MGrad: Optimizing Conflicting Gradients in Multilingual Machine Translation

Tianjian Li
Johns Hopkins University

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)).

Our work aims to answer the following research questions:

- In Multilingual Machine Translation, does the conflicting gradients between different translation pairs leads to negative transfer?
- Can we utilize language distance to design an optimization algorithm that encourage positive transfer and reduce negative transfer between language pairs?
- Can we decide what language pairs are the best to train on if we have a specific translate task to perform, in the absence of parallel text for that specific task?

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

Current Multi-Task optimization algorithms mainly falls into two categories: 1) aligning the magnitude of mismatching gradients, and 2) aligning the direction of conflicting gradients.

There has been a line of work that aims to align the magnitude of gradients in MTL to ensure each task get trained in a same rate: **GradNorm** (Chen et al. (2018)) views the convergence rate of each task as a signal of learning progress, and dynamically update the weight of each task is trained to a fixed rate. **IMTL-G** (Liu et al. (2021))

aims to make the projections of the gradients of each individual task onto the collective gradient equal.

Another line of work aims 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.

Recently, 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)).

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 individual gradients onto that direction.

4 Problem Formulation

We focus on the dataset containing on centric language scenario, where we train translation pairs either containing the centric langauge as a source language, or the containing the centric langauge as a target. Then we evaluate both in our supervised directions and unsupervised direction using the centric language as a pivot. Formally, we are given training corpus in n languages $\{S_1, S_2, ..., S_n\}$, and a centric language C. Our multilingual translation model learns a function f that either translates sentence $s_i \in S_i$ in the source languages to the centric language C or learns to translate sentences in the centric language to one of our n languages. For the cases when the centric is the target:

$$\theta := \arg\max_{\theta} p_\theta(C|s_i)$$

and for the cases when the centric language is the source, the model learns:

$$\theta := \arg\max_{\theta} p_\theta(S|c_i)$$

We can view each translation pair $S_i \rightarrow C$ or $C \rightarrow S_i$ as an individual task and view the entire training process as an example of **Multi-Task Learning**(MTL). Each individual translation task has a gradient g_i. Our goal is to find the optimal gradient that aligns the individual gradients that results in the best translation performance both in trained settings $S_i \rightarrow C$, $C \rightarrow S_j$ and zero-shot settings $S_i \rightarrow S_j$

$$g_{\text{final}} = \text{align}_{i=1}^{n}(g_i)$$

5 Experiments

5.1 Dataset

We follow previous work (Aharoni et al. (2019)) and use the multi-way parallel Ted Talk dataset (Duh (2018)). We train our model on four typologically and resource diverse languages: French, German, Chinese, Czech, English. Following the current multilingual machine translation systems, we use English as our centeric language, meaning that each training direction either contains English as a source or as a target. Table 1 contains the detailed directions of our multilingual model.

Supervised	Zero-Shot
De-En	De-Fr
Cs-En	De-Zh
En-Fr	Cs-Fr
En-Zh	Cs-Zh

Table 1: Training and Zero-shot translation directions

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 2: Preliminary Results, Metric: BLEU4

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 → English is closer to German → English, and further to Korean → English.

Then, we aim to evaluate ablate the effect of magnitude and direction on Multilingual Machine Translation by 1) only aligning the magnitude using IMTL-G (Liu et al. (2021)), and 2) only aligning the direction using PCGrad (Yu et al. (2020)) and Gradient Vaccine (Wang et al. (2021)).

Depending on the previous results, we aim to design an novel gradient optimization algorithm that either encourages magnitude and directional alignment among similar languages and dis-encourages alignment between dissimilar languages, and verify that our method improves upon simple multi-task optimization techniques.

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.

5.3 Baseline Results

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).

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.

Chen, Z., V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich (2018, 10–15 Jul). GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In J. Dy and A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, Volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR.

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.

Javaloy, A. and I. Valera (2022). Rotograd: Gradient homogenization in multitask learning. In *In-*

ternational 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.

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.

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.

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*.

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.