When does CLIP generalize better than unimodal models? When judging human-centric concepts

Romain Bielawski
ANITI, UT & CerCo, CNRS, France
romain.bielawski@univ-tlse3.fr

Benjamin Devillers
ANITI, UT & CerCo, CNRS, France
benjamin.devillers@univ-tlse3.fr

Tim Van De Cruys
Université de Louvain, Belgique
tim.vandecruys@kuleven.be

Rufin VanRullen
ANITI, UT & CerCo, CNRS, France
rufin.vanrullen@cnrs.fr

Abstract

CLIP, a vision-language network trained with a multimodal contrastive learning objective on a large dataset of images and captions, has demonstrated impressive zero-shot ability in various tasks. However, recent work showed that in comparison to unimodal (visual) networks, CLIP's multimodal training does not benefit generalization (e.g. few-shot or transfer learning) for standard visual classification tasks such as object, street numbers or animal recognition. Here, we hypothesize that CLIP's improved unimodal generalization abilities may be most prominent in domains that involve human-centric concepts (cultural, social, aesthetic, affective...); this is because CLIP's training dataset is mainly composed of image annotations made by humans for other humans. To evaluate this, we use 3 tasks that require judging human-centric concepts: sentiment analysis on tweets, genre classification on books or movies. We introduce and publicly release a new multimodal dataset for movie genre classification. We compare CLIP's visual stream against two visually trained networks and CLIP's textual stream against two linguistically trained networks, as well as multimodal combinations of these networks. We show that CLIP generally outperforms other networks, whether using one or two modalities. We conclude that CLIP's multimodal training is beneficial for both unimodal and multimodal tasks that require classification of human-centric concepts.

1 Introduction

Vision-language pretraining in neural networks is gaining popularity due to the growing interest in multimodal tasks such a Visual Question Answering or Image Captioning (Anderson et al., 2017; Lu et al., 2019; Li et al., 2019; Singh et al., 2019), but also to the availability of online resources that allow to build large-scale training datasets without manual annotations (Radford et al., 2021; Jia et al., 2021). In theory, training a model on multimodal data should help improve its representation of data from each of the modalities. For an image-text model, for instance, the image features could be enriched by the abstraction of the linguistic data –the semantic grounding property, and inversely, the linguistic features could gain informativeness through visual grounding (Harnad, 1990).

Unfortunately, this does not always happen in practice. Recently, Devillers et al. (2021) evaluated the visual generalization abilities of CLIP (Radford et al., 2021), a popular network trained with a contrastive learning objective on more than 400M image-caption pairs scraped from the web, and other multimodal models (Sariyildiz et al., 2020; Desai and Johnson, 2020). They showed that for standard object classification tasks (e.g. digit, fashion item or natural image classification), multimodal networks like CLIP underperformed compared to other unimodal (vision-only) models like BiT-M (Kolesnikov et al., 2019) in transfer learning, few-shot learning and unsupervised learning settings. Here, we revisit this question using datasets focusing on more "human-centric" concepts.

Human learning generally involves interacting with multimodal data. Thus, one could expect that CLIP's representations of images and text should be somewhat closer to human representations than those learned by unimodal models. Moreover, given that CLIP was trained on image-caption pairs from a variety of sources from the Internet (including social networks), we can assume that an important part of its training captions was written by humans for other humans. This is different from standard vision datasets, in which labels or annotations are sometimes human-generated (e.g. through Amazon's Mechanical Turk), but always produced for machine-learning purposes. Again, this difference should bring CLIP's representations closer to human ones when compared to unimodal models.

Proceedings of the 7th Workshop on Representation Learning for NLP, pages 29 - 38

Thus, there should exist at least *some specific tasks* for which CLIP's multimodal training provides advantages over unimodal models. As an example, consider the task of assigning a genre to a movie based on its poster and title. This requires retrieving fine-grained information about, among other things, the artistic, emotional or stylistic aspect of an image or a piece of text (or both). This can only be properly achieved if the model's training offered appropriate exposure to such human-centric concepts. Here, we use the term *human-centric* whenever a concept refers to cultural, social, aesthetic and/or affective components of the world.

We thus make the hypothesis that CLIP should perform better than unimodal models in generalization tasks where human-centric concepts are involved. We evaluate this hypothesis on three tasks involving such human-centric concepts: sentiment analysis on tweets; genre classification of books; genre classification of movies. All tasks can be performed based on visual data (images), text data (tweet, book or movie title, movie plot summary), or both. For the movie genre classification, we introduce a new, large-scale multimodal dataset obtained by a crawling on The Movie Database (TMDb). As detailed below, we find that CLIP outperforms unimodal models in both vision and text-based classification, as well as pairwise combinations of these unimodal models in the case of multimodal (image+text) classification. Consequently, CLIP establishes a new SOTA on these tasks.

We provide our code for reproducibility[1].

2 Models

We compare CLIP (trained contrastively on both images and text) against several unimodal models. For fairer comparisons, all the vision models are ResNet50 (He et al., 2015) based architectures and all the text models are transformer encoders.

CLIP was trained using a contrastive loss on a large (400M) set of image-text pairs. The training of CLIP consists in creating a joint (multimodal) embedding space. For one batch of image-text pairs, the objective of the network is that the embedding of an image (through a ResNet50 backbone, here simply referred to as CLIP) and the embedding of its text description (through a transformer backbone, here referred to as CLIP-T) are as close as possible, while the embedding of an image and the embeddings of text descriptions of other images in the batch are as far as possible. After training, the text encoder and the image encoder can be used as single-modality encoders.

For unimodally trained vision networks, we use two pretrained ResNet50-based models: the standard ResNet50 that was trained for classification on ImageNet-1K (here referred to as RN50), and BiT-M that was trained on ImageNet-22K (Deng et al., 2009).

For unimodal text embeddings, we test two standard text encoders against CLIP's: Bert-large and Bert-base (Devlin et al., 2018). We use the Bert sentence transformer version (Reimers and Gurevych, 2019), based on Bert's [CLS] token and fine-tuned on SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). Among the transformer encoders provided in the HuggingFace (Hug) repository at the time our experiments were conducted, these were the two best-performing across several text classification tasks, and are now still close to SOTA. These versions of Bert-large and Bert-base are fine-tuned on downstream text classification tasks, but we refer to them in this paper simply as Bert-large and Bert-base.

Although all 3 text encoders are transformer encoders (Vaswani et al., 2017), they do not have the same number of parameters. Bert-large has 300M, Bert-base has 110M, and CLIP-T has 80M parameters. This gives a structural disadvantage to CLIP-T, which only strengthens our conclusions, as we found CLIP-T to be the overall best-performing text model.

We consider both unimodal tasks (classification of images or text), as well as multimodal tasks (classification of image-text pairs). When performing a unimodal task, the encoding of the image (resp. the text) is used directly by the corresponding classifier. When performing a multimodal task (image-text based classification), the encoding of an image by a visual model and the encoding of the corresponding text by a textual model are simply concatenated to create the multimodal vector that is used for the classification.

For BiT-M and RN50, we use the last layer output before the classification head used for their training, which counts 2048 dimensions. For CLIP, we use the latent vector in the multimodal space generated by the visual pipeline, counting 1024 dimensions; for CLIP-T, the one generated by the textual pipeline (1024 dimensions); and for the

[1]Link not displayed here to preserve anonymity

two Bert models, we use the vectors directly provided by the Sentence Transformer pipelines (1024-dimensional for Bert-large and 768-dimensional for Bert-small).

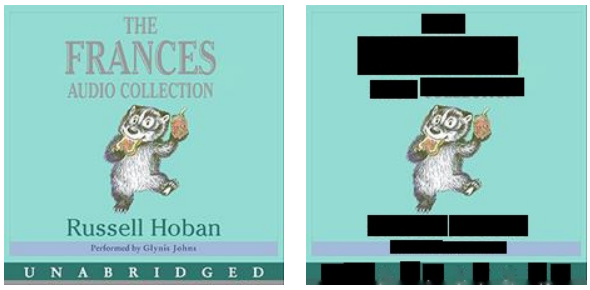

Figure 1: An original cover from the Book Cover dataset (left) and the associated masked cover (right). The title, the name of the author and parts of the text have been blacked out by the EAST algorithm, while the white text was incompletely detected, but subsequently blurred by the second algorithm. This sample belongs to the "Children's books" genre. Its title is: "Frances Audio Collection CD (I Can Read Level 2)". This image is copyright from Amazon.com, Inc. and used here for academic purpose only.

3 Datasets

We evaluate the models on three datasets composed of labelled image and text data, that can be inputted as pairs for multimodal classification tasks, or used as single inputs for unimodal classification tasks. The language part of all these datasets is in English.

3.1 MVSA

MVSA or "Multi-View Sentiment Analysis" (Niu et al., 2016) is a dataset of pairs of images and associated text from Twitter, labelled with three possible sentiments (Positive, Neutral or Negative). Each image and each piece of text has three labels given by three different users, adding up to 6 labels for each image-text pair. We assign a score for each label (Positive: 2, Neutral: 1, Negative: 0) and we compute the rounded average score for each pair. By doing so, we get only one label per image-text pair that we can then use for single-label classification across modalities.

3.2 Book covers

The Book Covers dataset was introduced by Iwana et al. (2017). It consists of 57k images of book covers scraped from the Amazon website, with their title as text information. Each pair of cover+title is labelled with one genre among 30 possibilities. A cleaner version of the dataset, removing one genre and grouping two similar ones, with only 28 classes and 55.1k images, was later introduced by Lucieri et al. (2020). This is the dataset we use for our experiments.

3.3 Plotster and TMDb

We introduce and publicly release the *Plotster* dataset[2], obtained by crawling TMDb (www.themoviedb.org) using their provided API. It consists of 207,902 triplets of {poster, title, plot} (split in 189,185 train samples and 18,717 test samples), with each having several potential labels among 19 genres. A representative sample from this dataset is shown in Figure 2. Typically, each movie has between 1 and 6 genres, with an average of 1.7. Each poster is an RGB image of 900×600 pixels (height$\times$width). Plots have an average length of 310.8 characters, and titles an average length of 18.6 characters. For text input, in unimodal or multimodal settings, we can choose either plot or title. The results of both configurations were computed and are displayed in this paper.

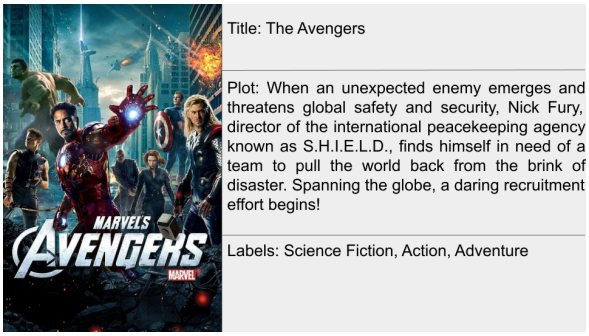

Figure 2: A data sample from *Plotster*. The image displayed here is property of The Walt Disney Company / Marvel Entertainment and under the CC BY-SA 2.0 license.

A previous crawling on TMDb had been made by Mangolin et al. (2020). It contained only 10,594 movies, as the authors aimed to retrieve other pieces of data such as trailer video clips and subtitles. They had not included titles in their dataset. From these movies, 10,554 (i.e., 99.6%) can also be found in *Plotster*. For comparison, we isolated the posters and plots from this dataset, and verified that our results obtained on the full *Plotster* were still valid on this subset.

In another control experiment, we verified that CLIP's improved performance on the *Plotster*

[2]Link not displayed here to preserve anonymity

Vision \ Text	None	Bert-base	Bert-large	CLIP-T
None	$\emptyset$	63.33 ± 0.18	64.02 ± 0.74	$\underline{64.60} \pm 0.30$
RN50	55.17 ± 0.37	63.93 ± 0.36	63.92 ± 0.55	64.13 ± 0.37
BiT-M	60.0 ± 1.46	61.93 ± 2.05	63.16 ± 2.82	62.77 ± 0.72
CLIP	$\mathbf{63.07} \pm 0.23$	$\underline{\mathbf{66.03}} \pm 0.15$	$\underline{\mathbf{66.03}} \pm 0.6$	65.58 ± 0.38

Table 1: Accuracies for the MVSA dataset. CLIP is the best vision model, CLIP-T the best text model. All text models perform similarly in both unimodal and multimodal setting, except when paired with CLIP (which yields the best performance of each column).

Vision \ Text		None	Bert-base	Bert-large	CLIP-T
None		$\emptyset$	54.70 ± 0.25	54.92 ± 0.43	$\underline{57.28} \pm 0.27$
Standard	RN50	10.04 ± 4.33	54.85 ± 0.52	55.53 ± 0.39	57.20 ± 0.49
	BiT-M	29.33 ± 0.92	50.11 ± 0.57	50.49 ± 0.59	52.60 ± 0.46
	CLIP	$\mathbf{53.75} \pm 0.23$	60.38 ± 0.34	60.62 ± 0.27	$\underline{\mathbf{60.66}} \pm 0.26$
Masked	RN50	10.41 ± 2.43	54.26 ± 0.25	55.11 ± 0.17	57.26 ± 0.29
	BiT-M	24.87 ± 0.99	48.93 ± 0.77	50.09 ± 0.71	52.08 ± 0.59
	CLIP	$\mathbf{33.04} \pm 0.21$	57.86 ± 0.45	58.47 ± 0.40	$\underline{\mathbf{59.54}} \pm 0.28$

Table 2: Accuracies for the Book Cover dataset (standard images on top, masked images on the bottom). CLIP and CLIP-T are the best performing models of each unimodal test, and together provide the best multimodal combination for both standard and masked images. Masks diminish the performance of all models (and their combinations), but the advantage for CLIP (and CLIP-T) remains.

dataset was not a result of specific movie posters, plots and titles from TMDb having been included in CLIP's training (as the training set is not public, there is no direct way to determine this). For our control experiment, we crawled TMDb again, looking for movies with a release date later than January 5, 2021, date of the OpenAI blog post introducing CLIP. We thus assume that most of this data could not have been included in CLIP's training dataset. The new crawl resulted in 20,280 movies, only 93 of which had been present in the original *Plotster* dataset. We tested on these 20,280 new samples the classifiers trained on *Plotster* (only in unimodal settings), and report the corresponding results.

3.4 Masking

CLIP has been found to have an ability to "read" text inside images (Goh et al., 2021). As most of the images in the Book Cover dataset and in *Plotster* have text on them, and as this text could be informative about the genre of the book or movie, we worried that this ability could give CLIP an unfair advantage over other vision models. To minimize this possibility, we created alternative versions of these two datasets by applying a masking procedure on the images (see Figure 1). We used the EAST algorithm (Zhou et al., 2017) to generate bounding boxes around text; if the score given to a text detection reached a certain threshold, a black rectangle was applied over the corresponding bounding box. On top of that, a second algorithm detects the remaining small white text using a thresholding method, a saturation filter and a size filter, and then does a Telea inpainting (Telea, 2004) to remove it. On 10 randomly selected posters, we verified that this algorithm masked or blurred 84% of the readable characters. It masked or blurred 95% of them on 10 randomly selected book covers as well.

The results on the datasets with masks are reported along with those of the originals.

4 Results

To compare the generalization capabilities of our text, vision, and multimodal models, we focus on transfer learning and few-shot learning settings.

4.1 Transfer learning

Our first experiment is transfer learning. We use the pretrained networks (see Section 2) with frozen weights as encoders, and train a new classification head for each of our datasets in unimodal or multimodal settings.

For transfer learning in single-label classification (sentiment on MVSA, book genre), we plug on top of the frozen feature vector encoder one dense layer (ReLu activations) bringing the dimensions down

Vision \ Text		None	Title			Plot		
			Bert-base	Bert-large	CLIP-T	Bert-base	Bert-large	CLIP-T
None		$\emptyset$	.314 ± .01	.323 ± .01	.397 ± .00	.582 ± .00	.599 ± .01	.612 ± .00
Standard	RN50	.090 ± .01	.338 ± .01	.363 ± .01	.393 ± .02	.578 ± .01	.599 ± .01	.599 ± .01
	BiT-M	.415 ± .01	.490 ± .01	.499 ± .01	.507 ± .01	.625 ± .01	.637 ± .01	.631 ± .01
	CLIP	**.526** ± .01	.559 ± .01	.558 ± .01	**.593** ± .01	.672 ± .00	.683 ± .00	**.687** ± .00
Masked	RN50	.070 ± .01	.335 ± .02	.352 ± .01	.383 ± .02	.576 ± .01	.597 ± .01	.596 ± .01
	BiT-M	.372 ± .00	.457 ± .02	.480 ± .01	.490 ± .01	.617 ± .01	.631 ± .01	.621 ± .01
	CLIP	**.449** ± .01	.525 ± .01	.534 ± .01	**.564** ± .00	.658 ± .00	.667 ± .00	**.676** ± .00

Table 3: f1-scores for the *Plotster* dataset. CLIP is the best model in vision, CLIP-T the best in text whether titles or plots are given as input, and CLIP+CLIP-T is the best multimodal combination in all cases. The masking doesn't affect the advantage for CLIP.

to 256, and then another dense layer (softmax activation) for the classification. We then train only the weights of these 2 layers on the classification task with a Cross-Entropy Loss; therefore the network learns to output a probability density over the classes.

For multi-label classification (movie genres) the loss is a Binary Cross-Entropy Loss, and therefore the second dense layer outputs a number between 0 and 1 for each class. As the ground-truth label vector for one sample is an 19-dimensional one-hot vector, we round the 19-dimensional prediction of the network to get a binary predicted label vector. A f1-score (Pedregosa et al., 2011) comparing the predicted label vector to the ground-truth vector is reported, as raw accuracy is not a reliable measurement for multi-label classification. The f1-score is computed for each movie, and subsequently averaged over the test set of each dataset. For f1-scores, as for accuracy, the higher the better.

Tables 1 and 2 show the results on the single-label datasets: MVSA and Book Cover. The first column corresponds to the result of the vision-only experiment, the first line to those of the text-only experiments, and the other cells display the results of the multimodal ones. Table 3 shows the results for the multi-label dataset (Plotster). In all tables, the best vision-only performance is highlighted in **bold**, the best text-only is underlined and the best multimodal one is **both underlined and bold**. The standard deviation is calculated over five experiments with different random seeds and random initialization of the weights of the classifiers.

On MVSA (Table 1), CLIP is the best performing vision-only model and CLIP-T the best text-only model. The best multimodal combinations are CLIP+Bert-base and CLIP+Bert-large, with CLIP+CLIP-T near the same level (less than 0.5 percentage point behind). This is not unexpected, as CLIP-T counts much fewer parameters than Bert-base or Bert-large (see section 2).

For the Book Cover dataset (Table 2), CLIP is by far the best performing vision model, both with the standard covers and with the masked covers as input. The difference between CLIP's accuracy (53.8%) and the other two (RN50: 10.0%; BiT-M: 29.3%) remains high in the masked configuration (with CLIP at 33.0% and the other two below 25%), even though CLIP has lost the ability to read the text on the covers. This indicates that CLIP's reading ability is not the sole explanation for its advantage over vision-only models. CLIP-T is again the best text-only model. Here, the best multimodal combination is CLIP+CLIP-T for both standard and masked configurations. Finally, compared to previously established SOTA performance on the Book Cover dataset by Lucieri et al. (2020), CLIP easily beats the previous visual SOTA (27.8 % accuracy), CLIP-T the previous textual SOTA (55.6%), and CLIP+CLIP-T the previous bimodal SOTA (55.7%).

Concerning our new *Plotster* dataset (Table 3), similar conclusions emerge. In vision-only conditions, RN50 performs relatively poorly; in the standard dataset, CLIP largely outperforms BiT, and this difference decreases but remains in the masked dataset. In text-only conditions, CLIP-T is the best model, both with titles and plots as input. Finally, in the multimodal settings, CLIP+CLIP-T is always the best-performing combination, whether using standard or masked images, title or plot as textual inputs. As before, the prevalence of CLIP in all task settings, even when text has been removed from the movie posters, indicates that its superior-

ity in our movie genre transfer learning task is not solely due to its reading ability. We surmise that this advantage reflects a form of semantic grounding resulting from CLIP's multimodal training.

We also tested CLIP, CLIP-T and their combination on a subset of *Plotster* corresponding to the dataset of Mangolin et al. (2020), in order to compare with previous SOTA values. We found that CLIP beats the previously established visual SOTA (f1-score of 0.603 against 0.409), CLIP-T the textual SOTA (f1-score of 0.589 against 0.488) and CLIP+CLIP-T the bimodal SOTA (0.670 against 0.628).

In a separate control experiment, we tested all our models (trained on the entire *Plotster* training set) on a new set of movies, all released after OpenAI's initial blogpost introducing the CLIP model. On this new test set, CLIP's f1-score changes from 0.526 to 0.439, BiT's goes from 0.415 to 0.318 and RN50's from 0.090 to 0.020. CLIP-T's (with title as text input) goes from 0.397 to 0.276, Bert-large from 0.323 to 0.237 and Bert-base from 0.314 to 0.229. The general diminution of the f1-score across all networks is probably due to the fact that features trained to classify older movies do not work equally well when they are applied to more recent movies. Nevertheless, CLIP and CLIP-T remain the top-performing models; as it is unlikely that these recent movie posters and captions had been included in CLIP's training dataset, we conclude that CLIP's high transfer-learning performance on *Plotster* is not a consequence of prior exposure to these stimuli, but a true form of generalization.

In general, we see that in all the unimodal settings, CLIP outperforms the other vision models, and CLIP-T the other text models. This is true, even though CLIP has roughly the same number of parameters than RN50 or BiT-M, and fewer dimensions in its latent space (and thus, less parameters in its classifier head). Similarly, CLIP-T counts much fewer parameters than Bert-base or Bert-large (although it has a higher-dimensional latent space than Bert-small). In most of the multimodal settings, changing from one visual model to CLIP or from one textual model to CLIP-T improves performance (the only exceptions are for CLIP-T on MVSA and on *Plotster* with plots as text inputs). The best multimodal models always involve CLIP, and also involve CLIP-T in all cases except MVSA. This makes the CLIP + CLIP-T combination the best overall multimodal model in our experiments.

4.2 Few-shot learning

The second experiment we conduct is a visual few-shot learning task: we measure test classification accuracy based on exposure to a small number of randomly chosen training samples (or "prototypes") from each class. We can thus compare the results for our datasets with those of Devillers et al. (2021), who also measured visual few-shot learning performance.

In their paper, Devillers et al. (2021) used a single prototype vector for each class, obtained by averaging the latent representation of the N randomly drawn training samples for that class. Here, we prefer to retain all N individual samples as prototypes, and use a 1-nearest-neighbor (1-NN) classifier (Pedregosa et al., 2011) to classify the new vectors. We verified that this method, when applied to the same datasets as in (Devillers et al., 2021), does not alter their conclusion (see the first plot of Figure 3). To select the class prototypes of *Plotster* (which is a multiclass dataset), we randomly choose movies with a given class label. For example, a movie with genres "adventure" and "action" could be randomly chosen as a prototype of either genre. Moreover, when predicting the genres of a movie using the 1-NN classifier, we predict all the genres of the closest prototype.

Figure 3 reports the few-shot accuracy on the Book Covers and MVSA datasets as well as the f1 score for the *Plotster* datasets. Contrary to the conclusion of Devillers et al. (2021) using standard visual datasets (see Figure 3, left), our results show a clear advantage to CLIP in our more "human-centric" visual tasks, even when masks are applied. For MVSA, the networks required more samples (between 20 and 100) to reach above-chance accuracy than for the other datasets (that use 1 to 10 samples). In that specific case, the three models are more difficult to distinguish, but CLIP still appears better than the other two visual models.

4.3 Summary

In the visual domain, CLIP systematically outperforms the unimodal vision models in transfer learning (Tables 1-3) and in visual few-shot learning (Figure 3), despite having a smaller embedding space than the other two ResNet50-based models. Part of CLIP's superiority may be due to its ability to read, but the advantage remains when text is removed from the images. This conclusion goes

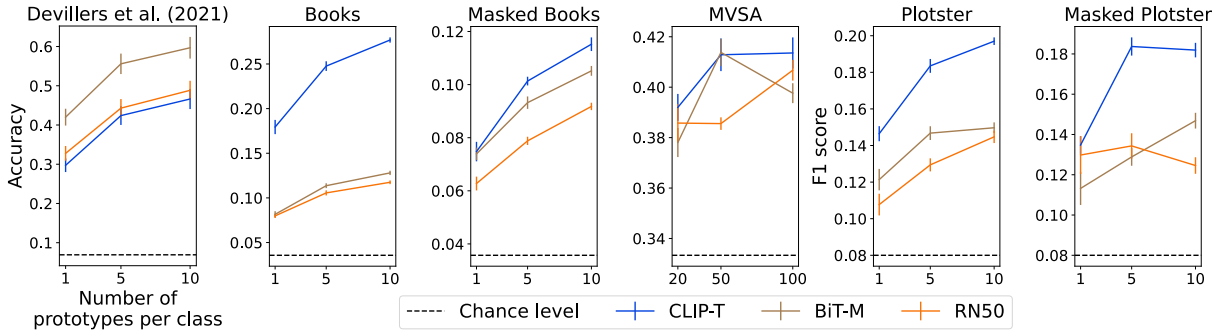

Figure 3: Few-shot learning accuracy (vision-only) over single label datasets (Book Covers, MVSA) and f1-score over the multilabel *Plotster* datasets. The leftmost panel reports average accuracy on 6 standard visual datasets used in Devillers et al. (2021) – namely CIFAR10, CIFAR100, CUB, FashionMNIST, MNIST and SVHN. Accuracy was recomputed using the same method as for our datasets; the conclusions are identical to those of Devillers et al. (2021): CLIP does not perform better than RN50 or BiT in this few-shot learning setting. On the contrary, for our datasets CLIP outperforms the two other vision models. The advantage is reduced but still present when masks are applied.

against the observations of Devillers et al. (2021) using standard visual datasets (including SVHN, a digit reading dataset), where CLIP was never better (and often slightly worse) than other ResNet50 based models, including RN50 and BiT-M. We explain this difference by the nature of the classification performed: our tasks involve human-centric concepts, as defined earlier.

In the text domain, CLIP-T, despite having been trained with fewer parameters than the other two transformers (Bert-small and Bert-large), is systematically the best performing model in transfer learning.

Across seven multimodal settings (MVSA dataset; Book Covers dataset [with / without masks]; *Plotster* with [titles / plots] $\times$ [with / without masks]), CLIP+CLIP-T was the best multimodal combination in six cases. In the remaining case (MVSA), it was a tie between CLIP+Bert-large and CLIP+Bert-small (two language models that count many more parameters than CLIP-T).

We think that the semantic grounding provided by linguistic inputs when training CLIP's visual stream, and respectively, the visual grounding provided by image features when training the CLIP-T language model, shaped their latent space in a way that makes it possible to better grasp the human-centric components of an image or a text.

5 Discussion and conclusion

CLIP's generalization abilities were originally described in the context of zero-shot learning (Radford et al., 2021), but they may also extend to other settings, including transfer learning and few-shot learning. Past work has revealed that this is not always the case (Devillers et al., 2021). Considering the latent representations learned by CLIP may help us better understand when multimodal training does or does not benefit generalization abilities, continuing the work of Hossain et al. (2019). In our case, it appears that one of the domains where the improvement is most significant is when human-centric concepts are being judged.

During their joint contrastive training, CLIP and CLIP-T have learned to extract common information between image and text modalities, so that the two streams would result in similar embedding vectors. This means that the representation of text in CLIP-T has been enriched with visual data, and symmetrically, that the representation of images in CLIP has been improved by semantic or linguistic enrichment. This is what is collectively referred to as the "semantic grounding" property (Harnad, 1990; Bender and Koller, 2020). However, another consequence of this multimodal contrastive training is that when learning a common ground between modalities, some relevant information could be lost. For text, what cannot be directly linked to images (including grammatical or syntactic properties); and for images, what is not directly relevant to the text description (including fine-grained visual details that are rarely mentioned in the corresponding caption). This information loss might be the reason why CLIP was found to perform worse than standard vision-only models in a unimodal setting with standard visual datasets (Devillers et al., 2021). For the same reason, one could actually expect that in a multimodal setting, the combination of CLIP's

vision and text streams (CLIP+CLIP-T) could lead to worse performance than other combinations (e.g. RN50+Bert). The unimodal networks are trained to capture the relevant features of their modality, and when combined, could cover the multimodal feature space more fully than CLIP, a network trained to discard information that is not redundant across modalities. Our results show that, at least in our human-centric classification tasks, this limitation was not consequential: CLIP, CLIP-T and their combination often performed optimally. This may be because human-centric information is particularly well captured by features expressed in *both* images and text, rather than in each modality independently. On the other hand, this same reasoning could explain why CLIP+Bert combinations performed slightly better than CLIP+CLIP-T on MVSA: Bert may have provided additional information not captured by CLIP, which was lacking in CLIP-T because of their redundant embeddings (or, this might simply be due to the fact that Bert has many more parameters than CLIP-T).

Our suggestion that CLIP (and CLIP-T) perform particularly well when judging human-centric concepts resonates with recent findings relating CLIP's representations to human brain representations. Goh et al. (2021) reported that some artificial neurons in CLIP's visual stream (but not in standard visual models like Inception or ResNet) are systematically activated by specific "concepts" such as a particular person, emotion, country, religion, etc. Furthermore, these neurons could be equally activated by visual features (e.g., a photograph or drawing of the person's face) or by written text (e.g., the person's name). The authors related this multimodal invariance to properties of specific biological neurons found in the human hippocampus and temporal medial lobe, called "concept cells": these cells would also systematically activate when presented with a picture, drawing or written word representing a specific concept, such as a photograph of the actress Jennifer Aniston or her written name (Quiroga et al., 2005; Reddy and Thorpe, 2014). Indeed, more recently Choksi et al. (2021) compared brain fMRI representations in the human hippocampus with the patterns of representations measured in various vision models. They found that CLIP and other networks trained with multimodal objectives were more similar to human hippocampus representations than standard vision models (including RN50 and BiT-M). This could explain why a multimodal network like CLIP performs better when judging "human-centric concepts".

To conclude, we think that it is crucial to investigate the specific domains in which a multimodal training such as CLIP's can (or cannot) improve generalization. Our work indicates that *multimodality* will be key for developing algorithms designed for human-centric tasks (even for *unimodal* tasks) such as detecting emotions, analyzing personality, conducting a conversation or, more generally, when human-machine interactions are involved.

6 Acknowledgements

All book cover images and book titles are copyright Amazon.com, Inc. The display of the images are transformative and are used as fair use for academic purposes.

All posters displayed here are property of The Walt Disney Company / Marvel Entertainment and under the CC BY-SA 2.0 license.

This research was supported by ANITI ANR grant ANR-19-PI3A-0004 and COCOBOTS ANR-DLR bilateral French-German project ANR-21-FAI2-0005.

References

Huggingface sentence transformer repository. https://huggingface.co/sentence-transformers. Accessed: 2022-02-21.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, abs/1707.07998.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Bhavin Choksi, Milad Mozafari, Rufin Vanrullen, and Leila Reddy. 2021. Multimodal neural networks better explain multivoxel patterns in the hippocampus. In *Neural Information Processing Systems (NeurIPS)*

conference: 3rd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM 2021).

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Karan Desai and Justin Johnson. 2020. Virtex: Learning visual representations from textual annotations. *CoRR*, abs/2006.06666.

Benjamin Devillers, Bhavin Choksi, Romain Bielawski, and Rufin VanRullen. 2021. Does language help generalization in vision models? In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 171–182, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*. Https://distill.pub/2021/multimodal-neurons.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6).

Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. 2017. Judging a book by its cover.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, abs/2102.05918.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2019. Large scale learning of general visual representations for transfer. *CoRR*, abs/1912.11370.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265.

Adriano Lucieri, Huzaifa Sabir, Shoaib Ahmed Siddiqui, Syed Tahseen Raza Rizvi, Brian Kenji Iwana, Seiichi Uchida, Andreas Dengel, and Sheraz Ahmed. 2020. Benchmarking deep learning models for classification of book covers. *SN Computer Science*, 1(3):139.

Rafael B. Mangolin, Rodolfo Miranda Pereira, Alceu S. Britto Jr., Carlos Nascimento Silla Jr., Valéria Delisandra Feltrim, Diego Bertolini, and Yandre M. G. Costa. 2020. A multimodal approach for multi-label movie genre classification. *CoRR*, abs/2006.00654.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling*, page 15–27.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. 2005. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Leila Reddy and Simon J Thorpe. 2014. Concept cells through associative learning of high-level representations. *Neuron*, 84(2):248–251.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Mert Bülent Sariyildiz, Julien Perez, and Diane Larlus. 2020. Learning visual representations with caption annotations. *CoRR*, abs/2008.01392.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. *CoRR*, abs/1904.08920.

Alexandru Telea. 2004. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. EAST: an efficient and accurate scene text detector. *CoRR*, abs/1704.03155.