Improving the Quality Trade-Off for Neural Machine Translation Multi-Domain Adaptation

Eva Hasler∗ Tobias Domhan∗ Jonay Trenous Ke Tran Bill Byrne Felix Hieber
Amazon AI Translate
Berlin, Germany
{ehasler,domhant,trenous,trnke,willbyrn,fhieber}@amazon.com

Abstract

Building neural machine translation systems to perform well on a specific target domain is a well-studied problem. Optimizing system performance for multiple, diverse target domains however remains a challenge. We study this problem in an adaptation setting where the goal is to preserve the existing system quality while incorporating data for domains that were not the focus of the original translation system. We find that we can improve over the performance trade-off offered by Elastic Weight Consolidation with a relatively simple data mixing strategy. At comparable performance on the new domains, catastrophic forgetting is mitigated significantly on strong WMT baselines. Combining both approaches improves the Pareto frontier on this task.

1 Introduction

The quality of Neural Machine Translation (NMT) has improved considerably in recent years, mostly due to improvements in model architecture (Bahdanau et al., 2015; Cho et al., 2014; Vaswani et al., 2017; Chen et al., 2018). Training NMT models typically involves collecting parallel training data from multiple sources to achieve high translation quality and generalize well to unseen data (Barrault et al., 2019). However, translation quality depends strongly on the relevance of the training data to the input text, which is why performance varies across target domains (Koehn and Knowles, 2017a).

A popular method for domain adaptation of NMT models is fine-tuning generic models on in-domain data to yield a domain-specific model (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016). When high quality output on more than one target domain is required, multi-domain adaptation methods aim to produce a single system that performs well on multiple domains (Britz et al., 2017; Pham et al., 2019; Currey et al., 2020).

∗Equal contributions.

Our goal is to train a single NMT system per language pair that performs well across many different domains. This is motivated by simplified deployment and maintenance in an industrial setting with hundreds of supported language pairs. At any point in the deployment cycle, new parallel training data – often significantly smaller than the original training data – may become available for an additional domain that the system has not yet been optimized for. Depending on the size of this additional data, fully retraining the NMT system may not be practical as it would require costly experimentation to find the right level of upsampling which might in turn lead to overfitting on that data. In addition, as system stability is desirable in an industrial setting, we want to maintain the status-quo performance – or generic domain performance – of our models which is easier to control in an adaptation setting.

In this paper, we explore the following research question: given a strong general-purpose model, how can we optimize the performance on multiple new, diverse domains of interest without compromising on generic domain performance?

A naive strategy would be to fine-tune on the new domain data and stop as soon as performance starts to decrease on the generic test set(s). However, this method allows for limited gains on the new domains as we quickly start observing catastrophic forgetting: performance on previously learned tasks degrades while increasing on the newly learned tasks (Kirkpatrick et al., 2017).

We therefore experiment with Elastic Weight Consolidation (EWC) to preserve the generic performance of our model during adaptation (Kirkpatrick et al., 2017). We corroborate the finding of Thompson et al. (2019) and Saunders et al. (2019) that EWC helps to reduce catastrophic forgetting in machine translation adaptation. However, we find the quality trade-off in our multi-domain setting to be unfavourable: when preserving most of the generic performance, the gains on the new domains
with EWC are limited. We further experiment with data mixing strategies to mitigate catastrophic forgetting and find that they are surprisingly effective. In summary, we make the following contributions:

- We provide a thorough comparison between data mixing and EWC to prevent catastrophic forgetting in a multi-domain adaptation setup.
- We show that combining EWC and data mixing outperforms EWC and provides a knob for regulating the performance trade-off with data mixing. Combining both approaches improves the Pareto frontier, thus striking a better balance than adaptation with EWC alone.
- We provide a theoretical analysis showing that regularization in data space and in parameter space are complementary within the Bayesian formulation of continued learning.

2 Related work

Most previous work on multi-domain adaptation focuses on a scenario with fixed training data. For example, Currey et al. (2020) use knowledge distillation to build a single model from expert models optimized for the training domains, Britz et al. (2017) train models that better distinguish between the training domains and Pham et al. (2019) learn domain-specific word embeddings for the domains present in the training data. In contrast, we focus on the adaptation setting where additional domain data becomes available over time.

Thompson et al. (2019) apply EWC for adaptation to a single new domain while Saunders et al. (2019) use it for sequentially adapting to two new domains. Both report positive results but at the same time show a performance trade-off which our work tries to address further.

Mixing out-of-domain and in-domain data for fine-tuning was proposed by Chu et al. (2017) who use tags to distinguish domains at test time while our models are domain-agnostic. Data mixing is also related to work on Episodic Memories for continual learning. For example, Chaudhry et al. (2019) show that a random sample of previous task data can outperform EWC for image recognition.

3 Multi-domain adaptation

Our goal is to optimize translation quality for several new domains represented by small amounts of parallel data while maintaining the performance of a high-quality, general-purpose NMT model. We focus on the scenario where only small amounts of additional data are available since it is suitable for an adaptation setup. For large amounts of additional data, retraining the model from scratch might be a more suitable approach.

3.1 Elastic Weight Consolidation

Kirkpatrick et al. (2017) study the problem of catastrophic forgetting in sequential machine learning settings. They propose EWC as a method to preserve model performance during sequential learning of task $B$ by selectively slowing down learning on the weights that are important for the original task $A$ learned by the model. This goal is achieved by adding a loss term to the training objective as shown in Equation 1:

$$ L = L_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta_A^*)^2, \quad (1) $$

where $\theta$ is a set of model parameters, $L_B(\theta)$ is the loss for task $B$ and task $A$ is represented by the parameters $\theta_A^*$ and the diagonal of the Fisher information matrix $F$. The strength of the regularization is controlled by $\lambda$ which can be used to balance the performance on task $A$ versus task $B$. Intuitively, this loss encourages updates to the model in a direction that improves the performance on task $B$ without altering the crucial parameters for task $A$ too much. In our setting, task $A$ represents the generic training, while task $B$ represents the specific domains we adapt to.

3.2 Data mixing

A simple, data-driven strategy to counteract catastrophic forgetting is to interleave weight updates according to the new domain gradients with weight updates according to the original training data gradients. This can be implemented by combining the domain-specific adaptation set with a sample of the original training data. We can increase the importance of the training data sample by increasing its size, thereby changing the ratio of training data and domain data to influence the trade-off between generic and domain performance. Conceptually, data mixing is similar to Episodic Memories where a memory of examples from all previous tasks is kept during continual learning (Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2019). Different from the mixed fine-tuning of Chu et al. (2017), the domain is not known at test time in our case.
3.3 EWC + data mixing

Combining EWC and data mixing is motivated by the need to improve the quality trade-off offered by EWC while retaining the ability to control a hyperparameter that does not affect the size of the adaptation set and thereby the number of training steps in an epoch. From a theoretical perspective, this can be justified as follows: EWC approximates \( \log p(\theta | A, B) \) under the strict conditional independence assumption \( P(B | A, \theta) = P(B | \theta) \), i.e. \( A \) and \( B \) are conditionally independent given \( \theta \). This may be too harsh for the case where \( A \) and \( B \) are language domains. Suppose that \( A \) is partitioned into two sets \( A_1 \) and \( A_2 \) where \( A_1 \) is a random sample of \( A \), much smaller than \( A_2 \). This allows the more relaxed conditional independence approximation \( P(B | A_1, A_2, \theta) = P(B | A_1, \theta) \), which assumes the sample \( A_1 \) says enough about the generic domain \( A \) that \( A_2 \) can be discarded given \( \theta \) and \( A_1 \). It can be shown that under this assumption the EWC objective becomes

\[
\mathcal{L}' = \mathcal{L}_B(\theta) + \mathcal{L}_{A_1}(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta_{A_1,i})^2
\]

which is equivalent to mixing the sampled set \( A_1 \) into the new domain data \( B \) as described here. See Appendix A for the full derivation.

4 Experiments

We evaluate multi-domain adaptation on top of two strong WMT baselines: German→English (DE→EN) and English→French (EN→FR).

4.1 Experimental setup

Train details We train Transformer models using the Sockeye 2 toolkit (Domhan et al., 2020) in the big variant with six encoder and decoder layers (Vaswani et al., 2017), using Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.06325 and a linear warmup over 4000 training steps. We use the constrained data settings from WMT20 (Barrault et al., 2020) and WMT15 (Stanojević et al., 2015) respectively (for EN→FR, we add newstest2008-2013 as additional training data) and train until convergence determined on a held-out validation set. We remove noisy pairs based on heuristics (length ratio > 1.5, > 70% token overlap, > 100 BPE tokens) and those where source or target language does not match according to LangID (Lui and Baldwin, 2012). We tokenize the data using sacremoses\(^1\), truecase the data, then apply Byte Pair Encoding (BPE) (Sennrich et al., 2016) with 32,000 merge operations. For EN→FR, we apply an additional normalization step after detokenization replacing single curly quotes surrounded by spaces with a single straight quote. This is to avoid conflating the actual domain translation quality gains with punctuation differences.

The baseline performance is 42.7 BLEU on newstest2019 and 41.8 BLEU on newstest2020 for our DE→EN system. Our EN→FR system yields 41.2 BLEU on newstest2014 and 39.2 BLEU on newstest2015. We evaluate using SacreBLEU (Post, 2018)\(^2\) on detokenized outputs.

<table>
<thead>
<tr>
<th>Domain</th>
<th>adapt</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE→EN</td>
<td>TED</td>
<td>9355</td>
<td>500</td>
</tr>
<tr>
<td>EMEA</td>
<td>10000</td>
<td>500</td>
<td>3000</td>
</tr>
<tr>
<td>EN→FR</td>
<td>law</td>
<td>10000</td>
<td>500</td>
</tr>
<tr>
<td>IT</td>
<td>10000</td>
<td>500</td>
<td>3000</td>
</tr>
</tbody>
</table>

Table 1: Number of (subsampled) adaptation, development and test examples for multi-domain adaptation.

---

\(^1\)https://github.com/alvations/sacremoses

\(^2\)With identifier BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+tok.13a+version.1.4.14.
4.2 Experimental results

Figure 1a shows DE→EN adaptation results where the adapted performance on the additional domains is represented as mean BLEU score across all target domains (x-axis) and generic performance is represented as mean BLEU score across newstest2018 and newstest2019 test sets (y-axis). Adapt denotes vanilla fine-tuning and for $\lambda \rightarrow 0$, EWC approaches vanilla fine-tuning. Although EWC succeeds in mitigating catastrophic forgetting, as seen by the reduced drop in BLEU on the news test sets, this comes at a considerable cost in terms of domain quality. In comparison, data mixing with a 1:1 ratio of train sample/adaptation data allows for high quality on the adapted domains while retaining substantially higher generic performance than
eWC (rightmost point on the data mixing curve). However, generic performance is not fully restored when increasing the ratio from 10:1 to 100:1, thus, altering the ratio does not reliably interpolate between generic and adapted domain performance. Thanks to the strength parameter $\lambda$, the combination of EWC and data mixing is able to provide this interpolation and yields an improved Pareto frontier for this task. For similar BLEU scores on the adapted domains (40.0 vs 40.2), EWC + data mixing with a 1:1 training sample/domain data ratio yields an improvement of 2 BLEU on news over EWC with $\lambda=10^{-5}$ (44.0 vs 42.0).

The EN→FR results in Figure 1b follow a similar trend. Here the improvement of EWC + data mixing over EWC is 0.8 BLEU on news (39.5 vs 38.7) for similar scores on the adapted domains of 47.2 (data mixing + EWC) and 47.1 (EWC) BLEU.

One explanation for this could be that for large data sets, finding a representative data sample is more difficult.
Table: BLEU scores on the generic sets.

<table>
<thead>
<tr>
<th>Domain</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TED</td>
<td>42.0</td>
</tr>
<tr>
<td>EMEA</td>
<td>38.5</td>
</tr>
<tr>
<td>IT</td>
<td>42.0</td>
</tr>
<tr>
<td>Law</td>
<td>38.5</td>
</tr>
</tbody>
</table>

We also show the effect of increasing or decreasing the learning rate of $2e^{-5}$ for EWC with and without data mixing. As expected, increasing the learning rate (lr=$2e^{-4}$) yields more forgetting on the generic sets while decreasing it (lr=$2e^{-6}$) yields smaller improvements on the adapted domains. The improvement of EWC + data mixing is robust to those changes, though, as the setting with 20x upsampling and lr=$2e^{-5}$ still yields the best results compared to EWC with different learning rates.

5 Conclusion

We investigated techniques to mitigate catastrophic forgetting during NMT model adaptation in order to optimize for new domains while maintaining the quality of already deployed systems. We found that data mixing provides a favourable quality trade-off and improves the Pareto frontier when combined with EWC. We showed that data mixing is robust to random sampling and sample size and that our reported gains persist for different learning rates.
References


A Combining Elastic Weight Consolidation (EWC) and data mixing

EWC

• EWC attempts to maximize $\log p(\theta | A, B)$ for sets $A$ and $B$, assuming $B$ follows $A$.

• EWC uses the Laplace approximation $p(\theta | A) \propto N(\theta; \theta^*, F^{-1})$ so that, ignoring terms that do not depend on $\theta$,

$$\log p(\theta | A) = \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta_{*A,i})^2$$

• EWC makes the assumption that $p(B | A, \theta) = p(B | \theta)$, i.e. that $B$ is conditionally independent of $A$ given $\theta$

$$p(\theta | A, B) = \frac{p(B, A, \theta)}{p(A, B)} = p(B | A, \theta) \frac{p(\theta) p(A)}{p(A, B)} = p(B | \theta) \frac{p(\theta) p(A)}{p(A, B)} \text{ assuming } p(B | A, \theta) = p(B | \theta)$$

• Ignoring terms that do not depend on $\theta$, the EWC criterion is

$$\mathcal{L}(\theta) = \log p(\theta | A, B) = \log P(B | \theta) + \log p(\theta | A) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta_{*A,i})^2$$

EWC + data mixing

• Suppose $A$ is partitioned into $A_1, A_2$, via a random sampling, with $A_1$ much smaller than $A_2$.

• Replacing $A$ by $A_1, A_2$ in the EWC derivation above leads to

$$p(\theta | A_1, A_2, B) = \frac{p(B, A_1, A_2, \theta)}{p(A_1, A_2, B)} = \frac{p(B, A_1, A_2, \theta)}{p(A, B)}$$

$$= p(B | A_1, A_2, \theta) p(A_1 | A_2, \theta) p(\theta | A_2) \frac{p(A_2)}{p(A, B)}$$

$$= p(B | A_1, \theta) p(A_1 | A_2, \theta) p(\theta | A_2) \frac{p(A_2)}{p(A, B)} \text{ assuming } p(B | A_1, \theta) = p(B | A_1, A_2, \theta)$$

$$= p(B | A_1, \theta) p(A_1 | \theta) p(\theta | A_2) \frac{p(A_2)}{p(A, B)} \text{ assuming } p(A_1 | A_2, \theta) = p(A_1 | \theta) \text{ (i.i.d. over } A)$$

$$= p(B, A_1 | \theta) p(\theta | A_2) \frac{p(A_2)}{p(A, B)} \text{ assuming } p(\theta | A_2) = p(\theta | A)$$

• Ignoring terms that do not depend on $\theta$, the EWC + mixing criterion is

$$\mathcal{L}'(\theta) = \log p(B, A_1 | \theta) + \log p(\theta | A) = \mathcal{L}_{B,A_1}(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta_{*A,i})^2$$