

Representation Collapse in Machine Translation Through the Lens of Angular Dispersion.

Evgeniia Tokarchuk, Maya K. Nachesa, Sergey Troshin and Vlad Niculae

Language Technology Lab, University of Amsterdam

evgeniia@tokarch.uk, {m.k.nachesa, s.troshin, v.niculae}@uva.nl

Abstract

Modern neural translation models based on the Transformer architecture are known for their high performance, particularly when trained on high-resource datasets. A standard next-token prediction training strategy, while widely adopted in practice, may lead to overlooked artifacts such as representation collapse. Previous works have shown that this problem is especially pronounced in the representation of the deeper Transformer layers, where it often fails to efficiently utilize the geometric space. Representation collapse is even more evident in end-to-end training of continuous-output neural machine translation, where the trivial solution would be to set all vectors to the same value. In this work, we analyze the dynamics of representation collapse at different levels of discrete and continuous NMT transformers throughout training. We incorporate an existing regularization method based on angular dispersion and demonstrate empirically that it not only mitigates collapse but also improves translation quality. Furthermore, we show that quantized models exhibit similar collapse behavior and that the benefits of regularization are preserved even after quantization.

1 Introduction

Text representations learned by Transformer-based models not only influence the performance of the target task, such as neural machine translation (NMT), but can also be extracted and used as a backbone for building retrieval-based models, such as retrieval-augmented generation (Lewis et al., 2020), k -Nearest Neighbors machine translation (Khandelwal et al., 2021), and continuous-output machine translation (Kumar and Tsvetkov, 2019; Tokarchuk and Niculae, 2022). Effective training of Transformer models for sequential language tasks (such as NMT) is, however, known to be difficult, in large part due to various forms of collapse of the internal representations learned (Gao et al.,

2019; Godey et al., 2024; Gerasimov et al., 2025; Barbero et al., 2024; Voita et al., 2019).

Representation collapse can be split into two phenomena: complete collapse and dimensional collapse (Hua et al., 2021; Jing et al., 2022). A complete collapse occurs when a trivial solution to the optimization problem is found by collapsing all vectors into a single point. It can be observed, for example, in continuous-output neural machine translation (CoNMT, introduced by Kumar and Tsvetkov (2019) and further described in §2.1), as well as in latent variable models (Chen et al., 2020). In contrast, dimensional collapse is partial, and occurs when a high-dimensional representation space is underutilized and all representations end up lying in a lower-dimensional space. While complete collapse is clearly visible from the training loss and automatic quality metrics, such as BLEU or COMET, dimensional collapse is not that evident, but studies have captured it for decoder-only Transformers (Barbero et al., 2024).

A popular approach to mitigate representation collapse in domains with continuous representation, such as images or videos, is to employ contrastive learning (Chen et al., 2020; Jing et al., 2022; Wang and Isola, 2020), which commonly relies on data augmentation and negative sampling to achieve diversity in the representations. Although data-augmentation of text data was also successfully applied in contrastive-learning setups (Wei and Zou, 2019; Shen et al., 2020; Su et al., 2021; Arefin et al., 2025), defining contrastive pairs for textual data is non-trivial due to the discrete nature of text. Despite being a competitive approach, one of the downsides of contrastive learning is that it can lead to higher variance of the objective function, in addition to typically requiring larger batch sizes to achieve a good performance (Guo et al., 2025). Moreover, performance depends heavily on the quality of the negative samples. Wang and Isola (2020) show that if we take into account

the spherical geometry of the representation, contrastive learning, in fact, can be decomposed into two separate components. The first is *alignment*, which enforces semantically close representations to be close in a vector space, and the second is *uniformity*, which ensures better coverage of the space. Following this idea, we can focus on uniformity or angular dispersion (Tokarchuk et al., 2025a) as a means to avoid representation collapse.

In this work, we focus on machine translation and revisit representation collapse in Transformer models. We analyze how collapse occurs in training and make a connection between representation collapse, the dimensionality of the model, and angular dispersion. We show that incorporating angular dispersion into the training of the model not only improves diversity of the representation but also leads to overall better MT quality.

2 Background

2.1 Next Token Prediction for Machine Translation

In NMT, given a source sequence $\mathbf{x} = (x_1, \dots, x_m)$, target sequence $\mathbf{y} = (y_1, \dots, y_n)$, and a finite vocabulary of type (sub)words $V = \{t_1, \dots, t_{|V|}\}$ such that $x_i, y_i \in V$, we aim to learn the mapping between the source and target sentences via the multi-class classification loss over V which amounts to next-token prediction:

$$\begin{aligned} L_{\text{NMT}}(y_i = t; \mathbf{y}_{<i}, \mathbf{x}) \\ = -\log p(y_i = t \mid \mathbf{y}_{<i}, \mathbf{x}). \end{aligned} \quad (1)$$

If we denote the latent representation produced by decoder as $\mathbf{H} = \mathbf{h}(\mathbf{x}, \mathbf{y}_{<i})$ and embeddings of the vocabulary token t as $\mathbf{E} = \mathbf{e}(t)$ Equation (1) can be expressed via Euclidean dot products as follows:

$$\begin{aligned} -\log p(y_i = t \mid \mathbf{y}_{<i}, \mathbf{x}) \\ = -\langle \mathbf{E}, \mathbf{H} \rangle + \log \sum_{t' \in V} \exp \langle \mathbf{E}', \mathbf{H} \rangle, \end{aligned} \quad (2)$$

where t is a token index, V is the vocabulary, $\mathbf{e} : V \rightarrow \mathbb{R}^d$ is an embedding lookup, and \mathbf{H} is a transformer hidden state calculated in terms of \mathbf{x} and the output prefix $\mathbf{y}_{<i}$.

The continuous alternative to next-token prediction, introduced by Kumar and Tsvetkov (2019), replaces classification with regression and can be expressed using cosine similarity, which replaces the more costly log-sum-exp:

$$L_{\text{CoNMT}}(y_i = t; \mathbf{y}_{<i}, \mathbf{x}) = 1 - \cos(\mathbf{E}, \mathbf{H}). \quad (3)$$

CoNMT was proposed as an efficient way to eliminate the training cost of the last classification layer and log-sum-exp, which scales linearly with vocabulary size. Instead, the regression loss does not depend on vocabulary size, but is vulnerable to collapse. All previous studies in this area use fixed embeddings, differing only in their specific configurations (Kumar and Tsvetkov, 2019; Tokarchuk and Niculae, 2022, 2024).

2.2 Measure of Representation Collapse

Previous works show that Transformer-based (Vaswani et al., 2017) models trained with next token prediction loss are prone to dimensional representation collapse, which manifests as a lack of diversity in the tokens' representations (Arefin et al., 2025; Barbero et al., 2024). In the continuous case the matter gets even worse with complete collapse, *i.e.* trivial global optima of eq. (3), achieved by setting all $\mathbf{e}(t)$ to the same vector for all t . Because of this, the target representations are typically fixed during training.

Representation collapse can be quantified using several methods, the most prominent of which we describe below.

Average cosine similarity. Given a matrix of representations $\mathbf{Z}^{(L)} \in \mathbb{R}^{N \times d}$, where \mathbf{Z}_{ij}^L is the activation of token i at position j of layer L , the average cosine similarity is: (Godey et al., 2024; Tokarchuk and Niculae, 2024).

$$S_{\text{avg}}(\mathbf{Z}^{(L)}) := \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \cos(\mathbf{Z}_i^{(L)}, \mathbf{Z}_j^{(L)}) \quad (4)$$

Abnormally high cosine similarities that cannot be explained by semantic similarity may serve as a signal of complete representation collapse.

Matrix entropy. Another common approach is to examine the covariance matrix (or Gram matrix) of the representations (Arefin et al., 2025; Gerasimov et al., 2025; Razdaibiedina et al., 2023; Skean et al., 2024), defined as $\mathbf{G} = \mathbf{Z}^{(L)}(\mathbf{Z}^{(L)})^\top \in \mathbb{R}^{d \times d}$. The challenge is however to find a basis-invariant metric, since rotations of the space should lead to an identical metric value. To this end, Skean et al. (2024) propose using the distribution of the eigenvalues of the Gram matrix. Denote by $\lambda_k(\mathbf{G})$ its k th eigenvalue, which measures the variance of the data along the direction spanned by the k th eigenvector, with $1 \leq k \leq d$. If the data is entirely constant in some dimension k then

$\lambda_i(\mathbf{G}) = 0$, and implicitly the data lies exactly on a low-dimensional subspace. More generally, since the eigenvalues are non-negative we may interpret the spectrum as a distribution by normalizing $\tilde{\lambda}_k(\mathbf{G}) = \lambda_k(\mathbf{G}) / \sum_{k'} \lambda_{k'}(\mathbf{G})$. Ideally, the distribution would be close to uniform, with high entropy, so as the variance is spread evenly across all dimensions. In cases of partial collapse the entropy becomes low as the distribution peaks on a few dimensions. This leads to the Rényi entropy of order α (Rényi, 1961) applied to the eigenvalue distribution, sometimes called the Matrix Rényi entropy (Sanchez Giraldo et al., 2015):

$$S_\alpha(\mathbf{Z}^{(L)}) = \frac{1}{1-\alpha} \log \left[\sum_k \tilde{\lambda}_k^\alpha(\mathbf{G}) \right]. \quad (5)$$

Note that in the limit of $\alpha \rightarrow 1$, S_α recovers the standard Shannon entropy of the eigenvalue distribution, and indeed in this work we keep $\alpha = 1$, despite the more general definition given by Skean et al. (2024). In a healthy model, S_α is high, high, which indicates that the empirical covariance covers all dimensions more uniformly. In contrast, when the data only varies in a linear subspace of small dimensions (a form of dimensional collapse), S_α will be low.

Spherical variance. Another helpful way to measure representation geometry is through the distribution of the directions of the vectors. Consider the mean vector of all (normalized) directions of a representation matrix:

$$\mathbf{M}(\mathbf{Z}^{(L)}) = \frac{1}{N} \sum_i \frac{\mathbf{z}_i^{(L)}}{\|\mathbf{z}_i^{(L)}\|_2}.$$

As this vector is the mean of points on the sphere, it will fall in the interior of the sphere (including the surface). The closer it is to the origin, the more spread out the directions are. Its length is therefore known as the spherical variance: (Jammalamadaka and Sengupta, 2001; Mardia, 1975)

$$\text{svar}(\mathbf{Z}^{(L)}) = 1 - \|\mathbf{M}(\mathbf{Z}^{(L)})\|. \quad (6)$$

The main advantage of spherical variance over the methods discussed above is its computational efficiency, which makes it well suited for being reported during training.

Other methods. The token separability test, introduced by Voita et al. (2019) can also be used

to test if contextual representations carry distinguishing information. By taking identical tokens (*e.g.*, the word “is”) in various contexts, one can examine whether the model’s representations for these occurrences are different or all “collapsed” to the same point. This method has been adopted in recent works. For instance, Gerasimov et al. (2025) used a similar approach to show representation collapse in Transformer layers.

2.3 Angular Dispersion

Representation collapse is often measured in terms of cosine similarity for language tasks, with the underlying assumption that the directions in a d -dimensional space can be represented as points on the sphere $\mathbb{S}_d \subset \mathbb{R}^d$. Unlike the entirety of \mathbb{R}^d , the sphere is compact and has many computationally attractive properties that allow us to quantify and optimize angular dispersion. We can view optimal dispersion as a counterpart of a representation collapse on the \mathbb{S}_d , and finding an optimally-dispersed configuration is known as the Tammes problem (Tammes, 1930). It is generally not tractable in a setup with a large amount of high-dimensional points, so several optimization approaches evolved to encourage dispersion (Wang and Isola, 2020; Wang et al., 2020; Liu et al., 2018; Tokarchuk et al., 2025a).

Unlike most approaches based on pairwise distances or kernels, *Sliced dispersion* (Bonet et al., 2023; Tokarchuk et al., 2025a) is an efficient alternative which avoids the quadratic complexity by making use of the fact that optimal dispersion is trivial on a circle (on $\mathbb{S}_2 \subset \mathbb{R}^2$). In this special case, any perfectly-dispersed configuration is made up of equidistant angles and given any input set of angles, the nearest dispersed configuration can be efficiently found. In $d = 2$, given a vector of angles Φ , let the total distance between it and a perfectly-dispersed configuration of angles be written $\delta(\Phi)$. As this distance cannot be calculated in higher dimension, we slice our data along great circles. On a sphere \mathbb{S}_d , the great circles correspond to pairs of orthogonal directions $C(\mathbb{S}_d) := \{(\mathbf{P}, \mathbf{Q}) : \mathbf{Q} \in \mathbb{S}_d, \mathbf{Q} \perp \mathbf{P}\}$. Let $\bar{\mathbf{Z}}$ denote a configuration of directions (*e.g.*, normalized representation vectors). If we denote by $\bar{\mathbf{Z}}_{PQ}$ the projection of the directions in $\bar{\mathbf{Z}}$ onto the great circle (\mathbf{P}, \mathbf{Q}) , sliced dispersion optimizes

$$R_{\text{sliced}}(\bar{\mathbf{Z}}) := \mathbb{E}_{\mathbf{P}, \mathbf{Q}} [\delta(\bar{\mathbf{Z}}_{PQ})], \quad (7)$$

where the expectation is over the uniform distribu-

tion on the $C(\mathbb{S}_d)$. In words, this objective minimizes the expected distance to an optimally-sliced configuration along any great circle.

2.4 Efficiency

While recent LLMs have changed the machine learning landscape, their performance comes with a trade-off in size and energy usage (Shterionov and Vanmassenhove, 2023), as well as the need for a wireless connection to run models in the cloud. Smaller models are still preferable where they need to be run on edge devices, have a lower energy footprint, may need to be run on devices without a wireless connection, or need to be trained on a smaller set of private data (Menghani, 2023). There are several ways to increase the efficiency of a model, without necessarily sacrificing its performance, either during or after training, during inference, or both. These include such methods as distillation, pruning, and, the focus here, quantization (Menghani, 2023). One particular way of doing quantization involves compressing the model by modifying the data type from (typically) float32 precision to a lower precision, such as half-precision floating-point float16 or int8 (Tang et al., 2024). Since quantization necessarily involves representing less information, it must strike a balance between efficiency and performance (Menghani, 2023). Despite its usefulness in increasing training efficiency and widespread application for training large neural networks (Grattafiori et al., 2024; Alves et al., 2024), Barbero et al. (2024) show that low-precision training further amplifies the issue of representation collapse. Quantization in this space could push the model to use the low-precision space more effectively.

3 Regularized Transformer

In order to prevent representation collapse and promote diversity in NMT representations, we propose simple yet efficient regularization on top of the next token prediction objective shown in Equation (1)

$$L_{\text{RNMT}}(y_i = t; \mathbf{y}_{<i}, \mathbf{x}) = L_{\text{NMT}} + \gamma R(\mathbf{Z}^L), \quad (8)$$

where $R(\mathbf{Z}^L)$ is a sliced dispersion regularization discussed in §2.3 over an output of a layer Z^L and γ is a dispersion weight. $R(\mathbf{Z}^L)$ can be, in principle, any latent representation of the source and target sequences. We propose to focus on the decoder’s latent representations in the transformer,

namely the decoder output and decoder embeddings. Gerasimov et al. (2025) showed recently that deeper layers of Transformers exhibit a higher degree of collapse. Therefore, we will apply regularization to the decoder output \mathbf{H} . Tokarchuk and Niculae (2024) showed that the embeddings of rare tokens collapse to similar representations, making it difficult to distinguish between them, so we apply dispersion on top of the embeddings representation \mathbf{E} . Additionally, we propose to apply dispersion on the encoder output representation \mathbf{F} .

Similar to the CoNMT case, regularization is applied to the target embeddings, as our primary goal is to prevent the collapse of target representations.

$$L_{\text{RCoNMT}}(y_i = t; \mathbf{y}_{<i}, \mathbf{x}) = L_{\text{CoNMT}} + \gamma R(\mathbf{Z}^{(l)} = \mathbf{E}). \quad (9)$$

Note that calculating regularization over the embedding matrix \mathbf{E} can be costly. Therefore, given the insights from previous studies (Tokarchuk and Niculae, 2024), we randomly subsample from the pool of rare tokens, *i.e.*, embeddings with a rank higher than the half of the vocabulary size.

4 Experimental Results

4.1 Data and Evaluation

We provide results on the English↔German language pair in both directions, namely en-de and de-en. We use 34M cleaned training sentences from publicly available WMT19 dataset (Barrault et al., 2019) provided by HuggingFace,¹ and tokenize it with BPE subword units² (Sennrich et al., 2016). We evaluate the models using sacrebleu (Post, 2018) and COMET (Rei et al., 2020) on newstest2018 and newstest2019. Statistics of the training and test data can be found in Table 4 of the Appendix A.1. We additionally report results for the Romanian→English language pair in Appendix C.

4.2 Models and Training

We study representation collapse in the Transformer-big and Transformer-base models defined by Vaswani et al. (2017), using the fairseq framework (Ott et al., 2019). Additionally, we train a CoNMT model with the same architecture as Transformer-big. Each model

¹<https://huggingface.co/datasets/wmt/wmt19>

²<https://github.com/glample/fastBPE>

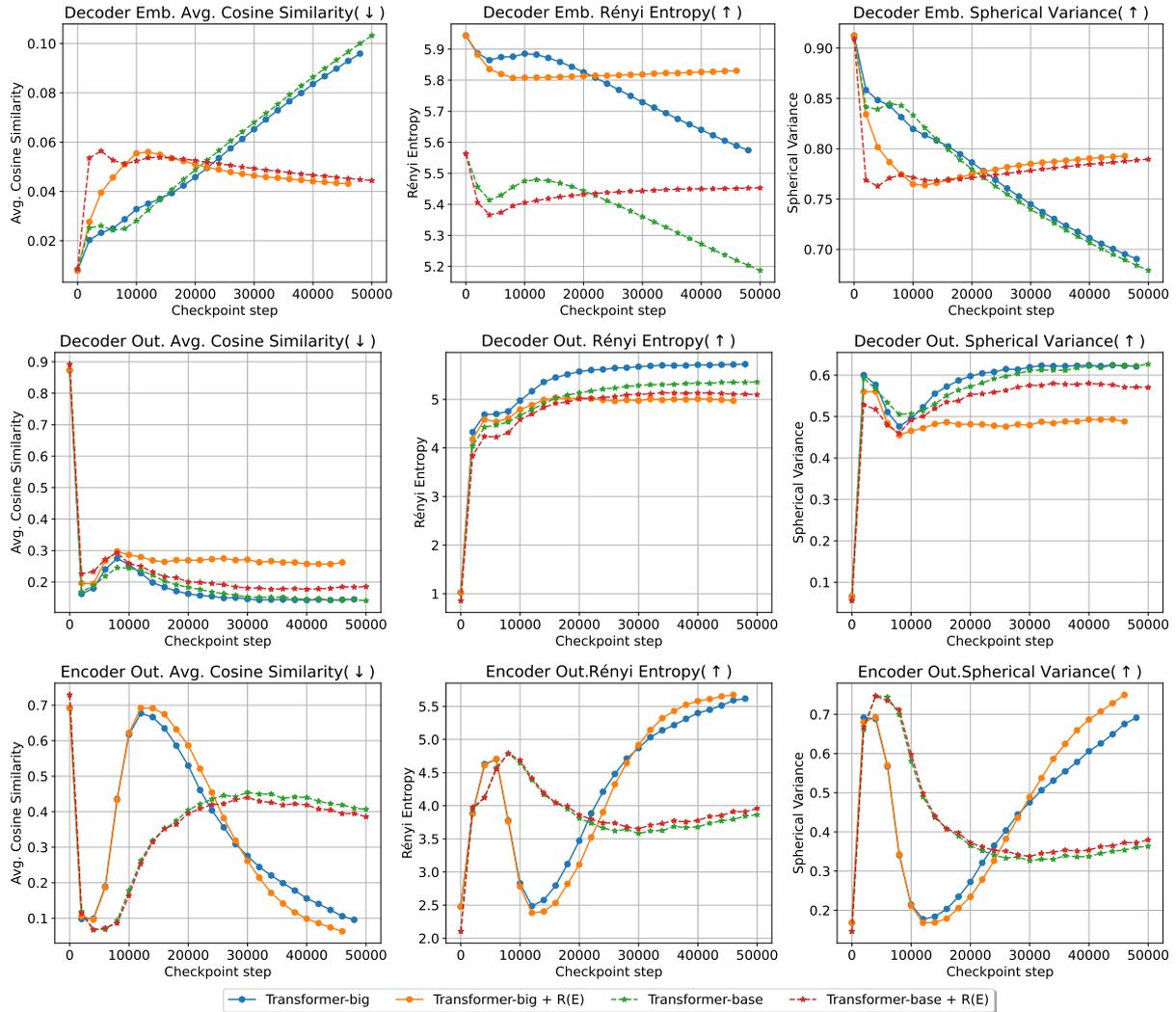


Figure 1: Average cosine similarity, Rényi entropy and spherical variance for decoder output, decoder embeddings and encoder outputs for Transformer-big and Transformer-base models.

is trained for 50,000 steps on a single NVIDIA H100 GPU using the Adam optimizer (Kingma and Ba, 2014). The learning rate is set to $5 \cdot 10^{-4}$ with 10,000 warm-up steps. The weight of the γ parameter is tuned on the development set for each component based on the Transformer-base model in the range $\{10^k : k \in \{-2, -1, 0, 1, 2\}\}$. We used Sliced regularizer for all models with regularization due to its efficiency.

4.3 Post-training quantization

As applying dispersion to a model with training-aware quantization is somewhat involved, we instead opt for post-training quantization on the Transformer-big and the Transformer-big with dispersion. We apply post-training quantization using the CTranslate2 framework, introduced by Klein et al. (2020). It allows for quantization to various compression levels, and applies

other efficiency techniques, such as weights pre-packing, to allow for a fast runtime after quantizing the model. Quantization may either be performed for one compression type, e.g. float16, in which case both the weights are stored and all layers are run at that precision level. In a mixed precision setting, e.g. int8_float16, the weights of the embeddings and linear layers are quantized to int8, while the rest of the layers are run in float16. We quantized the Transformer-big model to float16 and int8_float16.³ The float16 model acts as a comparison to the quantization-aware trained Transformer-big model. Because the Transformer-base model is approximately four times smaller in terms of parameters than the Transformer-big model, we included

³<https://opennmt.net/CTranslate2/quantization.html>

int8_float16 as a way to compare two models with a similar size overall. See Appendix B.2 for a size comparison of the quantized models on disk. We run inference on the GPU.

4.4 Representation Collapse in NMT

First, to identify the collapse phenomenon, we examine how the collapse metrics defined in §2.1 evolve over the course of training for the de-en language pair. For each batch of 4,096 tokens, we compute the matrix-based Rényi entropy, spherical variance, and average cosine similarity for the decoder output H , decoder embeddings E , and encoder output F . Figure 1 presents the averaged metric values across all batches at each training step for the Transformer-big, Transformer-big+dispersion (embeddings), and Transformer-base models.

The Transformer-base model exhibits stronger collapse in the encoder output and decoder embeddings compared to the Transformer-big model. Interestingly, the decoder output does not display clear signs of collapse in either configuration. While the Transformer-big encoder output partially recovers during training, the Transformer-base encoder representations collapse more severely and remain less diverse. We hypothesize that dimensionality and dispersion have strong connection with each other, and larger dimensionality naturally promotes dispersion, which is also in line with observations in previous works (Tokarchuk et al., 2025b).

Applying dispersion regularization to the Transformer-big embeddings improves all collapse metrics. However, applying dispersion to the decoder embeddings adversely affects the representation diversity of the decoder output, as reflected by higher (average) cosine similarity and lower entropy and variance scores.

Translation Quality. In Table 1, we quantify the effect of dispersion regularization applied to different components of the model using standard translation quality metrics. Results for both translation directions show that applying dispersion to the decoder embeddings can further improve the performance of the strong Transformer-big baseline, consistent with the trends observed in Figure 1. Furthermore, applying dispersion regularization directly to the encoder embeddings of the Transformer-base model effectively mitigates encoder-side representation collapse.

We analyze translation quality at the token level

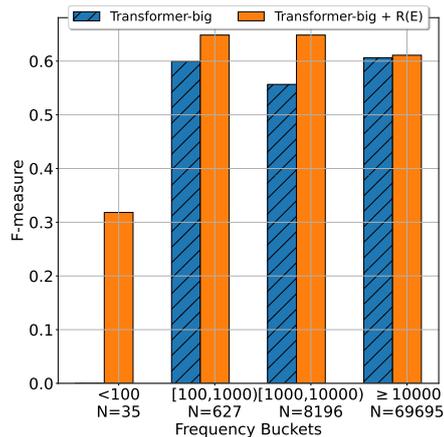


Figure 2: F1 score for each vocabulary token’s frequency bucket for en-de. Note that the F1 score for tokens with frequency < 100 for the model without regularization is 0.

using compare-mt (Neubig et al., 2019). Figure 2 presents the F1 score across token frequency buckets on the development set of the en-de model. The model with dispersion applied to the decoder embeddings clearly outperforms the non-dispersed baseline on rare tokens (frequency below 100), for which the latter achieves an F1 score close to zero. This improvement on rare tokens also translates into better overall performance. A similar trend is observed for the de-en direction, although the effect is less pronounced due to the smaller number of rare tokens. For detailed analysis on de-en, please refer to Appendix B.1.

Sensitivity to dispersion weight Varying the parameter γ in Equation (8) strongly influences the degree of dispersion and, consequently, the model’s performance. We conduct a sensitivity analysis on the Transformer-base model for the de-en language pair across different model representations, with results shown in Figure 3. Our experiments indicate that the decoder embeddings are less sensitive to the dispersion weight, whereas the encoder output is more susceptible to over-dispersion, which can degrade performance.

Quantization. To examine the model’s behavior with respect to representation collapse under half-precision training, we train the Transformer-big model in fp16. There is no noticeable differences in the representations of the decoder output or decoder embeddings between Transformer-big and Transformer-big-fp16 models. Encoder output metrics, however, indicate a slightly higher degree

model	en-de				de-en			
	newstest18		newstest19		newstest18		newstest19	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Transformer-big	41.9	0.803	31.7	0.790	43.0	0.811	39.3	0.776
+ $R(\mathbf{H})$	41.6	0.803	31.7	0.789	42.5	0.808	39.0	0.774
+ $R(\mathbf{E})$	42.6	0.806	32.6	0.793	43.2	0.812	39.2	0.776
Transformer-base	40.9	0.792	30.9	0.782	41.7	0.799	37.8	0.762
+ $R(\mathbf{H})$	40.4	0.788	30.6	0.777	41.0	0.797	37.5	0.759
+ $R(\mathbf{E})$	41.2	0.795	31.2	0.783	41.5	0.801	37.8	0.762
+ $R(\mathbf{F})$	40.7	0.791	31.1	0.781	42.0	0.801	37.8	0.764

Table 1: BLEU and COMET on newstest18 and newstest19 for en-de end en-de with different dispersion regularizers. Values in bold indicate results that are statistically significant compared to the baseline (p-value < 0.05)

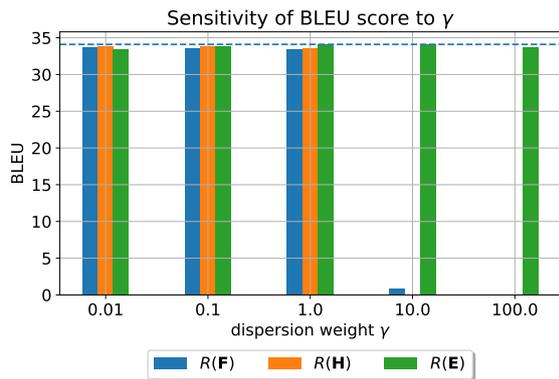


Figure 3: BLEU score for the de-en development set newstest17 of Transformer-base model for different parameters γ . Horizontal blue line indicates the BLEU score of the unregularized Transformer-base model.

of collapse for fp16 (Figure 4).

From Table 2 we can see that the both post-training quantization of the Transformer-big model in float16 and Transformer-big model trained in float16 perform similarly to the full precision Transformer-big model. However, quantization with int8_float16 hurts both models performance significantly, without clear evidence that dispersion of the decoder embeddings helps to preserve representation better. Even so, the benefit of dispersion is preserved even after quantization. Nonetheless, investigating representation collapse under reduced precision warrants further and more systematic study, especially in a dynamic setting.

4.5 Full Collapse in CoNMT

Table 3 demonstrates that dispersion regularization effectively mitigates representation collapse during end-to-end training of CoNMT. Notably, without applying regularization to the embeddings, the model collapses to a trivial solution and fails to achieve a BLEU score above zero unless the

prec.	reg.	BLEU	COMET
float32	\times	31.7	0.790
float32	$R(\mathbf{E})$	32.6	0.793
float16*	\times	31.7	0.787
float16	\times	31.7	0.845
float16	$R(\mathbf{E})$	32.6	0.847
int8_float16	\times	27.5	0.745
int8_float16	$R(\mathbf{E})$	28.0	0.813

Table 2: Comparison of quantization-aware training (float*) and post-training quantization on the Transformer-big en-de model with and without dispersion regularization on newstest19.

embeddings are frozen.

objective	init.	train \mathbf{E}	BLEU	COMET
L_{CoNMT}	rand.	\times	33.9	0.713
L_{CoNMT}	rand.	\checkmark	0.0	0.000
L_{CoNMT}	NMT	\times	32.9	0.704
L_{CoNMT}	NMT	\checkmark	0.0	0.000
L_{CoNMT}	RNMT	\times	36.6	0.749
L_{CoNMT}	RNMT	\checkmark	0.0	0.000
L_{RCoNMT}	rand.	\checkmark	33.2	0.708
L_{RCoNMT}	NMT	\checkmark	31.3	0.700
L_{RCoNMT}	RNMT	\checkmark	30.9	0.694

Table 3: de-en translation scores on newstest2019.

Looking more closely at the dynamics of representations, we can clearly see a complete representation collapse in Figure 5, where the spherical variance and matrix-based entropy drop to 0, and the average cosine similarity reaches 1. The model with regularization successfully prevents this collapse, achieving stability comparable to the model with fixed (frozen) embeddings. Note, that changing initialization of target representation from unregularized model or random embeddings to the embeddings from regularized model achieves the best overall performance for CoNMT.

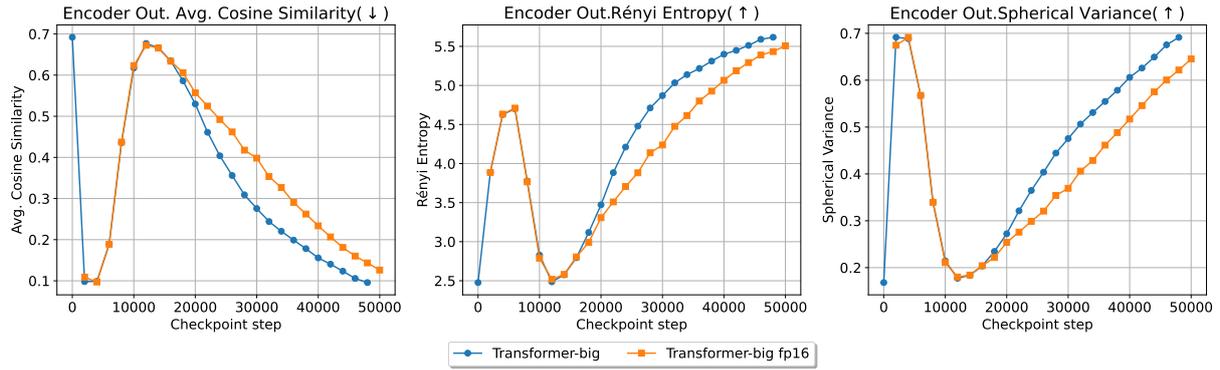


Figure 4: Comparison of the encoder outputs representation collapse metrics for Transformer-big and Transformer-big fp16 de-en.

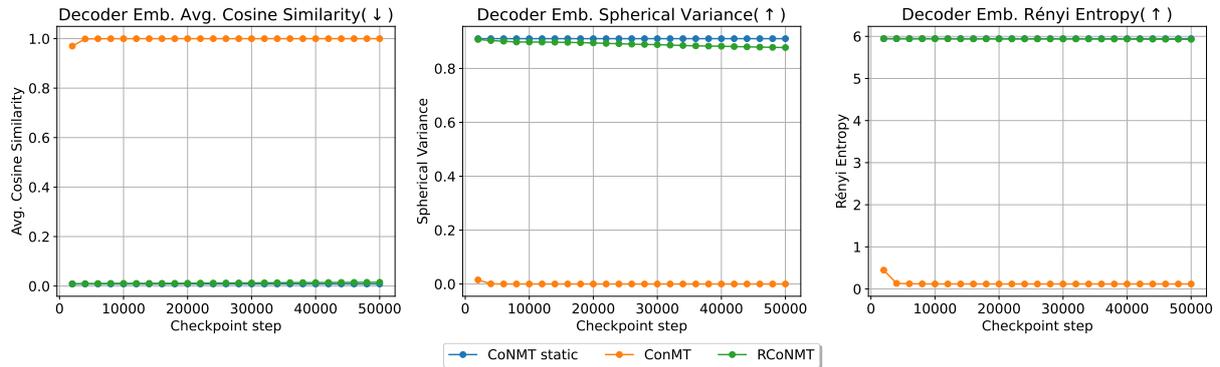


Figure 5: Representation collapse in target representations of the CoNMT models. CoNMT static refers to the model trained with frozen target embeddings.

5 Conclusion

In this work, we investigated the phenomenon of representation collapse in neural machine translation. Through experiments with the Transformer-big model, we demonstrated that promoting dispersion in representations can further enhance an already strong baseline. We observe that this benefit holds even after quantizing the model. Moreover, our results indicate that dispersion plays a critical role in preventing full collapse in continuous-output NMT, highlighting its potential as a simple yet effective strategy for stabilizing representation learning in Transformer-based models.

Limitations

Limited evaluation scenarios. In this work, we focused on the representations of three specific layers within the Transformer model and proposed a regularization strategy to prevent collapse. However, representation collapse may also occur in other layers, and further investigation would provide a more comprehensive understanding of the phenomenon. Moreover, we applied only one reg-

ularization component at a time. Our experiments suggest that regularizing one layer can negatively affect the representational diversity of another. A combined or coordinated regularization approach across multiple layers may therefore offer a more effective solution and warrants further study.

Low-resource tasks. While WMT19 English–German is a widely used translation benchmark, our results indicate that dispersion yields the greatest benefits for rare tokens, an effect that is likely to be more pronounced in low-resource language pairs. Extending our approach to such settings, as well as to multilingual translation scenarios, would provide stronger empirical support for the proposed method.

Acknowledgements

We thank members of the UvA Language Technology Lab for discussion and feedback on the manuscript. This work is supported by the Dutch Research Council (NWO) via VI.Veni.212.228. Maya K. Nachesa is supported by the ROBUST project (number KICH3.LTP.20.006) which is

(partly) financed by the Dutch Research Council (NWO), RTL, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023. Sergey Troshin is partially supported by “Hybrid Intelligence: augmenting human intellect” (<https://hybrid-intelligence-centre.nl>) with project number 024.004.022 of the research program “Gravitation” which is (partly) financed by the Dutch Research Council (NWO). The authors also thank SURF (www.surf.nl) for the support in using the National Supercomputer Snellius.

References

- Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. [Tower: An open multilingual large language model for translation-related tasks](#). In *First Conference on Language Modeling*.
- Md Rifat Arefin, Gopeshh Subbaraj, Nicolas Gontier, Yann LeCun, Irina Rish, Ravid Shwartz-Ziv, and Christopher Pal. 2025. [Seq-VCR: Preventing collapse in intermediate transformer representations for enhanced reasoning](#). In *The Thirteenth International Conference on Learning Representations*.
- Federico Barbero, Andrea Banino, Steven Kaptrowski, Dharshan Kumaran, João Guilherme Madeira Araújo, Alex Vitvitskiy, Razvan Pascanu, and Petar Veličković. 2024. [Transformers need glasses! information over-squashing in language tasks](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- 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.
- Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh Tan Pham. 2023. [Spherical Sliced-Wasserstein](#). In *The Eleventh International Conference on Learning Representations*.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. [A simple framework for contrastive learning of visual representations](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.
- Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tiejun Liu. 2019. [Representation degeneration problem in training natural language generation models](#). In *International Conference on Learning Representations*.
- Gleb Gerasimov, Yaroslav Aksenov, Nikita Balagansky, Viacheslav Sini, and Daniil Gavrilov. 2025. [You do not fully utilize transformer’s representation capacity](#). *Preprint*, arXiv:2502.09245.
- Nathan Godey, Éric Clergerie, and Benoît Sagot. 2024. [Anisotropy is inherent to self-attention in transformers](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48, St. Julian’s, Malta. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Zhaohan Daniel Guo, Bernardo Avila Pires, Khimya Khetarpal, Dale Schuurmans, and Bo Dai. 2025. [Representation learning via non-contrastive mutual information](#). *Preprint*, arXiv:2504.16667.
- Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. 2021. [On feature decorrelation in self-supervised learning](#). In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9578–9588.
- S Rao Jammalamadaka and Ambar Sengupta. 2001. *Topics in circular statistics*. Series on multivariate analysis; vol. 5. World Scientific, Singapore.
- Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. 2022. [Understanding dimensional collapse in contrastive self-supervised learning](#). In *International Conference on Learning Representations*.
- Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. [Nearest neighbor machine translation](#). In *International Conference on Learning Representations*.
- Diederik P. Kingma and Jimmy Ba. 2014. [Adam: A method for stochastic optimization](#). *CoRR*, abs/1412.6980.
- Guillaume Klein, Dakun Zhang, Clément Chouteau, Josep M Crego, and Jean Senellart. 2020. Efficient and high-quality neural machine translation with opennmt. In *Proceedings of the fourth workshop on neural generation and translation*, pages 211–217.
- Taku Kudo and John Richardson. 2018. [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*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Sachin Kumar and Yulia Tsvetkov. 2019. [Von Mises-Fisher loss for training sequence to sequence models with continuous outputs](#). In *International Conference on Learning Representations*.
- Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). *ArXiv*, abs/2005.11401.
- Weiyang Liu, Rongmei Lin, Z. Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. 2018. [Learning towards minimum hyperspherical energy](#). In *Neural Information Processing Systems*.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Kantilal Varichand Mardia. 1975. Statistics of directional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 37(3):349–371.
- Gaurav Menghani. 2023. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys*, 55(12):1–37.
- Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, and John Wieting. 2019. [compare-mt: A tool for holistic comparison of language generation systems](#). *CoRR*, abs/1903.07926.
- 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.
- Anastasia Razdaibiedina, Ashish Khetan, Zohar Karnin, Daniel Khashabi, and Vivek Madan. 2023. [Representation projection invariance mitigates representation collapse](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14638–14664, Singapore. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Alfréd Rényi. 1961. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press.
- Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C. Principe. 2015. [Measures of entropy from data using infinitely divisible kernels](#). 61(1):535–548. Publisher Copyright: © 2014 IEEE. Copyright: Copyright 2015 Elsevier B.V., All rights reserved.
- 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.
- Dinghan Shen, Ming Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. [A simple but tough-to-beat data augmentation approach for natural language understanding and generation](#). *ArXiv*, abs/2009.13818.
- Dimitar Shterionov and Eva Vanmassenhove. 2023. The ecological footprint of neural machine translation systems. In *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*, pages 185–213. Springer.
- Oscar Skean, Md Rifat Arefin, and Ravid Shwartz-Ziv. 2024. [Does representation matter? exploring intermediate layers in large language models](#). In *Workshop on Machine Learning and Compression, NeurIPS 2024*.
- Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. [Whitening sentence representations for better semantics and faster retrieval](#). *ArXiv*, abs/2103.15316.
- Pieter Merkus Lambertus Tammes. 1930. *On the origin of number and arrangement of the places of exit on the surface of pollen-grains*. Ph.D. thesis, University of Groningen. Relation: <http://www.rug.nl/> Rights: De Bussy.
- Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhijun Tu, Kai Han, Hailin Hu, and Dacheng Tao. 2024. A survey on transformer compression. *arXiv preprint arXiv:2402.05964*.
- Evgeniia Tokarchuk, Hua Chang Bakker, and Vlad Niculae. 2025a. [Keep your distance: learning dispersed embeddings on \$\mathbb{S}_m\$](#) . *Transactions on Machine Learning Research*.
- Evgeniia Tokarchuk and Vlad Niculae. 2022. [On target representation in continuous-output neural machine translation](#). In *Proceedings of the 7th Workshop on*

Representation Learning for NLP, pages 227–235, Dublin, Ireland. Association for Computational Linguistics.

Evgeniia Tokarchuk and Vlad Niculae. 2024. [The unreasonable effectiveness of random target embeddings for continuous-output neural machine translation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 653–662, Mexico City, Mexico. Association for Computational Linguistics.

Evgeniia Tokarchuk, Sergey Troshin, and Vlad Niculae. 2025b. [Angular dispersion accelerates \$\\$k\\$\$ -nearest neighbors machine translation](#). In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. [The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives](#). 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)*, pages 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Tongzhou Wang and Phillip Isola. 2020. [Understanding contrastive representation learning through alignment and uniformity on the hypersphere](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Zhennan Wang, Canqun Xiang, Wenbin Zou, and Chen Xu. 2020. [MMA regularization: Decorrelating weights of neural networks by maximizing the minimal angles](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 19099–19110. Curran Associates, Inc.

Jason Wei and Kai Zou. 2019. [EDA: Easy data augmentation techniques for boosting performance on text classification tasks](#). 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)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

	lang.	#sent.	#toks.	#vocab.
train	de	34M	916M	
newstest2017	de	3004	78553	42024
newstest2018	de	2998	81875	
newstest2019	de	2000	47035	
train	en	34M	888M	
newstest2017	en	3004	75011	42024
newstest2018	en	2998	78925	
newstest2019	en	2000	46261	

Table 4: Data statistics for training and test splits.

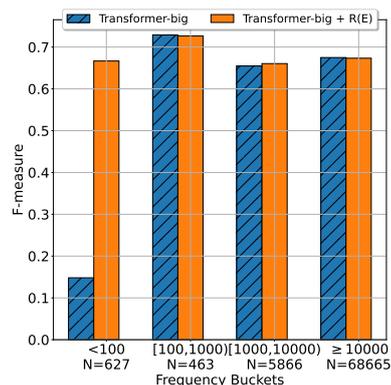


Figure 6: F1 score for de-en for each vocabulary token’s frequency bucket.

A Training Details

A.1 Data Statistics

B Additional Results

B.1 Frequency-based Analysis

Figure 6 provides frequency-based token analysis for de-en Transformer-big model. Similar to the Figure 2, the performance of the model with dispersion regularization on rare tokens is significantly better than the baseline model.

B.2 Quantization

architecture	precision	size
transformer-big	float32	100%
transformer-big	float16	50%
transformer-big	int8_float16	25%
transformer-base	float32	30%

Table 5: Transformer-big size on disk for the original models and quantized versions. Percentages based on disk storage.

Table 5 shows the model size on disk both for the original Transformer-big model, as well as various compression levels in CTranslate2. This was done as a proxy for parameter counts, as CTranslate2 returns a compressed model file. The

float32 model compressed with CTranslate2 is included as a reference point, such that the differences in file size are only due to the compression level. Note that the int8_float16 model is approximately four times smaller than the float32 compressed model. This is roughly equivalent to the difference in size between a Transformer-base and a Transformer-big.

C Additional Results

We report additional results for model training with dispersion regularization on the WMT16 Romanian→English (ro-en) benchmark. The dataset contains 612K training sentence pairs. For subword tokenization, we use the same SentencePiece model (Kudo and Richardson, 2018) across all language pairs, specifically the one employed in the mBART multilingual model (Liu et al., 2020). The vocabulary size is 27K on the target side and 28K on the source side. All hyperparameters are identical to those used for de-en, except for dropout, which is set to 0.3. Table 6 shows

model	ro-en			
	newsdev16		newstest16	
	BLEU	COMET	BLEU	COMET
Transformer-big	32.6	0.786	31.2	0.790
+ $R(\mathbf{H})$	32.7	0.785	31.0	0.788
+ $R(\mathbf{E})$	33.0	0.787	31.4	0.791
Transformer-base	33.2	0.793	31.8	0.795
+ $R(\mathbf{H})$	33.0	0.789	31.9	0.791
+ $R(\mathbf{E})$	33.6	0.794	32.3	0.799
+ $R(\mathbf{F})$	33.1	0.789	32.1	0.793

Table 6: BLEU and COMET on newsdev and newstest for ro-en with different dispersion regularizers. Values in bold indicate results that are statistically significant compared to the baseline (p-value < 0.05)

that the Transformer-base model benefits substantially from embedding dispersion, whereas for Transformer-big only marginal improvements are observed. Overall, these observations are consistent with our main results and further support the conclusion that dispersion is a beneficial property for the model.