IGLUE
A Benchmark for Transfer Learning across Modalities, Tasks, and Languages
Bugliarello, Emanuele; Liu, Fangyu; Pfeiffer, Jonas; Reddy, Siva; Elliott, Desmond; Ponti, Edoardo Maria; Vuli, Ivan

Published in:
Proceedings of the 39th International Conference on Machine Learning

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
Unspecified

Citation for published version (APA):
IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages

Emanuele Bugliarello 1 2  Fangyu Liu 3  Jonas Pfeiffer 4 5  Siva Reddy 2 6  Desmond Elliott 1  
Edoardo Maria Ponti 2 6  Ivan Vulić 3

Abstract

Reliable evaluation benchmarks designed for replicability and comprehensiveness have driven progress in machine learning. Due to the lack of a multilingual benchmark, however, vision-and-language research has mostly focused on English language tasks. To fill this gap, we introduce the Image-Grounded Language Understanding Evaluation benchmark. IGLUE brings together—by both aggregating pre-existing datasets and creating new ones—visual question answering, cross-modal retrieval, grounded reasoning, and grounded entailment tasks across 20 diverse languages. Our benchmark enables the evaluation of multilingual multimodal models for transfer learning, not only in a zero-shot setting, but also in newly defined few-shot learning setups. Based on the evaluation of the available state-of-the-art models, we find that translate-test transfer is superior to zero-shot transfer and that few-shot learning is hard to harness for many tasks. Moreover, downstream performance is partially explained by the amount of available unlabelled textual data for pretraining, and only weakly by the typological distance of target–source languages. We hope to encourage future research efforts in this area by releasing the benchmark to the community.

1. Introduction

Until recently, advances in multimodal vision-and-language (V&L) modelling have predominantly focused on English or a few high-resource Indo-European languages (Elliott et al., 2016), disregarding the enormous diversity of the world’s languages and cultures (Bender, 2011; Ponti et al., 2019; Liu et al., 2021), which exacerbated the Anglo-centric bias (Jauhar et al., 2018; Ponti et al., 2020) of V&L models.

Nonetheless, this trend is reversing by virtue of a series of independent efforts in the the V&L research community: on top of expanding multilingual image–sentence retrieval to lower-resource languages (Srinivasan et al., 2021) and additional modalities (Armitage et al., 2020), we have recently witnessed pioneering work on multimodal machine translation (Yao & Wan, 2020; Huang et al., 2020), multilingual visual question answering (Pfeiffer et al., 2022), multilingual text-to-video search (Huang et al., 2021; Lei et al., 2021), multilingual visual reasoning (Liu et al., 2021), and cross-modal retrieval (Jain et al., 2021). Yet, multilingual multimodal models (Ni et al., 2021; Jain et al., 2021) are typically tested only on the task of cross-modal retrieval.

In order to incentivise and guide future research in multilingual V&L research, we introduce the Image-Grounded Language Understanding Evaluation (IGLUE) benchmark. We create IGLUE by collating current research threads in this area and extending them with two datasets for cross-lingual visual entailment (XVNLI) and image–text retrieval (xFlickr&CO), for an even more comprehensive coverage of tasks. IGLUE is the first evaluation suite for multilingual multitask V&L modelling, comprising five datasets across four structurally different tasks that require different levels of syntactic-semantic V&L understanding in cross-lingual setups: cross-modal retrieval, visual question answering, natural language inference, and visual reasoning (Figure 1).

Aiming at a representative selection of languages, IGLUE covers 20 typologically diverse languages, spanning 11 language families, 9 scripts, and 3 WALS-defined (Dryer & Haspelmath, 2013) geographical macro-areas.

Similar to previous text-based, cross-lingual transfer benchmarks, such as XGLUE (Liang et al., 2020) and XTREME (Hu et al., 2020), our V&L benchmark is designed to evaluate zero-shot transfer scenarios, where annotated training data is provided in English, but none in the target language. In addition, contrary to the above-mentioned datasets, IGLUE also provides standardised data splits to guide cross-lingual few-shot learning experiments, which
can help reduce the gap between zero-shot and supervised performance (Lauscher et al., 2020; Zhao et al., 2021). In this setup, a small number of task-annotated examples for fine-tuning are available for a given target language. We also release machine-translated versions of the test sets to enable the evaluation of ‘translate test’ cross-lingual transfer.

By virtue of the newly created IGLUE benchmark, we also run the first systematic comparative evaluation of cutting-edge multilingually pretrained V&L architectures (Ni et al., 2021; Zhou et al., 2021; Liu et al., 2021), as well as a series of representative monolingual V&L models combined with machine translation (Lu et al., 2019; Chen et al., 2020, inter alia) across a range of diverse V&L tasks and languages.

This evaluation offers new insights on the models’ core strengths and current limitations. Foreshadowing, we showcase large gaps in performance between English and transfer performance, where (to a limited extent) the gaps are more prominent in lower-resource languages and languages more distant from English. In addition, leveraging target-language in-task few shots is remarkably arduous: whereas in previous text-only experiments (Lauscher et al., 2020), ceteris paribus there are huge gains of few-shot scenarios over their zero-shot counterparts, we demonstrate that current V&L models often require thousands of examples before showing signs of improvement. Finally, performance also seems correlated with task difficulty: for instance, NLI shows the smallest gaps between English and other languages and QA benefits the most from few-shot adaptation. On the other hand, visually grounded reasoning and cross-modal retrieval appear to be harder in both respects.

**Contributions.** 1) In order to guide and inspire more work in the area of multilingual V&L research, we present a first evaluation benchmark for cross-lingual transfer learning for V&L tasks, spanning 20 languages, 5 datasets, and 4 different tasks. 2) In the process of benchmark creation, we complement existing datasets with new training and evaluation data in several languages to increase diversity and enable few-shot learning, and introduce a first multilingual dataset for visually grounded cross-lingual NLI. 3) We conduct systematic evaluations of representative V&L architectures in zero-shot and few-shot cross-lingual transfer scenarios, offering standard data splits and empirical baselines for future research. 4) Our results and additional analyses take stock of the current gaps and challenges in cross-lingual V&L research. 5) To further facilitate replicable research in this area, we re-implement the existing multilingual V&L pretrained encoders in a unified framework (VOLTA; Bugliarello et al., 2021), which also provides access to five English V&L BERTs and 12 V&L tasks. We provide data and code for the evaluation of multilingual V&L models at https://iglue-benchmark.github.io/.

2. Related Work and Motivation

**Multilingual Multimodal Learning.** Multilingual V&L research focuses on collecting resources, developing models, and evaluating systems that need to jointly reason over multilingual text and multimodal inputs, this way combining two areas of crucial importance: multilingual (Snyder & Barzilay, 2010; Ponti et al., 2019) and multimodal learning (Bernardi et al., 2016; Baltrusaitis et al., 2019). A natural overlap of the two areas is language grounding in perception (typically vision: Deng et al., 2009; Kiela et al., 2018), where the perceptual input can be considered as an inherent cross-lingual signal (Kiela et al., 2015; Gella et al., 2017; Caglayan et al., 2021). Many other research goals lie at the intersection of these two areas, such as transfer learning and modularity (Ponti et al., 2021b; Ansell et al., 2022).

However, until recently the pace of progress in multilingual multimodal learning has not gained as much momentum as V&L in monolingual settings, mostly due to the scarcity of resources for training and evaluation. Texts in most multimodal datasets are usually only available in English, or in high-resource languages (e.g., Chinese and a few major Indo-European languages; Elliott et al., 2016; 2017; Barrault et al., 2018; Wang et al., 2019b). Consequently, models trained on such datasets do not take into account linguistic diversity (Ponti et al., 2020) or cross-cultural nuances (Armitage et al., 2020; Liu et al., 2021; Yin et al., 2021).\(^1\)

The need to expand V&L research towards more languages

---

\(^1\)Some notable exceptions, which did reach beyond English, were still limited to the image–text retrieval task and mostly Indo-European languages (Rotman et al., 2018; Wehrmann et al., 2019; Huang et al., 2019; Kim et al., 2020; Su et al., 2021).

**Figure 1.** Overview of the tasks in IGLUE, which include grounded natural language inference, visual question answering, grounded reasoning, and cross-modal retrieval. Each task is associated with an example of input and output (English translations at the bottom).
has been recognised by 1) the recent creation of multilingual training and evaluation data across diverse V&L tasks and languages (Armitage et al., 2020; Srinivasan et al., 2021; Su et al., 2021; Pfeiffer et al., 2022; Liu et al., 2021; Wang et al., 2021, *inter alia*), as well as 2) the emergence of the first large multilingual-multimodal pretrained models (Ni et al., 2021; Zhou et al., 2021; Liu et al., 2021) and monolingual V&L models adapted to multiple languages (Chen et al., 2020; Pfeiffer et al., 2022). In this work, we merge and expand on these two research threads, aiming to highlight current achievements and challenges in this area and to facilitate comparative evaluations, thus bringing together the above-mentioned collective research efforts.

**Multi-Task Evaluation Benchmarks in NLP.** IGLUE has been inspired by recent text-only multi-task benchmarks for natural language understanding: these benchmarks have been proven invaluable as key drivers of recent steep performance progress of NLP. The creation of such benchmarks in monolingual settings, sparked by the pioneering and now omnipresent English-only GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019a), has been extended to other languages such as Indonesian (Wilie et al., 2020), Korean (Park et al., 2021), Russian (Shavrina et al., 2020), and Romanian (Dumitrescu et al., 2021).

IGLUE is even more related to multi-task benchmarks developed for *cross-lingual transfer settings*: XTREME (Hu et al., 2020), XGLUE (Liang et al., 2020), and XTREME-R (Ruder et al., 2021). They have brought in the spotlight the necessity to evaluate not only on a diverse set of tasks, but also on a diverse set of languages, in order to incentivise research on models that forgo monolingual limitations and generalise well in multilingual settings.

With IGLUE, we make the first leap into multimodal cross-lingual evaluation. Driven by analyses of text-only benchmarks and lessons learned from them (Ethayarajh & Jurafsky, 2020), IGLUE aims to extend sheer accuracy-driven evaluation towards other crucial aspects such as fine-tuning efficiency, sample efficiency and adaptation to low-data scenarios, and enhance multilingual inclusivity and diversity.

### 3. IGLUE Benchmark

**Design Principles.** Aiming for a comprehensive resource, we take inspiration from the best practices and design principles of existing multilingual benchmarks (Hu et al., 2020; Liang et al., 2020; Ruder et al., 2021), and adapt them to account for the unique challenges of V&L tasks. In particular, we abode by the following principles:

1. **Task Diversity.** We selected a wide spectrum of multimodal tasks that reflect multiple facets of V&L learning.
2. **Language Diversity.** The datasets included in IGLUE should cover many languages that are diverse in terms of family, geographic area, typological features, and script.

(3) **Accessibility.** The data should be released under a licence that permits use and redistribution for research purposes.

Compared to previous text-only benchmarks, V&L tasks are more time- and compute-intensive by nature. To widen usability, we created IGLUE by carefully choosing datasets and training setups that would enable quick development even by practitioners constrained by limited resources.

#### 3.1. Tasks and Datasets

Challenges in V&L learning are multi-faceted and entangled with distinct abilities: e.g., drawing inferences, reasoning over and comparing images, answering questions, and retrieving images or their captions. In IGLUE we thus represent all these abilities together with their corresponding tasks, see Table 1 and Figure 1 for an overview.

<table>
<thead>
<tr>
<th>Task</th>
<th>NLI</th>
<th>QA</th>
<th>Reason</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train/Dev data</td>
<td>SNLI</td>
<td>GQA</td>
<td>NLR2</td>
<td>Flickr30K</td>
</tr>
<tr>
<td>Test data</td>
<td>XVNLI</td>
<td>xGQA</td>
<td>MaRVL</td>
<td>xFlickr&amp;CO</td>
</tr>
</tbody>
</table>

**Table 1.** IGLUE statistics. \(^1\)Aggregated across languages. Language counts include English. MaxS: Max-Shot. F: Flickr; VG: Visual Genome; N: NLVR2; M: MaRVL; C: COCO; W: Wikipedia

We propose the new task of Cross-lingual Visual Natural Language Inference (XVNLI). It requires the model to predict if a text-hypothesis ‘entails’, ‘contradicts’, or is ‘neutral’ to an image-premise. We combine the text-only dataset SNLI (Bowman et al., 2015), with its multimodal (Xie et al., 2019) and cross-lingual (Agić & Schluter, 2018) counterparts.\(^2\) We provide new train, development, and test splits such that the test split consists of images covered by the underlying cross-lingual text-only dataset. We discard all text examples of images in the test set for which no cross-lingual version exists. The remaining images for which only English hypothesis-examples exist are randomly split into the train and development set. To mitigate data leakage between the splits, we sample based on the images; all text

\(^2\)Vu et al. (2018) formulate grounded textual entailment as the task where the image supplements the textual premise and hypothesis. We follow the established setup of Xie et al. (2019).
examples corresponding to one image are added to one split.

**xGQA.** To solve the Cross-lingual Grounded Question Answering task (xGQA; Pfeiffer et al., 2022), a model must answer several types of **structured** questions about an image. The evaluation data in 7 target languages are manually translated from the validation set of GQA (Hudson & Manning, 2019), whereas the training data are sourced from the English training set of GQA. All images were originally sampled from Visual Genome (Krishna et al., 2017). In particular, we use the English balanced training set to train our models, and evaluate on the few-shot evaluation sets defined by Pfeiffer et al. (2022) to allow for direct comparison between zero-shot and few-shot experiments.

**MaRVL.** The Multicultural Reasoning over Vision and Language dataset (Liu et al., 2021) requires to determine whether a textual description is true or false about a pair of images. This involves comparing their visual representations and reasoning about the facts in the description. The creation of MaRVL is entirely driven by native speakers. As a consequence, the descriptions are written from scratch and images are selected to be culturally relevant. The NLVR2 data (Suhr et al., 2019) in English are used for training.

**xFlickr&CO.** We create a new, multilingual evaluation set for retrieval by combining 1,000 images from the Flickr30K (Young et al., 2014) and 1,000 from the COCO (Lin et al., 2014) test split defined by Karpathy & Fei-Fei (2015) (henceforth, Karpathy split). Each image is then associated with a single caption. The English evaluation set is obtained by sampling one caption per image from the existing datasets. With the exception of Japanese (Yoshikawa et al., 2017; Nakayama et al., 2020), no other language covers the evaluation splits of both Flickr and COCO. We hence crowdsource image descriptions in 6 other languages by asking annotators to describe each image, rather than translating its English caption, to prevent biasing their description (van Miltenburg et al., 2017; Frank et al., 2018). We ask annotators to follow the Flickr30K guidelines (see Appendix A.4 for details). For the Flickr30K subsplit in German, we sample captions from the Multi30K task 2 data (Elliott et al., 2016).

**WIT.** The Wikipedia-based Image Text dataset (Srinivasan et al., 2021) collected examples from Wikipedia in 108 languages. Similar to xFlickr&CO, the tasks consist in retrieving the correct image given a textual description (image retrieval) and vice versa (text retrieval). WIT represents a very diverse set of concepts and real world entities. Text fields in WIT often tend to be descriptive, verbose and use specific terminology, different from single-line captions of common words and objects in Flickr30K and COCO. This allows us to evaluate the high-level image–sentence understanding of multilingual V&L models across a wider range of languages and real-world entities than in xFlickr&CO. For training, we randomly sample a subset of 500K captions from the English training set of WIT. For evaluation, we use the WIT test data released as part of its corresponding Kaggle competition. In particular, as most of the test languages are Indo-European, we choose the 4 national languages with the lowest Wikipedia coverage (measured in number of articles as of Jan 2022), and 6 more languages to both cover different properties (language families, scripts) and overlap with other tasks in IGLUE. We ensure that each language has at least 500 image–caption pairs, and limit English and Japanese data to 1,000 samples (see also Appendix A.5).

### 3.2. Languages

IGLUE covers 20 languages (Table 2). They are extremely diverse, as they comprise 11 distinct language families, and span across 3 of the 5 geographic macro-areas defined by Dryer & Haspelmath (2013). Moreover, they are written in a variety of scripts: Arabic, Bengali–Assamese, Chinese characters, Cyrillic, Greek, Hangul, Kanji, Latin, and Tamil. The distribution of languages per task is also shown in Table 2.

We ensured that IGLUE includes languages with different levels of unlabelled data available (Joshi et al., 2020; Blasi et al., 2022). Thus, it allows for evaluating models with different data paucity regimes during pretraining. While not all languages are covered in each task—due to the (partly) independent selection of languages in the original datasets—

---

3We expand the Flickr30K evaluation set since state-of-the-art model performance is currently close to the ceiling of human performance (Jia et al., 2021). We show that this leads to a more challenging evaluation set which requires out-of-distribution generalisation, while maintaining an efficient training regime.

4Multilingual captions for Flickr30K also exist in Dutch (van Miltenburg et al., 2018) and Mandarin (Lan et al., 2017), but the Mandarin captions do not match the expected test split. COCO has also been (partially) translated to Hindi (Mishra et al., 2021), Italian (Scialetta et al., 2019), Mandarin (Li et al., 2019b), Spanish (García, 2020) and Vietnamese (Lam et al., 2020), with similar problems. The Mandarin captions have even been used to train and evaluate M^P and UC^C in a multilingual setup. We find that 882/1,000 test images are used as training examples in other languages, and that 148 of their training images are in the expected test split. In a similar vein, we find that 946/1,000 test samples in XTD (Aggarwal et al., 2021)—which contains COCO translations in seven languages (Italian, Korean, Mandarin, Portuguese, Russian, Spanish and Turkish)—are from the COCO training split, while only 7 of them are also in the Karpathy COCO test split. These overlaps lead to considerable leakage of visual and cross-lingual information during fine-tuning and evaluation.


6While Jain et al. (2021) define a set of 8 low-resource languages for evaluation on WIT, they were sampled from the training set. Moreover, we find that several of their test images contain pictures with identifiable people, which would complicate its public release due to regulations such as the GDPR in Europe. By only relying on the true test split, we allow researchers to easily adopt the full WIT training dataset in their experimental pipelines.
Table 2. Benchmark languages and tasks. English is only used used for training. Tasks legend: ✓ ✓ train and test sets available; ✓ test-only data available; * Japanese captions in xFlickr&CO are translations from English. Languages legend: A: Asiatic; C: Congo; E: European; T: Tibetan; Austron: Austronesian. Language codes are based on the ISO 639-3 international standard.

<table>
<thead>
<tr>
<th>Language</th>
<th>Code</th>
<th>Family</th>
<th>Script</th>
<th>NLI</th>
<th>QA</th>
<th>Reasoning</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>ENG</td>
<td>Indo-E</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Arabic</td>
<td>ARB</td>
<td>Indo-E</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Bengali</td>
<td>BEN</td>
<td>Indo-E</td>
<td>Bengali</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>BUL</td>
<td>Indo-E</td>
<td>Cyrillic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Danish</td>
<td>DAN</td>
<td>Indo-E</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Estonian</td>
<td>EST</td>
<td>Uralic</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>German</td>
<td>DEU</td>
<td>Indo-E</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Greek</td>
<td>ELL</td>
<td>Indo-E</td>
<td>Greek</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>French</td>
<td>FRA</td>
<td>Indo-E</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Indonesian</td>
<td>IND</td>
<td>Indo-E</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Japanese</td>
<td>JPN</td>
<td>Indo-E</td>
<td>Kanji</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Korean</td>
<td>KOR</td>
<td>Koreanic</td>
<td>Hangul</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mandarin</td>
<td>CMN</td>
<td>Sino-T</td>
<td>Hanzi</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Portuguese</td>
<td>POR</td>
<td>Indo-E</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Russian</td>
<td>RUS</td>
<td>Indo-E</td>
<td>Cyrillic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Spanish</td>
<td>SPA</td>
<td>Indo-E</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Swahili</td>
<td>SWA</td>
<td>Niger-C</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Tamil</td>
<td>TAM</td>
<td>Dravidian</td>
<td>Tamil</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Turkish</td>
<td>TUR</td>
<td>Turkic</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>VIE</td>
<td>Austro-A</td>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

For further details about the framework, we refer to Appendix A.

4. Evaluation Framework

The new IGLUE benchmark allows for running a series of unprecedented comparative experiments and analyses. Here, we provide an overview of our experimental setup. We evaluate all the existing multilingual V&L pretrained models released so far. In particular, in order to further facilitate research and development in multilingual V&L modelling, we re-implement them in a single framework based on VOLTA (Bugliarello et al., 2021)7 in PyTorch (Paszke et al., 2019), which also provides access to five English pretrained models and twelve downstream tasks.8 For further details about the framework, we refer to Appendix B.

Pretrained Models. The state-of-the-art multilingual V&L models that we evaluate follow a BERT-like architecture (Devlin et al., 2019). Their input is the concatenation of image region features extracted with a Faster R-CNN (Ren et al., 2015) object detector and textual tokens (Sennrich et al., 2016; Wu et al., 2016). The inputs are then processed by a single stack of Transformer layers (Vaswani et al., 2017) to obtain multimodal, contextualised representations. Each model is trained to minimise multiple objectives—with minimal variations across models—to learn how to understand text (masked language modelling), vision (masked region modelling) and their co-occurrence (image–text matching). We provide only condensed summaries here, and refer the reader to the original work for further details.

• mUNITER and xUNITER. Liu et al. (2021) extend the UNITER architecture (Chen et al., 2020) multilingually by following the base approach of Ni et al. (2021): a batch of multimodal English data from Conceptual Captions (CC; Sharma et al. 2018) is alternated with a batch of text-only multilingual Wikipedia data (sampled from the 104 languages used for mBERT). The two models mainly differ in their initialisation: mUNITER from mBERT (Devlin et al., 2019) and xUNITER from XLM-R (Conneau et al., 2020).

• M$^3$P. Ni et al. (2021) further introduce multimodal code-switched training tasks where words in English captions are randomly replaced with a translation with a certain probability. This setup allows the model to explicitly align images with non-English languages. In particular, the authors obtain word translations from bilingual dictionaries in 50 languages. In each multimodal batch, data from CC is fed to the model either fully in English or with code-switched words according

7https://github.com/e-bug/volta.
8Code to reproduce our results is available online at https://github.com/e-bug/iglue.
Table 3. Performance of multimodal models trained and evaluated on English test splits of the IGLUE tasks.

<table>
<thead>
<tr>
<th>Model</th>
<th>NLI</th>
<th>QA</th>
<th>Reasoning</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>XVNL</td>
<td>xGQA</td>
<td>MarVL</td>
<td>xFlickr&amp;CO</td>
</tr>
<tr>
<td>mUNITER</td>
<td>76.38</td>
<td>54.68</td>
<td>71.91</td>
<td>44.50</td>
</tr>
<tr>
<td>xUNITER</td>
<td>75.77</td>
<td>54.83</td>
<td>71.55</td>
<td>38.45</td>
</tr>
<tr>
<td>UC 2</td>
<td>76.38</td>
<td>55.19</td>
<td>70.56</td>
<td>37.40</td>
</tr>
<tr>
<td>M 3P</td>
<td>76.89</td>
<td>53.75</td>
<td>68.22</td>
<td>31.35</td>
</tr>
<tr>
<td>LXMERT</td>
<td>76.72</td>
<td>53.04</td>
<td>68.95</td>
<td>36.85</td>
</tr>
<tr>
<td>UNITER</td>
<td>77.15</td>
<td>55.59</td>
<td>71.11</td>
<td>43.05</td>
</tr>
<tr>
<td>ViLBERT</td>
<td>77.15</td>
<td>55.82</td>
<td>71.02</td>
<td>38.70</td>
</tr>
<tr>
<td>VisualBERT</td>
<td>77.15</td>
<td>53.67</td>
<td>71.18</td>
<td>43.95</td>
</tr>
<tr>
<td>VL-BERT</td>
<td>77.49</td>
<td>55.74</td>
<td>71.72</td>
<td>39.90</td>
</tr>
</tbody>
</table>

Experimental Setup. We fine-tune all models using the AdamW optimiser (Loshchilov & Hutter, 2019) relying on the same hyper-parameters as in the controlled setup of Bugliarello et al. (2021) in combination with ‘translate test’ transfer: LXMERT (Tan & Bansal, 2019), UNITER (Chen et al., 2020), ViLBERT (Lu et al., 2019), VisualBERT (Li et al., 2019a) and VL-BERT (Su et al., 2020).

In addition to evaluating all the released multilingual V&L encoders, we also benchmark representative English V&L encoders pretrained in a controlled setup (Bugliarello et al., 2021) in combination with ‘translate test’ transfer: mUNITER, xUNITER and UC 2; and up to 100 image regions using a ResNet-101 backbone (He et al., 2016) for M 3P—both trained on Visual Genome (Krishna et al., 2017; Anderson et al., 2018). We train all models on a single NVIDIA V100 (16GB) GPU card. To increase accessibility to the benchmark by practitioners with limited computing resources: 1) we set the number of epochs so that training a baseline can be completed in less than 12 hours per dataset (see Appendix B for details); 2) in addition to reporting performance at every number of shots for few-shot learning, we also define a max-shot evaluation setup to reduce the numbers of runs to one per dataset–language pair. For both fine-tuning and few-shot experiments, we evaluate the parameter sets that yield the best validation performance. For few-shot learning, models are initialised from the English fine-tuned parameter set and further trained on the data in the target language for the given task.

Table 4. Average zero-shot and ‘translate test’ performance of multimodal models trained in English and evaluated on target languages in IGLUE. The best model for each method is in bold.

<table>
<thead>
<tr>
<th>Model</th>
<th>NLI</th>
<th>QA</th>
<th>Reasoning</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>XVNL</td>
<td>xGQA</td>
<td>MarVL</td>
<td>xFlickr&amp;CO</td>
</tr>
<tr>
<td>mUNITER</td>
<td>53.69</td>
<td>9.97</td>
<td>53.72</td>
<td>8.06</td>
</tr>
<tr>
<td>xUNITER</td>
<td>58.48</td>
<td>21.72</td>
<td>54.59</td>
<td>14.04</td>
</tr>
<tr>
<td>UC 2</td>
<td>62.05</td>
<td>29.35</td>
<td>57.28</td>
<td>20.31</td>
</tr>
<tr>
<td>M 3P</td>
<td>58.25</td>
<td>28.17</td>
<td>56.00</td>
<td>12.91</td>
</tr>
</tbody>
</table>

5. Main Results and Discussion

5.1. Zero-shot Learning

The main zero-shot results, averaged across languages, are reported in Table 4. In order to measure the gap between cross-lingual transfer and supervised learning, we also provide the results for the same models trained and evaluated on English test data in Table 3. The performance metric is accuracy for all tasks except cross-modal retrieval, which uses Recall@1. Notably, the main bottleneck for current models is the retrieval task as this requires pairing each image with every caption (Geigle et al., 2022). In xFlickr&CO, it takes us ≈3 hours per language. Note that these metric are identical whenever image–caption pairs are unique. This is the case for all of xFlickr&CO data points and for more than 90% of the samples in WIT (Appendix A.5).
strate the superiority of ‘translate test’ transfer over zero-shot model transfer via multilingual encoders. The difference between the best models for each method reaches 12.1 points for XVNLI, 21.6 for xGQA, 6.9 for MaRVL, 21.3 for xFlickr&CO IR, 19.6 for xFlickr&CO TR, 7.1 for WIT IR, and 6.5 for WIT TR. Thus, gains are especially remarkable in cross-modal retrieval tasks and grounded question answering, where the performance is nearly doubled by virtue of ‘translate test’ transfer.

**Multilingual Models.** Even within each transfer method, we observe considerable variance among individual multilingual multimodal models. For instance, for zero-shot model transfer via multilingual encoders, UC² is consistently better across the board (except for WIT) and surpasses the other models by a large margin, especially in German, French, Japanese and Mandarin—languages in which the authors had translated CC and pretrained UC² on (Appendix C.2). Remarkably, this tendency is less accentuated in MaRVL, where pretraining Mandarin multimodal data is still insufficient to tackle the out-of-distribution nature of MaRVL’s culture-specific concepts (Table 15 in Appendix C.2). Nonetheless, these results prove how the simple ‘translate pretrain’ approach of UC² can be an effective baseline for multilingual transfer in multimodal pretrained models. As for ‘translate test’ transfer, no clear winner emerges across the monolingual models, whereas mUNITER generally performs better among multilingual models limited to cross-modal retrieval tasks. We also note that massively multilingual models (top rows) are often on a par with their monolingual counterparts (bottom rows).

**Transfer Gap.** Comparing the results on the English test set (Table 3) with those averaged across the other languages (Table 4) reveals an extremely large gap in performance due to cross-lingual transfer. Considering the best multilingual encoders for each task, the gap is 14.8 points for XVNLI, 26.5 for xGQA, 14.6 for MaRVL, 24.2/23.0 for xFlickr&CO IR/TR, and 10.7/11.9 for WIT IR/TR.

**Task Complexity.** There emerge stark contrasts among the model performances in each individual task. Nonetheless, we remind the reader that the number of classes to predict—and hence random baselines—vary across tasks: XVNLI 3, xGQA 1,842, and MaRVL 2. Moreover, some tasks include a collection of languages that are more homogeneous (e.g. XVNLI contains mostly Indo-European languages) whereas others are more diverse, such as WIT. Performance is expected to suffer the most in the latter group. Moreover, focusing on cross-modal retrieval tasks, scores are higher in xFlickr&Co than in WIT. This is arguably due to the fact that, while all models were pretrained on an image–text matching task, WIT captions are distributionally distant from standard captions as they mostly describe entities.

**Explanatory Variables.** Finally, we assess the impact of a series of explanatory factors on the performance in each individual target language. More specifically, we estimate its correlation 1) with the amount of each language’s unlabelled textual data available for pretraining in terms of Wikipedia articles (as of January 2022) in Figure 2(a); and 2) with the typological similarity between the source (English) and the target language in cross-lingual transfer in Figure 2(b). The latter is calculated as the cosine similarity between vectors of morpho-syntactic features extracted from WALS (Dryer & Haspelmath, 2013) by Littell et al. (2017). According to the received wisdom from text-only cross-lingual transfer (Ponti et al., 2020; Ruder et al., 2021), both these factors usually play a pivotal role in determining the downstream performance of each target language. While we find this to partially hold true for Wikipedia size (with the possible exception of WIT), we find that correlations of typological similarity are weaker (and even negative for MaRVL), implying that models may struggle comparatively on many target languages despite different similarities to English.

### 5.2. Few-Shot Learning

As we experimented with a range of data sizes for $k$-shot fine-tuning in the target language, we 1) plot these values against the corresponding performance in Figure 3 and 2) also report the exact scores for the maximum $k$ in Table 5.
which acts as an alternative benchmark for low-compute practitioners to evaluate few-shot learning. These results aid in determining the sample efficiency of the models, i.e., how much they benefit from observing each new data point in the target language. In addition, Table 18 (Appendix C.3) reports the Area Under the Curve (AUC) for accuracy across different \( k \) values, which instead measures how consistently well a model performs independently from data size.

### Gains from Few-shot Learning

Contrary to text-only multilingual tasks, where even few examples in the target language are sufficient to significantly improve the model performance (Lauscher et al., 2020), we find that this is largely not the case in multimodal settings. Figure 3 reveals that learning curves are flat for XVNLI, MaRVL, and xFlickr&CO. Moreover, inspecting the gap between zero-shot and max-shot performance in Table 19 (Appendix C.3), gains from few-shot learning are consistent but modest.

A notable exception is xGQA, where as few as 27 examples increase performance dramatically and accuracy continues to grow with the sample size. This difference is possibly due to the highly-structured nature of xGQA, where texts are generated from templates and are extremely short, and its larger label space. To investigate the latter, we follow Pfeiffer et al. (2022) and evaluate the accuracy for each structurally different question type. Figure 4 shows that increasing the number of examples has the least impact on Verify-type questions, which consist of (binary) answers/labels \( \{\text{Yes, No}\} \). These results are in line with our few-shot experiments on the other tasks, where few-shot tuning displays negligible impact on cross-lingual transfer when the label space is small. The largest impact in xGQA is observed on Choose-type and Query-type questions, where the diversity of the label space is large. This again shows that the multilingual misalignment of the prediction space negatively impacts zero-shot results, which can be mitigated by fine-tuning even on a few examples in the target language.

On the other hand, current models may require many more examples than the maximum of \( k \)-shot we established for the other tasks. We test this hypothesis in a few-shot learning experiment with more examples in JPN and DEU on xFlickr&CO. The results in Figure 5 demonstrate that examples in the order of the thousands may be needed to achieve significant gains. This leaves open an important challenge to make future models more sample efficient, in more resource-poor settings like those considered here.
We introduced the Image-Grounded Language Understanding Evaluation (IGLUE) benchmark, aiming to facilitate modelling and evaluation in multilingual vision-and-language research. IGLUE brings together 5 datasets across 4 structurally different tasks in 20 diverse languages, and provides standardised data splits and evaluation protocols to train and evaluate in zero-shot and few-shot transfer setups.

We set strong baselines and carry out analyses that shed new light into multilingual V&L models, now enabled by the creation of the IGLUE benchmark. In particular, we find that 1) ‘translate test’ transfer vastly outperforms zero-shot transfer via multilingual encoders; 2) few-shot learning requires thousands of examples before yielding gains; 3) unlabelled textual data size and typological distance from English are weaker predictors of models performance than usually observed in text-only tasks; 4) results vary significantly across tasks, with multicultural visual reasoning and cross-modal retrieval having the largest transfer gaps. Given the diversity and manifoldness of each task, we explicitly set IGLUE as a multi-faceted benchmark, without a single average score.

Along with putting forward an evaluation suite for current and future multilingual V&L models, IGLUE also opens up new opportunities for future research that goes beyond the scope of this paper: such as analysing single-source versus multi-source transfer for V&L tasks, conducting multi-task learning, or investigating staged multi-task transfer. A longer-term outlook concerns including other tasks and languages into the benchmark, as well as reaching beyond the image-only visual modality, towards multimodal tasks that rely on processing videos (Li et al., 2021) and speech (Yang et al., 2021) in different languages.

Acknowledgements

We are grateful to Heather Lent, Semih Yagcioglu, Ilker Kesen, and Mustafa Sercan Amac for their constructive feedback and help. We also thank Krishna Srinivasan for providing us with an early release of the WIT test data and further clarifications. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199 (`Emanuele Bugliarello). The work of Fangyu Liu has been supported by Grace & Thomas C.H. Chan Cambridge Scholarship. The work of Jonas Pfeiffer has been supported by the LOEWE initiative (Hesse, Germany) within the emergenCITY center. The work of Ivan Vulić has been supported by the ERC PoC Grant MultiConvAI (no. 957356), and a research donation from Huawei. The work of Edoardo Ponti and Siva Reddy has been supported by the Facebook CIFAR AI Chair program. The annotation cost is partially funded by Cambridge Digital Humanities Digitisation/Digital Resources Awards and the NSERC Discovery Grant.

References


6. Conclusion and Outlook

We introduced the Image-Grounded Language Understanding Evaluation (IGLUE) benchmark, aiming to facilitate modelling and evaluation in multilingual vision-and-language research. IGLUE brings together 5 datasets across 4 structurally different tasks in 20 diverse languages, and provides standardised data splits and evaluation protocols to train and evaluate in zero-shot and few-shot transfer setups.

We set strong baselines and carry out analyses that shed new light into multilingual V&L models, now enabled by the creation of the IGLUE benchmark. In particular, we find that 1) ‘translate test’ transfer vastly outperforms zero-shot transfer via multilingual encoders; 2) few-shot learning requires thousands of examples before yielding gains; 3) unlabelled textual data size and typological distance from English are weaker predictors of models performance than usually observed in text-only tasks; 4) results vary significantly across tasks, with multicultural visual reasoning and cross-modal retrieval having the largest transfer gaps. Given the diversity and manifoldness of each task, we explicitly set IGLUE as a multi-faceted benchmark, without a single average score.

Along with putting forward an evaluation suite for current and future multilingual V&L models, IGLUE also opens up new opportunities for future research that goes beyond the scope of this paper: such as analysing single-source versus multi-source transfer for V&L tasks, conducting multi-task learning, or investigating staged multi-task transfer. A longer-term outlook concerns including other tasks and languages into the benchmark, as well as reaching beyond the image-only visual modality, towards multimodal tasks that rely on processing videos (Li et al., 2021) and speech (Yang et al., 2021) in different languages.

Acknowledgements

We are grateful to Heather Lent, Semih Yagcioglu, Ilker Kesen, and Mustafa Sercan Amac for their constructive feedback and help. We also thank Krishna Srinivasan for providing us with an early release of the WIT test data and further clarifications. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199 (Emanuele Bugliarello). The work of Fangyu Liu has been supported by Grace & Thomas C.H. Chan Cambridge Scholarship. The work of Jonas Pfeiffer has been supported by the LOEWE initiative (Hesse, Germany) within the emergenCITY center. The work of Ivan Vulić has been supported by the ERC PoC Grant MultiConvAI (no. 957356), and a research donation from Huawei. The work of Edoardo Ponti and Siva Reddy has been supported by the Facebook CIFAR AI Chair program. The annotation cost is partially funded by Cambridge Digital Humanities Digitisation/Digital Resources Awards and the NSERC Discovery Grant.

References


IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages


hsien Huang, T., Tseng, W.-C., tik Lee, K., Liu, D.-R.,
Huang, Z., Dong, S., Li, S.-W., Watanabe, S., rahman
Mohamed, A., and yi Lee, H. Superb: Speech processing

Yao, S. and Wan, X. Multimodal transformer for multimodal
Meeting of the Association for Computational Linguistics,
pp. 4346–4350, Online, July 2020.

Yin, D., Li, L. H., Hu, Z., Peng, N., and Chang, K.-W.
Broaden the vision: Geo-diverse visual commonsense
2129, 2021.

Yoshikawa, Y., Shigeto, Y., and Takeuchi, A. STAIR cap-
tions: 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), pp. 417–421, Vancouver, Canada, July
2017.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From
image descriptions to visual denotations: New similarity
metrics for semantic inference over event descriptions.
Transactions of the Association for Computational Lin-

Zhao, M., Zhu, Y., Shareghi, E., Vulić, I., Reichart, R., Ko-
rhonen, A., and Schütze, H. A closer look at few-shot
crosslingual transfer: The choice of shots matters. In Pro-
ceedings of the 59th Annual Meeting of the Association
for Computational Linguistics and the 11th International
Joint Conference on Natural Language Processing (Vol-
ume 1: Long Papers), pp. 5751–5767, Online, August
2021.

Zhou, M., Zhou, L., Wang, S., Cheng, Y., Li, L., Yu, Z.,
and Liu, J. Uc2: Universal cross-lingual cross-modal
vision-and-language pre-training. In Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern
A. Datasets

In this section, we provide more details about the datasets and splits that we introduce in IGLUE, focusing on reproducibility of our experiments and any follow-up work that leverages the IGLUE benchmark. We also report how few-shot samples were collected for each dataset. In particular, we use the notion of ‘annotation context’ to define what a ‘shot’ means in each dataset (see details below). Table 6 lists the size of few-shot splits across all datasets.

A.1. XVNLI

Figure 6 illustrates statistical properties of XVNLI. In Figure 6(a), we see that the distribution of labels is uniform across splits. Figure 6(b) instead shows that the distribution of sentence lengths—in number of characters—in the test split is similar across languages, and often shorter for ARB.

![Graph showing distribution of labels across splits.](image)

(a) Uniform distribution of labels across splits.

![Graph showing similar test sentence length distribution across languages.](image)

(b) Similar test sentence length distribution across languages.

*Figure 6. Statistical properties of XVNLI.*

**Few-shot Data.** We extract few-shot samples from the original SNLI test split, which was translated by Agić & Schluter (2018). This was done prior to defining our test split, and by sampling all the examples corresponding to a given image, so as to avoid data leakage between training and evaluation data. In XVNLI, multiple samples (hypotheses and answers) are associated to a given image (premise). We hence consider an image as a single shot, as it serves as the context for an annotator to create multiple samples. Following Pfeiffer et al. (2022), we randomly sample up to 48 images as contexts, resulting in up to 168 hypotheses and answers.

A.2. xGQA

**Few-shot Data.** We reuse the few-shot splits defined by Pfeiffer et al. (2022), which also include development data in the target languages. The authors sampled 48 images, giving up to 1,490 questions translated for each language.

A.3. MaRVL

**Few-shot Data.** We follow the annotation protocol of Liu et al. (2021) to collect few-shot samples. For each language, we first ask native speakers to select 10 concepts among the ones identified by the authors in order to collect in-domain training data. The annotators are then required to retrieve at least 8 images per concept. When doing so, we further provide the MaRVL test images and require them to select different ones to avoid any visual leakage between few-shot and test samples. We finally sample 8 images per concept, randomly pair them and ask annotators to write a caption that is true for two pairs and false for the other two pairs. These 8 images then constitute an ‘annotation context’ and as such we consider the resulting data as one shot (i.e. one shot, eight images and one sentence giving four data points). As image collection is the most time-consuming step in the pipeline, we expand our few-shot data by re-shuffling and re-annotating each image pair (‘$10 \times 2$’ setup), resulting in 80 data points per language. Notably, due to the challenges in collecting few-shot data in low-resource languages (i.e. Swahili and Tamil), we only provide few-shot samples for three languages (i.e. Indonesian, Mandarin and Turkish).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th># shots</th>
<th># images</th>
<th># samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>XVNLI</td>
<td># shots</td>
<td>1 5 10 20 25 48</td>
<td>1 5 10 20 25 48</td>
<td>3 18 39 75 90 168</td>
</tr>
<tr>
<td>xGQA</td>
<td># shots</td>
<td>1 5 10 20 25 48</td>
<td>1 5 10 20 25 48</td>
<td>27 155 317 594 704 1,490</td>
</tr>
<tr>
<td>MaRVL*</td>
<td># shots</td>
<td>1 2 4 10 10×2</td>
<td>8 16 32 80 80</td>
<td>4 8 16 40 80</td>
</tr>
<tr>
<td>xFlickr&amp;CO</td>
<td># shots</td>
<td>1 5 10 25 50 100</td>
<td>1 5 10 25 50 100</td>
<td>1 5 10 25 50 100</td>
</tr>
</tbody>
</table>

*Table 6. Few-shot statistics. * for MaRVL, we collected few-shot data in three of the five languages due to challenges in annotations.*
A.4. xFlickr&CO

We follow the Flickr30K guidelines to collect our multilingual annotations, as shown in Figure 7(a), which we supplement with examples from Flickr30K (Figure 7(b)). In particular, we limit captions length based on previous work: 40 characters for CMN, 140 for TUR and 100 for the rest. Figure 7(c) shows the additional guidelines provided to validators. The resulting corpora were finally verified by native speakers known by the authors. Notably, the same guidelines were used to annotate both Flickr30K and COCO images. Figure 7 shows the distribution of caption lengths.

### Image Description Guidelines

- You must describe each of the following images with one sentence in your native language.
- Please provide an accurate description of the activities, people, animals and objects you see depicted in the image.
- Each description must be a single sentence under 120 characters. Write at least 8 words.
- Make sure your caption is precise and specific about this image, and not general to describe similar ones.
- Please pay attention to grammar and spelling.
- We will accept your results if you provide a good description for all images.

#### Examples in English

1. The dog is wearing a red collar. (1)
2. White dog wearing a red hat. (2)
3. The red hat is adorned with gold epaulettes. (3)
4. The dog is angry because he is hungry. (4)
5. The dog is eating a sandwich. (5)

#### Additional Flickr30K examples used to further guide annotators.

- A man holding a bottle while a bird is perching on a clothesline.
- A woman sitting on a bench while a child is playing with a toy.
- A woman is sitting on a bench with an umbrella and a handbag.
- A man sitting on a bench with a book and a cup of coffee.
- A man is sitting on a bench with his dog and a water bottle.

(c) Instructions given to native speakers to validate the captions.

Figure 7. Annotation guidelines for xFlickr&CO.

### Few-shot Data

We use the same guidelines illustrated in Figure 7 to annotate few-shot image descriptions from 100 images samples from the Flickr30K training set. German captions are extracted from Multi30K (Elliott et al., 2016), Mandarin ones from Flickr30K-cn (Mishra et al., 2021), and Turkish ones from TasvirEt (Unal et al., 2016) – all independently collected from the English captions.

### A.5. WIT

We extract 500,000 image captions from the caption_reference_description field of the English portion of the WIT dataset. This field corresponds to the captions shown on the Wikipedia pages directly below the images, and it tends to be the most topical and relevant description (Srinivasan et al., 2021). In particular, we sample among the images that were released by the Wikimedia Foundation for the corresponding Kaggle challenge, which do not include any identifiable personal information. This challenge also provides test data that was extracted from WIT prior to its public release. We use these data to define our multilingual test splits in order to let future practitioners evaluate models that were also trained on the WIT data. Per-language statistics are listed in Table 7. We note that a given image is mapped to a single caption more than 90% of the times in every language except for English (where this happens around 70% of the times and which is not used as an evaluation language in IGLUE). Figure 9 shows the overlap between test images across languages, and Figure 10 illustrates the distribution of caption lengths across languages in our WIT test splits.

#### Few-shot Data

Due to the demanding computational resources to evaluate image-text retrieval systems, we only provide few-shot splits for this task in xFlickr&CO.

---


15While we find no overlap between train and test image URLs, we find that a few base64 image encodings match between train and test splits: 1 image for ARB, ENG and JPN; 2 images for BUL, KOR and TUR; 3 images for IND and VIE; 0 for the other languages.
Figure 9. Number of images overlapping between any two languages in the WIT test splits.

Figure 10. Test sentence length distribution in WIT.

Table 7. WIT test data statistics. % unique measures the percentage of images that only have a single caption associated with them.

<table>
<thead>
<tr>
<th>Language</th>
<th>ENG</th>
<th>ARB</th>
<th>BUL</th>
<th>DAN</th>
<th>ELL</th>
<th>EST</th>
<th>IND</th>
<th>JPN</th>
<th>KOR</th>
<th>TUR</th>
<th>VIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>% unique</td>
<td>70.9</td>
<td>91.2</td>
<td>94.9</td>
<td>93.0</td>
<td>97.2</td>
<td>93.6</td>
<td>96.0</td>
<td>90.8</td>
<td>95.8</td>
<td>95.6</td>
<td>93.7</td>
</tr>
<tr>
<td># images</td>
<td>685</td>
<td>792</td>
<td>806</td>
<td>814</td>
<td>541</td>
<td>780</td>
<td>854</td>
<td>842</td>
<td>889</td>
<td>681</td>
<td>869</td>
</tr>
<tr>
<td># samples</td>
<td>1000</td>
<td>890</td>
<td>860</td>
<td>891</td>
<td>570</td>
<td>874</td>
<td>901</td>
<td>1000</td>
<td>931</td>
<td>721</td>
<td>946</td>
</tr>
</tbody>
</table>

A.6. Ethics Statement

For MaRVL and xFlickr&CO, we collected data from native speakers spread across the world in languages that are under-represented in current vision and language datasets. Our efforts were motivated by the belief that vision and language tasks need to better reflect the linguistic (and cultural) variety present in the world. We hired workers from the crowdsourcing platform prolific.co. The annotators’ identities were anonymous to us, and each of them was paid £10-20/hour (even if their data were not ultimately used).

B. Experimental Details

Pretrained Models. Table 8 summarises the key properties of the available pretrained multilingual V&L encoders. They are all single-stream architectures (Bugliarello et al., 2021), which mostly differ in the corpora they used for pretraining.

Experimental Settings. We train—both for zero-shot fine-tuning and for few-shot experiments—each encoder end-to-end by learning a task-specific ‘head’ network for each dataset on a single NVIDIA V100 (16GB) GPU card. For each task, we use the same hyperparameters as in the controlled study of Bugliarello et al. (2021) except for maximum sequence length, which we increase to accommodate the multilingual inputs (see details per task below). We use the AdamW (Loshchilov & Hutter, 2019) optimiser with parameters $\beta_1=0.9$, $\beta_2=0.999$ and $\epsilon=1e^{-6}$. The base learning rate depends on each task (see below) but it is always first linearly warmed up for 10% of the task-specific number of steps, and then linearly decayed until up to 20 epochs of training have been reached. We also apply a weight decay of 1e-4 and gradient clipping of 1.0. Before training, we extract, for each dataset, 36 image regions using a ResNet-101 backbone (He et al., 2016) for mUNITER, xUNITER and UC$^2$; and 10–100 image regions using a ResNeXt-101 backbone (Xie et al., 2017) for M$^3$P—both trained on Visual Genome (Krishna et al., 2017; Anderson et al., 2018).

Fine-Tuning Setup. To increase accessibility to the benchmark by practitioners with limited computing resources, we fix the number of epochs for which we train our baselines so that they can be trained in less than 12 hours per dataset. We use the AdamW (Loshchilov & Hutter, 2019) optimiser with parameters $\beta_1=0.9$, $\beta_2=0.999$ and $\epsilon=1e^{-6}$. The base learning rate depends on each task (see below) but it is always first linearly warmed up for 10% of the task-specific number of steps, and then linearly decayed until up to 20 epochs of training have been reached. We also apply a weight decay of 1e-4 and gradient clipping of 1.0. Before training, we extract, for each dataset, 36 image regions using a ResNet-101 backbone (He et al., 2016) for mUNITER, xUNITER and UC$^2$; and 10–100 image regions using a ResNeXt-101 backbone (Xie et al., 2017) for M$^3$P—both trained on Visual Genome (Krishna et al., 2017; Anderson et al., 2018).

• Visual Entailment. We treat XVNL as a three-way classification problem, and apply a linear layer from the pooled representation of each Transformer encoder. We fine-tune each model using a binary cross-entropy loss with a batch size of 128 for 10 epochs. We use a learning rate of 2e-5 and max token length of 80.

• Visual QA. As in Hudson & Manning (2019), we treat

16 Except for M$^3$P, which is twice as big as the other models.
GQA as a multi-label classification task by assigning a score to each answer according to its relevancy to the ground truth answer. We apply a two-layer MLP with a GeLU activation (Hendrycks & Gimpel, 2016) function in between on top of the pooled representation. We fine-tune on the balanced split of GQA using a binary cross-entropy loss against 1,842 labels with a batch size of 256 for 5 epochs. We use a learning rate of 4e-5 and max token length of 40.

- **Visual Reasoning.** We cast this task as a binary classification problem: Given the two pooled representations from encoding the textual description with each of the two images independently, we concatenate and feed them into an MLP (same as for GQA) to predict a True or False label. We fine-tune using a binary cross-entropy loss with a batch size of 64 for 20 epochs. We use a learning rate of 1e-5 and max token length of 80.

- **Image–Text Retrieval.** We train a four-way linear classifier against the pooled representations from the true image-caption pair and three hard negatives determined offline. We optimise our models using a cross-entropy loss, except for UC$^2$ for which we found the authors’ triplet loss to work better (see Appendix C). We train with a batch size of 64 for 10 epochs for xFlickr&CO; and for 2 epochs on the larger WIT. We use a learning rate of 2e-5 and max token length of 80.

**Few-shot Setup.** We follow the same setup as in the above described fine-tuning procedure, with the exception of searching across three values for the base learning rate {1e-5, 5e-5, 1e-4} and always training for 20 epochs. The batch size is adjusted according to the number of data points available for a given split (Table 9), and models are initialised from the English fine-tuned parameter set. For each run, we select the checkpoint that achieves the largest validation performance for evaluation on the test sets.

While the validation sets are in English for every dataset except for xGQA, we found no significant difference in using machine translations of each validation set for checkpoint selection. Therefore, we decided to report results using the original validation sets as this more closely matches the few-shot scenario. Nevertheless, we release the corresponding translations for each dataset for future extensions.

**C. Reproducibility and Additional Results**

We provide more details regarding our experimental setup in order to ensure reproducibility (§C.1). Further, we report additional per-language (§C.2) and few-shot (§C.3) results, which supplement the ones provided in the main paper.

**C.1. A Unified Implementation Framework**

In order to provide a more lightweight setup for all the experiments in IGLUE, which concerns the experiments we conducted for this work as well as any future experiments by other researchers, we reimplemented M$^3$P and UC$^2$ in a unified framework based on VOLTA (Bugliarello et al., 2021), which already implements mUNITER and xUNITER, along with five English pretrained encoders. In order to verify the correctness of the implementation, we evaluated each model on Multi30K (Elliott et al., 2016) in a zero-shot scenario. The results shown in Table 10 prove the correctness of our reimplementation: Performance of both models is within a couple of points from the one reported by the original authors. In particular, we obtain similar results as Geigle et al. (2022) for M$^3$P. Moreover, we found that using a cross-entropy loss for UC$^2$ led to a drop in performance of approximately 2.5 mean Recall points in each language. Similarly, using a triplet loss for M$^3$P led to a drop in performance of nearly 3 points.

In Table 11, we report the performance of our models on xGQA as a multi-label classification task by assigning a score to each answer according to its relevancy to the ground truth answer. We apply a two-layer MLP with a GeLU activation (Hendrycks & Gimpel, 2016) function in between on top of the pooled representation. We fine-tune on the balanced split of GQA using a binary cross-entropy loss against 1,842 labels with a batch size of 256 for 5 epochs. We use a learning rate of 4e-5 and max token length of 40.

- **Visual Reasoning.** We cast this task as a binary classification problem: Given the two pooled representations from encoding the textual description with each of the two images independently, we concatenate and feed them into an MLP (same as for GQA) to predict a True or False label. We fine-tune using a binary cross-entropy loss with a batch size of 64 for 20 epochs. We use a learning rate of 1e-5 and max token length of 80.

- **Image–Text Retrieval.** We train a four-way linear classifier against the pooled representations from the true image-caption pair and three hard negatives determined offline. We optimise our models using a cross-entropy loss, except for UC$^2$ for which we found the authors’ triplet loss to work better (see Appendix C). We train with a batch size of 64 for 10 epochs for xFlickr&CO; and for 2 epochs on the larger WIT. We use a learning rate of 2e-5 and max token length of 80.

**Few-shot Setup.** We follow the same setup as in the above described fine-tuning procedure, with the exception of searching across three values for the base learning rate {1e-5, 5e-5, 1e-4} and always training for 20 epochs. The batch size is adjusted according to the number of data points available for a given split (Table 9), and models are initialised from the English fine-tuned parameter set. For each run, we select the checkpoint that achieves the largest validation performance for evaluation on the test sets.

While the validation sets are in English for every dataset except for xGQA, we found no significant difference in using machine translations of each validation set for checkpoint selection. Therefore, we decided to report results using the original validation sets as this more closely matches the few-shot scenario. Nevertheless, we release the corresponding translations for each dataset for future extensions.

**C. Reproducibility and Additional Results**

We provide more details regarding our experimental setup in order to ensure reproducibility (§C.1). Further, we report additional per-language (§C.2) and few-shot (§C.3) results, which supplement the ones provided in the main paper.

**C.1. A Unified Implementation Framework**

In order to provide a more lightweight setup for all the experiments in IGLUE, which concerns the experiments we conducted for this work as well as any future experiments by other researchers, we reimplemented M$^3$P and UC$^2$ in a unified framework based on VOLTA (Bugliarello et al., 2021), which already implements mUNITER and xUNITER, along with five English pretrained encoders. In order to verify the correctness of the implementation, we evaluated each model on Multi30K (Elliott et al., 2016) in a zero-shot scenario. The results shown in Table 10 prove the correctness of our reimplementation: Performance of both models is within a couple of points from the one reported by the original authors. In particular, we obtain similar results as Geigle et al. (2022) for M$^3$P. Moreover, we found that using a cross-entropy loss for UC$^2$ led to a drop in performance of approximately 2.5 mean Recall points in each language. Similarly, using a triplet loss for M$^3$P led to a drop in performance of nearly 3 points.

In Table 11, we report the performance of our models on xGQA as a multi-label classification task by assigning a score to each answer according to its relevancy to the ground truth answer. We apply a two-layer MLP with a GeLU activation (Hendrycks & Gimpel, 2016) function in between on top of the pooled representation. We fine-tune on the balanced split of GQA using a binary cross-entropy loss against 1,842 labels with a batch size of 256 for 5 epochs. We use a learning rate of 4e-5 and max token length of 40.

- **Visual Reasoning.** We cast this task as a binary classification problem: Given the two pooled representations from encoding the textual description with each of the two images independently, we concatenate and feed them into an MLP (same as for GQA) to predict a True or False label. We fine-tune using a binary cross-entropy loss with a batch size of 64 for 20 epochs. We use a learning rate of 1e-5 and max token length of 80.

- **Image–Text Retrieval.** We train a four-way linear classifier against the pooled representations from the true image-caption pair and three hard negatives determined offline. We optimise our models using a cross-entropy loss, except for UC$^2$ for which we found the authors’ triplet loss to work better (see Appendix C). We train with a batch size of 64 for 10 epochs for xFlickr&CO; and for 2 epochs on the larger WIT. We use a learning rate of 2e-5 and max token length of 80.

**Few-shot Setup.** We follow the same setup as in the above described fine-tuning procedure, with the exception of searching across three values for the base learning rate {1e-5, 5e-5, 1e-4} and always training for 20 epochs. The batch size is adjusted according to the number of data points available for a given split (Table 9), and models are initialised from the English fine-tuned parameter set. For each run, we select the checkpoint that achieves the largest validation performance for evaluation on the test sets.

While the validation sets are in English for every dataset except for xGQA, we found no significant difference in using machine translations of each validation set for checkpoint selection. Therefore, we decided to report results using the original validation sets as this more closely matches the few-shot scenario. Nevertheless, we release the corresponding translations for each dataset for future extensions.

**C. Reproducibility and Additional Results**

We provide more details regarding our experimental setup in order to ensure reproducibility (§C.1). Further, we report additional per-language (§C.2) and few-shot (§C.3) results, which supplement the ones provided in the main paper.

**C.1. A Unified Implementation Framework**

In order to provide a more lightweight setup for all the experiments in IGLUE, which concerns the experiments we conducted for this work as well as any future experiments by other researchers, we reimplemented M$^3$P and UC$^2$ in a unified framework based on VOLTA (Bugliarello et al., 2021), which already implements mUNITER and xUNITER, along with five English pretrained encoders. In order to verify the correctness of the implementation, we evaluated each model on Multi30K (Elliott et al., 2016) in a zero-shot scenario. The results shown in Table 10 prove the correctness of our reimplementation: Performance of both models is within a couple of points from the one reported by the original authors. In particular, we obtain similar results as Geigle et al. (2022) for M$^3$P. Moreover, we found that using a cross-entropy loss for UC$^2$ led to a drop in performance of approximately 2.5 mean Recall points in each language. Similarly, using a triplet loss for M$^3$P led to a drop in performance of nearly 3 points.

In Table 11, we report the performance of our models on
Table 12. Zero-shot per-language results on WIT (IR and TR).

<table>
<thead>
<tr>
<th>Lang</th>
<th>Model</th>
<th>0</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>25</th>
<th>48</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARB</td>
<td>mUNITER</td>
<td>46.73</td>
<td>46.99</td>
<td>46.39</td>
<td>49.40</td>
<td>47.16</td>
<td>48.97</td>
<td>46.91</td>
</tr>
<tr>
<td>UNITER</td>
<td>51.98</td>
<td>52.32</td>
<td>52.32</td>
<td>54.81</td>
<td>54.55</td>
<td>53.95</td>
<td>54.04</td>
<td></td>
</tr>
<tr>
<td>UC^2</td>
<td>56.19</td>
<td>58.53</td>
<td>57.65</td>
<td>57.82</td>
<td>59.11</td>
<td>58.51</td>
<td>56.87</td>
<td></td>
</tr>
<tr>
<td>M^P</td>
<td>55.24</td>
<td>54.98</td>
<td>55.07</td>
<td>56.19</td>
<td>55.24</td>
<td>56.79</td>
<td>56.01</td>
<td></td>
</tr>
<tr>
<td>SPA</td>
<td>mUNITER</td>
<td>56.96</td>
<td>56.79</td>
<td>57.47</td>
<td>57.30</td>
<td>57.53</td>
<td>57.65</td>
<td>57.73</td>
</tr>
<tr>
<td>UNITER</td>
<td>58.94</td>
<td>59.02</td>
<td>58.94</td>
<td>60.74</td>
<td>59.88</td>
<td>59.11</td>
<td>60.22</td>
<td></td>
</tr>
<tr>
<td>UC^2</td>
<td>57.47</td>
<td>57.65</td>
<td>61.34</td>
<td>59.79</td>
<td>62.63</td>
<td>59.79</td>
<td>62.80</td>
<td></td>
</tr>
<tr>
<td>M^P</td>
<td>58.58</td>
<td>58.33</td>
<td>58.94</td>
<td>60.05</td>
<td>59.19</td>
<td>59.28</td>
<td>60.40</td>
<td></td>
</tr>
<tr>
<td>RUS</td>
<td>mUNITER</td>
<td>53.68</td>
<td>59.45</td>
<td>59.19</td>
<td>59.54</td>
<td>59.36</td>
<td>59.62</td>
<td>59.36</td>
</tr>
<tr>
<td>UNITER</td>
<td>63.32</td>
<td>63.49</td>
<td>64.26</td>
<td>63.83</td>
<td>64.26</td>
<td>64.09</td>
<td>64.52</td>
<td></td>
</tr>
<tr>
<td>UC^2</td>
<td>69.67</td>
<td>69.67</td>
<td>69.67</td>
<td>69.76</td>
<td>69.84</td>
<td>69.76</td>
<td>69.76</td>
<td></td>
</tr>
<tr>
<td>M^P</td>
<td>56.36</td>
<td>56.52</td>
<td>57.99</td>
<td>58.33</td>
<td>57.82</td>
<td>57.65</td>
<td>58.09</td>
<td></td>
</tr>
<tr>
<td>FRA</td>
<td>mUNITER</td>
<td>51.72</td>
<td>51.72</td>
<td>51.46</td>
<td>50.32</td>
<td>52.32</td>
<td>51.46</td>
<td>51.80</td>
</tr>
<tr>
<td>UNITER</td>
<td>59.71</td>
<td>59.54</td>
<td>59.54</td>
<td>61.94</td>
<td>61.34</td>
<td>63.40</td>
<td>63.40</td>
<td></td>
</tr>
<tr>
<td>UC^2</td>
<td>64.86</td>
<td>64.43</td>
<td>64.52</td>
<td>64.95</td>
<td>65.38</td>
<td>65.38</td>
<td>65.29</td>
<td></td>
</tr>
<tr>
<td>M^P</td>
<td>62.54</td>
<td>62.46</td>
<td>62.72</td>
<td>63.23</td>
<td>62.46</td>
<td>62.97</td>
<td>62.46</td>
<td></td>
</tr>
</tbody>
</table>

The zero-shot splits of xGQA, for a direct comparison with the standard GQA test split, is significantly better performance: our best model, UC^2 gains over 10 points compared to the best model from Pfeiffer et al. (2022). In addition, our M^P model is 13.4% more accurate than the one trained by the authors. We believe the main reason for this considerable performance gain is the deeper prediction head: While the authors place a linear classification head, our models use a non-linear two-layer feed-forward head on top of the Transformer representations.

Table 13. Full per-language results on XVNLI.

<table>
<thead>
<tr>
<th>XVNLI Lang</th>
<th>Model</th>
<th>0</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>25</th>
<th>48</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARB</td>
<td>mUNITER</td>
<td>46.73</td>
<td>46.99</td>
<td>46.39</td>
<td>49.40</td>
<td>47.16</td>
<td>48.97</td>
<td>46.91</td>
</tr>
<tr>
<td>UNITER</td>
<td>51.98</td>
<td>52.32</td>
<td>52.32</td>
<td>54.81</td>
<td>54.55</td>
<td>53.95</td>
<td>54.04</td>
<td></td>
</tr>
<tr>
<td>UC^2</td>
<td>56.19</td>
<td>58.53</td>
<td>57.65</td>
<td>57.82</td>
<td>59.11</td>
<td>58.51</td>
<td>56.87</td>
<td></td>
</tr>
<tr>
<td>M^P</td>
<td>55.24</td>
<td>54.98</td>
<td>55.07</td>
<td>56.19</td>
<td>55.24</td>
<td>56.79</td>
<td>56.01</td>
<td></td>
</tr>
<tr>
<td>SPA</td>
<td>mUNITER</td>
<td>56.96</td>
<td>56.79</td>
<td>57.47</td>
<td>57.30</td>
<td>57.53</td>
<td>57.65</td>
<td>57.73</td>
</tr>
<tr>
<td>UNITER</td>
<td>58.94</td>
<td>59.02</td>
<td>58.94</td>
<td>60.74</td>
<td>59.88</td>
<td>59.11</td>
<td>60.22</td>
<td></td>
</tr>
<tr>
<td>UC^2</td>
<td>57.47</td>
<td>57.65</td>
<td>61.34</td>
<td>59.79</td>
<td>62.63</td>
<td>59.79</td>
<td>62.80</td>
<td></td>
</tr>
<tr>
<td>M^P</td>
<td>58.58</td>
<td>58.33</td>
<td>58.94</td>
<td>60.05</td>
<td>59.19</td>
<td>59.28</td>
<td>60.40</td>
<td></td>
</tr>
<tr>
<td>RUS</td>
<td>mUNITER</td>
<td>53.68</td>
<td>59.45</td>
<td>59.19</td>
<td>59.54</td>
<td>59.36</td>
<td>59.62</td>
<td>59.36</td>
</tr>
<tr>
<td>UNITER</td>
<td>63.32</td>
<td>63.49</td>
<td>64.26</td>
<td>63.83</td>
<td>64.26</td>
<td>64.09</td>
<td>64.52</td>
<td></td>
</tr>
<tr>
<td>UC^2</td>
<td>69.67</td>
<td>69.67</td>
<td>69.67</td>
<td>69.76</td>
<td>69.84</td>
<td>69.76</td>
<td>69.76</td>
<td></td>
</tr>
<tr>
<td>M^P</td>
<td>56.36</td>
<td>56.52</td>
<td>57.99</td>
<td>58.33</td>
<td>57.82</td>
<td>57.65</td>
<td>58.09</td>
<td></td>
</tr>
<tr>
<td>FRA</td>
<td>mUNITER</td>
<td>51.72</td>
<td>51.72</td>
<td>51.46</td>
<td>50.32</td>
<td>52.32</td>
<td>51.46</td>
<td>51.80</td>
</tr>
<tr>
<td>UNITER</td>
<td>59.71</td>
<td>59.54</td>
<td>59.54</td>
<td>61.94</td>
<td>61.34</td>
<td>63.40</td>
<td>63.40</td>
<td></td>
</tr>
<tr>
<td>UC^2</td>
<td>64.86</td>
<td>64.43</td>
<td>64.52</td>
<td>64.95</td>
<td>65.38</td>
<td>65.38</td>
<td>65.29</td>
<td></td>
</tr>
<tr>
<td>M^P</td>
<td>62.54</td>
<td>62.46</td>
<td>62.72</td>
<td>63.23</td>
<td>62.46</td>
<td>62.97</td>
<td>62.46</td>
<td></td>
</tr>
</tbody>
</table>

C.2. Performance per Target Language

Tables 12 to 17 present language-specific results for each dataset and number of shots. Retrieval performance in WIT is relatively constant across languages, likely due to the different distribution of captions and images in the Wikipedia domain. Different from other datasets (see below), this is true even for UC^2 in the JPN split, even though the model was pretrained on large-scale multimodal data in Japanese.

In XVNLI and xGQA, performance peaks for French and German, respectively. While typologically more similar and
Table 16. Full per-language results on xFlickr&CO IR.

<table>
<thead>
<tr>
<th>Lang</th>
<th>xFlickr&amp;CO IR</th>
<th>Model</th>
<th>0</th>
<th>1</th>
<th>5</th>
<th>25</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEU</td>
<td>mUNITER</td>
<td>12.05</td>
<td>10.50</td>
<td>11.90</td>
<td>11.95</td>
<td>11.95</td>
<td>11.95</td>
<td>12.05</td>
</tr>
<tr>
<td></td>
<td>UC2</td>
<td>28.60</td>
<td>27.40</td>
<td>28.15</td>
<td>28.15</td>
<td>28.15</td>
<td>28.15</td>
<td>28.60</td>
</tr>
</tbody>
</table>

Table 17. Full per-language results on xFlickr&CO TR.

<table>
<thead>
<tr>
<th>Lang</th>
<th>xFlickr&amp;CO TR</th>
<th>Model</th>
<th>0</th>
<th>1</th>
<th>5</th>
<th>25</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEU</td>
<td>mUNITER</td>
<td>11.85</td>
<td>10.65</td>
<td>11.70</td>
<td>11.70</td>
<td>11.70</td>
<td>11.65</td>
<td>11.85</td>
</tr>
<tr>
<td></td>
<td>xUNITER</td>
<td>13.25</td>
<td>12.95</td>
<td>13.20</td>
<td>13.15</td>
<td>13.15</td>
<td>13.15</td>
<td>13.25</td>
</tr>
<tr>
<td></td>
<td>UC2</td>
<td>23.90</td>
<td>22.25</td>
<td>23.45</td>
<td>23.45</td>
<td>23.45</td>
<td>23.45</td>
<td>23.90</td>
</tr>
<tr>
<td></td>
<td>M4P</td>
<td>11.85</td>
<td>10.65</td>
<td>12.35</td>
<td>10.60</td>
<td>12.60</td>
<td>11.85</td>
<td>12.55</td>
</tr>
</tbody>
</table>

geographically closer to English, we find that the amount of multilingual data, shown in Figure 11, is also indicative of the final performance. For instance, both mUNITER and xUNITER—which were pretrained on the same data but with different multilingual sampling ratios during text-only and multimodal pretraining—perform on par or better in Russian than in Spanish and Portuguese, respectively. Moreover, UC2 consistently achieves the highest performance in French, German and Mandarin, which are 3/5 Conceptual Captions translations on which the model was pretrained.

On MaRVL, this is also the case for the Mandarin split, although UC2 now reaches lower gains relative to other models in the other datasets. This shows how the out-of-distribution nature of the images in MaRVL limits the efficacy of ‘translate train’ approaches for zero-shot transfer here. Interestingly, Mandarin is also the language in which the other models perform the worst, indicating that large-scale multimodal data in the target language can still benefit downstream performance. Overall, due to the domain shifts present in MaRVL, the performance across all languages and models is comparatively low.

Although being in-distribution, performance in xFlickr&CO is consistently low in every language, and especially in Turkish. More once, UC2 largely outperforms the other systems in DEU, JPN and CMN but is on par in the other languages.

All in all, the resulting differences between the best- and worst-performing languages are considerable, amounting 13.5 points for XVNLI, 22.9 points for xGQA, and 22.5 points for xFlickr&Co, even though the visual input is language-independent in these evaluation sets.

C.3. Additional Results for Few-shot Learning

Area under the Curve. To compare the overall performance of different models, we compute a normalised Area Under Curve (AUC) score for the few-shot setup (plotted in
Table 18. Few-shot AUC (in %) averaged across languages.

<table>
<thead>
<tr>
<th>Model</th>
<th>NLI</th>
<th>QA</th>
<th>Reasoning</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>XVNNLI</td>
<td>xGQA</td>
<td>MaRVL</td>
<td>xFlickr&amp;CO</td>
</tr>
<tr>
<td>mUNITER</td>
<td>54.10</td>
<td>33.67</td>
<td>54.68</td>
<td>8.36</td>
</tr>
<tr>
<td>xUNITER</td>
<td>59.78</td>
<td>38.10</td>
<td>56.82</td>
<td>14.05</td>
</tr>
<tr>
<td>UC²</td>
<td>63.43</td>
<td>40.43</td>
<td>57.11</td>
<td>19.24</td>
</tr>
<tr>
<td>M³P</td>
<td>59.07</td>
<td>38.89</td>
<td>49.96</td>
<td>12.90</td>
</tr>
</tbody>
</table>

Table 19. Difference in accuracy between the most-shots and zero-shot setups averaged across languages. For MaRVL, accuracy is averaged across the 3/5 languages with few-shot training data.

<table>
<thead>
<tr>
<th>Model</th>
<th>NLI</th>
<th>QA</th>
<th>Reasoning</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>XVNNLI</td>
<td>xGQA</td>
<td>MaRVL</td>
<td>xFlickr&amp;CO</td>
</tr>
<tr>
<td>mUNITER</td>
<td>↑ 0.26</td>
<td>↑ 27.24</td>
<td>↓ 0.31</td>
<td>↑ 0.48</td>
</tr>
<tr>
<td>xUNITER</td>
<td>↑ 2.07</td>
<td>↑ 18.96</td>
<td>↑ 2.87</td>
<td>↑ 0.26</td>
</tr>
<tr>
<td>UC²</td>
<td>↑ 1.63</td>
<td>↑ 13.60</td>
<td>↑ 1.04</td>
<td>↑ 0.52</td>
</tr>
<tr>
<td>M³P</td>
<td>↑ 1.11</td>
<td>↑ 12.87</td>
<td>↓ 6.21</td>
<td>↓ 0.30</td>
</tr>
</tbody>
</table>

Figure 3. For a given model $f$ and dataset $D$ with $N_D$ possible shots, each with $s_k$ examples, we approximate AUC according to the trapezoidal rule:

$$\text{AUC} (f, D) = \frac{1}{s_ND} \sum_{k=1}^{N_D} \frac{f(s_{k-1}) + f(s_k)}{2} \Delta s_k.$$ (1)

The results are listed in Table 18. In summary, UC² is by far the strongest model across all tasks; xUNITER comes second and it outperforms M³P (which ranks the third) on reasoning and retrieval; mUNITER performs the worst in general, except for being better than M³P on reasoning.

Gap between Zero-shot and Few-shot Learning. In contrast to previous results on text-only tasks, we find few-shot learning for multilingual multimodal tasks to be especially challenging, possibly due to the complex nature of the respective tasks combined with the sparsity of the few-shot learning data. As plotted in Figure 3, we observe no clear improvement as the number of shots increase in all tasks, the only exception being xGQA. This suggests that cost-efficiently hand-labelling $\sim$ 100 shots in order to improve the models performance on the target language, unlike in text-only setups, is insufficient to make meaningful progress on V&L tasks such as XVNNLI, MaRVL, and xFlickr&Co. This also stresses the importance of developing methods for improved zero-shot transfer in future work, as V&L models still require sufficient amounts of expensive task-annotated data to benefit from in ‘non-zero-shot’ setups (Table 19).

In contrast to XVNNLI, MaRVL and xFlickr&Co, we find that few-shot experiments on xGQA yield large performance gains in the target language. We attribute this to the structured nature of the data and the large label space of the task (1,842 classes) resulting in a misaligned multilingual embedding space in zero-shot setups, as suggested by the original authors. We analyse these results further in §D.

Increasing the Number of Shots. In the few-shot experiments, we noticed no observable improvement or clear trend of performance increase on 3 out of the 4 tasks. In order to identify if, and at what point, increasing the number of shots would start being beneficial, we extend our few-shot experiments on the retrieval task using the xFlickr&CO dataset. In Figure 5, we plotted DEU and JPN performance with up to 1,500 shots¹⁷ (instead of the 100 shots used in our standard few-shot setup for this dataset) for the two best performing models in this task. Indeed, we observe an overall trend of increasing performance with more training examples, although with noticeable instability issues (e.g. UC² accuracy in JPN drops when more than 500 shots are used). These results demonstrate that more training examples may indeed further improve the overall performance in the target language, but the large number of examples required questions whether this setup can even be dubbed ‘few-shot’.

More Images versus More Captions. An important question, which concerns data collection efficiency for the target languages in image–text retrieval scenarios, is whether it is more useful, for a given budget, to collect (1) more captions per image or (2) more images with a single caption. To investigate this question, we trained xUNITER and UC² on JPN and DEU data with different number of shots and different ratios of captions per image (Table 20). As indicated by the results, there is no clear trend suggesting that one setting is better than the other. We leave a larger-scale investigation related to this question for future work.

D. Further Analyses

We now provide a series of further and finer-grained analyses, which include performance of the ‘translate train’ transfer approach and domain-specific results in xFlickr&Co, as well as an analysis of performance over structurally different question types of xGQA in few-shot setups.

¹⁷ DEU and JPN captions are from Multi30K (Elliott et al., 2016) and Nakayama et al. (2020), respectively.
Figure 12. Image retrieval performance averaged across languages per xFlickr&CO subset.

Figure 13. UC$^2$ image retrieval performance across languages per xFlickr&CO subset.

Table 21. English image retrieval performance by evaluation subset in xFlickr&CO after fine-tuning.

<table>
<thead>
<tr>
<th>xFlickr&amp;CO subset</th>
<th>mUNITER</th>
<th>xUNITER</th>
<th>UC$^2$</th>
<th>M$^3$P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flickr</td>
<td>51.3</td>
<td>46.6</td>
<td>44.0</td>
<td>37.1</td>
</tr>
<tr>
<td>COCO</td>
<td>37.7</td>
<td>30.3</td>
<td>30.8</td>
<td>25.6</td>
</tr>
</tbody>
</table>

**Translate Train Performance.** We further compare the few-shot results with ‘translate train’ performance: We directly fine-tune the pretrained UC$^2$ and xUNITER checkpoints on the full, human-translated Flickr30K training set. As shown in Figure 5, UC$^2$ achieves remarkable zero-shot and few-shot performance: 100 or fewer data points are enough for it to match ‘translate train’ performance, which it surpasses with 1,000 or more data points. Similar results can also be observed in Japanese, although it needs 500 data points to match ‘translate train’ accuracy. xUNITER—which was not pretrained on German nor Japanese multimodal data—also reaps benefits from 500 or more data points, although still being far from its ‘translate train’ counterpart (despite being initialised from English fine-tuned weights). Finally, we note that ‘translate train’ performance is still far from the corresponding English performance. This result is rather surprising for UC$^2$, which was solely pretrained on Conceptual Captions in five languages, each of equal size.

**Retrieval Performance by Evaluation Subset.** Our xFlickr&CO dataset is a composition of Flickr and COCO data (each of 1,000 image–sentence pairs). As models are fine-tuned on Flickr, the COCO evaluation subset might require out-of-distribution generalisation due to its different visual domain. Indeed, this is what we observe in English (Table 21). However, Figure 13 shows that this is not the case for the multilingual encoders, which perform similarly in both subsets. A closer look at UC$^2$ further shows that this behaviour is language-independent. Given the relatively poor performance of current models, we cannot assess whether the domain shift we observe in English does not happen in xFlickr&CO due to its unified captioning guidelines, or whether domain generalisation will be a challenge in xFlickr&CO for future, tightly-multilingual models.