Residual Dropout: A Simple Approach to Improve Transformer's Data Efficiency

Carlos Escolano, Francesca De Luca Fornaciari , Maite Melero
Barcelona Supercomputing Center
{carlos.escolano, francesca.delucafornaciari, maite.melero}@bsc.es

Abstract

Transformer models often demand a vast amount of training data to achieve the desired level of performance. However, this data requirement poses a major challenge for low-resource languages seeking access to high-quality systems, particularly in tasks like Machine Translation. To address this issue, we propose adding Dropout to Transformer's Residual Connections. Our experimental results demonstrate that this modification effectively mitigates overfitting during training, resulting in substantial performance gains of over 4 BLEU points on a dataset consisting of merely 10 thousand examples.

Keywords: machine translation, low resource, transformers

1. Introduction

Neural Machine Translation (NMT) has revolutionized the field by achieving unprecedented results compared to previous methods. However, this progress has come at a cost—the escalating data requirements for training such systems. Currently, it is common practice to train models on millions of parallel sentences, a luxury only available for a limited number of high-resource languages. On the other hand, most languages lack access to this wealth of data and must settle for lower-quality translations or rely on generic multilingual models that are ill-suited to their specific linguistic nuances.

The primary factor contributing to this phenomenon is overfitting, wherein neural networks with millions of parameters tend to memorize training examples rather than actually learning the task at hand. Overfitting leads to poor generalization on unseen data, making models impractical. This issue exacerbates when training on a limited amount of data, as in the case of low-resource Neural Machine Translation.

The Transformer architecture, widely adopted in NMT, addresses overfitting by incorporating Dropout regularization and Batch Normalization at the output of attention blocks and feedforward layers. However, Residual Connections—wherein the output of previous layers is directly added without regularization—have received less attention in this regard. Yet recent research has underscored the significance of Residual Connections in preserving positional and semantic information across different attention layers.

This work aims to highlight the crucial role of Residual Connections in Neural Machine Translation. We explore the impact of incorporating Dropout regularization into all Residual Connections within the Transformer architecture. Our findings reveal that this approach effectively delays overfitting, particularly in scenarios with extremely limited resources, leading to noteworthy improvements in translation quality of over 4 BLEU points on average across diverse datasets encompassing various languages and domains.

2. Related Work

The Transformer architecture (Vaswani et al., 2017) has become the standard approach for various tasks, particularly Neural Machine Translation. It has demonstrated remarkable effectiveness not only in Natural Language Processing (NLP) tasks like Language Modeling (OpenAI, 2023) and Question Answering (Anil et al., 2023), but also in other domains such as Computer Vision (Liu et al., 2023) and Speech (Di Gangi et al., 2019).

At the core of this architecture lies the attention block, which consists of two main components: multi-head scaled dot-product attention and a feedforward layer. These elements work together to capture patterns and dependencies among different positions in a sequence. The attention mechanism can be applied within a sequence (self-attention) or between source and target sequences (cross-attention). The outcome is a contextual representation of the sequence tokens, enriched with information from other tokens and their positional relationships.

Previous studies (Geva et al., 2021) have highlighted the significance of the Transformer's feedforward networks as key-value memories that allow the model to capture novel patterns from the input data.

The outputs of both the attention and feedforward blocks are then normalized using Layer Normalization and added to the input of the block through a Residual Connection. This connection prevents the

SIGUL2024 Workshop, pages 294–299
21-22 May, 2024.

model from experiencing vanishing gradients, enabling the stacking of multiple Transformer blocks. Recent research (Ferrando et al., 2022) has emphasized the importance of Residual Connection in propagating information between layers. It has demonstrated that certain layers may have low attribution to all tokens in the sequence, relying on the Residual Connection to provide information to subsequent layers. The impact of Residual Connections has been particularly evident in Multilingual Machine Translation (Liu et al., 2021). When a Residual Connection is removed from the multilingual encoder, the models rely less on positional information, leading to a reduction in spurious correlations between trained languages. As a result, zero-shot translation improves.

One aspect that is often overlooked in the Transformer architecture is the utilization of Dropout (Srivastava et al., 2014). Transformer models typically have millions or even billions of parameters, which makes them prone to overfitting when insufficient training data is provided. Dropout helps mitigate this issue by randomly masking a percentage of the layer's outputs as 0. This delay in overfitting allows the models to generalize better to unseen data. In the Transformer architecture, Dropout is applied to both the attention and feedforward networks, during both self-attention and cross-attention operations.

3. Methodology

Residual Connections are integral to the flow of information within the Transformer's layers. However, during training, these connections lack regularization, making it easier for models to memorize patterns from them. Consequently, models are prone to overfitting, particularly in low-resource scenarios.

To address this issue, we propose the introduction of Residual Dropout. In addition to applying Dropout to the outputs of both the attention and feedforward networks, we suggest applying it to the input utilized during the Residual Connection (He et al., 2016). By incorporating this additional step, we aim to mitigate the overfitting tendency observed in standard Transformer models.

Figure 1 illustrates our proposed modification, highlighting the inclusion of Residual Dropout. It is worth noting that the proposed modification does not add any new trainable parameters to the model, hence does not affect its hardware requirements.

Our approach holds dual importance. Firstly, by randomly removing information from the Residual Connection, it forces the model to not rely exclusively on the most salient features. This variation helps delay overfitting and facilitates the learning of more robust representations. Secondly, by reducing the reliance on positional information, our models become more adaptable and robust, particularly in scenarios with limited available data.

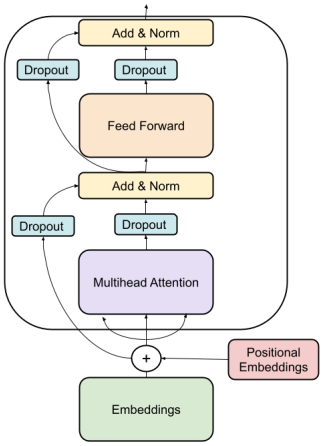

Figure 1: Transformer's attention block diagram with Residual Dropout.

4. Experimental Details

To assess the applicability of our method, we conducted a series of experiments focused on low-resource Machine Translation. Given the challenging nature of such datasets, particularly for extremely low-resource language pairs, we conducted our experiments using approximately 100k examples for standard evaluation. To analyze the impact of our method under different data conditions, we defined a range of training corpora sizes, ranging from 5k to 1M sentences.

Training corpora: Our training data comprised several datasets from diverse language families. For standard evaluation, we utilized the IWSLT 2017 (Nguyen et al., 2017) German-English corpus, which consists of 135 thousand sentences. Additionally, we chose the Tatoeba (Tiedemann, 2012) corpus, containing approximately 168 thousand sentences, for the Turkish-English translation task. To test a moderate-to-poor resource scenario, we randomly sampled 1M sentences from an in-house corpus that includes Europarl v7 (Koehn, 2005), CoVost 2 (Wang et al., 2021), CCAligned (El-Kishky et al., 2020), OpenSubtitles (Lison and Tiedemann, 2016), Wikimatrix (Schwenk et al., 2021), and Wikimedia. [1] For the size experiments, we randomly sampled subsets from this corpus. All datasets are tokenized using *Sentencepiece* with a subword vocabulary of 8000 tokens.

Evaluation corpora: To ensure comprehensive comparisons, we evaluated all translation directions

[1] Full disclosure of the datasets used can be found here

on both the FLORES (Goyal et al., 2022) dev and devtest sets. Furthermore, for the English-Catalan translation, we conducted tests on multiple test sets from different domains, including the Spanish Constitution and United Nations (Ziemski et al., 2016) from the administrative domain, WMT 19 from the biomedical domain, and WMT newstest 2013 from the news domain. All results are reported using *SacreBleu*'s (Post, 2018) standard configuration.

Implementation: In all our experiments, we adopted the standard "en-de-iswslt" Transformer configuration from *Fairseq* (Ott et al., 2019). This architecture consists of 6 Transformer layers in both the encoder and decoder. Each layer is equipped with 4 attention heads, a hidden size of 512 dimensions, and a feedforward size of 2048. We trained all models using 0.1 Dropout and the Adam optimizer (Kingma and Ba, 2015) with betas (0.9, 0.98) and a learning rate of $5e-4$." If not stated otherwise, Residual Dropout is applied on all encoder and decoder layers.

5. Results

When incorporating Dropout into a model, it is crucial to consider the tradeoff between regularization and the potential delay in overfitting, as well as the extent to which information is removed from the model. An excessively high Dropout value may prevent the model's ability to fully learn the task or even impair its overall performance. To determine the optimal value for our experiments, we conducted tests on the English-Catalan translation direction using 100 thousand sentences, exploring a range of values from 0.1 to 0.4.

Table 1 demonstrates that setting the Residual Dropout to 0.1 resulted in an average performance improvement of 3 BLEU points over the baseline. Remarkably, this improvement was consistently observed across all domains, including the biomedical domain, which was not present in the training data. Increasing the Dropout to 0.2 reduced the improvement to 0.6, and further increasing it led to a significant decline in the model's performance.

Furthermore, we observed that introducing Residual Dropout exclusively to either the encoder (RD 0.1 Enc) or decoder (RD 0.1 Dec) layers resulted in performance improvements. Upon comparing both models, we noted greater improvements when Residual Dropout was applied only to the decoder, particularly in the biomedical domain. However, when Residual Dropout was added to both the encoder and decoder layers, the overall performance improvement was even higher. Hence, for all subsequent experiments, we will employ a value of 0.1 on both encoder and decoder.

In order to test whether the gains observed in the EN-CA pair can be replicated in the other direction and for other language pairs, we chose our best value of RD and applied it to the training of three linguistically diverse models. Table 2 presents the results for the different models trained on datasets of approximately 100 thousand sentences. These results demonstrate that across all tested translation directions, the incorporation of Residual Dropout yields a consistent performance improvement of +2 BLEU points on both FLORES dev and devtest datasets.

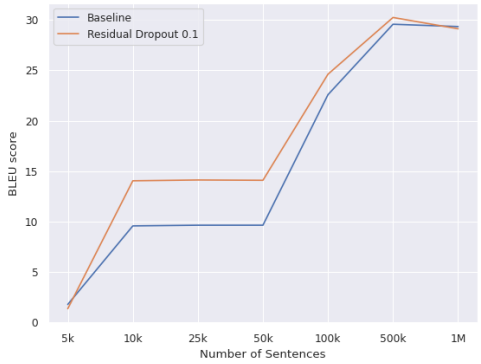

Figure 2: Performance comparison (BLEU) at different corpora sizes at 100k updates. In blue, baseline system, in orange, Residual Dropout at 0.1

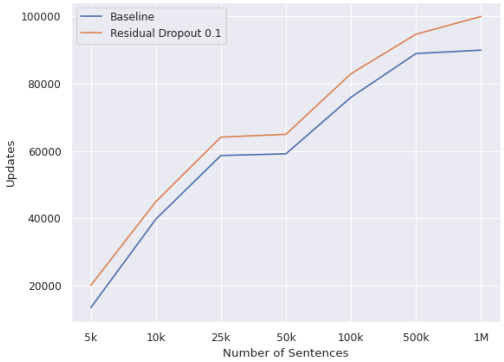

Figure 3: Number of updates until best checkpoint at different corpora sizes at 100k updates. In blue, baseline system, in orange, Residual Dropout at 0.1

Our hypothesis for the observed performance improvement, regardless of the translation direction or language pair, is that delayed overfitting plays a crucial role in enhancing translation quality. To substantiate this hypothesis, we conducted experiments training multiple English-Catalan models using varied corpus sizes ranging from 5 thousand to 1 million sentences. Each model underwent

Dataset	Baseline	RD 0.1	RD 0.2	RD 0.3	RD 0.4	RD 0.1 Enc	RD 0.1 Dec
Spanish Constitution	21.9	**23.7**	22.1	17.2	0.0	22.4	22.4
United Nations	24.7	**28.3**	26.5	20.3	0.0	26.4	27.1
FLORES dev	23.0	**27.7**	24.3	18.3	0.0	25.3	26.1
FLORES devtest	23.2	**26.9**	24.1	18.1	0.0	25.4	26.0
WMT 19 Biomedical	12.7	**13.9**	12.1	9.6	0.0	13.2	14.4
WMT 13 news	22.0	**25.6**	23.1	18	0.0	23.6	24.3
Average	21.4	**24.4**	22.0	16.9	0.0	22.7	23.35

Table 1: English-Catalan translation performance for different Residual Dropout values. All results are measured using BLEU.

Dataset	EN-CA		CA-EN		EN-DE		EN-TR	
Model	Baseline	RD 0.1	Baseline	RD 0.1	Baseline	RD 0.1	Baseline	RD 0.1
FLORES dev	23	**27.7**	25.5	**28.3**	20.4	**23.4**	12.0	**14.0**
FLORES devtest	23.2	**26.9**	25.3	**27.5**	18.7	**22.4**	11.3	**14.1**

Table 2: Translation results for all tested translation directions. All results are measured using BLEU.

training for 100 thousand updates, with the best checkpoint determined based on the lowest validation loss.

Figure 2 illustrates the translation quality achieved with the different corpus sizes. Notably, the most significant improvements were obtained with smaller corpora, showcasing a consistent enhancement of nearly 5 BLEU points between 10 and 50 thousand sentences. A special case is observed with only 5 thousand sentences, where both baseline and proposed models struggle to learn the task effectively. As the dataset size increases, the disparities between the two systems diminish, and they become almost identical when trained on 1 million sentences. Furthermore, examining the updates until the best checkpoint, as depicted in Figure 3, we observe that models employing Residual Dropout consistently require more updates to reach their peak performance.

6. Conclusions

Our research provides further evidence supporting the significance of Residual Connections in enhancing the performance of Transformer models. The introduction of Residual Dropout presents a straightforward and transparent approach to improving Transformer models, particularly in extremely low-resource scenarios. The experimental results demonstrate that our proposed modification can significantly enhance translation performance. For instance, on a dataset consisting of just 10 thousand sentences, our approach achieves an improvement of over 4 BLEU points over a standard Transformer configuration. Moreover, across multiple language pairs and a dataset of 100 thousand examples, the proposed modification yields a gain of more than 2 BLEU points.

As a potential future research, Residual Dropout can be applied to a wide range of tasks involving Transformers. The modification is agnostic to modalities, making it applicable across different domains.

7. Limitations

Our findings clearly demonstrate that the benefits achieved through the inclusion of Residual Dropout are closely linked to the postponement of overfitting. It is important to note that in high-resource scenarios or with models that do not exhibit pronounced signs of overfitting, e.g, model finetuning, the observed improvement may be significantly smaller or, in some cases, due to the model getting stuck on local minima.

8. Ethical Statement

The proposed method primarily emphasizes enhancing the data efficiency of the Transformer architecture, specifically in the domain of Machine Translation. Although the technique does not introduce any new ethical considerations into the architecture itself, it is important to note that it does not address the mitigation of societal biases or potential harms that may arise from such architectures.

Furthermore, it is essential to take into account the environmental implications of training neural models. The addition of Residual Dropout, while beneficial in delaying overfitting, also leads to an increase in the average number of updates required until convergence by approximately 10.75%. This increase in training iterations subsequently results in higher power consumption and CO_2 emissions.

By considering both ethical aspects and environmental impact, we can foster a more holistic approach to the development and deployment of Transformer architectures in Machine Translation and other domains.

9. Bibliographical References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019. Adapting transformer to end-to-end spoken language translation. In *Proceedings of INTERSPEECH 2019*, pages 1133–1137. International Speech Communication Association (ISCA).

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8756–8769. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, pages 630–645, Cham. Springer International Publishing.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13-15, 2005*, pages 79–86.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot

translation by disentangling positional information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1259–1273. Association for Computational Linguistics.

Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. 2023. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21.

Thai-Son Nguyen, Markus Müller, Matthias Sperber, Thomas Zenkel, Sebastian Stüker, and Alex Waibel. 2017. The 2017 KIT IWSLT speech-to-text systems for English and German. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 60–64, Tokyo, Japan. International Workshop on Spoken Language Translation.

OpenAI. 2023. Gpt-4 technical report.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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, Belgium, Brussels. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 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.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. Covost 2 and massively multilingual speech translation. In *Interspeech*, pages 2247–2251.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.