Sequential Domain Adaptation through Elastic Weight Consolidation for Sentiment Analysis

Avinash Madasu, Anvesh Rao Vijjini
Samsung R&D Institute, Bangalore, India
m.avinash@samsung.com, a.vijjini@samsung.com

Abstract

- Elastic Weight Consolidation (EWC) is a technique used in overcoming catastrophic forgetting between successive tasks
- Domain Adaptation (DA) aims to build algorithms that leverage information from source domains to facilitate performance on an unseen target domain.
- We propose a model-independent framework - Sequential Domain Adaptation (SDA).
- SDA draws on EWC for training on successive source domains to move towards a general domain solution, thereby solving the problem of domain adaptation.
- We test SDA on convolutional, recurrent, and attention-based architectures. Our experiments show that the proposed framework enables simple architectures such as CNNs to outperform complex state-of-the-art models in domain adaptation of SA.
- In addition, we observe that the effectiveness of a harder first Anti-Curriculum ordering of source domains leads to maximum performance.

Sequential Domain Adaptation

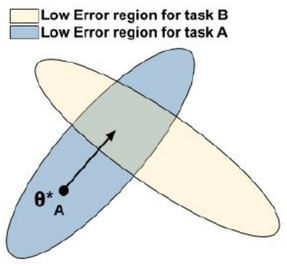


Fig. 1: EWC.

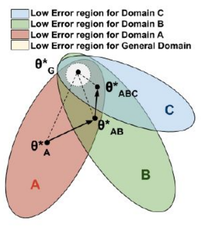


Fig. 2: Proposed framework SDA that facilitates Domain Adaptation through EWC.

1. (kirkpatrick et al. 2017) suggest that due to over parameterization, there exists a low error region in the parameter space for any task. All parameters within this space are equally optimal. When training on task B, using EWC we move to the common intersection with low error region of task B.
2. Each task could be interpreted as a domain. If there exists a general domain low error space, then successive EWC training across various domains will push the solution closest to this general domain low error space. G is the ideal general domain for which we lack any actual data. Such a general domain solution will maximize performance on unseen domains solving the problem of domain adaptation.
3. Since the framework only uses custom loss function from EWC, it is independent of the model architecture itself.

Experimental Setup

Datasets: Experiments on the standard Multi-Domain Sentiment Dataset. It contains reviews from 4 domains, namely Books (B), DVD (D), Electronics (E), and Kitchen (K). Each domain has 2000 reviews.

Architectures: CNN, LSTM, Attention LSTM (ALSTM) and Transformer Encoder (TE).

SoTA Domain Adaptation baselines:

1. PBLM: Pivot Based Language Model (PBLM)
2. DSR: Domain-Specific Representations
3. BLSE: Bilingual Sentiment Embeddings
4. DAS: Domain Adaptive Semi-supervised learning
5. ACAN: Adversarial Category Alignment Network

References

1) J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska et al., "Overcoming catastrophic forgetting in neural networks," Proceedings of the national academy of sciences, vol. 114, no. 13, pp. 3521–3526, 2017.

Visualizing SDA

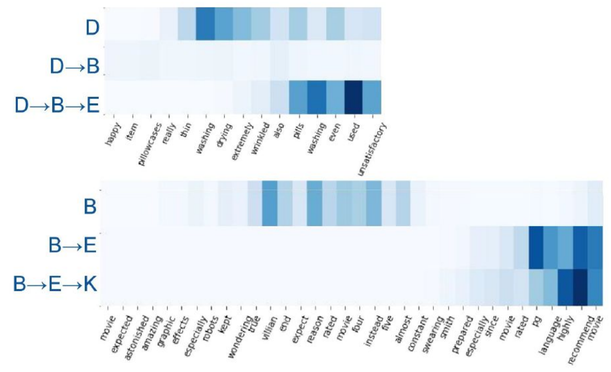


Fig. 3: Attention visualizations from ALSTM model.

In both examples, we see how the model with sequential EWC training, moves from arbitrary domain-specific attention to learning more general domain knowledge, helping it predict a sample of an unseen domain. Without the framework, the model fails to generalize from a single source domain and ends up making incorrect predictions.

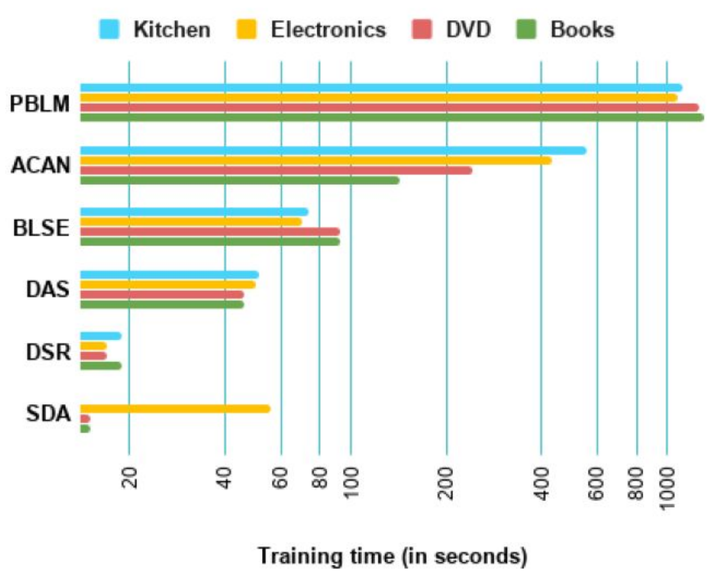
Results And Discussion

	Kitchen (K)	Electronics (E)	DVD (D)	Books (B)
PBLM [6]	68.95	59.4	59.55	61.4
DSR [10]	56	55	56.3	52
BLSE [18]	80	74.25	**75.25**	71.25
DAS [9]	70.75	72.35	60.85	62.6
ACAN [8]	79.55	73.65	70.8	**73.95**
EWC-CNN	**80.25**	77.25	**73.8**	**72.35**
EWC-LSTM	79.7	**78.25**	71.35	70.8
EWC-ALSTM	75	75.2	69.05	67.95
EWC-TE	71.35	65.25	65.45	66.6

TABLE II: Comparison of proposed framework SDA with state-of-the-art architectures on the Multi-Domain Sentiment Dataset. For the SotA, best performing source domain is chosen, while for SDA anti-curriculum source domain ordering is chosen.

1. Proposed outperforms all.
2. Some use Domain adaptation in a semi-supervised setting. In other words, they use large quantities of target domain unlabelled data Yet SDA outperforms all.
3. DSR, similar to the proposed framework, uses multiple source domains for learning domain representations.
4. Their model relies heavily on domain classification. However, training a robust domain classifier requires much more data,
5. BLSE outperforms SDA in one target domain, they utilize labelled target domain data for training their architecture, whereas as SDA keeps target domain strictly unseen.Compared to standard CNN on different text classification datasets.

Results and Discussion



1. Since we enable even simple CNN to outperform state-of-the art models, we also get an advantage in training time.
2. Apart from DSR, all architectures use only a single target domain and yet take much more time to train.
3. CNN not only performs best as we saw in Table II, but also is the quickest to train by a large margin.
4. Comparatively on Electronics domain, we see LSTM takes five times more time than CNN. However, it is still five to thirty times less than PBLM and ACAN.
5. This shows the efficiency of SDA at empowering low parameter models such as CNN to match large architectures.