BOOTSTRAPPING PARALLEL ANCHORS FOR RELATIVE REPRESENTATIONS

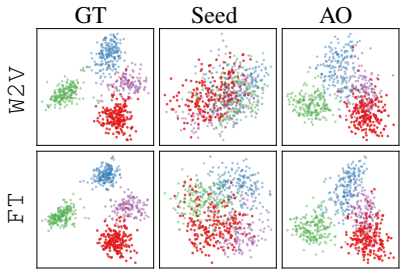
Irene Cannistraci[1] **Luca Moschella**[1] **Valentino Maiorca**[1]
Marco Fumero[1] **Antonio Norelli**[1] **Emanuele Rodolà**[1]
[1]Sapienza University of Rome

ABSTRACT

The use of relative representations for latent embeddings has shown potential in enabling latent space communication and zero-shot model stitching across a wide range of applications. Nevertheless, relative representations rely on a certain amount of parallel anchors to be given as input, which can be impractical to obtain in certain scenarios. To overcome this limitation, we propose an optimization-based method to discover new parallel anchors from a limited known set (*seed*). Our approach can be used to find semantic correspondence between different domains, align their relative spaces, and achieve competitive results in several tasks.

1 INTRODUCTION

Over the past few years, several studies have acknowledged how successful neural networks typically learn comparable representations regardless of their architecture, task, or domain (Li et al., 2016; Kornblith et al., 2019; Vulić et al., 2020). In line with this trend, Moschella et al. (2023) introduced the concept of relative representation, aiming to generate comparable latent spaces and enable zero-shot stitching to handle new, unseen tasks without requiring additional training. The approach consists in representing each data sample through latent similarities with respect to a set of training samples, denoted as *anchors*. This procedure transforms the absolute reference frame to a relative coordinate system defined by the anchors. To enable tasks like multimodal learning, this approach requires a semantic connection between the anchors of two data domains, denoted as *parallel anchors*. This correspondence, which must be provided as input, allows domain comparison and links their respective latent spaces (Norelli et al., 2022). However, obtaining a sufficient number of parallel anchors in specific applications can be challenging or impossible, hindering the use of relative representations. We focus on the scenario where there are only a very limited number of parallel anchors available, called *seed*, and we aim to expand this initial set through an Anchor Optimization (AO) process. Our method achieves competitive performance in NLP and Vision domains while significantly reducing the number of required parallel anchors by *one order of magnitude*.



	Src	Tgt	Jaccard ↑	MRR ↑	Cosine ↑
GT	FT	W2V	0.34 ± 0.01	0.94 ± 0.00	0.86 ± 0.00
	W2V	FT	0.39 ± 0.00	0.98 ± 0.00	0.86 ± 0.00
Seed	FT	W2V	0.06 ± 0.01	0.11 ± 0.01	0.85 ± 0.01
	W2V	FT	0.06 ± 0.01	0.15 ± 0.02	0.85 ± 0.01
AO	FT	W2V	0.52 ± 0.00	0.99 ± 0.00	0.94 ± 0.00
	W2V	FT	0.50 ± 0.01	0.99 ± 0.00	0.94 ± 0.00

Table 1: Qualitative (*left*) and quantitative (*right*) evaluation of the AO method in the retrieval task.

2 METHOD

Let us be given two domains $\mathcal{X}$ and $\mathcal{Y}$ and corresponding learned embedding functions $E_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}^n$ and $E_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R}^m$, where possibly $n \neq m$. Given two sets of anchors $\mathcal{A}_{\mathcal{X}} \subset \mathcal{X}$ and

$\mathcal{A}_\mathcal{Y} \subset \mathcal{Y}$, we define *parallel anchors* a subset of pairs $\mathcal{A}_p \subseteq \mathcal{A}_\mathcal{X} \times \mathcal{A}_\mathcal{Y}$ in semantic correspondence, e.g., images and captions as in Norelli et al. (2022). The relative representations for a sample $x \in \mathcal{X}$ (same for $\mathcal{Y}$) is computed as follows: $rr(x, \mathcal{A}_\mathcal{X}) = E_\mathcal{X}(x)\mathbf{A}_\mathcal{X}^T$, where $\mathbf{A}_\mathcal{X} = \bigoplus_{a \in \mathcal{A}_\mathcal{X}} E_\mathcal{X}(a)$, and $\bigoplus$ denotes the row-wise concatenation operator. We assume all embeddings are rescaled to unit norm, i.e., $\forall x \; ||E(x)|| = 1$. This corresponds to the choice of cosine similarity as a similarity function, according to the setting of Moschella et al. (2023).

In this work, we introduce an optimization procedure that reduces the required number of parallel anchors by one order of magnitude. Our method does not require complete knowledge of $\mathcal{A}_p$ but only of few initial *seed* anchors, denoted as $\mathcal{L} = \mathcal{L}_\mathcal{X} \times \mathcal{L}_\mathcal{Y} \subseteq \mathcal{A}p$, where $|\mathcal{L}| \ll |\mathcal{A}p|$. With no prior knowledge of $\mathcal{A}_\mathcal{Y}$, we initialize the optimization process by approximating $\mathbf{A}_\mathcal{Y} \approx \widetilde{\mathbf{A}}_\mathcal{Y}$ with the known seed $\mathbf{A}_{\mathcal{L}_\mathcal{Y}} = \bigoplus_{a \in \mathcal{L}_\mathcal{Y}} E_\mathcal{Y}(a)$ concatenated with $|\mathcal{A}_p| - |\mathcal{L}|$ random embeddings, i.e. $\widetilde{\mathbf{A}}_\mathcal{Y} = \mathbf{A}_{\mathcal{L}_\mathcal{Y}} \oplus \mathbf{N}$, with $\mathbf{N} \sim \mathcal{N}(0, \mathbf{I})$. We define the following objective function optimizing over $\widetilde{\mathbf{A}}_\mathcal{Y}$:

$$\underset{\widetilde{\mathbf{A}}_\mathcal{Y} \text{ s.t. } ||a||_2 = 1 \; \forall a \in \widetilde{\mathbf{A}}_\mathcal{Y}}{\arg\min} \sum_{y \in \mathcal{Y}} MSE(rr(\Pi(y), \mathcal{A}_\mathcal{X}), E_\mathcal{Y}(y)\widetilde{\mathbf{A}}_\mathcal{Y}^T) \tag{1}$$

where $\Pi : \mathcal{Y} \to \mathcal{X}$ is a correspondence estimated at each optimization step by the Sinkhorn (Cuturi, 2013) algorithm exploiting the initial seed and the current approximation of the remaining anchors: $\Pi = sinkhorn_{(x,y) \in \mathcal{X} \times \mathcal{Y}}(rr(x, \mathcal{A}_\mathcal{X}), E_\mathcal{Y}(y)\widetilde{\mathbf{A}}_\mathcal{Y}^T)$. After convergence, $\widetilde{\mathbf{A}}_\mathcal{Y}$ is discretized into $\widetilde{\mathcal{A}}_\mathcal{Y} \subseteq \mathcal{Y}$ considering the nearest embeddings in $E_\mathcal{Y}(\mathcal{Y})$. Further details in Appendix A.2.

3 Experiments

This section assesses the effectiveness of our AO method. We utilize 15 anchor to approximate 300 parallel anchors that serve as ground truth in all downstream tasks. Specifically, we compare the performance of our method against two different baselines: (1) *GT*, the Ground Truth employs all the anchors that our method aims to semantically approximate, (2) *Seed*, exploits only the seed anchors. For more information on the implementation please refer to the Appendix A.2 and the code[1].

Our method effectively discovers parallel anchors in the NLP and Vision domains, as demonstrated in Tables 1, 4, 5 and 7. Specifically, we explore different word embeddings and pre-trained foundational visual encoders, and assess the quality of the discovered anchors through a retrieval task. Results demonstrate that, when given the same number of starting anchors, our method outperforms the approach that relies solely on those without optimizing. Moreover, our results are *comparable or superior* to those obtained with all the ground truth parallel anchors. Furthermore, Table 6 demonstrates that our method can discover parallel anchors across different domains: the method finds aligned Amazon reviews in different languages with unavailable ground truth. Using only 15 *OOD* (Moschella et al., 2023) parallel anchors, our method enables zero-shot stitching, allowing to train a classifier on one language and perform predictions on another one.

Table 2: Cross-lingual zero-shot stitching performance evaluation.

		GT		Seed		AO	
Dec.	Enc.	Fscore	MAE	Fscore	MAE	Fscore	MAE
en	es	0.51 ± 0.01	0.67 ± 0.02	0.44 ± 0.01	0.80 ± 0.01	0.48 ± 0.01	0.70 ± 0.02
es	en	0.50 ± 0.02	0.72 ± 0.04	0.41 ± 0.01	0.92 ± 0.02	0.46 ± 0.01	0.76 ± 0.02

4 Conclusions, future works, and limitations

In this paper, we presented a novel method to compute robust relative representations even in scenarios where only a reduced number of parallel anchors is available. The method expands semantic correspondence between data domains without prior knowledge and achieves comparable results with *one order of magnitude fewer* parallel anchors. This approach has notable implications for latent space communication across domains with limited knowledge about semantic correspondence. Future research is needed to remove the need for an initial parallel seed.

[1]Fully reproducible codebase at: `https://github.com/icannistraci/bootstrapping-ao`

ACKNOWLEDGEMENTS

This work is supported by the ERC Grant no.802554 "SPECGEO" and PRIN 2020 project no.2020TA3K9N "LEGO.AI".

URM STATEMENT

The first author meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Federico Bianchi an. Fantastic embeddings and how to align them: Zero-shot inference i. *ArXiv preprint*, abs/2007.14906, 2020. URL https://arxiv.org/abs/2007.14906.

Niccolo Biondi an. Cores: Compatible representations via stationarity. *ArXiv preprint*, abs/2111.07632, 2021. URL https://arxiv.org/abs/2111.07632.

Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8332–8344. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/464074179972cbbd75a39abc6954cd12-Paper.pdf.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 225–236, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html.

Serguei Barannikov, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev. Representation topology divergence: A method for comparing neural network representations. *ArXiv preprint*, abs/2201.00058, 2022. URL https://arxiv.org/abs/2201.00058.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL https://aclanthology.org/Q17-1010.

Lisa Bonheme and Marek Grzes. How do variational autoencoders learn? insights from representational similarity. *ArXiv preprint*, abs/2205.08399, 2022. URL https://arxiv.org/abs/2205.08399.

Mario Lezcano Casado. Trivializations for gradient-based optimization on manifolds. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 9154–9164, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1b33d16fc562464579b7199ca3114982-Abstract.html.

Adrián Csiszárik, Péter Kőrösi-Szabó, Ákos K. Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations. *ArXiv preprint*, abs/2110.14633, 2021. URL https://arxiv.org/abs/2110.14633.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Michael Gygli, Jasper Uijlings, and Vittorio Ferrari. Towards reusable network components by learning compatible representations. *AAAI*, 35(9):7620–7629, 2021.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4563–4568, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.369. URL https://aclanthology.org/2020.emnlp-main.369.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL http://proceedings.mlr.press/v97/kornblith19a.html.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ruslan Kuprieiev, skshetry, Dmitry Petrov, Paweł Redzyński, Peter Rowlands, Casper da Costa-Luis, Alexander Schepanovski, Ivan Shcheklein, Batuhan Taskaya, Gao, Jorge Orpinel, David de la Iglesia Castro, Fábio Santos, Aman Sharma, Dave Berenbaum, Zhanibek, Dani Hodovic, daniele, Nikita Kodenko, Andrew Grigorev, Earl, Nabanita Dash, George Vyshnya, Ronan Lamy, maykulkarni, Max Hora, Vera, and Sanidhya Mangal. Dvc: Data version control - git for data & models, 2022. URL https://doi.org/10.5281/zenodo.7083378.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H196sainb.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 991–999. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298701. URL https://doi.org/10.1109/CVPR.2015.7298701.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John E. Hopcroft. Convergent learning: Do different neural networks learn the same representations? In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.07543.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692, 2019. URL https://arxiv.org/abs/1907.11692.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ArXiv preprint*, abs/1301.3781, 2013a. URL https://arxiv.org/abs/1301.3781.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013b. URL http://arxiv.org/abs/1309.4168.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 5732–5741, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a7a3d70c6d17a73140918996d03c014f-Abstract.html.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SrC-nwieGJ.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training. *ArXiv preprint*, abs/2210.01738, 2022. URL https://arxiv.org/abs/2210.01738.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 843–852. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.97. URL https://doi.org/10.1109/ICCV.2017.97.

Anton Tsitsulin, Marina Munkhoeva, Davide Mottin, Panagiotis Karras, Alexander M. Bronstein, Ivan V. Oseledets, and Emmanuel Müller. The shape of data: Intrinsic distance for data distributions. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HyebplHYwB.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3178–3192, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.257. URL https://aclanthology.org/2020.emnlp-main.257.

A APPENDIX

A.1 RELATED WORKS

In recent years, numerous studies (Lenc & Vedaldi, 2015; Mikolov et al., 2013b; Li et al., 2016; Lample et al., 2018; Morcos et al., 2018; Tsitsulin et al., 2020; Kornblith et al., 2019; Vulić et al., 2020; Antonello et al., 2021; Bonheme & Grzes, 2022; Barannikov et al., 2022; Norelli et al., 2022) have recognized that neural networks tend to learn comparable representations regardless of their architecture, task, or domain when trained on semantically similar data. This observation can be exploited to enable various applications, such as model stitching (Lenc & Vedaldi, 2015; Bansal et al., 2021; Csiszárik et al., 2021; Gygli et al., 2021; an, 2021; 2020), latent model comparison or supervision and more. In particular, Moschella et al. (2023) introduced the framework of relative representation, which aims to unify the representations learned from semantically similar data. Relative representations have demonstrated potential in facilitating communication within latent embeddings and enabling zero-shot stitching across various applications, relying on parallel anchors to link different domains. Our work aims to minimize the explicit supervision required for latent communication by reducing the reliance on parallel anchors to the minimum necessary and automatically expand the provided semantic correspondence between domains.

A.2 IMPLEMENTATION DETAILS

This section provides further details about the optimization procedure and the experiments.

Optimization Method Algorithm 1 outlines the pseudocode for the optimization procedure described in Section 2, while Table 3 details the hyperparameters. The method initializes $\widetilde{\mathbf{A}}_{\mathcal{Y}}$ and optimizes it iteratively. At each step, the Sinkhorn algorithm computes a rough estimate of the permutation between the two relative spaces. The loss function minimized in our optimization procedure is the MSE, with particular emphasis placed on ensuring that the optimized parameters $\widetilde{\mathbf{A}}_{\mathcal{Y}}$ adhere to unit norm using Casado (2019). This not only ensures the effectiveness of the optimization but also reduces the search space.

Algorithm 1 Anchor Optimization

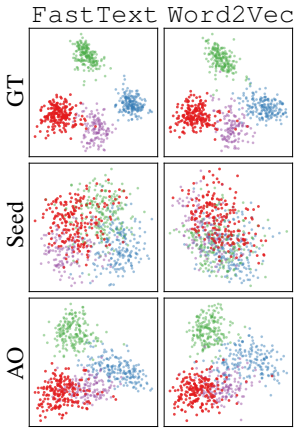
Initialize $\widetilde{\mathbf{A}}_{\mathcal{Y}} = \mathbf{A}_{\mathcal{L}_{\mathcal{Y}}} \oplus \mathbf{N}$, with $\mathbf{N} \in \mathcal{N}(0, \mathbf{I})$ and $|\mathbf{N}| = |\mathcal{A}_p| - |\mathcal{L}|$
Compute the relative representations of samples in $\mathcal{X}$ as $\mathbf{R}_{\mathcal{X}} = \bigoplus_{x \in \mathcal{X}} rr(x, \mathcal{A}_{\mathcal{X}})$
for K steps **do**
    Compute the relative representations of samples in $\mathcal{Y}$ as $\mathbf{R}_{\mathcal{Y}} = \bigoplus_{y \in \mathcal{Y}} E_{\mathcal{Y}}(y) \widetilde{\mathbf{A}}_{\mathcal{Y}}^T$
    Estimate the permutation between $\mathbf{R}_{\mathcal{Y}}$ and $\mathbf{R}_{\mathcal{X}}$ with $\Pi = sinkhorn(\mathbf{R}_{\mathcal{X}}, \mathbf{R}_{\mathcal{Y}})$
    Permute $\mathbf{R}_{\mathcal{Y}}$ according to Π
    Compute the error $MSE(\mathbf{R}_{\mathcal{X}}, \mathbf{R}_{\mathcal{Y}})$
    Optimize $\widetilde{\mathbf{A}}_{\mathcal{Y}}$ to minimize the error, while abiding to the constraint $||a||_2 = 1 \; \forall a \in \widetilde{\mathbf{A}}_{\mathcal{Y}}$
end for
return the nearest neighbours of $\widetilde{\mathbf{A}}_{\mathcal{Y}}$ in $E_{\mathcal{Y}}(\mathcal{Y})$

Retrieval Task We choose two English word embeddings trained on different data but with a partially shared vocabulary from which we extract $\approx$ 20K words: `FastText` (Bojanowski et al., 2017) and `Word2Vec` (Mikolov et al., 2013a). For testing the AO method, we select 15 seed anchors and shuffle the two embedding spaces to break their correspondence. Then, we choose 285 additional random anchors for one of the spaces while we use our optimization method to discover the associated 285 parallel anchors in the other one. Next, the absolute embeddings of each space are converted to their relative representations using the 300 optimized parallel anchors. For each word w, we consider its corresponding encodings x and y in the source and target space and validate their quality through a retrieval task. To facilitate a comparison with the relative representation baseline (Moschella et al., 2023), we employ the same evaluation metrics: (i) *Jaccard*: the discrete Jaccard similarity between the set of word neighbors of x in source and target; (ii) *Mean Reciprocal Rank (MRR)*: measures the (reciprocal) ranking of w among the top-k neighbors of x in the target space; (iii) *Cosine*: measures the cosine similarity between x and y. Results for the *GT* and *seed* methods

Table 3: Hyperparameter for the AO method in *retrieval* and *zero-shot stitching* tasks.

Hyperparameter	Retrieval	Zero-shot stitching
Random seed	0, 1, 2, 3, 4	0, 1, 2, 3, 4
Number of anchors to approximate	300	300
Number of seed anchors	15	15
Number of optimization steps	250	125
Learning Rate	0.02	0.05
Optimizer	Adam	Adam
Loss	MSE	MSE
Sinkhorn eps	1e-4	1e-4
Sinkhorn stop error	1e-5	1e-5
Number of Sinkhorn steps	1	1

are obtained by using all the given 300 anchors that our method aims to semantically approximate and only the 15 seed anchors, respectively.



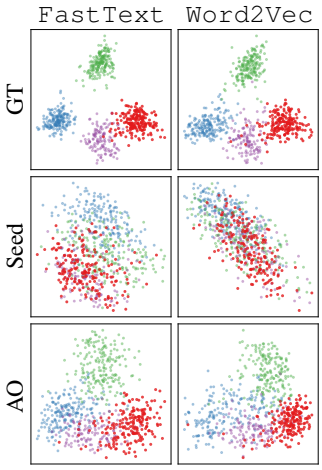
	Source	Target	Jaccard ↑	MRR ↑	Cosine ↑
GT	FT	FT	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		W2V	0.34 ± 0.01	0.94 ± 0.00	0.86 ± 0.00
	W2V	FT	0.39 ± 0.00	0.98 ± 0.00	0.86 ± 0.00
		W2V	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Seed	FT	FT	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		W2V	0.06 ± 0.01	0.11 ± 0.01	0.85 ± 0.01
	W2V	FT	0.06 ± 0.01	0.15 ± 0.02	0.85 ± 0.01
		W2V	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
AO	FT	FT	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		W2V	0.52 ± 0.00	0.99 ± 0.00	0.94 ± 0.00
	W2V	FT	0.50 ± 0.01	0.99 ± 0.00	0.94 ± 0.00
		W2V	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

Table 4: Complete results for the selected experiments are reported in Table 1. Qualitative (*left*) and quantitative (*right*) comparisons of the three methods when optimizing the `Word2Vec` space. All metrics are calculated with $K = 10$ averaged over 20k words across five random seeds.

Zero-shot stitching task We investigate the *Cross-lingual* text classification task on the multilingual `Amazon Reviews` dataset (Keung et al., 2020) to demonstrate a practical application of our method. We use the fine-grained formulation of the task, where the goal is to predict the star rating given a review (i.e., five classes to predict) and measure performance using FScore and Mean Absolute Error (MAE) metrics. To evaluate the effectiveness of our method, we utilize two different pre-trained language-specific RoBERTa transformers (Liu et al., 2019) and test their *zero-shot stitching performance on languages that were not seen during training*. Specifically, we evaluate our method using English and Spanish languages with `PlanTL-GOB-ES/roberta-base-bne` and `roberta-base`, respectively. Similar to the implementation details in word embeddings discussed in Section A.2, we begin by choosing 15 random parallel anchors as seed and then select an additional 285 random anchors for the Spanish space. We then apply our optimization method to discover the remaining 285 parallel anchors for the English space. Next, the absolute embeddings of each space are converted to relative representations using the 300 optimized parallel anchors. Table 6 presents the *Cross-lingual* zero-shot stitching performance of our approach, demonstrating its efficacy in learning to solve a downstream task on a specific language or transformer and making accurate predictions while relying on the discovered anchors.

Tools & Technologies We use the following tools in all the experiments presented in this work:

- *PyTorch Lightning*, to ensure reproducible results while also getting a clean and modular codebase;



	Source	Target	Jaccard ↑	MRR ↑	Cosine ↑
GT	FT	FT	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		W2V	0.34 ± 0.01	0.94 ± 0.00	0.86 ± 0.00
	W2V	FT	0.39 ± 0.00	0.98 ± 0.00	0.86 ± 0.00
		W2V	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Seed	FT	FT	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		W2V	0.06 ± 0.01	0.11 ± 0.01	0.85 ± 0.01
	W2V	FT	0.06 ± 0.01	0.15 ± 0.02	0.85 ± 0.01
		W2V	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
AO	FT	FT	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		W2V	0.49 ± 0.00	0.98 ± 0.00	0.93 ± 0.00
	W2V	FT	0.50 ± 0.00	0.99 ± 0.00	0.93 ± 0.00
		W2V	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

Table 5: Corresponding results to those reported in Table 4, illustrating the performance of the model when optimizing the anchors in the other latent space. Qualitative (*left*) and quantitative (*right*) comparisons of the three methods when optimizing the `FastText` space. All metrics are calculated with $K = 10$ averaged over 20k words across five random seeds.

Table 6: Complete results for the selected experiments are reported in Table 2. Cross-lingual zero-shot stitching performance evaluation. The table reports the mean weighted F1 and MAE on Amazon Reviews fine-grained dataset across five random seeds.

		GT		Seed		AO	
Decoder	Encoder	Fscore	MAE	Fscore	MAE	Fscore	MAE
en	en	0.64 ± 0.01	0.43 ± 0.01	0.50 ± 0.01	0.69 ± 0.01	0.62 ± 0.01	0.44 ± 0.01
	es	0.51 ± 0.01	0.67 ± 0.02	0.44 ± 0.01	0.80 ± 0.01	0.48 ± 0.01	0.70 ± 0.02
es	en	0.50 ± 0.02	0.72 ± 0.04	0.41 ± 0.01	0.92 ± 0.02	0.46 ± 0.01	0.76 ± 0.02
	es	0.60 ± 0.01	0.45 ± 0.01	0.48 ± 0.01	0.70 ± 0.01	0.61 ± 0.01	0.44 ± 0.01

- *GeoTorch* Casado (2019), to constrain optimized anchor vectors to have unit norm;
- *Fast, Memory-Efficient Approximate Wasserstein Distances*, to optimize anchor vectors;
- *Transformers by HuggingFace*, to get ready-to-use transformers for both text and images;
- *Datasets by HuggingFace*, to access most of the NLP datasets and CIFAR10 (Krizhevsky et al., 2009) for CV;
- *DVC* (Kuprieiev et al., 2022), for data versioning;

A.3 Additional Experiments

Building upon the methodology presented in the word embeddings experiment introduced in Section 3 and detailed in Appendix A.2, we generalize the retrieval results from the NLP to the Vision domain. To achieve this, we first extract $\approx 20K$ images from `CIFAR-10`. We then encode these images using two variants of the `VIT` transformer model: the `VIT_base_patch16` model, which is pre-trained on `JFT-300M` (Sun et al., 2017) and `ImageNet` (Deng et al., 2009), and the `VIT_small_patch16` model, which is pre-trained solely on `ImageNet`. The two models have respective encoding dimensions of 768 and 384. We follow the same experimental setting, comparing our model against two different baselines (*GT* and *Seed* methods), and we evaluate the performance with *Jaccard*, *MRR* and *Cosine* metrics. Results are reported in Table 7.

Table 7: Generalization of the results described in Section 3, from word embeddings to images using the `CIFAR-10` dataset. The table reports the mean results for each metric and its standard deviation across five different random seeds.

Mode	Type	Source	Target	Jaccard ↑	MRR ↑	Cosine ↑
GT	Absolute	`ViT-base`	`ViT-base`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			`ViT-small`	-	-	-
		`ViT-small`	`ViT-base`	-	-	-
			`ViT-small`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	`ViT-base`	`ViT-base`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			`ViT-small`	0.11 ± 0.00	0.27 ± 0.01	0.97 ± 0.00
		`ViT-small`	`ViT-base`	0.10 ± 0.00	0.28 ± 0.01	0.97 ± 0.00
			`ViT-small`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Seed	Absolute	`ViT-base`	`ViT-base`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			`ViT-small`	-	-	-
		`ViT-small`	`ViT-base`	-	-	-
			`ViT-small`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	`ViT-base`	`ViT-base`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			`ViT-small`	0.03 ± 0.00	0.03 ± 0.01	0.97 ± 0.00
		`ViT-small`	`ViT-base`	0.03 ± 0.00	0.04 ± 0.01	0.96 ± 0.00
			`ViT-small`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
AO	Absolute	`ViT-base`	`ViT-base`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			`ViT-small`	-	-	-
		`ViT-small`	`ViT-base`	-	-	-
			`ViT-small`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	`ViT-base`	`ViT-base`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			`ViT-small`	0.10 ± 0.01	0.23 ± 0.03	0.97 ± 0.00
		`ViT-small`	`ViT-base`	0.10 ± 0.00	0.28 ± 0.01	0.97 ± 0.00
			`ViT-small`	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00