arXiv:2303.00721v2 [cs.LG] 1 Jun 2023

为相对表示引导并行锚点

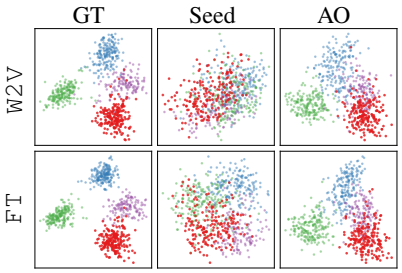
Irene Cannistraci[1] **Luca Moschella**[1] **Valentino Maiorca**[1] **Marco Fumero**[1] **Antonio Norelli**[1] **Emanuele Rodol`a**[1]
[1]罗马智慧大学

摘要

将相对表示用于潜在嵌入，有望在广泛的应用中实现潜在空间通信和零样本模型拼接。然而，相对表示依赖一定数量的平行锚点作为输入，而在某些场景中，获取这些锚点可能并不现实。为克服这一局限，我们提出一种基于优化的方法，从有限的已知集合（种子）中发现新的平行锚点。我们的方法可用于发现不同领域之间的语义对应关系、对齐其相对空间，并在多项任务中取得有竞争力的结果。

1 引言

过去几年，若干研究指出，无论神经网络的架构、任务或域如何，它们通常都能学习到可比的表征（Li 等人，2016；Kornblith 等人，2019；Vuli´c 等人，2020）。顺应这一趋势，Moschella et al. (2023) 提出了相对表征的概念，旨在生成可比的潜在空间，并实现零样本拼接，从而无需额外训练即可处理新的、未见过的任务。该方法通过相对于一组训练样本的潜在相似度来表征每个数据样本，这组训练样本称为锚点。这一过程将绝对参考系转换为由锚点定义的相对坐标系。为了支持多模态学习等任务，该方法要求两个数据域的锚点之间存在语义联系，这些锚点称为平行锚点。这种对应关系必须作为输入提供，它能够实现域比较，并连接各自的潜在空间（Norelli et al., 2022）。然而，在特定应用中，获取足够数量的平行锚点可能颇具挑战，甚至不可能，从而妨碍相对表示的应用。我们关注的情形是可用的平行锚点极少，这些锚点称为种子；我们的目标是通过锚点优化（AO）过程扩充这一初始集合。我们的方法在自然语言处理和视觉领域取得了有竞争力的性能，同时将所需平行锚点的数量减少了一个数量级。



	Src	Tgt	Jaccard ↑	平均倒数排名 ↑	余弦相似度 ↑
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

表1：AO 方法在检索任务中的定性（左）和定量（右）评估。

2 方法

给定两个域 $\mathcal{X}$ 和 $\mathcal{Y}$ 以及相应的已学习嵌入函数$E_{\mathcal{X}}$：$\mathcal{X} \to \mathbb{R}^n$ 和 $E_{\mathcal{Y}}$：$\mathcal{Y} \to \mathbb{R}^m$，其中 $n \neq m$可能成立。给定两组锚点 $\mathcal{A}_{\mathcal{X}} \subset \mathcal{X}$和

$\mathcal{A}_\mathcal{Y} \subset \mathcal{Y}$，我们将语义对应的一组配对 $\mathcal{A}_p \subseteq \mathcal{A}_\mathcal{X} \times \mathcal{A}_\mathcal{Y}$ 定义为平行锚点，例如 Norelli et al. (2022) 中的图像和说明文字。样本 $x \in \mathcal{X}$（$\mathcal{Y}$同理）的相对表示按如下方式计算：$rr(x, \mathcal{A}_\mathcal{X}) = E_\mathcal{X}(x)\mathbf{A}_\mathcal{X}^T$，其中 $\mathbf{A}_\mathcal{X} = \bigoplus_{a \in \mathcal{A}_\mathcal{X}} E_\mathcal{X}(a)$，$\bigoplus$ 表示按行拼接运算符。我们假设所有嵌入都已重新缩放至单位范数，即 $\forall x \; ||E(x)|| = 1$。根据 Moschella et al. (2023) 的设定，这对应于选择余弦相似度作为相似度函数。

在这项工作中，我们提出了一种优化过程，可将所需平行锚点的数量减少一个数量级。我们的方法不需要完整了解 $\mathcal{A}_p$，只需了解少量初始种子锚点，记作 $\mathcal{L} = \mathcal{L}_\mathcal{X} \times \mathcal{L}_\mathcal{Y} \subseteq \mathcal{A}p$，其中 $|\mathcal{L}| \ll |\mathcal{A}p|$。在对 $\mathcal{A}_\mathcal{Y}$ 毫无~~先~~验知识的情况下，我们通过将已知种子$\mathbf{A}_{\mathcal{L}_\mathcal{Y}} = \bigoplus_{a \in \mathcal{L}_\mathcal{Y}} E_\mathcal{Y}(a)$与 $|\mathcal{A}_p| - |\mathcal{L}|$ 个随机嵌入拼接，来近似$\mathbf{A}_\mathcal{Y} \approx \mathbf{A}\mathcal{Y}$，从而初始化优化过程，即$\mathbf{A}\mathcal{Y} = \widetilde{\mathbf{A}}_{\mathcal{L}_\mathcal{Y}} \oplus \mathbf{N}$，其中$\mathbf{N} \sim \mathcal{N}(0, \mathbf{I})$。我们定义以下目标函数，对$\mathbf{A}\mathcal{Y}$进行优化：

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

其中 $\Pi: \mathcal{Y} \to \mathcal{X}$ 是通过 Sinkhorn（Cuturi，2013）算法在每个优化步骤中估计的对应关系，该算法利用初始种子以及对其余锚点的当前近似：$\Pi = sinkhorn_{(x,y) \in \mathcal{X} \times \mathcal{Y}}(rr(x, \mathcal{A}_\mathcal{X}) \; E_\mathcal{Y}(y) \; \widetilde{\mathbf{A}}_\mathcal{Y}^T)$。收敛后，将 $\widetilde{\mathbf{A}}\mathcal{Y}$离散化为 $\mathcal{A}_\mathcal{Y} \subseteq \mathcal{Y}$，方法是选取 $E_\mathcal{Y}(\mathcal{Y})$ 中距离最近的嵌入。更多细节见附录 A.2。

3 实验

本节评估 AO 方法的有效性。我们使用 15 个锚点来近似 300 个平行锚点，这些锚点在所有下游任务中充当真实值。具体而言，我们将本方法的表现与两种不同的基线方法进行比较：(1) 真实值，即使用本方法旨在通过语义近似的所有锚点；(2) 种子，即仅利用种子锚点。有关实现的更多信息，请参阅附录 A.2 和代码[1]。

如表 1、4、5 和 7 所示，我们的方法能够有效发现自然语言处理和视觉领域中的平行锚点。具体而言，我们探索了不同的词嵌入和预训练基础视觉编码器，并通过检索任务评估所发现锚点的质量。结果表明，在起始锚点数量相同的情况下，我们的方法优于仅依赖这些锚点而不进行优化的方法。此外，我们的结果与使用全部真实值平行锚点所得结果相当或更优。此外，表6表明，我们的方法能够发现不同域之间的平行锚点：即使没有可用的真实值，我们也能找到不同语言中对齐的亚马逊评论。仅使用 15 个分布外（Moschella 等，2023）平行锚点，我们的方法便能实现零样本拼接，从而可以在一种语言上训练分类器，并在另一种语言上进行预测。

表2：跨语言零样本拼接性能评估。

		GT		Seed		AO	
Dec.	Enc.	F分数	MAE	F分数	MAE	F分数	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 结论、未来工作与局限性

本文提出了一种新方法，即使在仅有少量平行锚点的情况下，也能计算出稳健的相对表示。该方法无需先验知识即可扩展数据域之间的语义对应关系，并且仅需少一个数量级的平行锚点，便能取得相当的结果。这种方法对于在语义对应关系知识有限的域之间进行潜在空间通信具有重要意义。未来还需要开展研究，以免除对初始平行种子的需求。

[1]完整可复现的代码库：https://github.com/icannistraci/bootstrapping-ao

致谢

本研究得到 ERC资助（编号 802554，项目“SPECGEO”）和 PRIN 2020 项目（编号 2020TA3K9N，项目“LEGO.AI”）的支持。

代表性不足群体声明

第一作者符合 ICLR 2023 短论文专题的代表性不足群体标准。

参考文献

Richard Antonello、Javier S Turek、Vy Vo 和 Alexander Huth。语言表征空间中的低维结构反映在脑反应中。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan（编），神经信息处理系统进展，第 34 卷，第 8332–8344 页。Curran Associates, Inc.，2021 年。网址：https://proceedings.neurips.cc/paper/2021/file/464074179972cbbd75a39abc6954cd12-Paper.pdf。

Yamini Bansal、Preetum Nakkiran 和 Boaz Barak。重新审视模型拼接以比较神经表征。载于 Marc’Aurelio Ranzato、Alina Beygelzimer、Yann N. Dauphin、Percy Liang 和 Jennifer Wortman Vaughan（编），神经信息处理系统进展第*34*卷：神经信息处理系统*2021*年度会议，*NeurIPS 2021*，*2021*年*12*月*6*日至*14*日，线上，第225–236页，2021年。网址：https://proceedings.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html。

Serguei Barannikov、Ilya Trofimov、Nikita Balabin 和 Evgeny Burnaev。表征拓扑发散：一种比较神经网络表征的方法。*arXiv*预印本，abs/2201.00058，2022年。网址：https://arxiv.org/abs/2201.00058。

Federico Bianchi、Jacopo Tagliabue、Bingqing Yu、Luca Bigon 和 Ciro Greco。《奇妙的嵌入及其对齐方法：多店铺场景中的零样本推理》。*arXiv* 预印本，abs/2007.14906，2020年。网址：https://arxiv.org/abs/2007.14906。

Niccolo Biondi、Federico Pernici、Matteo Bruni 和 Alberto Del Bimbo。《Cores：通过平稳性实现兼容表征》。*arXiv* 预印本，abs/2111.07632，2021年。网址：https://arxiv.org/abs/2111.07632。

Piotr Bojanowski、Edouard Grave、Armand Joulin 和 Tomas Mikolov。《利用子词信息丰富词向量》。计算语言学协会汇刊，5：135–146，2017年。doi：10.1162/tacl a 00051。网址：https://aclanthology.org/Q17-1010。

Lisa Bonheme 和 Marek Grzes。变分自编码器如何学习？来自表征相似性的洞见。*arXiv* 预印本，abs/2205.08399，2022年。网址：https://arxiv.org/abs/ 2205.08399。

Mario Lezcano Casado。流形上基于梯度的优化的平凡化。载于 Hanna M. Wallach、Hugo Larochelle、Alina Beygelzimer、Florence d’Alch´e-Buc、Emily B. Fox 和 Roman Garnett（编），神经信息处理系统进展第*32*卷：*2019*年神经信息处理系统大会年会，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，第9154–9164页，2019年。网址：https://proceedings.neurips.cc/paper/2019/hash/1b33d16fc562464579b7199ca3114982-Abstract.html。

´Adri´an Csisz´arik, P´eter K˝or¨osi-Szab´o, Akos K. Matszangosz, Gergely Papp 和 D´aniel Varga。《神经网络表征的相似性与匹配》，2021 年。网址：https://arxiv. org/abs/2110.14633。

Marco Cuturi。Sinkhorn距离：最优传输的高速计算。载于Christopher J. C. Burges，L´eon Bottou、Zoubin Ghahramani和Kilian Q. Weinberger（编），神经信息处理系统进展*26*：*2013*年第*27*届神经信息处理系统大会。会议论文集，会议于*2013*年*12*月*5*日至*8*日在太浩湖、内华达州、美国举行，第2292–2300页，2013年。
URLhttps://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html。

Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li和Fei-Fei Li。ImageNet：大型分层图像数据库。载于*2009*年*IEEE*计算机学会计算机视觉与模式识别会议（*CVPR 2009*），*2009*年*6*月*20*日至*25*日，美国佛罗里达州迈阿密，第248–255页。IEEE计算机学会，2009年。doi：10.1109/CVPR.2009.5206848。URL https://doi.org/10. 1109/CVPR.2009.5206848。

Michael Gygli、Jasper Uijlings 和 Vittorio Ferrari。通过学习兼容表征，迈向可复用网络组件。*AAAI*，35(9):7620–7629，2021年。

Phillip Keung、Yichao Lu、Gy¨orgy Szarvas 和 Noah A. Smith。多语言亚马逊评论语料库。载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第 4563–4568 页，线上举办，2020年。计算语言学协会。doi：10.18653/v1/2020.emnlp-main.369。URL https://aclanthology.org/2020. emnlp-main.369。

Simon Kornblith、Mohammad Norouzi、Honglak Lee 和 Geoffrey E. Hinton。重新审视神经网络表征的相似性。载于 Kamalika Chaudhuri 和 Ruslan Salakhutdinov（编），第 *36* 届国际机器学习会议（*ICML 2019*），*2019* 年 *6* 月 *9* 日至 *15* 日，美国加利福尼亚州长滩，第 97 卷机器学习研究论文集，第 3519–3529 页。PMLR，2019年。URL http://proceedings.mlr.press/v97/kornblith19a.html。

Alex Krizhevsky、Geoffrey Hinton 等人。通过微型图像学习多层特征。2009年。

Ruslan Kuprieiev、skshetry、Dmitry Petrov、Paweł Redzy´nski、Peter Rowlands、Casper da Costa- Luis、Alexander Schepanovski、Ivan Shcheklein、Batuhan Taskaya、Gao、Jorge Orpinel、David de la Iglesia Castro、F´abio Santos、Aman Sharma、Dave Berenbaum、Zhanibek、Dani Hodovic、daniele、Nikita Kodenko、Andrew Grigorev、Earl、Nabanita Dash、George Vyshnya、Ronan Lamy、maykulkarni、Max Hora、Vera 和 Sanidhya Mangal。《DVC：数据版本控制——数据和模型的Git》，2022年。网址：https://doi.org/10.5281/zenodo.7083378。

Guillaume Lample、Alexis Conneau、Marc'Aurelio Ranzato、Ludovic Denoyer 和 Herv´e J´egou。无需平行数据的词语翻译。载于第六届国际学习表征会议（*ICLR 2018*），温哥华，不列颠哥伦比亚省，加拿大，*2018*年*4*月*30*日至*5*月*3*日，会议论文集。OpenReview.net，2018年。网址：https://openreview.net/forum?id=H196sainb。

Karel Lenc 和 Andrea Vedaldi。通过测量图像表征的等变性和等价性来理解图像表征。载于*IEEE*计算机视觉与模式识别会议（*CVPR 2015*），美国马萨诸塞州波士顿，*2015*年*6*月*7*日至*12*日，第991–999页。IEEE计算机学会，2015年。doi：10. 1109/CVPR.2015.7298701。URL https://doi.org/10.1109/CVPR.2015.7298701。

Yixuan Li、Jason Yosinski、Jeff Clune、Hod Lipson 和 John E. Hopcroft。趋同学习：不同的神经网络会学习相同的表征吗？载于 Yoshua Bengio 和 Yann LeCun（编），第四届国际学习表征会议（*ICLR 2016*），波多黎各圣胡安，*2016*年*5*月*2*日至*4*日，会议论文集，2016年。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 和 Veselin Stoyanov。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*arXiv* 预印本，abs/1907.11692，2019 年。网址：https://arxiv.org/abs/1907. 11692。

Tomas Mikolov、Kai Chen、Greg Corrado 和 Jeffrey Dean。高效估计向量空间中的词表示。*arXiv* 预印本，abs/1301.3781，2013a 年。网址：https://arxiv.org/ abs/1301.3781。

Tom´as Mikolov、Quoc V. Le 和 Ilya Sutskever。利用语言之间的相似性进行机器翻译。*CoRR*，abs/1309.4168，2013b。网址：http://arxiv.org/abs/1309. 4168。

Ari S. Morcos、Maithra Raghu 和 Samy Bengio。通过典型相关分析洞察神经网络中的表征相似性。载于 Samy Bengio、Hanna M. Wallach、Hugo Larochelle、Kristen Grauman、Nicol`o Cesa-Bianchi 和 Roman Garnett（编），神经信息处理系统进展第*31*卷：神经信息处理系统年度会议 *2018*，*NeurIPS 2018*，*2018*年*12*月*3*日至*8*日，蒙特利尔´，加拿大，第5732–5741页，2018年。网址：https://proceedings.neurips.cc/paper/2018/hash/a7a3d70c6d17a73140918996d03c014f-Abstract.html。

Luca Moschella、Valentino Maiorca、Marco Fumero、Antonio Norelli、Francesco Locatello 和 Emanuele Rodol`a。相对表示支持零样本潜在空间通信。在国际学习表征会议，2023 年。网址：https://openreview. net/forum?id=SrC-nwieGJ。

Antonio Norelli、Marco Fumero、Valentino Maiorca、Luca Moschella、Emanuele Rodol`a 和 Francesco Locatello。Asif：耦合数据将单模态模型转变为无需训练的多模态模型。*arXiv* 预印本，abs/2210.01738，2022 年。网址：https://arxiv.org/abs/2210.01738。

Chen Sun、Abhinav Shrivastava、Saurabh Singh 和 Abhinav Gupta。重新审视深度学习时代数据的非理想有效性。在*IEEE*国际计算机视觉会议，*ICCV 2017*，意大利威尼斯，*2017*年*10*月*22*日至*29*日，第 843–852 页。IEEE计算机学会，2017 年。doi：10.1109/ICCV.2017.97。网址：https://doi.org/10.1109/ICCV.2017.97。

Anton Tsitsulin、Marina Munkhoeva、Davide Mottin、Panagiotis Karras、Alexander M. Bronstein、Ivan V. Oseledets 和 Emmanuel M¨uller。《数据的形状：数据分布的内在距离》。载于第八届国际学习表征会议，*ICLR 2020*，亚的斯亚贝巴，埃塞俄比亚，*2020*年*4*月*26*日至*30*日。OpenReview.net，2020年。网址：https://openreview.net/ forum?id=HyebplHYwB。

Ivan Vuli´c、Sebastian Ruder 和 Anders Søgaard。所有优质词向量空间都同构吗？载于*2020*年自然语言处理实证方法会议论文集 *(EMNLP)*，第 3178–3192 页，线上，2020年。计算语言学协会。doi：10.18653/v1/2020.emnlp-main.257。网址：https://aclanthology.org/2020.emnlp-main.257。

A 附录

A.1 相关工作

近年来，众多研究（Lenc & Vedaldi, 2015; Mikolov 等人, 2013b; Li 等人, 2016; Lample 等人, 2018; Morcos 等人, 2018; Tsitsulin 等人, 2020; Kornblith 等人, 2019; Vuli´c 等人, 2020; Antonello 等人, 2021; Bonheme & Grzes, 2022; Barannikov 等人, 2022; Norelli 等人, 2022）发现，当神经网络使用语义相似的数据进行训练时，无论其架构、任务或域如何，往往都会学习到可比较的表征。这一观察可用于实现多种应用，例如模型拼接（Lenc & Vedaldi, 2015; Bansal 等人, 2021; Csisz´arik 等人, 2021; Gygli 等人, 2021; Biondi 等人, 2021; Bianchi 等人, 2020）、潜在模型比较、监督等。具体而言，Moschella et al. (2023) 提出了相对表征框架，旨在统一从语义相似的数据中学习到的表征。相对表示展现出促进潜在嵌入之间交流，以及在各种应用中实现零样本拼接的潜力；它们依靠平行锚点来关联不同的域。我们的工作旨在最大限度地减少潜在交流所需的显式监督，将对平行锚点的依赖降至最低，并自动扩展所提供的域间语义对应关系。

A.2 实现细节

本节进一步介绍优化过程和实验的详细信息。

优化方法 算法 1 概述了第 2 节所述优化过程的伪代码，而表3列出了超参数的详细信息。该方法初始化 $\mathbf{A}_{\mathcal{Y}}$，并通过迭代对其进行优化。在每一步中，Sinkhorn 算法都会粗略估计两个相对空间之间的置换。优化过程中最小化的损失函数是均方误差，同时特别注重确保优化后的参数 $\mathbf{A}_{\mathcal{Y}}$使用 Casado (2019) 的方法满足单位范数约束。这不仅确保了优化的有效性，也缩小了搜索空间。

算法 1锚点优化

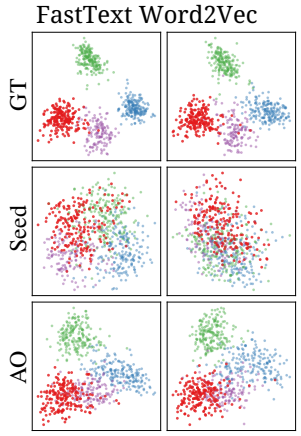
~1 $\mathcal{Y} = \mathcal{L}_{\mathcal{Y}} \oplus \in \mathcal{N}\ 0,$ $| \quad | = |\mathcal{A}_p| - |\mathcal{L}|$：初始化 A A N, 其中 N 2 (I) 且 N：计算 中样本的相对表征，得到 $\mathrm{R}_{\mathcal{X}} = \bigoplus_{x \in \mathcal{X}} rr(x, \mathcal{A}_{\mathcal{X}})$ 3：执行 K 步 4：计算 $\mathcal{Y}$ 中样本的相对表征，得到 $\mathrm{R}_{\mathcal{Y}} = \bigoplus_{y \in \mathcal{Y}} E_{\mathcal{Y}}(y)$ ~$\mathrm{A}_{\mathcal{Y}}^{T}$5：使用 $\Pi = sinkhorn(\mathrm{R}_{\mathcal{X}}, \mathrm{R}\mathcal{Y})$ 估计 R$\mathcal{Y}$ 和 $\mathrm{R}_{\mathcal{X}}$ 之间的置换 6：根据 Π 对 R$\mathcal{Y}$ 进行置换 7：计算误差 $MSE(\mathrm{R}_{\mathcal{X}}, \mathrm{R}\mathcal{Y})$ ~ ~8：优化 A$\mathcal{Y}$ 以最小化误差，同时满足约束 $||a||_2 = 1\ \forall a \in$ A$\mathcal{Y}$9：结束循环~10：返回 A$\mathcal{Y}$ 在 $E_{\mathcal{Y}}(\mathcal{Y})$ 中的最近邻

检索任务 我们选择了两个使用不同数据训练、但词汇表部分重叠的英语词嵌入，并从中提取 ≈ 20 K 个词：FastText（Bojanowski 等，2017）和 Word2Vec（Mikolov 等，2013a）。为测试 AO 方法，我们选择 15 个种子锚点，并对两个嵌入空间进行打乱，以破坏它们之间的对应关系。然后，我们为其中一个空间选择另外 285 个随机锚点，同时使用我们的优化方法，在另一个空间中发现与之关联的 285 个平行锚点。接下来，利用这 300 个优化后的平行锚点，将每个空间中的绝对嵌入转换为相对表征。对于每个词w，我们考察其在源空间和目标空间中对应的编码 x 和 y，并通过检索任务验证其质量。为便于与相对表示基线（Moschella 等，2023）进行比较，我们采用相同的评估指标：（i）*Jaccard*：源空间和目标空间中 x 的词邻居集合之间的离散 Jaccard 相似度；（ii）平均倒数排名（*MRR*）：衡量 w 在目标空间中 x 的前 k 个邻居中的（倒数）排名；（iii）余弦相似度：衡量 x 与 y 之间的余弦相似度。*GT* 方法和种子 方法的结果

表3：检索和零样本拼接任务中 AO 方法的超参数。

超参数	检索	零样本拼接
随机种子	0, 1, 2, 3, 4	0, 1, 2, 3, 4
待近似锚点数	300	300
种子锚点数	15	15
优化步数	250	125
学习率	0.02	0.05
优化器	Adam	Adam
Loss	MSE	MSE
Sinkhorn 正则化系数	1e-4	1e-4
Sinkhorn 停止误差	1e-5	1e-5
Sinkhorn 步数	1	1

分别通过使用给定的全部 300 个锚点（我们的方法旨在对其进行语义近似）以及仅使用这 15 个种子锚点获得。



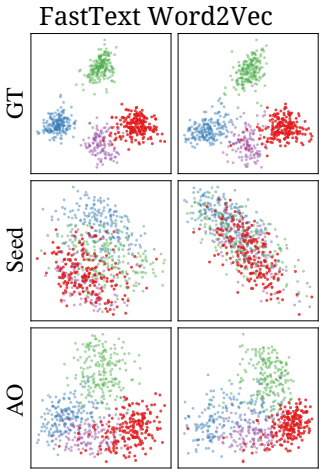
	源	目标	Jaccard ↑	平均倒数排名 ↑	余弦相似度 ↑
GT	FT	FT	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		W2V	0.34 ± 0.01	0.94 ± 0.00	0.86 ± 0.00
	W2V	FT	0.39 ± 0.00	0.98 ± 0.00	0.86 ± 0.00
		W2V	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Seed	FT	FT	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		W2V	0.06 ± 0.01	0.11 ± 0.01	0.85 ± 0.01
	W2V	FT	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

表4：所选实验的完整结果见表1。优化 Word2Vec 空间时三种方法的定性（左侧）和定量（右侧）比较。所有指标均以 $K = 10$为依据，在五个随机种子下对 2 万个词取平均值计算。

零样本拼接任务 我们研究多语言亚马逊评论数据集（Keung 等人，2020年）上的跨语言文本分类任务，以展示我们方法的实际应用。我们采用该任务的细粒度表述，目标是根据评论预测星级评分（即预测五个类别），并使用 F分数和平均绝对误差（平均绝对误差）指标衡量性能。为了评估我们方法的有效性，我们使用两种不同的预训练特定语言 RoBERTa Transformer（Liu 等人，2019年），并测试它们在训练期间未见过的语言上的零样本拼接性能。具体而言，我们使用英语和西班牙语，以及 PlanTL-GOB-ES/roberta-base-bne 和 roberta-base，分别评估我们的方法。与 A.2 节中讨论的词嵌入实现细节类似，我们首先选择 15 个随机平行锚点作为种子，然后为西班牙语空间再选择 285 个随机锚点。接着，我们应用优化方法，为英语空间找出其余 285 个平行锚点。随后，使用这 300 个优化后的平行锚点，将每个空间中的绝对嵌入转换为相对表示。表6展示了我们方法的跨语言 零样本拼接性能，证明该方法能够有效地学习如何在特定语言或 Transformer 上解决下游任务，并依靠发现的锚点做出准确预测。

工具与技术 我们在本文所呈现的所有实验中使用以下工具 rk:

- *PyTorch Lightning*，以确保结果可复现，同时保持代码库整洁且模块化；



	源	目标	Jaccard ↑	平均倒数排名 ↑	余弦相似度 ↑
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

表5：与表4所报告结果相对应的结果，展示了在另一个潜在空间中优化锚点时模型的表现。在优化 FastText 空间时，对三种方法进行定性（左）和定量（右）比较。所有指标均使用 $K = 10$ 在五个随机种子下对 2 万个词取平均值计算。

表6：所选实验的完整结果见表2。跨语言零样本拼接性能评估。该表报告了亚马逊评论细粒度数据集在五个随机种子下的加权 F1 均值和平均绝对误差。

		GT		Seed		AO	
解码器	编码器	F分数	MAE	F分数	MAE	F分数	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年），用于约束优化后的锚点向量具有单位范数；
- 快速、内存高效的近似 *Wasserstein* 距离，用于优化锚点向量；
- *HuggingFace* 的 *Transformers*，用于获取可直接使用的文本和图像 Transformers；
- *HuggingFace* 的 *Datasets*，用于访问大多数自然语言处理数据集，以及用于计算机视觉的 CIFAR10（Krizhevsky 等，2009年）；
- *DVC*（Kuprieiev 等，2022年），用于数据版本管理；

A.3 其他实验

基于第 3 节介绍并在附录 A.2 中详述的词嵌入实验方法，我们将自然语言处理领域的检索结果推广到视觉领域。为此，我们首先从 CIFAR-10 中提取 $\approx 20K$ 图像。随后，我们使用 ViT Transformer 模型的两个变体对这些图像进行编码：ViT base patch16 模型，该模型在 JFT-300M（Sun 等，2017）和 ImageNet（Deng 等，2009）上进行了预训练；以及 ViT small patch16 模型，该模型仅在 ImageNet 上进行了预训练。这两个模型的编码维度分别为 768 和 384。我们采用相同的实验设置，将我们的方法与两种不同的基线方法（*GT* 和种子 方法）进行比较，并使用*Jaccard*、平均倒数排名 和余弦相似度 指标评估性能。结果见表 7。

表 7：将第3节所述结果从词嵌入泛化到使用 CIFAR-10 数据集的图像。表中报告了各项指标的平均结果及其在五个不同随机种子下的标准差。

Mode	Type	源	目标	Jaccard ↑	平均倒数排名 ↑	余弦相似度 ↑
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