MURAL: Multimodal, Multitask Retrieval Across Languages

Aashi Jain Mandy Guo Krishna Srinivasan Ting Chen Sneha Kudugunta
Chao Jia Yinfei Yang Jason Baldridge
Google Research
{aashijain, xyguo, krishnaps, iamtingchen, snehakudugunta,
chaojia, yinfeiy, jasonbaldridge}@google.com

Abstract

Both image-caption pairs and translation pairs provide the means to learn deep representations of and connections between languages. We use both types of pairs in MURAL (MUltimodal, MUltitask Representations Across Languages), a dual encoder that solves two tasks: 1) image-text matching and 2) translation pair matching. By incorporating billions of translation pairs, MURAL extends ALIGN (Jia et al., 2021)–a state-of-the-art dual encoder learned from 1.8 billion noisy image-text pairs. When using the same encoders, MURAL's performance matches or exceeds ALIGN's cross-modal retrieval performance on well-resourced languages across several datasets. More importantly, it considerably improves performance on under-resourced languages, showing that text-text learning can overcome a paucity of image-caption examples for these languages. On the Wikipedia Image-Text dataset, for example, MURAL-BASE improves zero-shot mean recall by 8.1% on average for eight under-resourced languages and by 6.8% on average when fine-tuning. We additionally show that MURAL's text representations cluster not only with respect to genealogical connections but also based on areal linguistics, such as the Balkan Sprachbund.

1 Introduction

Multilingual captions for images provide indirect but valuable associations between languages (Gella et al., 2017). Burns et al. (2020) exploit this to scale multimodal representations to support more languages with a smaller model than prior studies. More recent work learns cross encoder models with multitask training objectives (Ni et al., 2021; Zhou et al., 2021); in these, a single multimodal encoder attends to both inputs and exploits deep associations between images and captions. Unfortunately, such models do not support efficient retrieval (Geigle et al., 2021), and they use object detection, machine translation, bilingual dictionaries and many losses. In contrast, multimodal dual encoders can be learned directly on noisy, massive image-caption datasets using a simple loss based on in-batch bidirectional retrieval (Jia et al., 2021; Radford et al., 2021). These support efficient retrieval via approximate nearest neighbors search (Guo et al., 2020) and can predict similarity within and across modalities (Parekh et al., 2021).

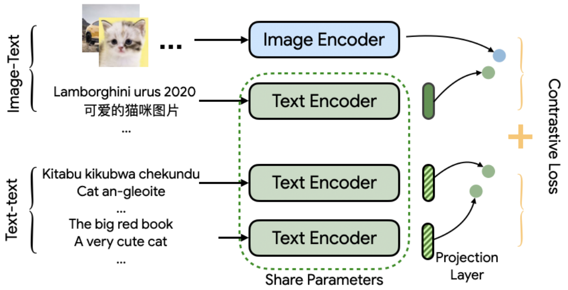


Figure 1: MURAL learns encoders for both language and images by combining both image-text matching and text-text matching tasks, using scalable dual encoder models trained with contrastive losses.

With **MURAL**: MUltimodal, MUltitask Representations Across Languages (Fig. 1), we explore dual encoder learning from both image-caption and translation pairs at massive scale: 6 billion translation pairs (Feng et al., 2020) and 1.8 billion image-caption pairs (Jia et al., 2021). We particularly seek to improve performance for under-resourced languages. Addressing this was infeasible until now because existing multilingual image-text datasets—Multi30k (Elliott et al., 2016)), STAIR (Yoshikawa et al., 2017), and XTD (Aggarwal and Kale, 2020)–support only high-resource languages. However, the recent Wikipedia Image-Text (WIT) dataset (Srinivasan et al., 2021), which covers 108 languages, addresses this gap.

Our results, as a whole, demonstrate that ALIGN, a state-of-the-art multimodal dual encoder, is improved by adding a bitext ranking objective (Yang et al., 2019a) (=MURAL). The latter matches

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3449–3463
November 7–11, 2021. ©2021 Association for Computational Linguistics

Name	Train-I	Train-T	Dev-I	Dev-T	Test-I	Test-T	#Langs
EOBT Pairs	-	500m	-	-	-	-	124
MBT Pairs†	-	6b	-	-	-	-	109
CC12m	12m	12m	-	-	-	-	1
Alt-Text†	1.8b	1.8b	-	-	-	-	110
XTD	-	-	-	-	1k	1k	7
Multi30k	29k	145k	1k	5k	1k	5k	4
MS-COCO	82k	410k	5k	25k	5k	25k	1
STAIR	82k	410k	5k	25k	5k	25k	1
WIT	11.4m	16m	5/3/1k	5/3/1k	5/3/1k	5/3/1k	108

Table 1: Dataset statistics. Counts are per language, except that Alt-Text and WIT training counts aggregate over all languages. WIT text counts are for reference descriptions. (*Key*: I=Image, T=Text; †: indicates internal datasets); see Section 2 for abbreviations and further details on each dataset.)

zero-shot image-text retrieval performance on well-resourced languages, and it dramatically improves performance on under-resourced languages. For XTD, MURAL improves recall@10 by 4% on average. On WIT zero-shot, MURAL improves mean recall by 1.7% on average for nine well-resourced languages, and by 8.1% for eight under-resourced ones. After fine-tuning on WIT, MURAL mean recall is 1.8% and 6.8% better than ALIGN, on average, for well-resourced and under-resourced languages, respectively.

We also show that the resulting dual encoder model can outperform more complex cross-encoder baseline models by a wide margin, thus obtaining stronger performance from models that support scalable retrieval. Our largest model, MURAL-LARGE, improves mean recall for zero-shot retrieval by 47.7% on average for four languages in Multi30k over M3P (Ni et al., 2021). It improves mean recall by 5.9% over UC2 (Zhou et al., 2021) for the fine-tuning setting of Multi30k. MURAL-LARGE also improves over a strong *translate-test* baseline on WIT in a zero-shot setting for well-resourced languages by 13.2% and for under-resourced ones by 9.6%.

We report results on Crisscrossed Captions (CxC) (Parekh et al., 2021), which additionally provides image-text, text-text, and image-image similarity ratings. MURAL-LARGE obtains the highest scores to date on CxC text→text and image→image retrieval. Our small ALIGN model and MURAL-LARGE model tie for best Semantic Image Similarity, which measures the correlation between model rankings and human rankings over image-image pairs.

Finally, we show that multilingual representations learned in MURAL form clusters which are influenced from areal linguistics and contact linguistics, in addition to previously shown genealogical relationships (Kudugunta et al., 2019).

2 Data

For training, we use both publicly available datasets and internal ones that are much larger. We evaluate on many publicly available image captioning datasets. Table 1 summarizes their statistics.

2.1 Training datasets

Conceptual 12M (CC12M) Changpinyo et al. (2021) is a publicly available image captioning dataset in English with 12 million pairs obtained from web images and their corresponding alt-text descriptions. CC12M loosens the strong quality filters on the earlier Conceptual Captions (CC3M) dataset (Sharma et al., 2018) to obtain greater scale.

The multilingual version of **Alt-Text** (Jia et al., 2021) is a noisy dataset with 1.8 billion images and their alt-text descriptions, covering 110 languages. Alt-Text is minimal filtered; this increases the scale and diversity of both images and languages. Fig. 2, which gives the distribution over all languages: over half the captions are English, and the top fifth of languages covers 95% of captions, so many languages still have *relatively* fewer examples.

We create an **Ensemble of Open Bilingual Translation (EOBT) Pairs** dataset by combining publicly available datasets, including Europarl (Koehn, 2005), Paracrawl (Esplà et al., 2019), Wikimatrix (Schwenk et al., 2021), and JW300 (Agić and Vulić, 2019)—see Appendix A.2 for a full list. EOBT has ≈500 million pairs across all languages.

Feng et al. (2020) mine translations from the

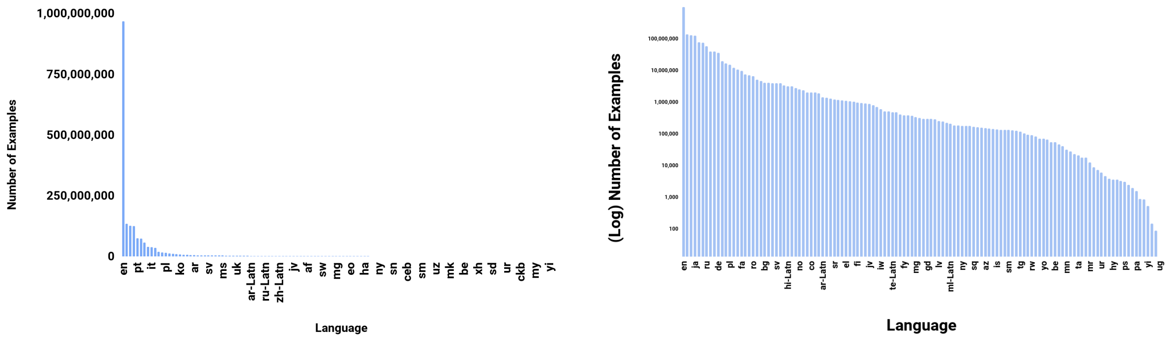


Figure 2: Alt-Text language distribution: (left) linear scale, which clearly conveys the skew toward well-resourced languages; (right) log-scale, which provides a better view of under-represented languages.

web; we call their dataset as **Mined Bilingual Translation (MBT) Pairs**. It has 6 billion pairs (up to 100 million per language) for 109 languages.

2.2 Evaluation datasets

Flickr30K (Young et al., 2014) has 31k images, with five English captions per image. **Multi30K** extends Flickr30k with German, French, and Czech captions. Elliott et al. (2016) introduces German annotations by 1) translating some Flickr30k English captions and 2) crowdsourcing new German captions for Flickr30K images. Following prior work (Burns et al., 2020), we report results on the independent 5 captions/image split. Elliott et al. (2017) and Barrault et al. (2018) further extend the dataset by collecting human translations of English Flickr30k captions to French and Czech.

MS-COCO (Lin et al., 2014) also has five human generated English captions per image. We report results on both the 1k and 5k splits defined by Karpathy and Li (2015). The **STAIR** dataset (Yoshikawa et al., 2017) adds human crowdsourced Japaneses captions for MSCOCO images.

XTD Aggarwal and Kale (2020) created the Cross-lingual Test Dataset for evaluating multimodal retrieval models. XTD does not include any training examples, but it supports retrieval evaluation on seven diverse languages.

The large-scale **Wikipedia Image Text (WIT)** dataset (Srinivasan et al., 2021) is mined from Wikipedia, covering 108 languages. The validation and test splits for WIT are not publicly available, so we partition the training data to construct new splits for WIT.[1] For most languages, we use 5k image-text pairs each for validation and test, but for less well-resourced languages, we use 3k or 1k pairs. See Appendix A.3 for details.

Crisscrossed Captions (CxC) (Parekh et al., 2021) extends the English MSCOCO 5k dev and test sets with human similarity annotations for both intra- and inter- modal tasks. As such, CxC supports evaluation for both inter-modal (image-text) and intra-modal (text-text, image-image) retrieval tasks, and correlation measures that compare model rankings with rankings derived from human similarity judgments (again, for image-text, image-image and text-text comparisons).

3 Models

ALIGN (Jia et al., 2021) is a family of multimodal dual encoders that learn to represent images and text in a shared embedding space. ALIGN's encoders are trained *from scratch* on image-text pairs via an in-batch normalized softmax loss (contrastive learning). This loss encourages the model to encode positive image-text pairs closer to each other while pushing away in-batch negative pairs.

ALIGN delivers state-of-the-art results for several datasets; however, the *Alt-Text* data used to train it is heavily skewed towards well-resourced languages (see Fig. 2). This imbalance reduces ALIGN's ability to represent under-resourced languages; we address that here by using more representative text-text translation pairs mined at scale from the web.

3.1 MURAL

MURAL extends ALIGN with a *multitask* contrastive learning objective that adds text-text contrastive losses to the image-text ones. MURAL is

[1] https://github.com/google-research-datasets/wit

trained simultaneously with two tasks of image-text (i2t) matching and text-text (t2t) matching. The text encoder is shared between these two tasks to allow transfer of multilingual learning from the text-text task to cross-modal representations. The resulting loss function is the sum of losses from both tasks.

Weighting of i2t and t2t tasks in the loss function (Parekh et al., 2021) allows the tasks to be balanced. We experiment with different weights for both tasks; our main focus is cross-modal retrieval, so we weigh the image-text task higher than the text-text task. We use the following loss function:

$$\mathcal{L} = w_{i2t} * (\mathcal{L}_{i2t} + \mathcal{L}_{t2i}) + w_{t2t} * (\mathcal{L}_{r2l} + \mathcal{L}_{l2r})$$

Each loss is an in-batch softmax of the form:

$$\mathcal{L}_{i2t} = -\frac{1}{N}\sum_{i}^{N} \log \frac{\exp(\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{y}_i)/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{y}_j)/\tau)}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are embeddings of the image in the i-th pair and the text in the j-th pair, respectively. $\mathrm{sim}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{y}/\|\boldsymbol{x}\|\|\boldsymbol{y}\|$ denotes the dot product between ℓ_2 normalized $\boldsymbol{x}$ and $\boldsymbol{y}$ (cosine similarity). N is the batch size. τ is the temperature to scale the logits. We use a similar construction for $\mathcal{L}_{t2i}, \mathcal{L}_{r2l}$, and $\mathcal{L}_{l2r}$, where l is left-text and r is right-text. The softmax temperature is shared between $\mathcal{L}_{i2t}$ and $\mathcal{L}_{t2i}$, and is learned with initial value 1.0. In $\mathcal{L}_{r2l}$ and $\mathcal{L}_{l2r}$, the temperature is fixed to 0.01. Following Feng et al. (2020), we use additive margin 0.3 in $\mathcal{L}_{r2l}$ and $\mathcal{L}_{l2r}$.

Task-specific projection heads that transform encoder representations before computing cosine similarity between inputs can improve contrastive learning (Chen et al., 2020). Similar designs have also been used for a traditional multitask setting (Guo et al., 2019). In MURAL, we use two single-layer, task-specific projection heads above the text encoder: one transforms the text embedding for image-text contrastive loss, and the other for text-text contrastive loss (more details in A.1).

Fine-tuning: single-task vs. multi-task. Our primary goal with MURAL is to improve zero-shot performance by learning with both image-text *and* text-text pairs. Nevertheless, fine-tuning has a large impact on performance for any given dataset. After initial experiments, we find that single-task fine-tuning using image-text pairs performed slightly better than multitask finetuning using co-captions. For further discussion on this comparison, see Appendix A.1. For all models, we report results using single-task fine-tuning using any available training image-text pairs for a given dataset.

3.2 Model variants

Jia et al. (2021) trains a very large model, **ALIGN-L2**, that uses EfficientNet-L2 (Tan and Le, 2019) as image encoder and BERT-Large (Devlin et al., 2019) as the text encoder. It was trained on English-only Alt-Text data. We explore smaller models and fewer training epochs to study various strategies more efficiently. For this, we use directly comparable **ALIGN-BASE** and **MURAL-BASE** models: both use EfficientNet-B5 for image encoding and BERT-Base for text. MURAL-BASE also uses text-text learning and an additional projection head for the image-text task (see Sect. 3.1). We also consider **MURAL-LARGE**, which uses EfficientNet-B7 and BERT-Large. ALIGN-BASE and MURAL-BASE have ≈300M parameters, MURAL-LARGE has ≈430M, and ALIGN-L2 has ≈840M parameters. Appendix A.1 gives more details.

Following ALIGN (Jia et al., 2021), we use LAMB optimizer (You et al., 2020) with a weight decay ratio of 1e-5. For ALIGN-BASE and MURAL-BASE, we train our models on 128 Cloud TPU V3 cores with a global batch size of 4096. The image-text task uses a learning rate of 1e-3 and the text-text task uses 1e-4. Both learning rates are linearly warmed up from zero to their final values in 10k steps and then decayed linearly to zero in 600k steps. This corresponds to only around 1.4 epochs of the Alt-Text dataset and 0.4 epochs of the MBT dataset. MURAL-LARGE is trained on 512 TPU cores (4x larger samples used in training).

We build a 250k word-piece vocabulary from the Alt-Text training data,[2] which is kept the same in all our experiments to control the changing factors.

3.3 Baseline Strategies

Our main goal is to explore the potential of large, diverse translations pairs for learning better multimodal encoders, including a *single* multilingual text encoder. We compare this strategy to the well-established, effective baselines of **translate-train** and **translate-test** using a strong Neural Machine Translation (NMT) system[3] (Yang et al., 2019b).

Translate-train: To reduce the heavy bias toward English and to support other languages for models training only on image-text pairs (e.g. for

[2] The vocabulary is built using the standard wpm library from tensorflow_text.

[3] `https://cloud.google.com/translate`

					Multi30K				MSCOCO 1K		MSCOCO 5K	
		Model	**Data**	**Type**	en	de	fr	cs	en	ja	en	ja
Zero-shot	(1)	M3P	CC3m+Wiki	CE	57.9	36.8	27.1	20.4	63.1	33.3	-	-
	(2)	ALIGN-BASE	TrTrain(AT-en)	DE	82.0	75.2	74.7	68.2	77.1	70.6	55.9	46
	(3)	ALIGN-BASE-EN	AT-en→translate-test	DE	84.3	78.9	78.3	71.1	80.0	71.5	60.6	51.9
	(4)	ALIGN-BASE	AT	DE	83.3	75.0	74.2	47.9	79.5	70.9	59.6	53.9
	(5)	MURAL-BASE	TrTrain(CC12m)+EOBT	DE	80.9	76.0	75.7	68.2	78.1	72.5	58.0	49.7
	(6)	MURAL-BASE	AT+MBT	DE	82.4	76.2	75.0	64.6	79.2	73.4	59.5	54.4
	(7)	MURAL-LARGE	AT+MBT	DE	89.2	**83.5**	**83.1**	**77.0**	**84.4**	**81.3**	67.7	**64.6**
	(8)	ALIGN-L2	AT-en	DE	**92.2**	-	-	-	-	-	**70.9**	-
Fine-tuned	(9)	SMALR	*no pretraining*	DE	74.5	69.8	65.9	64.8	81.5†	77.5†	-	-
	(10)	M3P	CC3m+Wiki	CE	87.7	82.7	73.9	72.2	88.7†	87.9†	-	-
	(11)	UC2	TrTrain(CC3m)	CE	88.2	84.5	83.9	81.2	88.1†	87.5†	-	-
	(12)	ALIGN-BASE	TrTrain(AT-en)	DE	92.2	88.5	88.1	84.5	89.0	87.5	74.8	72.5
	(13)	ALIGN-BASE	AT	DE	92.3	88.3	78.8	81.4	89.2	86.7	76.1	74.1
	(14)	MURAL-BASE	TrTrain(CC12m)+EOBT	DE	91.0	87.3	86.4	82.4	89.4	87.4	73.7	71.9
	(15)	MURAL-BASE	AT+MBT	DE	92.2	88.6	87.6	84.2	88.6	88.4	75.4	74.9
	(16)	MURAL-LARGE	AT+MBT	DE	93.8	**90.4**	**89.9**	**87.1**	**92.3**	**91.6**	81.2	**81.3**
	(17)	ALIGN-L2	AT-en	DE	**96.0**	-	-	-	-	-	**83.4**	-

Table 2: Mean recall on standard datasets. †: Numbers from UC2 paper; these were fine-tuned on MSCOCO-CN (Li et al., 2019), which has a different split than *en* and *ja*, resulting in possible train/test infiltration. SMALR MSCOCO 1K results use a different test split. (*Key*: AT=Alt-Text dataset, DE=Dual Encoder, CE=Cross Encoder, TrTrain=translate-train)

ALIGN), we artificially create image-text pairs by using the NMT system to translate English texts to other languages.[4] These additional pairs are then used to train the model – a core strategy used in UC2 (Zhou et al., 2021).

Translate-test: An alternative strategy is to train a high-performing English model and then translate non-English inputs into English, which are then encoded for cross-modal retrieval at test time.

Both strategies are highly dependent on the quality of NMT system, the languages it supports, while also incurring additional cost and complexity [5].

4 Results

We focus on:

1. Evaluating the impact of MURAL's text-text loss by comparing ALIGN-BASE and MURAL-BASE, **especially for under-resourced languages**.
2. Understanding the impact of training data scale by comparing Alt-Text+MBT to CC12M+EOBT.
3. Situating our best model, MURAL-LARGE, with respect to previous work.

We number the rows in our results tables to ease reference in our discussion and across tables.

Multi30k and MSCOCO. Table 2 compares MURAL and previous results (Burns et al., 2020; Ni et al., 2021; Zhou et al., 2021; Jia et al., 2021) in both zero-shot and fine-tuned settings.

The additional text-text task used by MURAL-BASE improves zero-shot performance on Czech, a relatively lower-resourced language, by a large margin over ALIGN-BASE (4 vs 6), 47.9 → 64.6, while nearly matching or somewhat exceeding performance on higher-resource languages.

Large, noisy pre-training greatly reduces the need for fine-tuning. M3P sees huge performance gains by fine-tuning[6] (1 vs 10), sometimes 3x the zero-shot performance. Both ALIGN-BASE and MURAL-BASE see large gains, but their zero-shot performance is already near M3P's fine-tuned performance for highly resourced languages. MURAL-LARGE's zero-shot (7) actually exceeds M3P's fine-tuned performance (10) and almost matches UC2's fine-tuned performance (11).

Even with far less data than AT+MBT, MURAL-BASE trained on CC12M+EOBT (5) has much stronger zero-shot performance than M3P (1). With fine-tuning, MURAL-BASE (CC12M+EOBT) improves on both fine-tuned M3P and UC2 (14 vs 10,11), except for Japanese. Though MURAL benefits from four times more image-text pairs than the others (CC12m > CC3M), both M3P and UC2 are more complex cross-encoder models that require

[4] Refer to appendix A.4 for more details.

[5] Translating a text query with 10 tokens adds additional latency of upto 400ms in run on CPU with a batch size of 1,

[6] Fine-tuned on Multi30k and MSCOCO combined, trained for 40k steps and learning rate sweeping of 1e-5, 5e-5, and 1e-4. Other hyperparameters are kept the same.

			Well-resourced									Under-resourced						
		Model	en	de	fr	cs	ja	zh	ru	pl	tr	tg	uz	ga	be	mg	ceb	ht	war
Zero-shot	(3)	ALIGN-BASE-EN	46.5	33.9	42.3	32.4	29.9	36.2	40.1	39.2	40.5	30.0	23.4	26.1	27.3	33.6	34.9	41.6	n/a*
	(4)	ALIGN-BASE	46.7	33.5	45.0	26.5	33.6	35.2	30.9	29.9	31.4	21.2	15.6	12.9	8.9	23.9	31.0	33.1	24.0
	(6)	MURAL-BASE	46.4	33.9	44.8	31.5	34.3	35.6	33.7	33.2	34.7	35.3	24.1	20.8	21.4	33.0	35.7	39.1	26.1
	(7)	MURAL-LARGE	**60.7**	**46.1**	**60.0**	**43.6**	**48.1**	**49.9**	**45.7**	**45.8**	**49.8**	**45.7**	**33.7**	**30.8**	**33.4**	**45.6**	**45.6**	**52.4**	**37.7**
Fine-tuned	(21)	ALIGN-BASE-EN	66.4	48.8	58.5	44.7	40.2	48.2	55.2	52.0	58.0	47.0	29.6	32.7	37.7	44.2	48.4	53.5	n/a*
	(18)	ALIGN-BASE	75.6	69.2	76.2	65.5	64.4	78.2	68.3	68.3	75.0	53.0	36.3	35.8	50.3	45.0	72.4	62.5	78.1
	(19)	MURAL-BASE	77.1	70.0	77.2	68.4	64.8	79.6	70.8	70.7	78.2	64.2	44.1	41.9	59.3	55.1	76.4	67.6	79.0
	(20)	MURAL-LARGE	**82.4**	**76.3**	**83.3**	**74.5**	**71.9**	**86.7**	**77.4**	**77.4**	**85.7**	**72.9**	**53.5**	**51.4**	**69.8**	**62.3**	**82.3**	**76.7**	**84.2**

Table 3: Mean Recall on WIT for English (en); German (de); French (fr); Czech (cs); Japanese (ja); Chinese (zh); Russian (ru); Polish (pl); Turkish (tr); Tajik (tg); Uzbek (uz); Irish (ga); Belarusian (be); Malagasy (mg); Cebuano (ceb); Haitian (ht); Waray-Waray (war); *: Translation system not available

other resources. M3P uses several different losses and it relies on a synthetic code-switched data generation process and a pretrained Faster-RCN model to obtain object bounding boxes and labels. MURAL is simpler: it is a dual encoder using just two loss types, and it works directly on raw text and pixels.

The *translate-train* strategy works well compared to using only multilingual image-text pairs (2 vs 4; 12 vs 13) and versus text-text training (2 vs 6; 12 vs 15). Given this, using translate-train (2) to increase language diversity in image-text pairs *combined* with text-text pair training (6) may yield even more gains. As a zero-shot strategy, *translate-test* also works well . This suggests that SMALR's combination of multilingual encoding and translate-test (Burns et al., 2020) may improve zero-shot performance further with MURAL (i.e., 3+6+SMALR).

Like others before, we find that training larger models on data of this scale produces remarkable gains: MURAL-LARGE obtains big improvements even over MURAL-BASE. MURAL-LARGE's results are state-of-the-art for all languages except English (where the larger, English-only ALIGN-L2 is best). MURAL-LARGE does this while–as a dual encoder–also supporting efficient retrieval. This makes a huge difference when retrieving from billions of items rather than the 1k to 5k examples of Multi30k's and MS-COCO's test sets (for which expensive, exhaustive comparisons can be performed with cross-encoders). See Geigle et al. (2021) for extensive discussion and experiments around the computational cost of cross-encoders versus dual encoders for retrieval.

Wikipedia Image Text Results. We extracted two subsets of WIT for evaluation: 1) well-resourced languages and 2) under-resourced languages (more details in Appendix A.3). There are no prior results; here, we compare MURAL with ALIGN-BASE and ALIGN-BASE-EN using the translate-test baseline. Table 3 shows MURAL-BASE achieves slightly better zero-shot performance compared to ALIGN-BASE on well-resourced languages, and a large boost on the under-represented ones. These results confirm our hypothesis of combining two tasks to address data scarcity in cross modal pairs. For WIT, MURAL-LARGE again shows that increasing model capacity improves zero-shot performance dramatically (row 7).

	Model	it	es	ru	zh	pl	tr	ko
–	mUSE+M3L	78.9	76.7	73.6	76.1	71.7	70.9	70.7
(4)	ALIGN-BASE	87.9	88.8	82.3	86.5	79.8	73.5	76.6
(6)	MURAL-BASE	88.4	89.6	83.6	88.3	86.1	84.8	82.4
(7)	MURAL-LARGE	**91.8**	**92.9**	**87.2**	**89.7**	**91.0**	**89.5**	**88.1**

Table 4: XTD zero-shot Text→Image Recall@10.

With WIT, the translate-test strategy again proves effective (row 3). It is comparable to both MURAL-BASE and ALIGN-BASE in a zero-shot setting– each wins some contests. Nevertheless, translate-test fails for the extremely under-resourced Waray-Waray language because the NMT system lacks support for it. In all, we found that 27 of WIT's 108 languages lacked NMT support. Thus, we cannot fully rely on translation systems for many under-represented languages; this further bolsters exploration into pivoting on images to overcome data scarcity. Furthermore, simple dual-encoder models are fast and simple at test-time, and thus scale better than translate-test.

Finally, both ALIGN-BASE and MURAL models benefit from fine-tuning on in-domain multilingual image-text training pairs,[7] when available; both obtain very large gains across all languages, and also easily beat the translate-test baseline fine-tuned on

[7] We fine-tune on WIT training split for 300K steps with initial learning rate 1e-4. Other hyper-parameters are the same as pre-training.

		Image → Text				Text → Image				Text → Text				Image → Image			
	Model	R@1	R@5	R@10	avg r	R@1	R@5	R@10	avg r	R@1	R@5	R@10	avg r	R@1	R@5	R@10	avg r
(22)	DE-T2T+I2T	55.9	84.2	91.8	-	41.7	72.3	83.0	-	42.4	64.9	74.0	-	38.5	73.6	84.9	-
(13)	ALIGN-BASE	67.1	89.0	94.2	3.6	50.0	77.3	85.9	11.5	43.5	64.7	73.5	45.4	42.6	76.6	86.2	16.0
(15)	MURAL-BASE	65.8	89.1	94.3	3.2	49.7	77.5	86.0	11.0	43.9	64.9	73.9	**44.9**	43.9	76.7	86.5	16.1
(16)	MURAL-LARGE	74.6	92.8	96.6	**2.3**	57.8	83.1	90.0	**9.4**	**46.5**	**67.5**	**76.1**	47.8	**50.3**	**81.8**	**90.1**	**12.4**
(17)	ALIGN-L2	**78.1**	**94.3**	**97.4**	-	**61.8**	**84.9**	**91.1**	-	45.4	66.8	75.2	-	49.4	81.4	89.1	-

Table 5: CxC Image↔text (left), Text→Text (middle), and Image→Image (right) retrieval results. DE-T2T+I2T is the strongest model of Parekh et al. (2021). DE-T2T+I2T and ALIGN-L2 are fine-tuned on MSCOCO data, while ALIGN-BASE, MURAL-BASE, and MURAL-LARGE are fine-tuned on both Multi30K and MSCOCO data).

	Model	STS	SIS	SITS
		avg ± std	avg ± std	avg ± std
(22)	DE-T2T+I2T	**74.5** ± **0.4**	74.5 ± 0.9	61.9 ± 1.3
(13)	ALIGN-BASE	72.7 ± 0.4	**80.4** ± **0.7**	63.7 ± 1.3
(15)	MURAL-BASE	73.9 ± 0.4	80.0 ± 0.7	64.0 ± 1.2
(16)	MURAL-LARGE	74.1 ± 0.4	**80.4** ± **0.7**	67.1 ± 1.3
(17)	ALIGN-L2	72.9 ± 0.4	77.2 ± 0.8	**67.6** ± **1.2**

Table 6: Semantic Simliarity using CxC.

WIT-en (18, 19, 20 vs 21).

XTD. As shown in Table 4, both ALIGN and MURAL obtain massive gains over the best strategy reported by Aggarwal and Kale (2020)—mUSE (Yang et al., 2020) with a multimodal metric loss (M3L). MURAL-LARGE shows especially strong performance across all languages. Note that we only obtained these scores after all experimentation was done on other datasets—this is methodologically important as there is neither training data nor development data for XTD.

Crisscrossed Captions. For CxC image-text retrieval (Table 5), ALIGN-L2 scores highest across all metrics; it is the largest model and was trained only on English Alt-Text. ALIGN-BASE also beats MURAL-BASE for image-text retrieval, but the latter comes back with better text-text and image-image scores. This indicates that MURAL's text-text task balances both encoders better than a loss focused only on image-text pairs. Similarly, MURAL-LARGE beats ALIGN-L2 for both text-text and image-image retrieval, despite the fact that ALIGN-L2 uses a much larger image encoder.

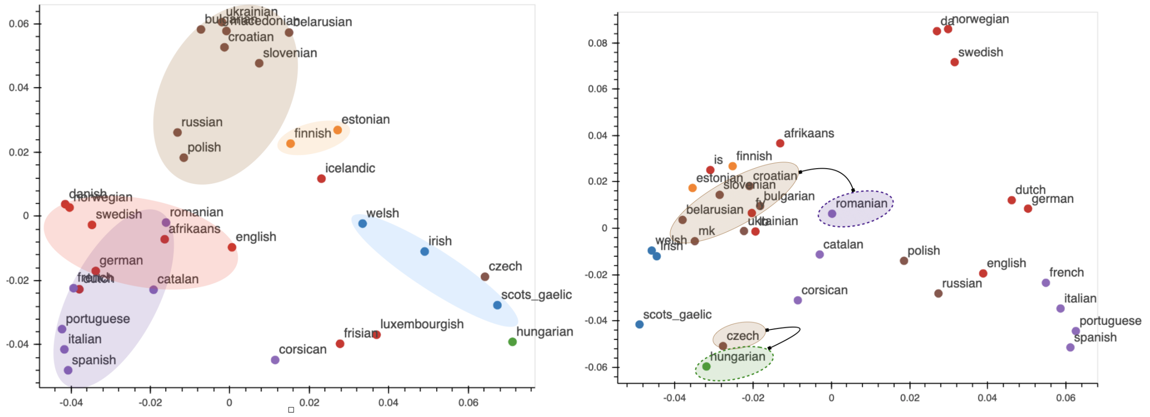
The correlation results given in Table 6 tell an interesting story. Contrary to intuition and retrieval results, Semantic Image Similarity (SIS) seems connected with multilinguality, as all Alt-Text models (ALIGN-BASE, MURAL-BASE, MURAL-LARGE) perform nearly the same (and better). DE-T2T+I2T scores the highest on Semantic Text Similarity (STS) followed closely by MURAL-LARGE. It is worth noting that DE-T2T+I2T was trained with MSCOCO co-captions which could explain its high correlation. Semantic Image-Text Similarity (SITS) agrees with Image-Text retrieval results the most, with both MURAL-LARGE and ALIGN-L2 performing considerably better than others. However, with the SITS metric, the gap between both these models diminishes, indicating that ALIGN-L2 is probably more focused on getting positive matches while MURAL-LARGE captures non-matches more effectively.

The combined retrieval and correlation lens of CxC indicates there is much more to evaluating multimodal representations than the predominant cross-modal retrieval tasks. Ranking a set of items in a manner consistent with human similarity judgments is arguably a harder task than getting a single paired item to be more similar than nearly all others. These two perspectives may reveal useful tensions in finer-grained semantic distinctions. In fact, it is with these correlation measures that we expect cross-encoders to shine compared to the retrieval-oriented dual encoders.

5 Analysis

Embedding Visualization. We visualize multilingual text representations using Singular Value Canonical Correlation Analysis (SVCCA) (Raghu et al., 2017), which allows similarity scores to be computed between languages. Using SVCCA scores computed for 100 languages, we plot a 2-dimensional visualization using Laplacian Eigenmaps (Belkin and Niyogi, 2003). Following Kudugunta et al. (2019), we do so for a subset of languages belonging to the Germanic, Romance, Slavic, Uralic, Finnic, Celtic, and Finno-Ugric language families (widely spoken in Europe and Western Asia). For a fair evaluation, we artificially create a multilingual aligned dataset by using Google's Translation system to translate 1K English captions from the Multi30K dataset to 100 languages.

Figure 3 plots the embedding in a 2-dimensional space for two models: 1) LaBSE, a multilingual *text-only* sentence representation model (Feng



(a) LaBSE representations

(b) MURAL-BASE representations

Figure 3: Visualization of text representations of LaBSE (Feng et al., 2020) and MURAL for 35 languages using laplacian eigen values and SVCCA scores. Languages are color coded based on their genealogical association.

et al., 2020) and 2) MURAL, a multingual *multimodal* model. It is evident from the visualization of LaBSE representations that embeddings group largely based on genealogical connections between languages, a phenomenon observed previously in Kudugunta et al. (2019). In addition to groupings informed by linguistic genealogy, the MURAL visualization interestingly shows some clusters which are in line with areal linguistics and contact linguistics. Notably, Romanian (ro) is closer to the Slavic languages like Bulgarian (bg), Macedonian (mk) in MURAL than it is for LaBSE, which is in line with the Balkan Sprachbund (Joseph, 1999). English (en) and French (fr) are also embedded closer to each other, reflecting their extensive contact (Haeberli, 2014). Another possible language contact brings Finnic languages, Estonian (et) and Finnish (fi), closer to the Slavic languages cluster.

The fact that MURAL pivots on images as well as translations thus appears to add an additional view on language relatedness as learned in deep representations, beyond the language family clustering observed in a text-only setting. This suggests potential future work to explore different linguistic phenomena in these representations. It also suggests that it may be worth trying to improve multimodal, multilingual representations for a given lower-resource language by pivoting on a well-resourced language that is linguistically related or which has been in significant contact with it– similar to previous studies for machine translation (Islam and Hoenen, 2013).

Retrieval Error Analysis. We analyzed zero-shot retrieved examples on WIT for ALIGN-BASE and MURAL-BASE for English (en), Hindi (hi), French (fr), and Portugese (pt). We list some examples here that indicate the value of using translation pair data for learning multilingual multimodal representations. See Appendix A.5 for more examples.

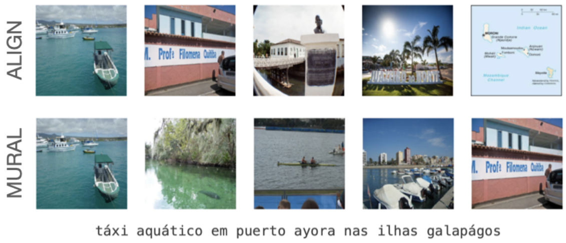


Figure 4: Portuguese: retrieval coherence. (*“Water taxi in Puerto Ayora in the Galapágos Islands.”*)

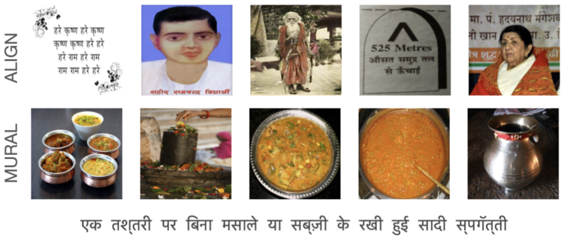


Figure 5: Hindi: Text→Image. (*“A bowl containing plain noodles without any spices or vegetables.”*)

Across languages, for both Image→Text retrieval and Text→Image, we observed that MURAL displays better fidelity to the concepts described in the image and text. For instance, in Fig. 4 ALIGN's top five results are somewhat scattered, whereas MURAL's results cohere better around boats with people (water taxis) near land (islands).

For under-resourced languages like Hindi, MURAL shows an improvement with respect to re-

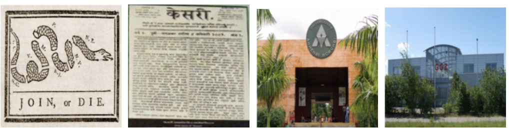


Figure 6: Image → Text examples where recognizing text in the input image would greatly help.

trieving results that are culturally more suited to the language (Fig. 5).

Finally, with both models, retrieval for some examples could greatly benefit from better recognition of words present in the images. Fig. 6 shows examples where extracting text from the images would make Image→Text almost trivial.

6 Conclusion

English provides a strong starting point for learning multilingual representations because it is so widespread and examples of English paired with other languages can be gathered well-beyond that of any other language, currently. We exploit this to train on translation pairs as a means to improve handling of multilingual inputs in cross-modal representations. With simple dual encoder models trained on large-scale datasets via contrastive learning, we obtain consistent, strong retrieval performance across all languages—especially under-resourced ones. Our error analysis also indicates that this helps increasing cultural specificity and diversity of the retrieved examples. The nuanced results we obtained for CxC also indicate that further improvements in such models might come from better calibration of the different tasks during learning. We also expect that more aggressive use of the *translate-train* strategy will straightforwardly yield further gains.

Embedding visualizations of MURAL's text representations also illustrates how languages cluster based on multimodal learning. Prior work has shown that English is not the ideal pivot language for many under-resourced languages (Mulcaire et al., 2019; Conneau and Lample, 2019). Our improvements for multilingual and multimodal models suggest further investigations into which well-resourced languages can be better pivots for learning representations for under-resourced languages. In addition to reflecting established language groupings, it also opens up possibilities of discovering new clusters. For instance, the proximity of Hungarian and Czech (Fig 3) for MURAL might be attributed to the geographical proximity of these languages, and warrants further analysis.

7 Ethics

Models trained on data collected from the web show strong results, and we are particularly encouraged by the fact that doing so leads to large improvements on under-resourced languages—and does so without requiring large amounts of (or any) image-text training data for those languages. Nevertheless, we should take utmost caution when using large datasets which went through minimal filtering processes. There could be potential biases both in the training data and models trained on them. Conscious research efforts should be made to detect and address such biases prior to releasing and using these models.

Fortunately, with prior research work in ethical AI research, it is possible to use findings from these areas to make the cross-modal models more accountable for their retrieval and broader use. We believe our findings and models can contribute positively to better understanding issues of and opportunities for addressing ethics, fairness, bias, and responsibility–especially with respect to cross-cultural issues–in language and images.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. We also thank Anosh Raj and Daphne Luong from Google for initial reviews of the manuscript. We thank Zarana Parekh for helping us setup evaluation on the CxC dataset, Orhan Firat for providing guidance on vocabulary coverage, Yuqing Chen and Apu Shah for assisting with latency metrics of the NMT models.

References

Pranav Aggarwal and Ajinkya Kale. 2020. Towards zero-shot Cross-lingual Image retrieval. *arXiv preprint arXiv:2012.05107*.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. 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, Belgium, Brussels. Association for Computational Linguistics.

Mikhail Belkin and P. Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396.

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A. Plummer. 2020. Learning to scale multilingual representations for vision-language tasks. In *The European Conference on Computer Vision (ECCV)*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *arXiv preprint arXiv:2102.08981*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.

Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulic, and Iryna Gurevych. 2021. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *CoRR*, abs/2103.11920.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845, Copenhagen, Denmark. Association for Computational Linguistics.

Hendrik J Groenewald and Liza du Plooy. 2010. Processing parallel text corpora for three South African language pairs in the Autshumato project. *AfLaT 2010*, page 27.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2019. AutoSeM: Automatic task selection and mixing in multi-task learning. 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)*, pages 3520–3531, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3887–3896. PMLR.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English.

Barry Haddow and Faheem Kirefu. 2020. PMIndia–a collection of parallel corpora of languages of India. *arXiv preprint arXiv:2001.09907*.

Eric Haeberli. 2014. When English meets French: A case study of language contact in Middle English. *Papers Dedicated to Jacques Moeschler*.

François Hernandez and Vincent Nguyen. 2020. The ubiqus English-Inuktitut system for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 213–217, Online. Association for Computational Linguistics.

Zahurul Islam and Armin Hoenen. 2013. Source and translation classification using most frequent words. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1299–1305, Nagoya, Japan. Asian Federation of Natural Language Processing.

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

Brian D Joseph. 1999. Romanian and the Balkans: Some comparative perspectives. *The Emergence of the Modern Language Sciences. Studies on the Transition from Historical-Comparative to Structural Linguistics in Honour of EFK Koerner*, 2:218–235.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Xirong Li, Chaoxi Xu, X. Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21:2347–2360.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. 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)*, pages 3912–3918, Minneapolis, Minnesota. Association for Computational Linguistics.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3977–3986.

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. 2021. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2855–2870, Online. Association for Computational Linguistics.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6076–6085.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wiki-Matrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. *arXiv preprint arXiv:2103.01913*.

Nimisha Srivastava, Rudrabha Mukhopadhyay, Prajwal K R, and C V Jawahar. 2020. IndicSpeech: Text-to-speech corpus for Indian languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6417–6422, Marseille, France. European Language Resources Association.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Francis M Tyers and Murat Serdar Alperen. 2010. South-East European Times: A parallel corpus of Balkan languages. In *Proceedings of the LREC workshop on exploitation of multilingual resources and tools for Central and (South-) Eastern European Languages*, pages 49–53.

Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019a. Improving multilingual sentence embedding using bidirectional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5370–5378. ijcai.org.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019b. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada. Association for Computational Linguistics.

Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training BERT in 76 minutes. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. 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.

A Supplementary Material

A.1 Modeling

Model variants We include further details about the main model variants we explore:

ALIGN-BASE: We use EfficientNet-B5 for the image encoder and BERT-Base Transformer for the text encoder which uses 12 layers, 12 attention heads resulting in an embedding of 768-dimensions. To match the image representation dimension of 512, we add an additional FC layer on top of the text encoder. The ALIGN-BASE model has 300M parameters in total, including 30M for EfficientNet-B5, 192M for the token embeddings, and 78M for the BERT Transformer. With this setting, we train on both the full multilingual Alt-Text dataset and the English subset, to get ALIGN-BASE and ALIGN-BASE-EN, respectively.

MURAL-BASE: The same as ALIGN-BASE, but also using text-text learning and the additional projection head for the image-text task (an FC layer that projects the text embedding from 768d to 512d). **MURAL-LARGE**: We use EfficientNet-B7 for the image encoder and BERT-Large Transformer[8] for the text encoder. To fit this model into memory, we use a 256-dimension token embedding size and project it to 1024 hidden size, which is then used by the large transformer encoder. The model uses 66M parameters for EfficientNet-B7, 64M for the token embeddings, and 300M for the BERT Transformer (=430M parameters total).

ALIGN-L2 uses an EfficientNet-L2 (=480M parameters) image encoder with a BERT-Large Transformer (300M parameters) as a text encoder. Along with the 64M parameters for token embeddings, ALIGN-L2 has 840M parameters.

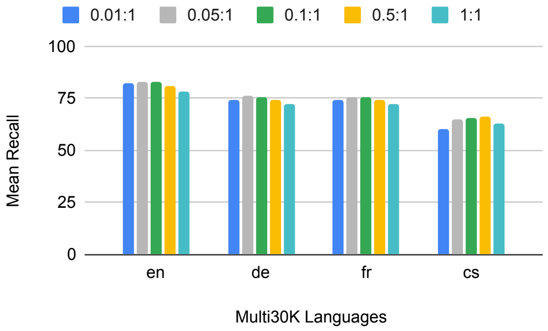


Figure 7: Zero-shot performance on Multi30K (val set) for different task weights (format: text-text weight : image-text weight). Overall, a ratio of 0.1:1 works best across all languages.

Projection Heads For MURAL, we experiment with different layers of projection heads, e.g. 1 Fully Connected (FC) layer and a Multi-Layer Perceptron with non-linearity in between the FC layers. Empirically, we find that MURAL learns better image-text representations when using single layer projection heads on top of the text-encoder, one per task.

Different Task Weights Figure 7 shows retrieval performance of models trained using different task weights in the loss function. We report zero-shot results on Multi30K val set for comparison. Weighing both t2i and i2t tasks equally (1:1) shows a consistent drop in cross-modal retrieval performance, which indicates that we need to weigh text-image task higher than the text-text task for optimal performance. From the figure we see that the ratios 0.1:1 and 0.05:1 achieve similar mean recall for t2t and i2t tasks across all Multi30K languages. In all our experiments, we use the ratio 0.1:1 for training MURAL.

Checkpoint Initialization. For MURAL, we either (1) initialize from a trained ALIGN checkpoint or (2) train both encoders from scratch. Our early experiments showed that the first strategy does not work as well. This is likely because ALIGN discards information about other languages early on because of English dominance in the Alt-Text dataset (2)–and as a result, performance on other languages is worse when training with a multitask objective. Since the model training with checkpoint initialization achieves a higher performance faster than the model trained on scratch, it offers a potential trade-off between performance and time for training. Given the early empirical results, in this paper, we always train MURAL from scratch unless otherwise stated. We stress that in the MURAL multitask model, the per-task layers on top of the text-encoders are trained from scratch in both the settings.

Finetuning Strategies: Single-task vs. Multi-task We experimented with the standard single-task fine-tuning using image-text pairs in downstream datasets like Multi30K. However, we also tried constructing text-text aligned pairs from the Multi30K dataset (e.g. by using co-caption pairs as text-text pairs), similar to the multitask strategy of Parekh et al. (2021). We found that including text-text fine-tuning slightly decreased cross-modal retrieval performance. This is may be because the

[8] 24 layers, 24 attention heads, and 1024 hidden size.

large pretrained MURAL model benefits little from seeing text-text pairs at the fine-tuning stage. This is interesting because this indicates that the training strategies at different stages affect the final performance differently. That said, it may just be that we lack the necessary evaluation data, such as multilingual variant of Crisscrossed Captions (Parekh et al., 2021) with non-English Semantic Textual Similarity scores.

A.2 Ensemble of Open Bilingual Translation (EOBT) Pairs

The complete list of open-sourced bilingual translation pairs dataset used in the construction of EOBT includes: Europarl (Koehn, 2005), Paracrawl (Esplà et al., 2019), TED57, Tanzil (Tiedemann, 2012), NewsCommentary, Wikimatrix (Schwenk et al., 2021), Wikititles, JW300 (Agić and Vulić, 2019), Opus100 (Zhang et al., 2020), SETimes (Tyers and Alperen, 2010), UNv1.0, Autshumato (Groenewald and du Plooy, 2010), PMIndia (Haddow and Kirefu, 2020), CVIT (Srivastava et al., 2020), Inuktitut (Hernandez and Nguyen, 2020), NLPC, JESC (Pryzant et al., 2018), KFTT (Neubig, 2011), ASPEC (Nakazawa et al., 2016), Flores (Guzmán et al., 2019). The data was processed in the same way as outlined in Siddhant et al. (2020).

A.3 Wikipedia Image-text Dataset

To maintain high quality text descriptions, all the splits in the WIT dataset uses the reference descriptions paired with the images. This is the text description underneath an image in a Wikipedia page. This also prevents any potential overlap with the Alt-Text training data. Similar to the Alt-Text data distribution across languages, WIT data distribution (8) is heavily skewed in favor of well-resourced languages. Refer to the Srinivasan et al. (2021) for more details on dataset collection and statistics. Since WIT's test set has been withheld for a competition, we use only the publicly available training set of approximately 37M image-text examples with 11M images. The actual available data is reduced because of our use of only reference description text as there are only about 16M reference descriptions in the WIT dataset. We split this into 108 individual language sets based on the language of the Wikipedia page. We observe that sometimes a particular language page might include a caption in an alternate language, especially an under-resourced language using a text in an well-resourced language. For e.g., an image in a Hindi page has a text caption in English. Each language set is further split into train, val and test sets. We maintain 5K image-text pairs for most of the languages but for the under-resourced we cut this down to 3K or 1K. For each language, we make sure that an image is only in one set (train, val, test).

Table 7: Image-Text data size distribution across languages for WIT and Alt-Text Datasets

# Examples	Alt-Text # Lang	WIT # Lang
$> 10^8$	4	-
$> 10^7$	11	-
$> 10^6$	22	2
$> 10^5$	37	29
$> 10^4$	18	52
$> 10^3$	12	25
$> 10^2$	4	-
$> 10^1$	2	-
Total	110	108

We also create two evaluation groups from WIT for well-resourced languages and under-resourced ones, ensuring they cover a broad range of language families and geographic areas:

- **well-resourced**: English (en), German(de), French (fr), Czech (cs), Japanese (ja), Chinese (zh), Russian (ru), Polish (pl), Turkish (tr)
- **under-resourced**: Tajik (tg), Uzbek (uz), Irish (ga), Belarusian (be), Malagasy (mg), Cebuano (ceb), Haitian (ht), Waray-Waray (war)

A.4 Translate-Train Languages

For translate-train baseline, we translate the English captions to some other well-resourced languages. For Alt-Text translation we translate English Alt-Text to German, French, Czech, Japanese, Korean, and Chinese. For CC12m dataset, we translate to languages present in the Multi30k and MSCOCO dataset namely, German, French, Czech, and Japanese. We augment the image-text pairs in English with these machine translated captions for training.

A.5 Error Analysis

We include more examples of retrieved images and text on the WIT dataset comparing ALIGN and MURAL. Some more observations-

Using color as pivots is displayed by both ALIGN and MURAL in retrieving examples, but is stronger in MURAL. For instance (Figure 11),

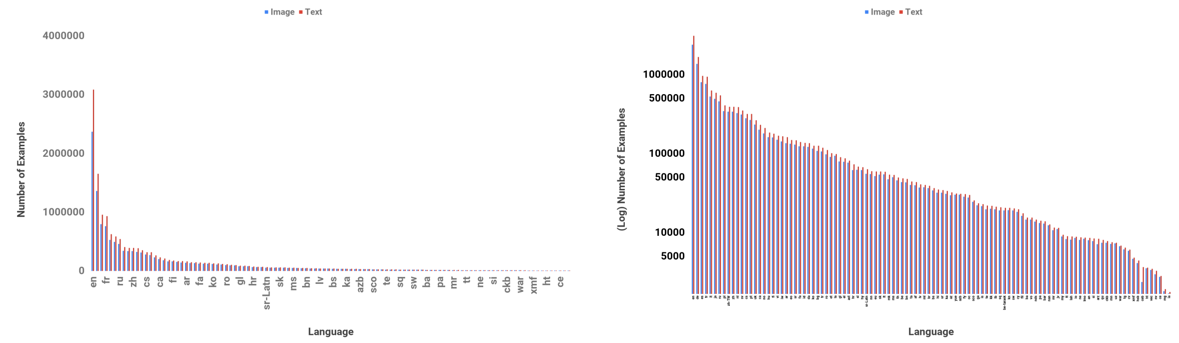


Figure 8: WIT language distribution: (left) linear scale, which clearly conveys the skew toward well-resourced languages; (right) log-scale, which provides a better view of under-represented languages.

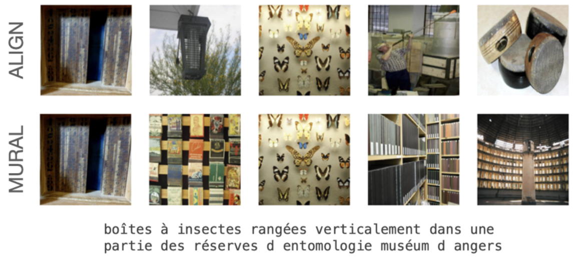

ALIGN
MURAL
boîtes à insectes rangées verticalement dans une
partie des réserves d entomologie muséum d angers

Figure 9: Fidelity to word ‘boîtes‘ (boxes) in a French caption

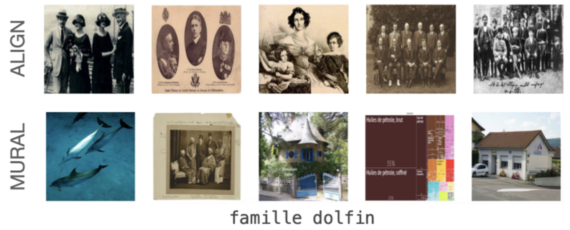


Figure 10: Fidelity to both words famille and dolfin with MURAL

identifying image of flour by its color. Also in Figure 12, ALIGN uses white and blue to retrieve captions mentioning those colors. This kind of backfires for ALIGN, because it retrieves "Blue colored lava lamp" as one of the captions. With MURAL we observe an increased object identification performance. In Figure 13, ALIGN fails to identify the sundial in the image, whereas MURAL retrieves the correct caption. We believe additional translation pairs helped MURAL learn the word for sundial in French.

For a relatively under-resourced language such as Hindi, both ALIGN and MURAL have a tendency to retrieve captions in English, which is comparatively high-resourced (Figure 14. However,

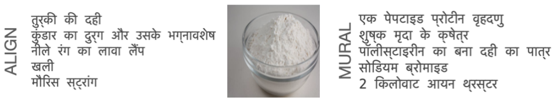


Figure 11: Color identification of the image to retrieve captions describing food that matches the white color represented in the image

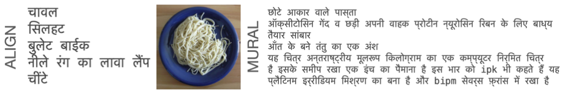

ALIGN
चावल
सिलहट
बुलेट बाईक
नीले रंग का लावा लैंप
चींटे
MURAL
छोटे आकार वाले पास्ता
ऑक्सीटोसिन गेंद व छड़ी अपनी वाहक प्रोटीन न्यूरोसिन रिबन के लिए बाध्य
तैयार सांबार
आँत के बने तंतु का एक अंश
यह चित्र अन्तराष्ट्रीय मूलरूप किलोग्राम का एक कम्प्यूटर निर्मित चित्र
है इसके समीप रखा एक इंच का पैमाना है इस भार को ipk भी कहते हैं यह
प्लैटिनम इर्रीडियम मिश्रण का बना है और bipm सेवर्स फ़्रांस में रखा है

Figure 12: Identifying the noodles by its color and shape to retrieve captions such as "rice".

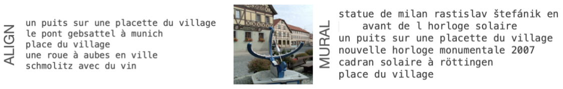

ALIGN
un puits sur une placette du village
le pont gebsattel à munich
place du village
une roue à aubes en ville
schmolitz avec du vin
MURAL
statue de milan rastislav štefánik en
avant de l horloge solaire
un puits sur une placette du village
nouvelle horloge monumentale 2007
cadran solaire à röttingen
place du village

Figure 13: MURAL learns to identify the sundial ("cadran solaire" in French) being displayed in the input image

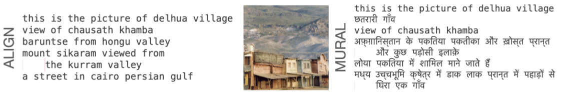

ALIGN
this is the picture of delhua village
view of chausath khamba
baruntse from hongu valley
mount sikaram viewed from
the kurram valley
a street in cairo persian gulf
MURAL
this is the picture of delhua village
छतरारी गाँव
view of chausath khamba
अफ़ग़ानिस्तान के पकतिया पकतीका और ख़ोस्त प्रान्त
और कुछ पड़ोसी इलाक़े
लोया पकतिया में शामिल माने जाते हैं
मध्य उच्चभूमि क्षेत्र में डाक लाक प्रान्त में पहाड़ों से
घिरा एक गाँव

Figure 14: For an input image, both ALIGN and MURAL tend to retrieve English captions than Hindi captions

in comparison to ALIGN, MURAL tends to infer characters and culture from the images and retrieve more Hindi captions.

Some of these observations hint us that there is definite value in using translation data to improve representations for which data is scarce. We see there are clear benefits of MURAL over ALIGN for languages other than English.