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).

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

[*] Equal contributions.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8470–8477
November 7–11, 2021. ©2021 Association for Computational Linguistics

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:

$$\mathcal{L} = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2, \qquad (1)$$

where θ is a set of model parameters, $\mathcal{L}_B(\theta)$ is the loss for task B and task A is represented by the parameters θ^*_A and the diagonal of the Fisher information matrix F. The strength of the regularization is controlled by λ 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 θ. 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 θ and A_1. It can be shown that under this assumption the EWC objective becomes

$$\mathcal{L}' = \mathcal{L}_B(\theta) + \mathcal{L}_{A_1}(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2 \quad (2)$$

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.

	Domain	adapt	dev	test
DE→EN	TED	9355	500	1305
	Tanzil	10000	500	3000
	WMT20chat	11279	500	2100
EN→FR	EMEA	10000	500	3000
	law	10000	500	3000
	IT	10000	500	3000

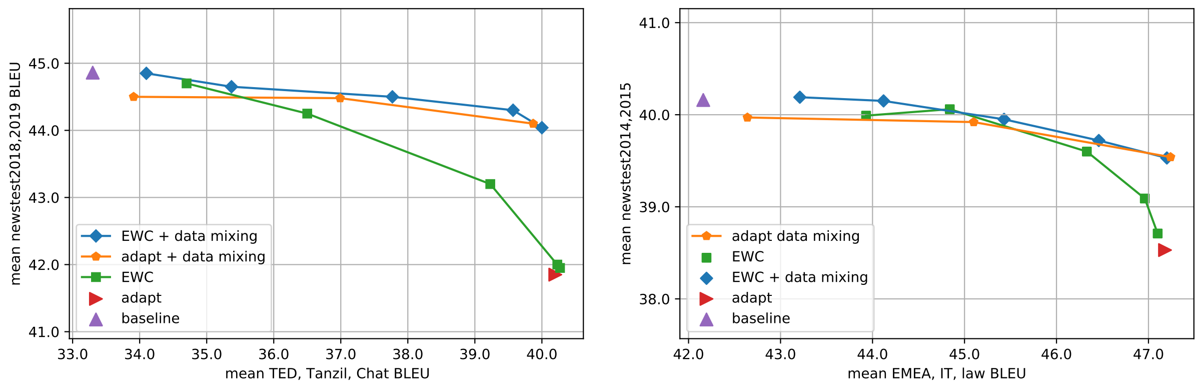
Table 1: Number of (subsampled) adaptation, development and test examples for multi-domain adaptation.

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.

Adaptation details We use TED (Cettolo et al., 2016), Tanzil (Tiedemann, 2012) and WMT20chat (Farajian et al., 2020) corpora as additional target domains for DE→EN and EMEA, law and IT corpora (Tiedemann, 2012) for EN→FR. IT is a combination of the GNOME, KDE, PHP, Ubuntu, and OpenOffice corpora (Koehn and Knowles, 2017b). For EMEA, law, IT and Tanzil we randomly sample 10k, 500 and 3k sentences for adaptation, development and test data, respectively. For TED we use 2010-2014 TED/TEDX development and test sets, except for test2014, for sampling adaptation and development sets and test on test2014. The adaptation sets consist of examples from all target domains, roughly balanced in size. We choose ∼10k examples to match our scenario of adaptation with little parallel data. Adaptation set sizes are shown in Table 1. For the adaptation step, we use dev set BLEU on the concatenation of domain-specific development sets for early stopping and checkpoint selection. After preliminary experiments, we chose a reduced initial adaptation learning rate of 2e-5 without warmup since adapta-

[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.



(a) DE→EN

(b) EN→FR

Figure 1: Adaptation results varying λ for EWC (left to right from 10^{-1} to 10^{-5}) and the train sample/domain data ratio (100:1, 10:1 and 1:1) for data mixing. For EWC + data mixing, the train sample/domain data ratio is 1:1.

tion starts from a fully trained model.

Training data samples For data mixing, we concatenate a sample from the training data of the baseline system to the adaptation data. This training sample is of equal size to the domain-specific set by default. In order to avoid overfitting to the training data sample during adaptation, we upsample the domain-specific adaptation set 20x and concatenate a training sample of the increased size for a 1:1 train sample and domain data ratio.

EWC We compute the diagonal of the empirical Fisher information matrix using accumulated, averaged gradients from the original training data over 200 training steps after convergence. We validated empirically that increasing the number of steps to 2,000 or 20,000 does not significantly change the results. The Fisher information values are normalized and we vary the strength of the EWC loss by setting $\lambda = \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$.

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[3]. Thanks to the strength parameter λ, 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 λ=10^{-5} (44.0 vs 42.0) .

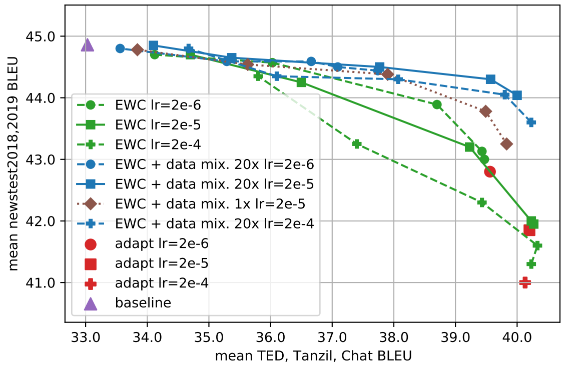
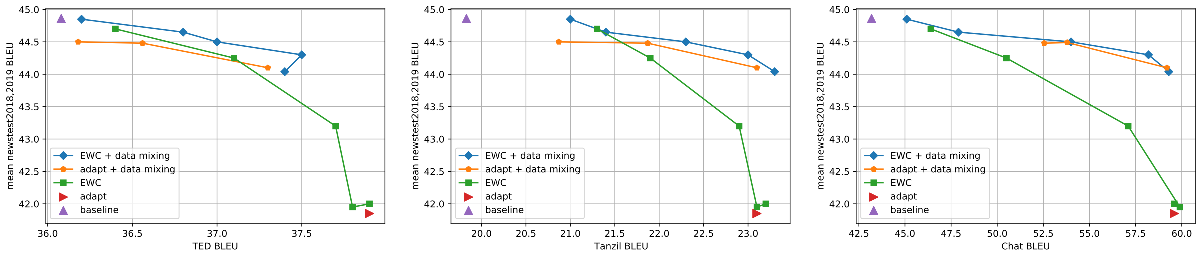


Figure 2: DE→EN adaptation results varying the learning rate (lr) and the size of the training data sample while maintaining a 1:1 train/domain ratio. The dots on each curve correspond to varying λ values.

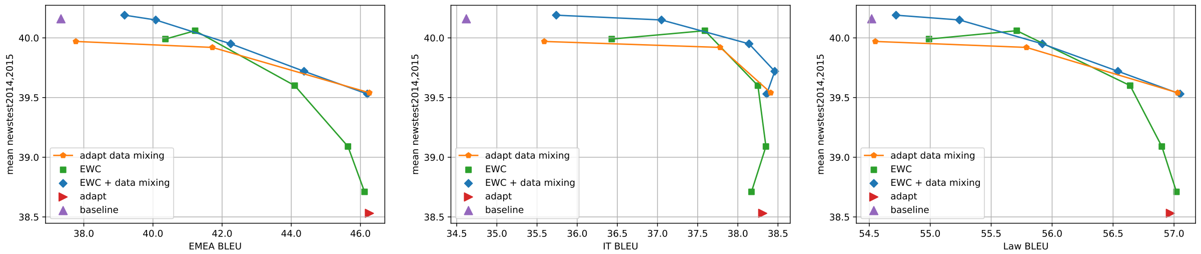
The EN→FR results in Figure 1b follow a similar trend. Here the improvement of EWC + data mixing over EWC is 0.8 BLEU on news (39.5 vs 38.7) for similar scores on the adapted domains of 47.2 (data mixing + EWC) and 47.1 (EWC) BLEU.

[3] One explanation for this could be that for large data sets, finding a representative data sample is more difficult.



(a) TED (b) Tanzil (c) Chat

Figure 3: DE→EN adaptation results per domain varying λ for EWC (decreasing from left to right) and the train sample/domain data ratio (100:1, 10:1 and 1:1) for data mixing. For EWC + data mixing, the train sample/domain data ratio is 1:1.



(a) EMEA (b) IT (c) Law

Figure 4: EN→FR adaptation results per domain varying λ for EWC (decreasing from left to right) and the train sample/domain data ratio (100:1, 10:1 and 1:1) for data mixing. For EWC + data mixing, the train sample/domain data ratio is 1:1.

Adaptation scores per domain Figures 3 and 4 show the results for each domain individually. Overall, the trends are similar across all domains, with the combination of EWC and data mixing offering the best trade-off between generic and domain performance. For all domains except TED for DE→EN we observe that the domain performance of EWC + data mixing is similar or better than vanilla adaptation while preserving more of the translation quality on news.

4.3 Robustness of data mixing & learning rate

Data mixing uses a random sample of the original training data. We check its robustness by sampling with different random seeds. The mean of the BLEU standard deviations across all generic and domain-specific test sets is 0.2, showing that the results are sufficiently robust to different random samples. Figure 2 shows the effect of upsampling the adaptation data and training sample 20x compared to 1x (no upsampling), i.e. using a smaller training sample that matches the original size of the adaptation data. While we achieve good results even without upsampling, it yields slightly higher scores on the generic sets.

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. For completeness, we also show that varying the learning rate for vanilla adaptation does not yield stronger results.

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Third International Conference on Learning Representations*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126, Copenhagen, Denmark. Association for Computational Linguistics.

M. Cettolo, Niehues Jan, Stüker Sebastian, L. Bentivogli, R. Cattoni, and Marcello Federico. 2016. The IWSLT 2016 evaluation campaign.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. 2019. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86, Melbourne, Australia. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Anna Currey, Prashant Mathur, and Georgiana Dinu. 2020. Distilling multiple domains for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4500–4511, Online. Association for Computational Linguistics.

Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The sockeye 2 neural machine translation toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 110–115, Virtual. Association for Machine Translation in the Americas.

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75, Online. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Third International Conference on Learning Representations*.

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.

Philipp Koehn and Rebecca Knowles. 2017a. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017b. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

David Lopez-Paz and Marc' Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

Minh Quang Pham, Josep-Maria Crego, François Yvon, and Jean Senellart. 2019. Generic and specialized word embeddings for multi-domain machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Danielle Saunders, Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2019. Domain adaptive inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 222–228, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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 $p(\theta|A) \propto \mathcal{N}(\theta\,;\,\theta^*, F^{-1})$ so that, ignoring terms that do not depend on θ,

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

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

$$\begin{aligned} p(\theta|A, B) &= \frac{p(B, A, \theta)}{p(A, B)} \\ &= p(B|A, \theta)\, p(\theta|A)\, \frac{p(A)}{p(A, B)} \\ &= p(B|\theta)\, p(\theta|A)\, \frac{p(A)}{p(A, B)} \qquad \text{assuming } P(B|A, \theta) = P(B|\theta) \end{aligned}$$

- Ignoring terms that do not depend on θ, 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

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

- Ignoring terms that do not depend on θ, the EWC + mixing criterion is

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