Contrastive Language-Image Pre-training for the Italian Language

Bianchi, Federico; Attanasio, Giuseppe; Pisoni, Raphael; Terragni, Silvia; Sarti, Gabriele; Balestri, Dario

Published in:
Proceedings of the 9th Italian Conference on Computational Linguistics

Contrastive Language–Image Pre-training for the Italian Language

Federico Bianchi[1], Giuseppe Attanasio[2], Raphael Pisoni[3], Silvia Terragni[4], Gabriele Sarti[5] and Dario Balestri[3]

[1]Stanford University, California, USA
[2]Bocconi University, Milan, Italy
[3]Independent Researcher
[4]Telepathy Labs, Zürich, Switzerland
[5]University of Groningen, The Netherlands

Abstract

Recently, multi-modal systems such as CLIP (Contrastive Language–Image Pre-training) were introduced to represent images and texts jointly in the same embedding space. These models are trained on massive amounts of image-caption pairs and show impressive performance on zero-shot image classification. However, their usage is limited to English due to their training data. Training the same model for different languages is non-trivial since the amount of natural data in those might not be sufficient, and automatic translations of original captions might not have sufficient quality, harming performance. In this paper, we present the first CLIP model for the Italian Language (CLIP-Italian), trained on more than 1.4 million image-text pairs. Results show that CLIP-Italian outperforms a multilingual CLIP model on image retrieval and zero-shot classification tasks for the Italian language.[1]

Sistemi multimodali come CLIP (Contrastive Language-Image Pre-training) sono stati proposti di recente al fine di ottenere rappresentazioni di immagini e testo in uno spazio latente condiviso. Questi modelli sono allenati su enormi quantità di immagini associate alle loro didascalie, e dimostrano abilità eccellenti nell'effettuare classificazioni "zero-shot". Ciononostante, il loro utilizzo è limitato all'inglese, la lingua utilizzata durante il loro addestramento. Ottenere modelli del genere per altre lingue non è cosa da poco, poiché la quantità di dati a disposizione per queste lingue potrebbe non essere sufficiente e la traduzione automatica delle didascalie inglesi originali potrebbe portare a risultati non soddisfacenti. In questo articolo presentiamo il primo modello CLIP per la lingua italiana (CLIP-Italian), addestrato con più di 1.4 milioni di immagini e rispettive didascalie. I risultati riportati dimostrano l'efficacia di CLIP-Italian per l'estrazione e la classificazione zero-shot in italiano, ottenendo risultati migliori di un modello CLIP multilingue.

Keywords

clip, italian, contrastive, language, image, pretraining, multimodal

1. Introduction

The recent interest in combining different source domains to incorporate broader context in the training process has led to a surge in multi-modal models spanning modalities like text and vision [1] or text and speech [2]. A multi-modal architecture learns by jointly optimizing its parameters on two or more input domains (e.g., images, texts, tabular data, or audio signals), with a cost function that may vary depending on the task.

Contrastive Language–Image Pre-training (CLIP) [1] is a multi-modal model for joint learning image and text representations. CLIP learns to pair visual concepts with descriptions in natural language by leveraging a contrastive loss that pushes images and their respective captions closer in an embedding space. CLIP is trained on a large-scale dataset of images and their corresponding captions. The dataset used in the seminal paper contains 400 million images collected from the web. In recent years, there have been many successful domain-specific implementations of CLIP [3, 4, 5, 6, 7, *inter alia*].

While the model shows impressive zero-shot performance across various supervised tasks, its capabilities are bounded to the language the model is trained in, i.e., English. Despite the ongoing efforts on training multilingual variants of CLIP, different works have shown that multilingual models often do not achieve the same level of performance as language-specific ones [8, 9, 10].

In this paper, we describe how to fine-tune a specialized version of CLIP in a language different than English, i.e., Italian. We dub this model CLIP-Italian.[1] Crucially, we collect for the task a dataset of 1.4M high-quality text-image pairs for Italian, the largest collection of this kind to date. We release our best-performing checkpoint, the modeling and training code, a CometML report with training longs and metrics, and a live demo to showcase

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

giuseppe.attanasio3@unibocconi.it (G. Attanasio)

[1]While Italian was selected for this study, the approach presented in this paper can be generalized to other languages and domains without loss of generality.

```
# using BERT and ViT to encode raw images and texts
encoded_images = clip.image_encoder(images)
encoded_texts = clip.text_encoder(images)

# nomalize the projections
embedded_images = l2_normalization(clip.image_projection(encoded_images))
embedded_texts = l2_normalization(clip.text_projection(encoded_texts))

logits = np.dot(embedded_images, embedded_texts.T) * logit_scale

labels = np.arange(n)  # correct image-text match is on the main diagonal
loss_images = cross_entropy_loss(logits, labels, axis=0)
loss_texts = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_images + loss_texts) / 2
```

Figure 1: Numpy-like pseudo code that describes the CLIP-Italian loss.

CLIP-Italian capabilities and streamline testing.[2]

Contributions. We create the largest publicly available multi-modal dataset for the Italian language. We use this dataset to train and release the first CLIP image-text model for the Italian language. We show that this model performs better than its multilingual counterpart in two well-established multi-modal tasks: image retrieval and zero-shot image classification. Moreover, we release the model checkpoint, code, and an online demo to showcase CLIP-Italian capabilities.

2. Contrastive Language–Image Pre-training

CLIP is trained to put images and captions in close positions in the vector space. Therefore, the model is taught to associate visual concepts and their natural language descriptions.

CLIP's architecture consists of two distinct encoders, one for images and one for texts. At training time, all images and texts in a mini-batch are each projected to a 512-dimensional space. Next, vector similarities are computed for each image-text pair, and cross-entropy loss is applied. Finally, the average loss along the image and text dimensions is used to update model parameters. The loss is used to align the two 512-dimensional projection spaces. Figure 1 briefly summarizes how the contrastive loss is computed in CLIP. We refer the reader to [1] for additional details.

After training, CLIP can be used without further training for a variety of different tasks. Since images and texts are embedded in the same space, CLIP embeddings can be used for zero-shot text-based image retrieval and zero-shot image classification by looking at the similarities between available texts and images.

CLIP-Italian differs from the original CLIP in that encoders are not trained from scratch. We continue training from checkpoints of other pre-trained models. This approach allows us to leverage pre-training knowledge of existing models and remap it to new lexical items to create CLIP-Italian. We extensively cover training details in Section 4.

3. Datasets

We describe the four data sources we used to train our CLIP-Italian model.

- WIT [11] is a multilingual image-caption dataset collected from Wikipedia. We pre-process and extract the Italian subset, selecting the Reference Description captions as captions of interest. While several possible captions are available, we select those described as the most topical and highest-quality captions in the original paper.
- MSCOCO-IT [12].[3] The captions of this dataset come from the original MSCOCO dataset [13] and are translated with Microsoft Translator. The 2017 MSCOCO training set contains more than 100K images. More than one caption is available for each image.
- Conceptual Captions (CC) [14].[4] In this dataset, there are more than 3 million image-caption pairs, collected from the web. All images with available URLs were downloaded, and their captions

[2] Model: https://huggingface.co/clip-italian/clip-italian, Logs: https://www.comet.ml/g8a9/clip-italian/reports/clip-italian-training-metrics, Demo: https://huggingface.co/spaces/clip-italian/clip-italian-demo

[3] https://github.com/crux82/mscoco-it

[4] https://github.com/google-research-datasets/conceptual-captions

Dataset	Ratio	Captions
WIT	38%	525,950
MSCOCO-IT	8%	116,195
CC	52%	712,890
ILPOST	2%	29,055
Total		1,384,090

Table 1
A summary of datasets used in this work with the number of captions collected per dataset.

were translated to Italian using DeepL.[5], totaling roughly 710K captions.

- La Foto del Giorno (ILPOST).[6] This image-caption dataset is collected from *Il Post*, a prominent Italian online newspaper. Starting from early 2011, every day, the editors at Il Post have selected several images picturing the most salient events in the world. Each photo comes along with an original Italian caption. The resulting collection contains almost 30K pairs of images-captions.

3.1. Translations

We used automatic translation to augment the training set due to the low amount of captioned images for Italian compared to the original CLIP training dataset. Instead of relying on open-source translators, we use the proprietary DeepL API to obtain readily available high-quality English captions. While this choice aims to minimize the noise in translated data, we know about the bias (e.g., gender and age) that translation systems introduce during translations [15]. Some of the captions are available in Figure 2.

To assess the translation quality, three native Italian speakers among the authors inspected a sample of 100 translations alongside their original English sources, rating translations with scores between 1 and 4. We adopt the following categorization for the provided scores: 1, the sentence has lost its meaning, or it is not possible to understand it; 2, it is possible to get the idea, but there is something wrong; 3, good, however, a native speaker might complain about some parts of the translation; 4, correct translation.

The average score was 3.78, suggesting that the translations were good on average. We also computed an inter-rater agreement with Gwet's AC1 using ordinal weighting, obtaining a value of 0.858. This value suggests a strong agreement between annotators.

[5] https://www.deepl.com/
[6] https://www.ilpost.it/foto-del-giorno/

3.2. Data Cleaning

Many of the captions in WIT describe encyclopedic facts (e.g., "Roberto Baggio in 1994"). We believe these descriptions will not be helpful in learning a good mapping between images and captions, as most of the information in the description is factual knowledge. To prevent polluting the data with overly specific factual captions, we used Part-Of-Speech (POS) tagging using spacy[7] on the text and removed all the captions that were composed for the 80% or more by proper nouns (around 10% of the total captions for WIT). This simple solution allowed us to retain much of the dataset without introducing noise. Captions like "Dora Riparia", "Anna Maria Mozzoni", "Joey Ramone Place", "Kim Rhodes", "Ralph George Hawtrey" which are proper nouns (`PROPN`) have been removed. For the dataset ILPOST, we used `langdetect`[8] to filter non-Italian captions, resulting in only 2% captions being removed.

4. Training

Our CLIP-Italian model is based on previous pre-trained state-of-the-art models for both the vision and textual parts. We use Vision Transformer (ViT) [16] and BERT-inspired [17] text encoder. We limit the sequence length to 96 tokens and use a local batch size of 128 for each of the 8 TPU cores we used. For the optimization procedure, we used the AdaBelief optimizer [18] with Adaptive Gradient Clipping (AGC) and a Cosine Annealing Schedule [19]. We run training for a maximum of 15 epochs, evaluate at the end of each epoch, and release the checkpoint with the best validation loss.

Data Augmentation Following standard practices in computer vision, we applied several augmentations to the available images. In particular, we used random affine transformations, perspective changes, occasional equalization, and random changes to brightness, contrast, saturation, and hue. Importantly, we made sure to keep hue augmentations limited to allow the model to learn color definitions.

Projection Layers Warmup Since pre-trained checkpoints were used as starting points for both the vision and the text encoders, we found it helpful to warm-up projection layers. To do so, we first train the entire network using frozen vision and text encoders until loss convergence. After this first phase, the rest of the model is unfrozen to perform end-to-end training. We always pick the model with the best evaluation loss, evaluating every 15 epochs.

[7] https://spacy.io/models/it#it_core_news_lg
[8] https://github.com/Mimino666/langdetect

English Caption	Italian Caption
an endless cargo of tanks on a train pulled down tracks in an empty dry landscape	un carico infinito di carri armati su un treno trascinato lungo i binari in un paesaggio secco e vuoto
person walking down the aisle	persona che cammina lungo la navata
popular rides at night at the county fair	giostre popolari di notte alla fiera della contea

Table 2
Examples of automatically translated captions from the Conceptual Captions dataset.

Starting Checkpoints We used an Italian BERT checkpoint[9] as text encoder and the original CLIP vision encoder.[10]

Logits Scaling Both images and texts are then projected to 512-dimensional vectors to which we apply the loss defined in CLIP using logit scaling equal to 20. We empirically observed that logit scaling has a strong positive impact on model performance, suggesting that the embeddings have similar Euclidian norms and that scaling their dot similarities helped the cross entropy.

5. Quantitative Evaluation

To our knowledge, CLIP-Italian is the first multi-modal system explicitly trained for the Italian language. Hence, to provide meaningful comparisons, we compare its performance to an available multilingual CLIP[11] model trained with multilingual knowledge distillation [20].

5.1. Image Retrieval

The image retrieval task is as follows. Given a caption, the task is to retrieve the correct image from a set of available images, where the correct image is the one that is described by the caption. This search can be done by embedding the caption and the images and selecting the nearest neighbors to the caption embedding. We use the MSCOCO-IT validation dataset left out for this purpose during the training procedure, containing a total of 2,000 image-caption pairs.

Metric We compare models on the standard Mean Reciprocal Rank (MRR) retrieval metric. The metric computes the rank assigned to each image to be retrieved (r, where $r = 1$ is best), takes its reciprocal, and averages it across all the dataset samples ($MRR = 1/|D| \cdot \sum_i^{|D|} 1/r_i$). We consider only the first k retrieved images for each sample's contribution. If the target image is not within them, we approximate $1/r_i$ to 0 (MRR@k).

Measure (↑)	CLIP-Italian	mCLIP
MRR@1	**0.3797**	0.2874
MRR@5	**0.5039**	0.3957
MRR@10	**0.5204**	0.4129

Table 3
Results on MSCOCO image retrieval task. Best result in bold.

Table 3 reports the results for the image retrieval task, in terms of MRR@k, where $k \in \{1, 5, 10\}$. CLIP-Italian outscores mCLIP across the board.

5.2. Zero-shot Classification

The zero-shot image classification task replicates the experiment run by Radford et al. [1] on ImageNet. We first used DeepL to translate the image labels in ImageNet automatically. Then we prepend all test set labels with determiners and translate them (e.g., *a cat* is translated into "un gatto") and then prepended with the text "una foto di" (a photo of) as in "una foto di un gatto" (a photo of a cat) to obtain the final caption. This procedure is simpler than the one adopted by Radford et al. [1], where different templates are tested and averaged. Given an input image and the so-generated captions, we generate the embeddings (both for the image and all captions) and compute the similarities, assessing whether the correct image label corresponds to the closest caption in the embedding space.

Metric We compare models on the standard accuracy. Similarly to MRR@k, we consider a "hit" if the predicted class is within the top k retrieved and a "miss" otherwise. Similarly to the image retrieval task, we compute accuracy at k (Accuracy@k) with $k \in \{1, 5, 10\}$.

Table 4 reports the results for the zero-shot classification task. CLIP-Italian outscores mCLIP across the board.

5.3. Discussion

Our results across two tasks confirm that CLIP-Italian is very competitive and outperforms mCLIP on the two

[9] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased
[10] https://huggingface.co/openai/clip-vit-base-patch32
[11] https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1

Measure (↑)	CLIP-Italian	mCLIP
Accuracy@1	**22.11**	20.15
Accuracy@5	**43.69**	36.75
Accuracy@10	**52.55**	42.91

Table 4
Results on ImageNet-1000 classification task. Best result in bold.

Figure 2: Result of the query "due cani sulla neve" (eng: two dogs on the snow) on Unsplash25K.

tasks we have been testing. Note that the performance for zero-shot ImageNet classification of the CLIP-Italian model (trained on 1.4 million image-text pairs) are lower than those shown in Radford et al. [1] (trained on 400 million image-text pairs). However, considering that our results align with those obtained by mCLIP, we think that the quality of the translated image labels most probably impacted the final scores.

6. Qualitative Evaluation

We examine some examples related to the image retrieval task on the Unsplash25K dataset.[12] Figure 2 shows the results of the query "due cani sulla neve" (two dogs on the snow), the model correctly finds the image, combining

[12] https://github.com/unsplash/datasets

Figure 3: Result of the query "una coppia al tramonto" (eng: a couple at the sunset) on Unsplash25K.

the concept of "snow" and the one of "two dogs".[13] We anecdotally find moderate numeracy capabilities during empirical evaluation, with sufficient ability to identify up to three distinct or repeated elements inside images, with a steep drop in coherence when more than three elements are present. Given the likely low number of training points depicting more than three subjects in a scene, we impute this finding to implicit bias in the training set. Figure 3 shows a similar performance for "una coppia al tramonto" (a couple during sunset), where the model could identify two people with sunlight in the background. A similar query, but with a mountain as a background, can be found in Figure 4. Despite the overall good performances, the model is inevitably subject to limitations and biases. For example, Figure 5 shows an image of a tiny hedgehog retrieved using the query "un topolino" (a tiny mouse). We leave a more thorough exploration of biases and stereotypes learned by the CLIP-Italian model to future work.

7. Conclusions

This paper presents the first CLIP model for the Italian language, trained on 1.4 million image-text pairs. The model shows promising zero-shot performance in two well-established tasks, suggesting many possible future applications.

Acknowledgments

This work was possible thanks to Hugging Face and Google which provided the computational resources to train CLIP-Italian. This project has also in part received

[13] Note, however, that compositional understanding in CLIP is limited, see [21]

Figure 4: Result of the query "una coppia in montagna" (eng: a couple in the mountains) on Unsplash25K.

Figure 5: Result of the query "un topolino" (eng: a small mouse) on the Unsplash25K dataset.

funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). We thank Sri Lakshmi for her help on the project.

Ethical Considerations and Limitations

Large-scale models are difficult and costly to train, and important considerations have to be taken into account when developing them [22, 23]. We computed the cost of the different experiments we ran, and we estimated a total of 2,688$ for each TPU used. This result comes from the hourly cost of the TPU (8$) for 14 days; Note that we had access to a second TPU VM for part of the project and that, in this estimate, we are ignoring storage and data transfer costs. Strubell et al. [22] describe how these models can have a substantial environmental impact. As described by Bianchi and Hovy [24], these computational needs are quickly becoming unfeasible for many universities.

Moreover, recent evidence has shown that large-scale multimodal vision and language models exhibit biases in portraying several sociodemographic groups [25, 26, 27, 28, *inter alia*]. Moreover, the datasets on which these models have been trained on often contain harmful content [29]. As we build on pretrained vision and language models, we cannot exclude the presence of such biases. However, we want to point out that our vision and language models were pretrained on different language data. Hence, "concepts" in embedding spaces are not aligned. While we cannot exclude that models pick up biases from *our* training data, starting from unaligned embedding spaces can reduce the risk of unwanted biased associations.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[2] S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: Unsupervised Pre-Training for Speech Recognition, in: Proc. Interspeech 2019, 2019, pp. 3465–3469. doi:10.21437/Interspeech.2019-1873.

[3] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhães, D. Gonçalves, C. Greco, J. Tagliabue, Contrastive language and vision learning of general fashion concepts, Scientific Reports 12 (2022). URL: https://api.semanticscholar.org/CorpusID:253387447.

[4] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, P. Rajpurkar, Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning, Nature Biomedical Engineering 6 (2022) 1399–1406.

[5] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, et al., Large-scale domain-specific pretraining for biomedical vision-language processing, arXiv preprint arXiv:2303.00915 (2023).

[6] G. Chen, L. Hou, Y. Chen, W. Dai, L. Shang, X. Jiang, Q. Liu, J. Pan, W. Wang, mclip: Multilingual clip via cross-lingual transfer, in: Proceedings of the 61st Annual Meeting of the Association for Compu-

tational Linguistics (Volume 1: Long Papers), 2023, pp. 13028–13043.

[7] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, J. Y. Zou, A visual–language foundation model for pathology image analysis using medical twitter, Nature Medicine 29 (2023) 2307 – 2316. URL: https://api.semanticscholar.org/CorpusID:260970273.

[8] D. Nozza, F. Bianchi, D. Hovy, What the MASK? making sense of language-specific bert models, arXiv preprint arXiv:2003.02912 (2020).

[9] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, I. Gurevych, How good is your tokenizer? on the monolingual performance of multilingual language models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3118–3135. URL: https://aclanthology.org/2021.acl-long.243. doi:10.18653/v1/2021.acl-long.243.

[10] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv abs/2203.03759 (2022).

[11] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, M. Najork, Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning, arXiv preprint arXiv:2103.01913 (2021).

[12] A. Scaiella, D. Croce, R. Basili, Large scale datasets for image and video captioning in italian, IJCoL. Italian Journal of Computational Linguistics 5 (2019) 49–60.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[14] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2556–2565. URL: https://www.aclweb.org/anthology/P18-1238. doi:10.18653/v1/P18-1238.

[15] D. Hovy, F. Bianchi, T. Fornaciari, "you sound just like your father" commercial machine translation systems include stylistic biases, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1686–1690. URL: https://www.aclweb.org/anthology/2020.acl-main.154. doi:10.18653/v1/2020.acl-main.154.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[18] J. Zhuang, T. Tang, Y. Ding, S. Tatikonda, N. Dvornek, X. Papademetris, J. S. Duncan, Adabelief optimizer: Adapting stepsizes by the belief in observed gradients, arXiv preprint arXiv:2010.07468 (2020).

[19] I. Loshchilov, F. Hutter, SGDR: stochastic gradient descent with warm restarts, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017. URL: https://openreview.net/forum?id=Skq89Scxx.

[20] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4512–4525. URL: https://www.aclweb.org/anthology/2020.emnlp-main.365. doi:10.18653/v1/2020.emnlp-main.365.

[21] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, J. Zou, When and why vision-language models behave like bags-of-words, and what to do about it?, in: The Eleventh International Conference on Learning Representations, 2022.

[22] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650. URL: https://www.aclweb.org/anthology/P19-1355. doi:10.18653/v1/P19-1355.

[23] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623.

[24] F. Bianchi, D. Hovy, On the gap between adoption and understanding in NLP, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 3895–3901. URL: https://aclanthology.org/2021.findings-acl.340. doi:10.18653/v1/2021.findings-acl.340.

[25] J. Wang, Y. Liu, X. Wang, Are gender-neutral queries really gender-neutral? mitigating gender bias in image search, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1995–2008. URL: https://aclanthology.org/2021.emnlp-main.151. doi:10.18653/v1/2021.emnlp-main.151.

[26] R. Wolfe, A. Caliskan, American== white in multimodal language-and-image ai, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022, pp. 800–812.

[27] A. S. Luccioni, C. Akiki, M. Mitchell, Y. Jernite, Stable bias: Analyzing societal representations in diffusion models, arXiv preprint arXiv:2303.11408 (2023).

[28] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, A. Caliskan, Easily accessible text-to-image generation amplifies demographic stereotypes at large scale, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 1493–1504.

[29] A. Birhane, V. U. Prabhu, E. Kahembwe, Multimodal datasets: misogyny, pornography, and malignant stereotypes, arXiv preprint arXiv:2110.01963 (2021).