Enhancing Context Through Contrast

Kshitij Ambilduke
VNIT, India
kshitij.a.vnit@gmail.com

Aneesh Shetye
VNIT, India
aneeshashetye@gmail.com

Diksha Bagade
VNIT, India
diskhabagade003@gmail.com

Rishika Bhagwatkar
VNIT, India
rishika.vnit@gmail.com

Khurshed Fitter
VNIT, India
khurshedpf@gmail.com

Prasad Vagdargi
Johns Hopkins University, USA
prasad@jhu.edu

Shital Chiddarwar
VNIT, India
shitalsc@mec.vnit.ac.in

Abstract

Neural machine translation benefits from semantically rich representations. Considerable progress in learning such representations has been achieved by language modelling and mutual information maximization objectives using contrastive learning. The language-dependent nature of language modelling introduces a trade-off between the universality of the learned representations and the model's performance on the language modelling tasks. Although contrastive learning improves performance, its success cannot be attributed to mutual information alone. We propose a novel Context Enhancement step to improve performance on neural machine translation by maximizing mutual information using the Barlow Twins loss. Unlike other approaches, we do not explicitly augment the data but view languages as implicit augmentations, eradicating the risk of disrupting semantic information. Further, our method does not learn embeddings from scratch and can be generalised to any set of pre-trained embeddings. Finally, we evaluate the language-agnosticism of our embeddings through language classification and use them for neural machine translation to compare with state-of-the-art approaches.

1 Introduction

The performance of Deep Learning models implicitly depends on the data representations [4], hence learning paradigms and metrics are defined in ways that optimize the model's capacity to extract useful features from the data. Contrastive Learning (CL) approaches focus on learning representations of data, generally in self-supervised settings [26]. The abundance of unlabeled visual data and the ease of introducing subtle yet effective augmentations are two main factors responsible for the success of CL models. The pivotal motivation is to maximize the Mutual Information (MI) between features extracted from augmented views of the data. Although CL paradigms achieve SOTA performance on a variety of tasks, their success cannot be attributed to the properties of MI alone [55].

In the lingual domain, representations are affected by the semantic as well as the temporal information present in the data [32]. Traditional approaches [39, 44] try to encode words into vectors according to their relative positions in the corpus, whereas recent approaches optimize performance on language modelling tasks to learn representations [46, 14, 7, 30, 37]. Lately, CL-based approaches have emerged for learning universal representations by introducing augmentations during pre-training [48, 35, 42, 10]. However, the discrete nature of languages makes it difficult to design label-preserving

35th Pre-registration workshop (NeurIPS 2021), Remote.

data augmentations [45]. Also, training paradigms like Multilingual Masked Language Modelling (MMLM) [14, 10, 29] and Translation Language Modelling (TLM) [28, 10] require the model to learn language-specific information too [10]. Due to mixed training objectives, there exists a trade-off between the universality of the learned representations which depends on language-agnosticism and the performance of the model on tasks that require language-specific information [10].

Neural Machine Translation (NMT) models aim to translate sentences from one language to another while preserving meaning. This requires models to focus on extracting the semantic and language-agnostic information over the language-specific information. For improving this, we propose a novel Context Enhancement (CE) step that leverages the Barlow Twins loss [62] to maximize MI and minimize redundancies between representations of parallel sentences. We do not explicitly augment the data and rather consider sentences as augmented views of their meaning. Further, we do not learn the embeddings from scratch and enhance pre-trained embeddings, increasing generalizability and reducing the compute footprint. Unlike similar works [28, 10, 29], our objective does not conflict with the primary training objective of NMT. We aim to validate our approach by evaluating performance on Language Classification and NMT by using the WMT-14 [5] En $\rightarrow$ De and En $\leftrightarrow$ Fr datasets to compare the performance with state-of-the-art (SOTA) approaches for NMT.

Our main contributions are:

1. Improving performance on NMT through a novel Context Enhancement step that maximizes MI by leveraging a contrastive loss, namely Barlow Twins, without explicitly augmenting the data.
2. We do not learn embeddings from scratch, hence our method and experiments can be generalised to any set of pre-trained embeddings.

2 Related works

2.1 Contrastive learning

Deep convolutional networks [54, 23, 51, 27, 49, 6, 18] and even Transformers [59, 16] have played a foundational role in learning reliable representations from labeled visual data. Owing to the abundance of unlabeled data, there has been a shift from supervised to self-supervised learning [63, 41, 43, 15, 20]. Recently, CL-based approaches [56, 9, 24, 21, 8, 62, 3] have gained popularity and have shown exceptional performance in a variety of downstream tasks.

Most CL objectives maximize a tractable estimate of the lower bound of MI between two augmented views of the same image [55]. Some approaches benefit from large batch sizes [9] and careful implementation tricks like momentum updates [24, 21] or asymmetric encoders [21] to prevent collapse. However, Barlow Twins and VICReg [62, 3] introduce loss functions that naturally avoid collapse and reduce the dependency on the number of negatives while maximizing MI.

2.2 Neural machine translation

Performance of Natural Language Processing (NLP) models inherently depends on the word and sentence embeddings. Models trained on large multilingual corpora learn embeddings that can be used for a variety of downstream tasks [14, 47, 38, 31]. Some models try to improve performance on multilingual tasks by focusing on learning language-agnostic components [1, 58, 10, 35, 42]. Although centroid subtraction displays signs of eradicating language-specific components [34], recent works leverage contrastive approaches for the same [10, 35, 42]. However, recent literature has shown that their success cannot be attributed to the properties of MI alone and rather it depends on the choice of feature extractor architectures and the parametrization of the employed MI estimators [55]. Due to random masking, MMLM and TLM require the model and embeddings to learn language-specific information [10]. This may lead to a trade-off between the universality of the learned embeddings and their performance on these tasks. Also, these paradigms require longer training durations [11] and may not push the model to learn meaningful language semantics by masking common words [61] or words with too many false negatives [22].

Most approaches for NMT use encoder-decoder architectures [50, 2, 19, 57]. The current SOTA methods [53, 52, 17, 33] introduce subtle yet effective changes in the architecture [53] and training method [52, 33] of the original Transformer [57]. Some methods even introduce augmentations by back-translation or by exchanging words with their synonyms and cognates [35, 42, 17, 36, 40]. However, directly augmenting languages may alter the semantic and syntactic correctness [35, 42].

Since NMT leverages joint information from two sentences, we improve NMT performance by maximizing mutual information and minimizing redundancies between representations of such sentences. Our method does not rely on explicitly augmenting the data and instead treats languages as inherent augmentations introduced in the process of representing abstract meaning. Further, we do not directly maximise any lower bound estimates on the MI but rather use an instantiation of the Information Bottleneck Principle through Barlow Twins [62]. In addition, our method does not learn embeddings from scratch but improves the language-agnosticism of pre-trained embeddings.

3 Approach

In a typical Transformer-based sequence-to-sequence translation model [57], the encoder $E(.;\theta_E)$ learns to map a sequence of n-dimensional word embeddings $\tilde{\mathbf{x}} = (x_1, x_2, \cdots, x_{t_1}) \in \mathbb{R}^{t_1 \times n}$ from the source language, to a sequence of h-dimensional latent representations $\tilde{\boldsymbol{\omega}} = (\omega_1, \omega_2, \cdots, \omega_{t_1}) \in \mathbb{R}^{t_1 \times h}$. This is followed by a decoder $D(.;\theta_D)$ which maps the sequence $\tilde{\boldsymbol{\omega}}$, to a sequence of tokens $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, ...\hat{y}_{t_2}) \in \mathbb{R}^{t_2 \times n}$ in the target language. The encoder uses masked self-attention whereas the decoder uses both cross-attention and self-attention. For a parallel corpus ξ, the loss function $\mathcal{L}_{trans}$, optimizes the objective $\mathcal{O}_{trans}$, that is the log probability of obtaining the correct translation $\tilde{\mathbf{y}}$, given the source sentence $\tilde{\mathbf{x}}$

$$\mathcal{O}_{trans} = -\frac{1}{|\xi|} \sum_{\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \xi} \log p(\hat{\mathbf{y}}|\tilde{\mathbf{x}}) \quad , \quad \mathcal{L}_{trans} = -\frac{1}{|\xi|} \sum_{\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \xi} \tilde{\mathbf{y}} \log \hat{\mathbf{y}} \tag{1}$$

Unlike recent works [35, 42], our method does not depend on explicitly augmenting the training data. Rather, we hypothesise corresponding sentences from parallel corpora $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ as different views of the same meaning Ω i.e. languages are linguistic transforms that map meaning to sentences.

$$\tilde{\mathbf{x}} = \Lambda_S(\Omega) \quad , \quad \tilde{\mathbf{y}} = \Lambda_T(\Omega) \tag{2}$$

where, Λ_S and Λ_T represent the linguistic transforms. The encoder tries to learn a transform Λ_s^*, that maps sentences to their meaning. An ideal encoder-decoder pair would learn the transforms Λ_s^* and Λ_T respectively, such that $\Lambda_s^*(\tilde{\mathbf{x}}) = \Lambda_s^*(\Lambda_S(\Omega)) = \Omega$ and $\Lambda_T(\Omega) = \tilde{\mathbf{y}}$.

We intend to improve NMT performance by maximizing MI between the representations of parallel sentences and minimizing the redundant information about the language-specific components. We propose an additional CE step for Transformer-based NMT models that focuses on enriching the language-agnostic features of the sentence embeddings by using a contrastive loss function $\mathcal{L}_{\mathcal{BT}}$ inspired by Barlow Twins [62].

Analogous to the original work [62], we use a Transformer encoder network that encodes two parallel sentences $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ from two different languages (S, T) into two sequences of latent representations $\tilde{\boldsymbol{\omega}}^S = E(\tilde{\mathbf{x}}; \theta_E)$ and $\tilde{\boldsymbol{\omega}}^T = E(\tilde{\mathbf{y}}; \theta_E)$. Then a pooling function $\phi(.)$ is used to obtain sentence embeddings $\boldsymbol{\sigma}^S = \phi(\tilde{\boldsymbol{\omega}}^S)$ and $\boldsymbol{\sigma}^T = \phi(\tilde{\boldsymbol{\omega}}^T) \in \mathbb{R}^{B \times h}$. The loss is calculated between batch normalized projections $\mathbf{Z}^S = \mathtt{BN}(\rho(\boldsymbol{\sigma}^S; \theta_\rho))$ and $\mathbf{Z}^T = \mathtt{BN}(\rho(\boldsymbol{\sigma}^T; \theta_\rho)) \in \mathbb{R}^{B \times d}$ where $\rho(.;\theta_\rho)$ represents the projection network.

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction term}} \tag{3}$$

where λ is a positive constant controlling the relative importance of the two terms. $\mathcal{C}$ is the empirical cross-correlation matrix computed between the two batches of projections:

$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z_{b,i}^S z_{b,j}^T}{\sqrt{\sum_b (z_{b,i}^S)^2} \sqrt{\sum_b (z_{b,j}^T)^2}} \tag{4}$$

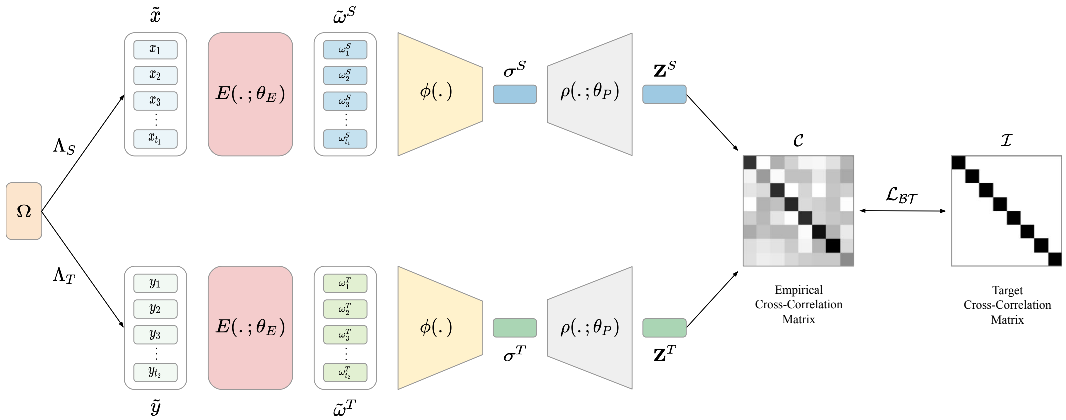

Figure 1: A block diagram of our proposed architecture for the CE step. The encoder maps sentences $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ to sequences of latent representations $(\tilde{\boldsymbol{\omega}}^S, \tilde{\boldsymbol{\omega}}^T)$. These are then aggregated to get sentence embeddings $\tilde{\boldsymbol{\sigma}}^S = \phi(\tilde{\boldsymbol{\omega}}^S)$ and $\tilde{\boldsymbol{\sigma}}^T = \phi(\tilde{\boldsymbol{\omega}}^T)$. The contrastive loss $\mathcal{L}_{\mathcal{BT}}$, is applied to batch normalized projections $\mathbf{Z}^S = \texttt{BN}(\rho(\boldsymbol{\sigma}^S; \theta_\rho))$ and $\mathbf{Z}^T = \texttt{BN}(\rho(\boldsymbol{\sigma}^T; \theta_\rho))$. The weights θ_E, are fine tuned for NMT after the CE step. Also, $\tilde{\boldsymbol{\omega}}^S$ is directly passed to the decoder while training on NMT.

where, z_b^S and z_b^T are the b^{th} batch samples and i, j indicate the projection network's output dimensions. The cross-correlation matrix $\mathcal{C} \in \mathbb{R}^{d \times d}$ consists of values between 1 and -1 representing ideal correlation and anti-correlation respectively.

We improve the pre-trained embeddings by optimizing the contrastive loss $\mathcal{L}_{\mathcal{BT}}$ on sentence embeddings with only the encoder during the CE step. For obtaining sentence embeddings, we use pooling as a substitute to the widely used `[CLS]` token [47]. Using the learned weights from the CE step, we attach a matching decoder to train on NMT by optimizing the loss $\mathcal{L}_{trans}$. However, during the latter step, the encoder's output $\tilde{\boldsymbol{\omega}}^S$, is directly passed to the decoder without applying pooling, projection or batch normalization.

4 Experiments

4.1 Datasets

WMT 2014 English-German: This dataset [5] contains about 4.5M En-De parallel sentences from Europarl, News Commentary and Common Crawl.

WMT 2014 English-French: This dataset [5] contains about 27.9M En-Fr parallel sentences from Europarl, News Commentary, Common Crawl and the 10^9 Word corpora.

Further, we will expand our evaluation to other language pairs from distant families following preliminary results.

4.2 Ablations

4.2.1 Context enhancement step

In the CE step, we use N encoder blocks of the Transformer to form a sentence encoder. We train the encoder using the Barlow Twins loss $\mathcal{L}_{\mathcal{BT}}$, for a relatively small number of epochs ≤ 1000 with different values of λ between 0 and 1 (in steps of 5×10^{-3}). We use pre-trained embeddings from mBERT [14], InfoXLM [10], XLM-RoBERTa [13] and XLM [29].

Encoder architecture: We vary the depth of the model N, from 8 to 24 (in steps of 2) and the model dimension h, from about 500 to 2,000 (in steps of ≈ 200), following the work on mBERT [14]. Further, we vary the number of attention heads from 4 to 12 (in steps of 2).

Pooling function: We experiment with two pooling functions, $\phi_{mean}(.)$ and $\phi_{max}(.)$, representing average pooling and max pooling respectively.

Projection network: The projection network has 3 linear layers, each having d output units. The first two layers are followed by batch normalization and rectified linear units. We study how the projection dimension d, affects the performance of our model on each evaluation task by varying it from 32 to 16,384 (as per the powers of 2), following the original work [62].

Batch size: We study the dependence of our method on the batch size B, by varying the batch size from 128 to 4,096 (as per the powers of 2), as proposed in the original work [62].

4.2.2 Translation

For each of the settings from Section 4.2.1, we fine-tune the model for NMT after the CE step. A decoder with the same number of layers and model dimension is jointly trained with the context enhanced encoder. Further, the decoder uses masked self-attention and cross-attention as opposed to only self-attention in the encoder [57]. However, during translation, the output of the encoder $\tilde{\omega}^S$, is passed directly to the decoder without using the pooling, projection or batch normalization layers.

Table 1: SACRE-BLEU [25] scores (higher is better) on WMT-14 dataset for En-De and En-Fr NMT († Represents methods that use augmentations)

Method	**En→De** (↑)	**En→Fr** (↑)
Transformer [57]	29.12	42.69
MUSE [12]	29.90	43.50
Depth Growing [60]	30.07	43.27
Transformer-Admin † [36]	30.10	43.80
Data-Diversification † [40]	30.70	43.70
BERT-Fused NMT [64]	30.75	43.78
Transformer + RD [33]	30.91	43.95
Ours (centroid subtracted)	-	-
Ours (after CE)	-	-

Table 2: Tokenized-BLEU scores (higher is better) on WMT-14 dataset for Fr-En NMT.

Method	**Fr→En** (↑)
Transformer-6 [57]	39.8
mRASP2 † [42]	39.3
mRASP † [35]	45.4
Ours (centroid subtracted)	-
Ours (after CE)	-

4.3 Evaluation

For the classification task, sentence embeddings are obtained by pooling the encoder's output. However, for translation, the entire output sequence $\tilde{\omega}^S$ is passed to the decoder without pooling.

Translation: We evaluate our model's performance on NMT before and after the CE step. We compare it with SOTA models as shown in Tables 1 and 2. Further, we also evaluate the performance after subtracting the centroid from pre-trained word embeddings [34].

Language classification: To evaluate the language-agnosticism of the embeddings learned by our model, we perform language classification on them. We compute the accuracy a_1 of a language classifier C_1 trained on sentence embeddings obtained from mBERT after pooling. Freezing the parameters of C_1, we evaluate it's accuracy a_2 on embeddings obtained after the CE step. Then, we train a language classifier C_2 on embeddings obtained after the CE step and compute it's accuracy a_3. For both word and sentence embeddings, the relation $a_2 < a_3 < a_1$ indicates the absence of language-specific components in the embeddings, validating an increase in language-agnosticism.

To compare our method with prior works [34], we compute a'_1, a'_2 and a'_3 before and after subtracting the centroid of all sentence embeddings. We compute a'_1 by training a classifier C'_1 on the sentence embeddings obtained from mBERT. Then, for computing a'_2, we evaluate C'_1 on the centroid subtracted sentence embeddings. Finally, we compute a'_3 as the accuracy of a model C'_2 trained on the centroid subtracted sentence embeddings. We extend this entire procedure for word embeddings too.

Qualitative analysis: We visualize the distribution of word and sentence embeddings using t-SNE plots of word and sentence embeddings before and after the CE step. To analyse the word-level redundancies, we plot the correlation matrices between corresponding word pairs at different stages of the CE step. Further, we plot the attention maps of every head of the encoder and decoder to evaluate how the CE step affects the attention mechanism.

5 Conclusion

We propose a novel Context Enhancement step to push neural machine translation performance using contrastive learning. Our method maximizes Mutual Information between two views of the same meaning by leveraging the Barlow Twins loss. Unlike most works, our method does not depend on explicit augmentations or implementation tricks. Further, our proposed objective pushes the model to learn language-agnostic features which directly improves neural machine translation performance.

References

[1] M. Artetxe and H. Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, 2019.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[3] A. Bardes, J. Ponce, and Y. LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2021.

[4] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives, 2014.

[5] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, and L. Specia, editors. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-33. URL `https://aclanthology.org/W14-3300`.

[6] A. Brock, S. De, S. L. Smith, and K. Simonyan. High-performance large-scale image recognition without normalization, 2021.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[8] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations, 2020.

[10] Z. Chi, L. Dong, F. Wei, N. Yang, S. Singhal, W. Wang, X. Song, X.-L. Mao, H. Huang, and M. Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training, 2021.

[11] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020.

[12] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data, 2018.

[13] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[15] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction, 2016.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[17] S. Edunov, M. Ott, M. Auli, and D. Grangier. Understanding back-translation at scale, 2018.

[18] I. Freeman, L. Roese-Koerner, and A. Kummert. Effnet: An efficient structure for convolutional neural networks, 2018.

[19] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning, 2017.

[20] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations, 2018.

[21] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[22] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang. Realm: Retrieval-augmented language model pre-training, 2020.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[25] https://github.com/awslabs/sockeye/.

[26] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A survey on contrastive self-supervised learning, 2021.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[28] G. Lample and A. Conneau. Cross-lingual language model pretraining, 2019.

[29] G. Lample and A. Conneau. Cross-lingual language model pretraining, 2019.

[30] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.

[31] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[32] Y. Li and T. Yang. Word embedding for understanding natural language: A survey. 2018.

[33] X. Liang, L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, and T.-Y. Liu. R-drop: Regularized dropout for neural networks, 2021.

[34] J. Libovický, R. Rosa, and A. Fraser. How language-neutral is multilingual bert?, 2019.

[35] Z. Lin, X. Pan, M. Wang, X. Qiu, J. Feng, H. Zhou, and L. Li. Pre-training multilingual neural machine translation by leveraging alignment information, 2021.

[36] X. Liu, K. Duh, L. Liu, and J. Gao. Very deep transformers for neural machine translation, 2020.

[37] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[38] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020.

[39] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[40] X.-P. Nguyen, S. Joty, W. Kui, and A. T. Aw. Data diversification: A simple strategy for neural machine translation, 2020.

[41] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles, 2017.

[42] X. Pan, M. Wang, L. Wu, and L. Li. Contrastive learning for many-to-many multilingual neural machine translation, 2021.

[43] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting, 2016.

[44] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

[45] Y. Qu, D. Shen, Y. Shen, S. Sajeev, J. Han, and W. Chen. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding, 2020.

[46] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2018. URL `https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf`.

[47] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[48] N. Rethmeier and I. Augenstein. A primer on contrastive pretraining in language processing: Methods, lessons learned and perspectives, 2021.

[49] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[50] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks, 2014.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, 2014.

[52] S. Takase and S. Kiyono. Rethinking perturbations in encoder-decoders for fast training, 2021.

[53] S. Takase and S. Kiyono. Lessons on parameter sharing across layers in transformers, 2021.

[54] M. Tan and Q. V. Le. Efficientnetv2: Smaller models and faster training, 2021.

[55] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning, 2020.

[56] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding, 2019.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.

[58] R. Vázquez, A. Raganato, J. Tiedemann, and M. Creutz. Multilingual nmt with a language-independent attention bridge. *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 2019. doi: 10.18653/v1/w19-4305. URL `http://dx.doi.org/10.18653/v1/W19-4305`.

[59] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.

[60] L. Wu, Y. Wang, Y. Xia, F. Tian, F. Gao, T. Qin, J. Lai, and T.-Y. Liu. Depth growing for neural machine translation, 2019.

[61] Z. Xu, L. Gong, G. Ke, D. He, S. Zheng, L. Wang, J. Bian, and T.-Y. Liu. Mc-bert: Efficient language pre-training via a meta controller, 2020.

[62] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021.

[63] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization, 2016.

[64] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T.-Y. Liu. Incorporating bert into neural machine translation, 2020.