

Bootstrapping a high quality multilingual multimodal dataset for Bletchley

Owais Khan Mohammed*

Kriti Aggarwal*

Qiang Liu

Saksham Singhal

Johan Bjorck

Subhojit Som

Microsoft

OWAIS.MOHAMMED@MICROSOFT.COM

KRAGGA@MICROSOFT.COM

QIANGLIU@MICROSOFT.COM

SAKSHAM.SINGHAL@MICROSOFT.COM

JOHANBJORCK@MICROSOFT.COM

SUBHOJIT.SOM@MICROSOFT.COM

Abstract

Vision-language models have recently made impressive strides, primarily driven by large-scale training on web data. While pioneering works such as CLIP and ALIGN show significant improvements, these are focused on English data as it is easy to source them from the web. Towards serving non-English-speaking demographics, we consider various methods for generating multilingual data and find that a simple bootstrapping mechanism works surprisingly well. Specifically, just using English image captions data and text-only multilingual translation pairs we train a fairly strong multilingual vision-language model and then leverage it to create a much cleaner version of the multilingual image captions dataset we collected. We demonstrate that this dataset which was used to train Bletchley result in a strong multi-modal and multilingual model which reaches strong performance across several multilingual zero-shot tasks. Specifically, Bletchley achieves state-of-the-art results on multilingual COCO, Multi30k sets, IGLUE WIT and xFlickr&CO datasets.

1. Introduction

Vision and language models are important in a lot of applications including image-retrieval (Wan et al., 2014), assistance for the sight-impaired (Gurari et al., 2018), and image generation (Ramesh et al., 2021). Inspired by large-scale pretraining in NLP, vision-language models have recently made impressive strides by pretraining on web-sourced image-text data. Models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) have significantly improved the state-of-the-art(SOTA) for vision-language tasks. These models focus on English, perhaps as a significant fraction of web data is written in English. Nonetheless, much of the world population does not speak English, and for serving this demographic, multilingual vision-language models are a necessity.

Motivated by this, we consider the construction of a large scale multilingual multimodal dataset. More specifically we focus on the dataset creation recipe that was used for training Bletchley (Tiwary, 2021) which achieves state-of-the-art results on various multimodal benchmarks. Data sourcing multilingual from crawled web data is challenging as English represents a significant fraction of the data. Furthermore, most NLP tools are models developed with English in mind. To avoid these issues, we consider a simple bootstrapping

*. Equal contribution

method – we first collect English-only data by leveraging publically available NLP tools. Specifically, we use a POS and entity tagger to filter the images and train a multilingual vision-language model on this data. Although the model was trained using English-only image-text pairs it was a multilingual model thanks to the multilingual language task included during the training. Once this model was trained, we use this model to filter the multilingual dataset to create a multilingual image-text dataset. Our recipe for filtering and bootstrapping is simple and relies on open-source tools, making our data filtering easy to replicate for researchers.

On the modeling side, we devise a simple method of utilizing pretrained multilingual models. We initialize the text encoder from a multilingual text-only model. In order to keep the multilingual properties of the text encoder strong, we add a contrastive task with translated pairs during the training.

With these recipes for modeling and dataset bootstrapping, we train 3 versions, one with English-only image-text data but TTM (Translated Text contrastive Matching) task, one with multilingual cleaned data & TTM task, and then scale the model to 2.5B parameters and train it on the cleaned multilingual data.

These models achieve good results across various vision and language tasks and shine in multilingual benchmarks. Specifically, our largest model achieves SOTA results on multiple multilingual retrieval tasks. Our results highlight how dataset bootstrapping, and careful modeling choices can lead to strong multilingual vision-language models, using only crawled web data. We summarize our contributions:

- We introduce a method for constructing a high-quality English-only image-text dataset by using open source entity tagging for filtering captions.
- We show that it is possible to bootstrap a high-quality multilingual image-text dataset from vision-language models trained on an English-only image-language dataset.
- We consider a simple method of ensuring the vision-language models retain multilingual capabilities when training on heavily English-skewed image-text data.
- We demonstrate that the resulting recipe yields a strong vision-language model which excels in multilingual settings, reaching SOTA results on multilingual retrieval.

2. Related Work

Vision datasets. Large scale model pretraining has recently been revolutionizing the NLP domain. The computer vision domain too has been catching up by following similar trends of increasing the magnitude of their datasets. From 14M Imagenet dataset (Deng et al., 2009), (Ridnik et al., 2021) to 300M JFT (Sun et al., 2017) dataset, to 1B noisy hashtag Instagram dataset (Mahajan et al., 2018). Although, all of these datasets were focused on only image classification and hence could not be easily utilized for multi-modal tasks.

Large scale Image-text datasets. The curation of datasets which require human labels tends to be very hard and time consuming. This has prompted a surge in methodologies that leverage noisy self-supervision for training various vision-language tasks. CLIP Radford et al. (2021) curated 400M noisy image alt-text samples for training multi-modal models

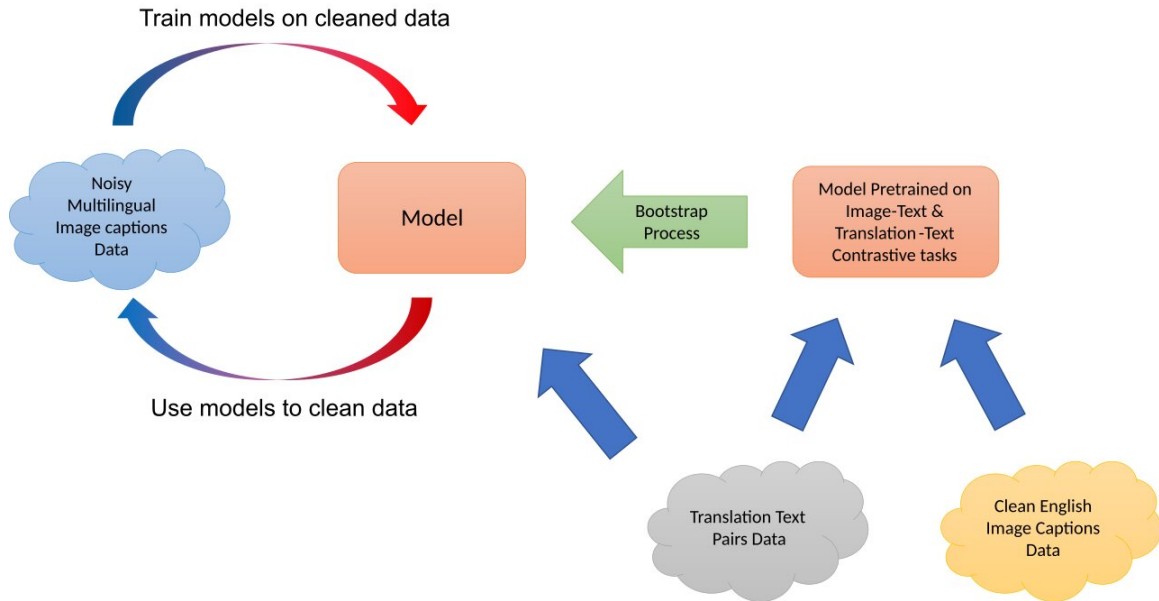


Figure 1: Model pretrained on Translated-Text Contrastive Task and Image-Text Contrastive Task using English only data, is used for bootstrapping multilingual vision-language models. The multilingual image-text data can then be filtered using this pretrained model and a new model can be trained using this filtered data. This step can be repeated to train better models and create cleaner data.

that align image and text embeddings. ALIGN Jia et al. (2021) went even further by curating a dataset of 1.8B image-text dataset which improved the performance of these models.

Image-Text Pretraining. Broadly speaking there are three major ways in which the image and text encoders are pretrained.

1. A single encoder is used for processing both image and text modalities.
2. A dual encoder architecture, where images and text are encoded by separate encoders.
3. Similar to 2, with the addition of a deep cross-modality encoder which operates on the outputs of the image and text encoders

An example of the first strategy is UNITER (Chen et al., 2020) which introduced using multiple training tasks on a single encoder to help learn joint multimodal embeddings. GLIP (Li et al., 2022) is an example of the third strategy which unified object detection with phrase grounding for its pretraining objective using a deep cross modality fusion. VLMo (Wang et al., 2021) trains both a dual encoder and a fusion encoder with a modular Transformer network.

ALIGN, CLIP and BASIC (Pham et al., 2021) fall in the second category which has shown that training a dual-encoder model (i.e., one model trained with two separate en-

coders) on image-text pairs using a contrastive learning loss works remarkably well when trained with large amounts of training data and even more importantly a very large batch size.

Multilingual Image-Text Pretraining. While most of the work has been limited to English-only models, some of the works which worked on multilingual multimodal training include, M3P (Ni et al., 2021) which encouraged the fine-grained alignment between images and multilingual text by adding an additional task of multimodal Code-Switched Training while using a fusion encoder. UC2 (Zhou et al., 2021) which used machine translation for augmenting existing English-only datasets with other languages while also adding two novel tasks Masked Region-To-Token Modeling and Visual Translation Language Modeling. MURAL (Jain et al., 2021) adds on top of ALIGN and CLIP by adding another contrastive translated text-text objective to make the model learn better multilingual text representations.

In this work, similar to MURAL, we have also trained Bletchley using two pretraining tasks, namely, the Image-Text Matching and Translated-Text Matching tasks. MURAL used a 1.8B noisy multilingual image-caption dataset with minimal filtering, presumably because the diversity in languages makes it nontrivial to clean multilingual datasets. In contrast, we provide a recipe for creating an English dataset using entity-based filtering which can then be used for bootstrapping of training multilingual multimodal models. We show that a model trained using English-only image alt-text pairs along with multilingual translation pairs can be used to clean a multilingual image captions dataset and the resulting model trained on this cleaned dataset shows significantly better performance compared to the noisy dataset.

3. Dataset Creation

3.1. English Image-Text Dataset

The image-text pairs used for creating the English dataset were derived from 1B documents sampled from web crawled data. The images were paired with both the alternate text and the caption in the document. Either the alternate text or the caption was randomly picked during training time. Using this technique, we had close to 3.8B multilingual image-text pairs.

Since the data consisted of multiple languages including English, we ran a language detection pipeline, namely BlingFire (Microsoft, 2021), on these alt-text and captions to detect the languages and filtered out any non-English samples. This step helped us identify 1.8B English samples from the total dataset.

3.1.1. ENTITY BASED FILTERING

To clean the data we first tagged the English data using a POS tagger (Loper and Bird, 2002), which helped us tag the nouns, pronouns, verbs, etc, in the alt-text and captions. We removed the sentences which had no pronouns/nouns. To filter the data further, we started with a list of the 1.5M most commonly used entities from the Google Freebase dataset (Google, 2013). The entities were bigrams/trigrams of names of people, nouns, verbs, adjectives, etc. For example: 'scarlet', 'kiwi fruit', 'paradise', 'take care', 'open air',

'quiet time', 'emergency services', etc. Out of these entities, we removed all stop words and stemmed the entities to combine entities with the same root. We further pruned this list by selecting only those entities which were linked in at least 100 documents. This process left us with 700k entities in total.

After tagging our data with these entities, we sorted the entities by the number of alt-text/captions in which they occurred in our data. From this list we removed the 5 most common entities namely *photo*, *stock*, *image*, *set* and *figure* which were applicable to any image and were generally present in images which had a template caption. We then analyzed the effect of filtering the data using different top-k entities. We found that with k=30,000 entities, we were able to preserve 80% of the data which had 78% matching rate of image to alt-text/caption pairs on manual inspection.

We then removed the image-caption pairs which did not have any one of those top 30,000 entities. On analyzing the distribution of the top 30,000 entities, we found that the data had a very heavy tail distribution, there were about 1M images with top 10 tagged entities (eg. 'art', 'design', 'people', etc.) and only 180 images for the 30,000th most frequent entities (eg. 'seaport', 'jag', 'miro', 'tuberculosis', etc.). Each image had 3.5 tagged entities on average.

On manual inspection, we found that entity filtering steps helped us in improving the quality of data. After this cleaning step, we were left with approximately 1B image-caption pairs out of 1.8B original English data pairs. Figure 2 shows some examples from the English dataset with the image, caption, and extracted entities after entity-based filtering.







Image	Caption	Extracted Entities	Filtered
	Men's UA ClutchFit Renegade Training Gloves	Men's, UA, Renegade, Training, Gloves	No
	This Tree Looks Like A Dragon	Tree, Dragon	No
	Thank You For Your Cooperation Sign	Thankyou, Cooperation, Sign	No
	Ccp 6	-	Yes, because no entity was extracted
	Ambystoma-macrodactylum	-	Yes, filtered because no entity was extracted because the caption is of a very specific
	Image	Image	Yes, filtered because the extracted entity is very common and does not tell us anything about the image. breed

Figure 2: Some examples from the English dataset containing the image-text pairs with the extracted entities and whether or not the sample was filtered.

Our original collected data consisted of English as well as 120 other languages. Since our entity tagging pipeline was only available in English, we first used only the English-only data for the first version of the model. An important note here is that we did not use any other heuristics for filtering such as the number of words in the text, removing boilerplate texts, removing misspelled words, removing dates, etc.

3.2. Translation Pairs Dataset

For the translation text contrastive task, we used a parallel corpus consisting of around 500 million translation pairs. This was the same dataset that was used for training InfoXLM (Chi et al., 2020). We also follow the same sampling strategy used for sampling across different languages when creating a mini-batch for doing the translation text contrastive task.

3.3. Multilingual Image-Text Dataset

We first trained a large variant of Bletchley on the English image captions data described in Section 3.1 and the multilingual translation pairs described in Section 3.2. The settings and evaluation metrics for this model are given in Section 4.2 and Section 5 respectively. We then used this model to score the multilingual image captions data we collected. More specifically we encoded the images and captions and took the cosine similarity of the resulting l_2 -normalized embeddings. In order to filter the scored data, we set a threshold such that 85% of the English image-text pairs that were above the threshold were good matching pairs per manual judgment. We scored both image-alt-text and image-caption pairs and chose a different threshold for them to maintain the relevance of 85% on English image-text pairs by manual judgment. At run time, if both of the fields were above the selected threshold we randomly picked either one of them. In this way, we were able to create a clean multilingual dataset with 2.3B samples. Figure 3 shows the distribution of languages in the dataset.

4. Pretraining

4.1. Pretraining Tasks

We pretrain our models on two major tasks, described as following.

4.1.1. IMAGE TEXT CONTRASTIVE TASK

For this task, we follow CLIP (Radford et al., 2021). More specifically we independently encode each image and caption in a batch and l_2 -normalize the generated representations. Next for each image embedding we calculate its dot product with all the other captions in the batch and apply the cross-entropy loss over the scores with the target being set to the corresponding caption for that image. We follow a similar process for each caption in the batch. If x_i is the normalized representation for the i^{th} image in the batch and y_j is the normalized representation for the j^{th} caption in the batch, then the loss is given by

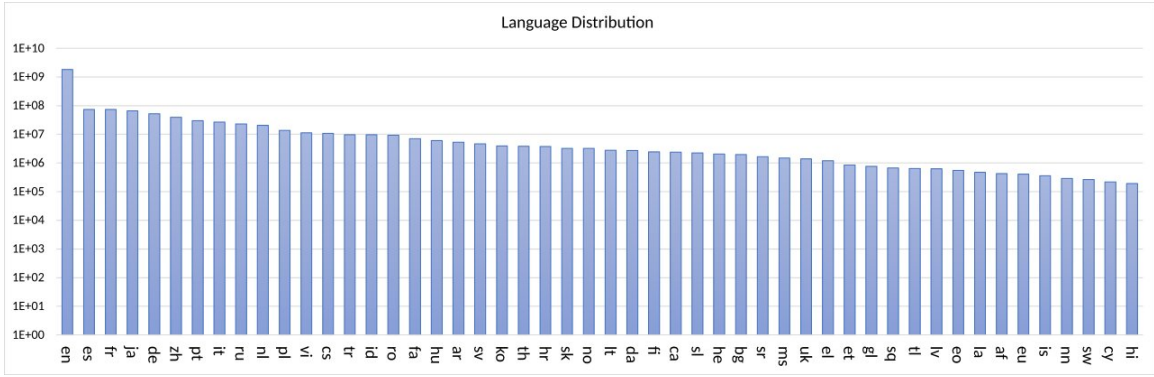


Figure 3: Distribution of languages in the alt-text present in the data for the top 50 languages. Note that the y-axis is in log scale. The plot shows that English is still the most predominant language in the data followed by Spanish, French and Japanese.

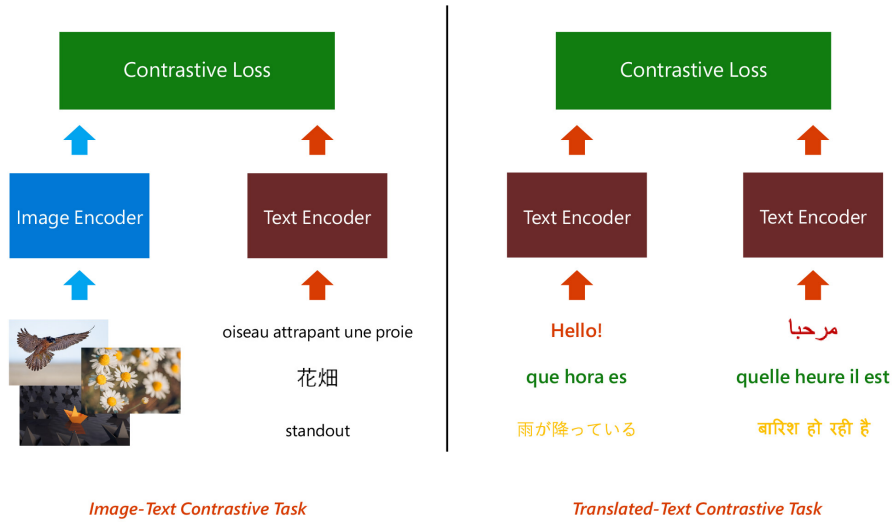


Figure 4: The two tasks the model was pretrained on. Images and captions were independently encoded by their corresponding encoders and contrastive loss was taken over the resulting embeddings

$$L = -\frac{1}{2N} \left(\sum_i \log \frac{\exp(x_i \cdot y_i/\tau)}{\sum_j \exp(x_i \cdot y_j/\tau)} + \sum_j \log \frac{\exp(x_j \cdot y_j/\tau)}{\sum_i \exp(x_i \cdot y_j/\tau)} \right) \quad (1)$$

The first term is the image to caption contrastive loss and the second term is the caption to image contrastive loss. τ here refers to the temperature which was a trainable parameter.

4.1.2. TRANSLATION TEXT CONTRASTIVE TASK

For the translation text contrastive task, if we train the model with a batch of N translation pairs, then there would be a total of $2N$ sentences in the batch. Each of these sentences is independently encoded by the text encoder and the resulting representation is l_2 -normalized. Next for each sentence embedding, we calculate its dot product with every other sentence in the batch giving a total of $2N - 1$ scores for each sentence. We apply the cross-entropy loss over these scores with the target being set to the corresponding translation. Let x_i be the normalized representation for the source sentence of the i^{th} translation pair and let $x_{p(i)}$ be the normalized representation of the target sentence. The loss for one sentence is given below, where the sum in the denominator goes over $2N - 1$ elements:

$$L = -\frac{1}{2N} \sum_i^{2N} \log \frac{\exp(x_i \cdot x_{p(i)}/\tau)}{\sum_{j \neq i} \exp(x_i \cdot x_j/\tau)} \quad (2)$$

4.2. Model Variants

We experiment with two different model scales, a large version that has around 860M parameters and an XL version with a total of 2.5B parameters. The model dimensions for the image and text encoder, and other relevant hyperparameters are provided in Appendix A. The language encoder is initialized from a pre-trained model trained following InfoXLM (Chi et al., 2020). The image encoder is trained from scratch. For the transformer implementation and training optimizations, we leverage the Deepspeed codebase (Rajbhandari et al., 2020). We train three different model variants as shown in the following table.

Table 1: Model variants. All models are multilingual. All models were trained using the image text contrastive task and the translation text contrastive task. The English and Multilingual denote the language of the Image-Text data used during training.

	Number of Paramters	Image Captions Dataset
Bletchley Large - English	860M	English
Bletchley Large - Multilingual	860M	Multilingual
Bletchley XL - Multilingual	2.5B	Multilingual

4.3. Pretraining Recipe

We pretrain all model variants for a total of 700k steps, out of which 500k steps are on the image-text contrastive task and 200k steps on the translation text contrastive task. The large version takes 10 days to train on 256 A100 GPUs and the XL version takes 30 days on 256 A100 GPUs. We use the Adam optimizer with a learning rate of 1e-4 for the weights initialized from scratch and a learning rate of 1e-5 for the pretrained weights in the language

encoder. The large model was trained using FP16, whereas for the XL model we found that using BFLOAT16 (Burgess et al., 2019) was more stable.

5. Experiments & Results

We focus our evaluation of the model’s performance on image retrieval and text retrieval tasks. To better understand the impact of using multilingual instead of English image captions we also evaluate the models on standard English retrieval benchmarks.

5.1. Multilingual benchmarks

To evaluate the model’s multilingual retrieval capabilities we use the COCO-CN (Li et al., 2019), COCO-JP (Merritt et al., 2020), Multi30k (Elliott et al., 2016) (Elliott et al., 2017) (Barrault et al., 2018), IGLUE (Bugliarello et al., 2022) WIT and IGLUE xFlickr&CO sets. For the zero-shot retrieval evaluation, we independently encode the images and captions using the image encoder and text encoder respectively. We then calculate the mean recall for Multi30k and COCO datasets and recall@1 for WIT and xFlickr&CO sets. We do not use any sort of prompt engineering for these tasks.

Table 2: Zeroshot Image and Text Retrieval on multilingual COCO and Multi-30k test sets. Bletchley XL sets a new zeroshot retrieval state-of-the-art on these sets.

Model	COCO		Multi30k		
	JP	CN	FR	CS	DE
M3P			27.1	20.4	36.8
ALIGN - Multilingual			84.9	63.2	84.1
MURAL	64.6		83.1	77.0	83.5
Bletchley Large - English	59.5	80.5	82.5	80.3	81.8
Bletchley Large - Multilingual	64.1	81.3	85.7	81.2	84.3
Bletchley XL - Multilingual	66.2	81.7	86.9	84.2	85.6

We see a significant improvement when moving from utilizing just the English image captions data to the multilingual image captions data, which shows that, only utilizing a pretrained language encoder or the translation text contrastive task isn’t enough to build a strong multilingual vision language encoder, but having clean multilingual image captions data is a key ingredient as well.

Evaluating Bletchley on zero-shot WIT and xFlickr&CO datasets from the IGLUE benchmark, we found that Bletchley sets new records on both of these datasets by outperforming all other models by huge margins, showing the generalizability of Bletchley across both high and low resource languages.

Table 3: Zeroshot Image and Text Retrieval on WIT. Results of all compared models are directly taken from the IGLUE benchmark (Bugliarello et al., 2022)

Retrieval@1	Model	Language									
		ARB	BUL	DAN	ELL	EST	IND	JPN	KOR	TUR	VIE
IR	mUNITER	7.74	8.26	10.66	8.95	7.67	10.88	9.00	5.91	9.57	13.00
	xUNITER	7.63	8.49	10.32	11.23	6.41	10.21	7.30	6.34	9.57	9.72
	UC2	6.62	8.84	9.43	8.77	4.69	9.88	9.80	4.30	7.49	8.46
	M3P	8.87	8.84	9.43	9.65	5.38	8.66	7.00	6.12	6.52	10.78
	Bletchley Large - English	49.77	38.37	55.22	49.12	36.73	59.16	34.4	38.45	59.92	60.57
	Bletchley Large - Multilingual	52.36	47.44	62.96	55.96	42.79	66.59	47.0	45.75	67.54	67.55
	Bletchley XL - Multilingual	58.54	51.16	67.34	58.77	48.28	71.81	50.6	49.09	71.01	71.78
TR	mUNITER	9.21	10.17	12.16	10.54	8.33	12.88	8.79	6.75	10.87	15.07
	xUNITER	9.08	10.30	9.34	12.38	7.82	10.66	10.10	6.97	9.69	11.74
	UC2	8.32	7.69	10.44	11.64	6.03	11.47	10.81	5.74	8.81	9.90
	M3P	8.32	9.80	11.79	12.02	8.21	10.89	8.43	7.09	10.57	12.66
	Bletchley Large - English	46.51	41.63	56.68	51.57	38.44	59.71	33.7	37.06	60.47	60.88
	Bletchley Large - Multilingual	57.86	53.84	66.78	61.58	45.42	69.37	49.0	51.02	70.73	71.67
	Bletchley XL - Multilingual	64.61	57.79	70.59	66.49	50.80	74.81	52.6	55.21	76.00	74.10

Table 4: Zeroshot Image and Text Retrieval on xFlickr&CO. Results of all compared models are directly taken from the IGLUE benchmark.

Retrieval@1	Model	Language						
		DEU	SPA	IND	JPN	RUS	TUR	CMN
IR	mUNITER	12.05	13.15	5.95	6.30	5.85	1.75	11.35
	xUNITER	14.55	16.10	16.50	10.25	15.90	9.05	15.95
	UC2	28.60	15.95	14.60	24.25	20.00	7.15	31.60
	M3P	13.35	13.40	13.20	10.30	15.95	7.75	16.45
	Bletchley Large - English	55.25	63.65	53.05	51.45	64.45	57.65	52.55
	Bletchley Large - Multilingual	57.70	67.05	54.65	53.65	65.50	58.90	54.50
	Bletchley XL - Multilingual	59.60	69.30	55.50	55.60	67.50	60.20	56.00
TR	mUNITER	11.85	13.05	7.55	7.70	6.80	3.25	11.85
	xUNITER	13.25	15.10	16.75	9.85	14.80	10.05	14.80
	UC2	23.90	15.30	13.60	22.40	16.75	10.05	26.30
	M3P	11.85	12.15	12.10	9.65	14.45	8.35	14.75
	Bletchley Large - English	50.90	59.50	50.25	52.70	62.60	55.55	58.80
	Bletchley Large - Multilingual	59.45	67.55	59.25	55.25	69.35	60.85	63.30
	Bletchley XL - Multilingual	62.80	69.70	61.95	61.65	72.40	64.35	64.40

For finetuning the model on retrieval tasks we increase the image resolution to 518x518 and interpolate the image position embeddings as done in ViT (Dosovitskiy et al., 2020). We reduce the batch size to 2048, use a learning rate of 1e-5 for the whole model and simply continue to train the model on the image-text contrastive task. To prevent any sort of leakage we do not combine the finetuning datasets as commonly done, rather we finetune only on the training sets corresponding to each of the test sets.

Table 5: Finetuned Retrieval on the multilingual COCO and Multi-30k test sets. Bletchley XL significantly outperforms previously SOTA methods.

Model	COCO		Multi30k		
	JP	CN	FR	CS	DE
SMALR (Burns et al., 2020)			65.9	64.8	69.8
M3P			73.9	72.2	82.7
UC2 (Zhou et al., 2021)			83.9	81.2	84.5
MURAL	81.3		89.9	87.1	90.4
Bletchley Large - English	82.0	86.4	90.1	89.3	90.3
Bletchley Large - Multilingual	85.9	88.9	93.9	92.6	93.9
Bletchley XL - Multilingual	87.1	91.7	94.6	93.7	94.2

As expected even when finetuned the large variant of Bletchley trained on multilingual image captions outperforms the variant trained on English image captions. Moreover, the XL variant of Bletchley sets a new SOTA on the multilingual retrieval benchmarks.

5.2. English Benchmarks

To understand the impact of using multilingual image captions instead of English image captions we also finetune the model for retrieval on the English COCO and Flickr-1k sets. We follow (Karpathy and Fei-Fei, 2015) for obtaining the train and test splits for these retrieval sets.

Table 6: Finetuned Image and Text Retrieval on English COCO and Flickr test sets. Bletchley XL is SOTA even though it’s trained on multilingual image captions data

Model	COCO		Flickr	
	T → I	I → T	T → I	I → T
ALIGN	59.9	77.0	84.9	95.3
FILIP(Yao et al., 2021)	61.2	78.9	87.1	96.6
Florence(Yuan et al., 2021)	63.2	81.8	87.9	97.2
Bletchley Large - English	62.5	79.0	86.0	96.2
Bletchley Large - Multilingual	62.0	79.4	86.1	96.1
Bletchley XL - Multilingual	65.0	82.1	88.7	97.5

We don’t see a significant improvement in retrieval performance when we go from the English data to the multilingual data, rather we see a slight drop on the COCO text to image retrieval task. When we train the model on multilingual data, the batch consists of not just English sentences but sentences from other languages as well, this might effectively make the contrastive task for English sentences easier compared to when the whole batch

consisted of just English sentences. Considering that the captions in COCO tend to be fairly descriptive, our current hypothesis is that fewer English sentences in a batch might be leading to worse retrieval performance. That being said, Bletchley XL still sets a new state-of-the-art on finetuned retrieval despite being trained on the multilingual captions dataset.

5.3. Impact of Dataset Cleaning

In order to evaluate the impact of cleaning the multilingual image captions data using Bletchley Large - English which was trained on English image captions data, we trained a base sized model for 100k steps on the filtered and unfiltered multilingual datasets.

Table 7: English zeroshot retrieval metrics on Flickr & COCO. We use Recall@1 as our evaluation metrics.

Model	COCO		Flickr	
	T \rightarrow I	I \rightarrow T	T \rightarrow I	I \rightarrow T
Without filtering	18.2	31.1	37.4	53.5
With filtering	21.5	35.9	42.4	59.9

Table 8: Multilingual zeroshot retrieval metrics on Multi-30k, COCO-CN & COCO-JP. We take mean recall as the evaluation metric.

Model	COCO		Multi30k		
	JP	CN	FR	CS	DE
Without filtering	34.8	57.3	52.0	45.1	49.9
With filtering	40.5	63.7	57.5	49.8	54.9

From table 7 we see that when evaluated on English benchmarks the filtered data gives about a 2-3 pts improvement in metrics. Whereas when evaluated on the multilingual benchmarks, from table 8, the difference is even starker with about a 5 pt improvement across sets, suggesting that despite being trained only on English image captions, the model can do a decent job of filtering non-English image-captions pairs.

6. Conclusion

We have introduced a method for bootstrapping multilingual image-text data from web-crawled data, leveraging English-only NLP infrastructure as a starting point. Using this data pipeline, we argue that a simple method of adding multilingual tasks during pretraining while initializing from a text-only model suffices to achieve strong multilingual performance. We empirically verify this and build several models which achieve strong results across

common vision-language benchmarks, while reaching state-of-the-art results for multilingual retrieval tasks. Our results highlight how dataset bootstrapping can be used to build strong multilingual vision-language models from web crawled data which is predominantly English.

7. Social Impacts

In this work, we curated large-scale datasets using web documents with minimal filtering. Although this work shows a very promising direction for creating multilingual large datasets which leads to huge improvements in multilingual benchmarks, additional analysis of the data and the model needs to be done before using the model in practice.

Considerable progress has been made in ethical AI research which makes the use of multilingual models more accountable. We hope that our research and findings can lead to a better understanding of issues including but not limited to fairness, accountability, ethics, and responsibility, especially in multi-modal domains.

References

- Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, 2018.
- Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. *arXiv preprint arXiv:2201.11732*, 2022.
- Neil Burgess, Jelena Milanovic, Nigel Stephens, Konstantinos Monachopoulos, and David Mansell. Bfloat16 processing for neural networks. In *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, pages 88–91, 2019. doi: 10.1109/ARITH.2019.00022.
- Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A Plummer. Learning to scale multilingual representations for vision-language tasks. In *European Conference on Computer Vision*, pages 197–213. Springer, 2020.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.
- Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*, 2020.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain

- Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-3210. URL <http://www.aclweb.org/anthology/W16-3210>.
- Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W17-4718>.
- Google. Freebase data dumps. <https://developers.google.com/freebase/data>, 2013.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.
- Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125*, 2021.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.
- Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.
- Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360, 2019.
- Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.
- Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

- Andrew Merritt, Chenhui Chu, and Yuki Arase. A corpus for english-japanese multimodal neural machine translation with comparable sentences, 2020.
- Microsoft. Blingfire. <https://github.com/microsoft/BlingFire>, 2021.
- Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3977–3986, 2021.
- Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv:2111.10050*, 2021.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.
- Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pre-training for the masses. *arXiv preprint arXiv:2104.10972*, 2021.
- Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.
- Saurabh Tiwary. Turing bletchley: A universal image language representation model by microsoft, Nov 2021. URL <https://aka.ms/bletchley-blog>.
- Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 157–166, 2014.
- Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.
- Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Mingyang Zhou, Luwei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165, 2021.

Appendix A. Appendix

A.1. Hyperparameters used for training Bletchley

Hyperparameters		Bletchley Large	Bletchley XL
Image Encoder	Hidden Size	1024	2048
	Depth	24	36
	Head Size	16	32
	Sequence Length	257	257
Text Encoder	Hidden Size	1024	1024
	Depth	24	24
	Head Size	16	16
	Sequence Length	128	128
Weight Decay		0.01	0.01
Dropout		0.0	0.0
Adam β_1		0.9	0.9
Adam β_2		0.999	0.999
Scratch Learning Rate		1e-4	1e-4
Pretrained Learning Rate		1e-5	1e-5