines.

Reproducibility: Code, datasets, and documentation are available at -
https://github.com/jitinkrishnan/Diversity-Based-Generalization

Acknowledgements

References

1. Alam, F., Joty, S., Imran, M.: Domain adaptation with adversarial training and graph embeddings. arXiv preprint arXiv:1805.05151 (2018)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

3. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: Advances in neural information processing systems. pp. 137–144 (2007)
4. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th annual meeting of the association of computational linguistics. pp. 440–447 (2007)
5. Blitzer, J., McDonald, R., Pereira, F.: Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 conference on empirical methods in natural language processing. pp. 120–128 (2006)
6. Chen, M., Xu, Z., Weinberger, K., Sha, F.: Marginalized denoising autoencoders for domain adaptation. arXiv preprint arXiv:1206.4683 (2012)
7. Cui, W., Zheng, G., Shen, Z., Jiang, S., Wang, W.: Transfer learning for sequences via learning to collocate. arXiv preprint arXiv:1902.09092 (2019)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495 (2014)
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
11. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 513–520 (2011)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
13. Karuna, P., Rana, M., Purohit, H.: Citizenhelper: A streaming analytics system to mine citizen and web data for humanitarian organizations. In: Eleventh International AAAI Conference on Web and Social Media. pp. 729–730 (2017)
14. Li, Z., Wei, Y., Zhang, Y., Yang, Q.: Hierarchical attention transfer network for cross-domain sentiment classification. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
15. Li, Z., Zhang, Y., Wei, Y., Wu, Y., Yang, Q.: End-to-end adversarial memory network for cross-domain sentiment classification. In: IJCAI. pp. 2237–2243 (2017)
16. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
17. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: The 49th Annual Meeting of the Association for Computational Linguistics (2011)
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
19. Pan, S.J., Ni, X., Sun, J.T., Yang, Q., Chen, Z.: Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of the 19th international conference on World wide web. pp. 751–760. ACM (2010)
20. Purohit, H., Castillo, C., Imran, M., Pandey, R.: Social-eoc: Serviceability model to rank social media requests for emergency operation centers. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 119–126. IEEE (2018)

21. Ruder, S., Plank, B.: Strong baselines for neural semi-supervised learning under domain shift. arXiv preprint arXiv:1804.09530 (2018)
22. Saito, K., Ushiku, Y., Harada, T.: Asymmetric tri-training for unsupervised domain adaptation. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 2988–2997. JMLR. org (2017)
23. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
25. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research **11**(Dec), 3371–3408 (2010)
26. Zhang, K., Zhang, H., Liu, Q., Zhao, H., Zhu, H., Chen, E.: Interactive attention transfer network for cross-domain sentiment classification (2019)
27. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 207–212 (2016)
28. Zhou, Z.H., Li, M.: Tri-training: Exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge & Data Engineering (11), 1529–1541 (2005)