Investigating Gradients in Multilingual Neural Machine Translation

Tianjian Li, Haoping Yu, Mingxuan Che
Johns Hopkins University

Abstract

Multilingual Neural Machine Translation (NMT) (Johnson et al. (2017); Aharoni et al. (2019); Aharoni et al. (2019)) has improved the performance of translation systems and also enabled zero resource translation, which models can translate in directions without parallel text. Multilingual NMT systems can be viewed as multi-task learning. While previous studies have applied multi-task gradient optimization methods (Liu et al. (2021); Yu et al. (2020)) to improve multilingual NMT systems (Yang et al. (2021); Wang et al. (2021)) little work has focused on the dynamics of gradients that their relatedness to linguistic features. In our work, we analyze the gradient similarities between different language pairs and reveal that gradient similarities reflect more of language order similarity than script and language similarity. Moreover, we reveal that current optimization tricks are sensitive to hyperparameters and does not clearly outperform weighted sum of the task losses (Xin et al., 2022).

1 Introduction

Neural methods have become the prevailing solution to machine translation. However, building a decent Neural Machine Translation (NMT) requires a large amount of text (Koehn and Knowles (2017)). Therefore, the quality of NMT systems on low resource language pairs are subpar and improving performance on low resource languages remains to be one of the major challenges of Neural Machine Translation.

To build more accurate NMT systems for low resource language pair, a promising direction is to utilize cross-lingual transfer leaning. Multilingual NMT systems trained on various language pairs can achieve better performance on a low resource language pair than a model explicitly trained on such language pair, even through the multilingual MT model does not see this pair during training (Aharoni et al. (2019)). Furthermore, such multilingual MT models can also be fine-tuned to achieve better performance on desired translation directions (Liu et al. (2020); Song et al. (2019); Johnson et al. (2017)).

The multilingual training paradigm, which trains a translation model on multiple language pairs, can be viewed as a Multi-Task Learning(MTL) method where each task is a different language pair. Numerous optimization algorithms that encourage gradient align between tasks have been proposed for the past few years (Yu et al. (2020); Wang et al. (2021); Liu et al. (2021); Lee et al. (2022)). However, recent studies have shown such optimization algorithm does not yield improvement on multilingual NMT systems compared to simply weighting the loss of different training language pairs (Xin et al. (2022)).

There has been work that aim to analyze representations in Multilingual Neural Machine Translation. However, work on NMT intepretability have focused on understanding the attention mechanism (Michel et al., 2019); the hidden representations (Kudugunta et al., 2019); the importance of individual words (He et al., 2019)); and the influence of different training objectives(one-to-many, many-to-one, many-to-many) (Chiang et al., 2022). In our work, we aim to understand the dynamics of gradient in different training directions of multilingual neural machine translation. We believe our research shed light on understanding the training dynamics and how should we choose training languages in multilingual machine translation.

Our work aims to answer the following research questions:

- How does the the gradient similarity between language pairs change as we train more iterations?
- What are the specific linguistic properties that are most reflected by the gradient similarities?
- Does current complex gradient optimization methods outperform simple optimization tricks?

2 Related Works

2.1 Multilingual Machine Translation

Neural Machine Translation have undergone the transition from dedicated bilingual models to Multilingual Models, which not only outperform bilingual baselines (Aharoni et al. (2019)), but also enables zero-shot translation (Johnson et al. (2017)), where a model performs translation task in a pair that it is not explicitly trained on. Another line of work reveals that large language models trained with a multilingual denoising objective (Liu et al. (2020)) or a machine translation objective can learn cross-lingual representations that enables zero-shot transfer.

2.2 Multi-Task Optimization

Gradient optimization techniques aim to reduce the discrepancy in directions of conflicting gradients: **PCGrad** (Yu et al. (2020)) aims to project the gradient of a task onto the orthogonal plane of the gradients of the other task, which explicitly aligns the gradient of different tasks. **Gradient Vaccine** (Wang et al. (2021)) points out that using PCGrad to optimize multilingual pre-training relies on the false assumption that all languages are equally related. To solve this issue, Gradient Vaccine aligns the similarity of the gradient to be the relatedness of the two languages. We mathematically formalize PCGrad and Gradient Vaccine at §4.1. Figure 1 depicts various optimization strategies and we leave the details of the algorithms at Appendix A.

Algorithms that both optimizes direction and magnitude of gradients has been proposed. Either by directly optimizing the cosine similarity of the un-normalized gradients (Lee et al. (2022)), or by iteratively optimizing the direction and magnitude (Javaloy and Valera (2022)). However, recent studies points out that such MTL optimization techniques does not yield significant improvement upon a weighted sum of individual task losses (Xin et al. (2022)).

2.3 Interpreting Multilingual NMT

Numerous work aim to interpret the word and sentence representations of multilingual NMT systems. Michel et al. (2019) investigated the usefulness of the multi-headed attention mechanism (Vaswani et al., 2017) by masking out individual attention heads. Kudugunta et al. (2019) investigated on the hidden representations of the encoder and decoder, finding out that the encoder output representations are clustered by language similarity. Wang et al. (2021) shows that the gradient similarity also reflects language family relatedness. To the best of our knowledge, our work is the first to study the importance of linguistic properties on gradient similarities.

3 Challenges in Optimizing Multilingual Machine Translation

In this section, we provide analysis of the challenges in multilingual machine translation systems. The most common issue in multilingual machine translation is that it often contains off-target translations (Anonymous (2022), which is more severe in multilingual MT systems that contains one centric langauge (Anonymous (2023)). Johnson et al. (2017) proposed to use target language tags to tell the model in which language it should generate. Numerous data augumentation techniques include back-translation (Zhang et al. (2020)), adopting pre-trained language models (Gu et al. (2019)), utilizing transliteration to add noise (Sun et al. (2022)) have been proposed to mitigate this issue.

The closest to our work is Yang et al. (2021), which proposes two additional methods to mitigate the off-target translation issue that often occurs in multilingual machine translation system with a centric language (Anonymous (2022)). 1) Adding an additional target language prediction task by the decoder's final hidden states into a linear classifier and 2) Constructing an oracle gradient update direction with the development set and use PCGrad (Yu et al. (2020)) to project all the in-

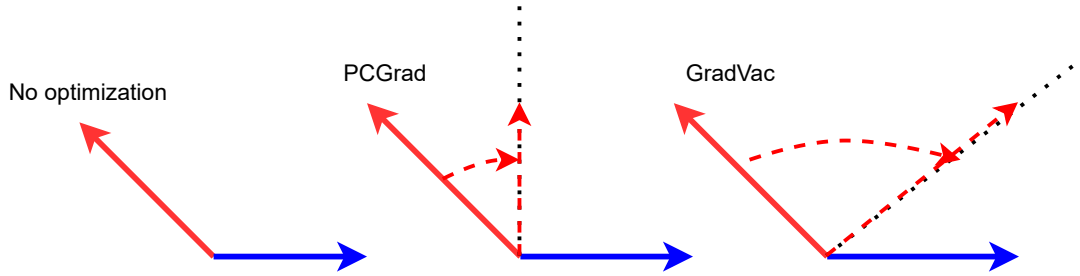

Figure 1: Figure of gradient vectors with and without gradient optimization, the red dashed arrow is the projected gradient ofter optimization.

dividual gradients onto that direction.

4 Problem Formulation

4.1 Setup

We focus on the one-to-many scenario, where the source language is only one language and we select two target languages to calculate the gradient similarity between them. Formally, our multilingual machine translation systems learns a mapping from text of the source language s to text of several target languages $t = \{t_1, t_2, ..t_n\}$. The model is trained with maximum likelihood estimation objective. Formally, the model learns a parameter θ that maximizes the conditional probability of the target sentences given the source sentences:

$$\theta := \arg\max_{\theta} \prod_{i=1}^{n} p_\theta(t_i|s)$$

We can view each translation pair $S \rightarrow t$ as an individual task and view the entire training process as an instance of **Multi-Task Learning**(MTL). Each individual translation task $S \rightarrow t_i$ has a gradient g_i. And the gradient similarity is defined to be the cosine similarity between the two gradient vectors:

$$\text{Sim}(g_i, g_j) = \frac{g_i \cdot g_j}{\|g_i\|\|g_j\|}$$

4.2 Gradient Optimization

In this section, we introduce the details of the two algorithms PCGrad (Yu et al., 2020) at §4.2.1 and Gradient Vaccine (Wang et al., 2021) at §4.2.2. And we describe our method of random projection at §4.2.3

4.2.1 PCGrad

Given two gradient vectors g_i and g_j, PCGrad projects g_i onto the orthogonal plane to g_j only if the angle between g_i and g_j is obtuse. This is done by subtracting the projection of g_i onto g_j from g_i. Formally,

$$\text{PCGrad}(g_i, g_j) = \begin{cases} g_i - \frac{g_i \cdot g_j}{\|g_j\|^2} & \text{if Sim}(g_i, g_j) < 0 \\ g_i & \text{otherwise} \end{cases}$$

4.2.2 Gradient Vaccine

Gradient Vaccine argues that PCGrad effectively views any two tasks with conflicting gradients as equal by artificially fixing the cosine similarity between them to be 0. (Wang et al., 2021) argues that the cosine similarity between translation direction reflect language relatedness, therefore we should encourage training directions with similar languages to have a higher cosine similarity.

To achieve relatedness between gradient similarity and language similarity. Wang et al. (2021) propose to use a predefined cosine similarity α at each layer, and manually fix the cosine similarity of the tasks to be α regardless of positive or negative value of the un-optimized raw cosine similarity. Formally, given two gradient vectors g_i and g_j, and a predefined cosine similarity α:

$$\text{GradVac}(g_i, g_j, \alpha) =$$

$$g_i + \frac{\|g_i\|(\alpha^\top \sqrt{1-\alpha^2} - \alpha\sqrt{1-\alpha^2})}{\|g_j\|\sqrt{1-\alpha^2}} \cdot g_j$$

Manually defining an α for every language pair, layer and training step is expensive, so the authors use the exponential moving average of gradient similarities of previous training steps for setting

the α of the current training step. Formally, if the current training step is t, the predefined cosine similarity of the current training step $\hat{\alpha}^t$ is:

$$\hat{\alpha}^t = \alpha^t + (1-\beta)\hat{\alpha}^{t-1}$$

where α_t is the calculated raw gradient similarity of the model at training step t, and β is a hyperparameter.

4.2.3 Random Projection

Our random projection method is to check the validity of Gradient Vaccine. We randomly choose an value $\alpha_{\text{rand}} \in (0, 1)$ at each time step t and project the gradient of g_i onto the direction such that $Sim(g_i, g_j) = \alpha_{\text{rand}}$ using the gradient vaccine method.

5 Experiments

5.1 Dataset

We use the multi-way parallel Ted Talk dataset (Duh (2018)). Following previous work (Aharoni et al. (2019)), we use a language tag [2xx] specifying the target language we want to translate into. For example, if we want to translate to french, the tag would be [2fr].

First, we aim to see if the language relatedness corresponds to the gradient similarities without any optimization techniques, i.e., if the gradient of French $\rightarrow$ English is closer to German $\rightarrow$ English, and further to Korean $\rightarrow$ English.

Then, we aim to evaluate ablate the effect of gradient optimization techniques on Multilingual Machine Translation by 1) Not using any optimization techniques, 2) Optimizing the gradients with Gradient Vaccine (Wang et al., 2021) and 3) randomly choosing an positive number α between 0 and 1 and force the cosine similarity to be α.

We use the official implementation of PCGrad (Yu et al. (2020)) and since we could not found an official implementation of Gradient Vaccine (Wang et al. (2021)), we implement it ourselves and evaluate on our multilingual machine translation model.

5.2 Model

The structure of our model is a standard Transformer (Vaswani et al. (2017)) but much smaller due to limited computational resources. We use a 3 layer encoder and 3 layer decoder with hidden dimension 256 and feed forward dimension of 512. We use the fairseq (Ott et al. (2019)) implementation of the transformer architecture and we modify upon this to implement the optimization methods. We use SentencePiece (Kudo and Richardson, 2018) to preprocess the text and train a joint dictionary of vocabulary size of 16000 for the source language and the two target languages. We modify the fairseq library to implement Gradient Vaccine (Wang et al., 2021) and random projection of gradients.

Directions	Bilingual	Multilngual	+/-
De-En	24.36	21.36	-3.0
Cs-En	16.32	19.65	+3.33
En-Fr	28.25	26.22	-2.03
En-Zh	6.98	8.10	+1.12

Table 1: Preliminary Results, Metric: BLEU4

5.3 Preliminary Results

Our preliminary results of training a multilingual model with four direction with English as a center language can be found at Table 1. We can see that overall, multilingual training benefit languages with less vocabulary overlap (Cs-En, En-Zh) at the cost of harming the performance of directions with higher subword overlap and are more similar (De-En, En-Fr).

5.4 Gradient Dynamics

We plot the gradient similarity of our multilingual translation model (En-De, Fr) at each training step at figure 2, and the gradient similarity of the model with Gradient Vaccine and random projection at figure 4 and figure 3. As we can see, regardless of using gradient optimization methods, the cosine similarity between gradients shows an dropping trend and converges to 0, indicating the training process renders the final direction of gradients to be orthogonal.

5.5 Linguistic Features

While Wang et al. (2021) shows that gradient similarity is positively correlated with language family, we aim to learn if the gradient similarity reflect more of script or order of language (SVO or SOV). Our results can be found at table 2. Echoing previous research, we observe a trend of languages with similar script and same order has the highest similarity (0.35). Moreover, we observe that languages with different script and

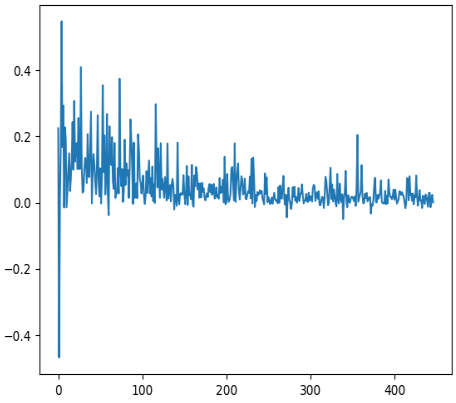

Figure 2: x-axis: training iterations, y-axis: gradient similarity of model without any optimization.

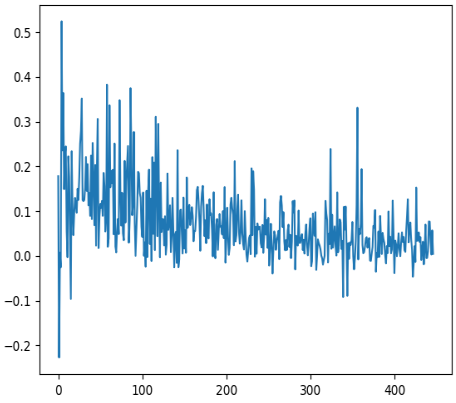

Figure 3: x-axis: training iterations, y-axis: gradient similarity of model with Gradient Vaccine (Wang et al., 2021).

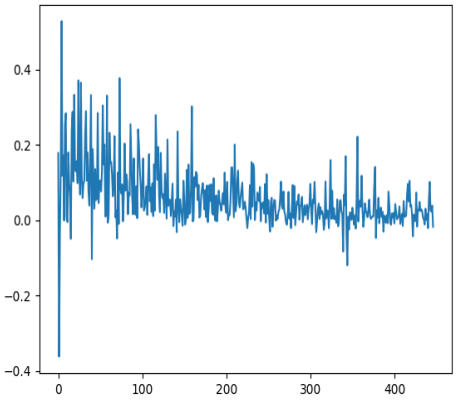

Figure 4: x-axis: training iterations, y-axis: gradient similarity of model with random projection.

Similar script, same order	
French, Portugese	0.35
Different script, same order	
French, Russian	0.13
French, Korean	0.22
Similar script, different order	
Chinese, Japanese	0.06
French, German	0.12
Different script, different order	
French, Japanese	0.06
Chinese, Korean	0.008

Table 2: Results on varying the script and order of target language. All experiments have the same parameter and the same source language (English).

Setting	BLEU (En-Fr/En-De)
Standard Training	27.66/24.20
Gradient Vaccine	**28.84**/24.91
Random Projection	28.35/**25.53**

Table 3: Results with and without gradient optimization. Best BLEU score is bolded.

same order have slightly higher $(0.13, 0.22)$ similarity than languages with similar script but different order $(0.06, 0.12)$. Not surprisingly, language pairs with the least similarity are languages that does not share script nor order $(0.06, 0.008)$.

We acknowledge that due to limited resources, we do not have a thorough analysis on various languages, therefore our claims might not be generalize well. However, as previous studies (Pires et al. (2019); K et al. (2020)) have also concurred that language order matters more than vocabulary overlap in cross-lingual transfer, we believe that word order is also critical when choosing source and target languages in multilingual NMT.

5.6 Random Projection

We study the validity and efficacy of Gradient Vaccine by randomly choosing an angle to project onto. Our results is at table 3. We do not observe significant improvement over randomly projecting to an acute angle. Indicating the gradient vaccine method is sensitive to the hyper-parameters and languages of the model.

6 Conclusion

In this work, we conclude that the gradients of different training directions in multilingual NMT converges orthogonally, regardless of using or not using gradient optimization techniques. We then probe into which linguistic feature is more reflected in the cosine gradient similarity, finding that in our experiments, the gradients reflect more of word order than sharing script. Finally, we investigated the validity of the Gradient Vaccine method, which does not show a clear advantage over randomly projecting onto a positive direction.

References

Aharoni, R., M. Johnson, and O. Firat (2019, June). Massively multilingual 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)*, Minneapolis, Minnesota, pp. 3874–3884. Association for Computational Linguistics.

Anonymous (2022). On the off-target problem of zero-shot multilingual neural machine translation. In *Submitted to ACL ARR 2022 October*. under review.

Anonymous (2023). On the shortcut learning in multilingual neural machine translation. In *Submitted to The Eleventh International Conference on Learning Representations*. under review.

Chiang, T.-R., Y.-P. Chen, Y.-T. Yeh, and G. Neubig (2022, May). Breaking down multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland, pp. 2766–2780. Association for Computational Linguistics.

Duh, K. (2018). The multitarget ted talks task. `http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/`.

Gu, J., Y. Wang, K. Cho, and V. O. Li (2019, July). Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 1258–1268. Association for Computational Linguistics.

He, S., Z. Tu, X. Wang, L. Wang, M. Lyu, and S. Shi (2019, November). Towards understanding neural machine translation with word importance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 953–962. Association for Computational Linguistics.

Javaloy, A. and I. Valera (2022). Rotograd: Gradient homogenization in multitask learning. In *International Conference on Learning Representations*.

Johnson, M., M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics 5*, 339–351.

K, K., Z. Wang, S. Mayhew, and D. Roth (2020). Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Koehn, P. and R. Knowles (2017, August). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver, pp. 28–39. Association for Computational Linguistics.

Kudo, T. and J. Richardson (2018, November). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium, pp. 66–71. Association for Computational Linguistics.

Kudugunta, S., A. Bapna, I. Caswell, and O. Firat (2019, November). Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 1565–1575. Association for Computational Linguistics.

Lee, S., H. B. Lee, J. Lee, and S. J. Hwang (2022). Sequential reptile: Inter-task gradient alignment

for multilingual learning. In *International Conference on Learning Representations*.

Liu, L., Y. Li, Z. Kuang, J.-H. Xue, Y. Chen, W. Yang, Q. Liao, and W. Zhang (2021). Towards impartial multi-task learning. In *International Conference on Learning Representations*.

Liu, Y., J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics 8*, 726–742.

Michel, P., O. Levy, and G. Neubig (2019). Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 32. Curran Associates, Inc.

Ott, M., S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Pires, T., E. Schlinger, and D. Garrette (2019, July). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 4996–5001. Association for Computational Linguistics.

Song, K., X. Tan, T. Qin, J. Lu, and T.-Y. Liu (2019). Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pp. 5926–5936.

Sun, S., A. Fan, J. Cross, V. Chaudhary, C. Tran, P. Koehn, and F. Guzmán (2022, May). Alternative input signals ease transfer in multilingual machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, pp. 5291–5305. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Wang, Z., Y. Tsvetkov, O. Firat, and Y. Cao (2021). Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations*.

Xin, D., B. Ghorbani, A. Garg, O. Firat, and J. Gilmer (2022). Do current multi-task optimization methods in deep learning even help?

Yang, Y., A. Eriguchi, A. Muzio, P. Tadepalli, S. Lee, and H. Hassan (2021, November). Improving multilingual translation by representation and gradient regularization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, pp. 7266–7279. Association for Computational Linguistics.

Yu, T., S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn (2020). Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*.

Zhang, B., P. Williams, I. Titov, and R. Sennrich (2020, July). Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 1628–1639. Association for Computational Linguistics.