IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages
Vision-and-Language Data
Vision-and-Language Data

- MS COCO, Flickr30K, Visual Genome
- Conceptual Captions, Conceptual 12M, RedCaps
- VQA, GQA, Visual7W, VizWiz, RefCOCO, NLVR2, …
Vision-and-Language Data

- MS COCO, Flickr30K, Visual Genome
- Conceptual Captions, Conceptual 12M, RedCaps
- VQA, GQA, Visual7W, VizWiz, RefCOCO, NLVR2, …
Vision-and-Language Data

- MS COCO, Flickr30K, Visual Genome
- Conceptual Captions, Conceptual 12M, RedCaps
- VQA, GQA, Visual7W, VizWiz, RefCOCO, NLVR2, …

But this trend is reversing

- Flickr30K-🇨🇳,🇯🇵 STAIR Captions (🇯🇵) COCO-🇨🇳,🇨🇳,🇪🇸,🇮🇹,🇮🇳
- Multi30K (🇩🇪,🇫🇷,🇨🇿) XTD (10 langs) GEM (20 langs) WIT (108 langs)
- MultiSubs (🇩🇪,🇫🇷,🇪🇸) MuCO-VQA (🇮🇳) xGQA (🇩🇪,🇫🇷,🇪🇸,🇮🇩,🇧🇩,🇷🇺,🇰🇷) MaRVL (🇨🇳,🇮🇩,🇮🇳,🇹🇷,🇰🇪)
Vision-and-Language Data

- MS COCO, Flickr30K, Visual Genome
- Conceptual Captions, Conceptual 12M, RedCaps
- VQA, GQA, Visual7W, VizWiz, RefCOCO, NLVR2, …

But this trend is reversing

- Flickr30K-{🇨🇳,🇯🇵} STAIR Captions (🇯🇵) COCO-{🇨🇳,🇯🇵,🇪🇸,🇮🇹,🇮🇳,🇻🇳}
- Multi30K (🇩🇪,🇫🇷,🇨🇿) XTD (10 langs) GEM (20 langs) WIT (108 langs)
- MultiSubs (🇩🇪,🇫🇷,🇪🇸,🇵🇹) MuCO-VQA (🇮🇳) xGQA (🇩🇪,🇫🇷,🇪🇸,🇵🇹,🇮🇩,🇧🇩,🇷🇺,🇰🇷) MaRVL (🇨🇳,🇮🇩,🇮🇳,🇹🇷,🇰🇪)

Mostly Indo-European languages  Mostly translations from English
Benchmarks have driven progress in machine learning

ImageNet

GLUE, SuperGLUE, IndoNLU, KLUE, RussianSuperGLUE, Liro

XGLUE, XTREME, XTREME-R
IGLUE: A Benchmark to the Rescue

Benchmarks have driven progress in machine learning

- ImageNet
- GLUE, SuperGLUE, IndoNLU, KLUE, RussianSuperGLUE, Liro
- XGLUE, XTREME, XTREME-R

IGLUE: Image-Grounded Language Understanding Evaluation

- 20 languages: 11 families, 9 scripts, 3/5 WALS macro-areas
- 4 V&L tasks requiring different levels of syntactic-semantic understanding
- 5 datasets, both pre-existing and new ones
- Zero-shot & Few-shot learning setups
IGLUE: Tasks & Datasets
IGLUE: Tasks & Datasets

NATURAL LANGUAGE INFERENCE

Given an image-premise, predict if a text-hypothesis entails, contradicts, or is neutral to it

XVNLI *

5 Languages: Arabic, French, Russian and Spanish

ENG: The basketball player shoots a three pointer
IGLUE: Tasks & Datasets

**NATURAL LANGUAGE INFERENCE**

Given an image-premise, predict if a text-hypothesis entails, contradicts, or is neutral to it.

XVNLI *

🌍 5 Languages: Arabic, French, Russian and Spanish

<table>
<thead>
<tr>
<th>NLI</th>
<th>QUESTION ANSWERING</th>
</tr>
</thead>
<tbody>
<tr>
<td>🇮🇷 لاعب كرة السلة يرمي كرة بثلاث نقط.</td>
<td></td>
</tr>
<tr>
<td>contradiction</td>
<td></td>
</tr>
<tr>
<td>ENG: The basketball player shoots a three pointer</td>
<td></td>
</tr>
<tr>
<td>🇬🇧 Given an image and question about it, predict the answer</td>
<td></td>
</tr>
<tr>
<td>🇸🇦 xGQA (Pfeiffer+, 2022)</td>
<td></td>
</tr>
<tr>
<td>🌍 8 Languages: Bengali, German, Indonesian, Korean, Mandarin, Portuguese, Russian</td>
<td></td>
</tr>
<tr>
<td>🇰🇷 감색이 아닌 음식은 어떤 것입니까?</td>
<td></td>
</tr>
<tr>
<td>vegetables</td>
<td></td>
</tr>
<tr>
<td>ENG: Which kind of food is not brown?</td>
<td></td>
</tr>
</tbody>
</table>
IGLUE: Tasks & Datasets

**Natural Language Inference (NLI)**
- Given an image-premise, predict if a text-hypothesis entails, contradicts, or is neutral to it.
- **XVNLI**
  - 5 Languages: Arabic, French, Russian and Spanish

**Question Answering (QA)**
- Given an image and question about it, predict the answer.
- **xGQA** (Pfeiffer+, 2022)
  - 8 Languages: Bengali, German, Indonesian, Korean, Mandarin, Portuguese, Russian

**Visual Reasoning**
- Given two images and a textual description, predict if the description applies to both images (true/false).
- **MaRVL** (Liu&Bugliarello+, 2021)
  - 6 Languages: Indonesian, Mandarin, Swahili, Tamil, Turkish

---

EMANUELE BUGLIARELLO | IGLUE: A BENCHMARK FOR TRANSFER LEARNING ACROSS MODALITIES, TASKS, AND LANGUAGES

ICML 2022
<table>
<thead>
<tr>
<th>TASK</th>
<th>DATASETS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Natural Language Inference</strong></td>
<td></td>
</tr>
<tr>
<td><strong>NLI</strong></td>
<td>Given an image-premise, predict if a text-hypothesis entails, contradicts, or is neutral to it</td>
</tr>
<tr>
<td>XVNLI *</td>
<td>5 Languages: Arabic, French, Russian and Spanish</td>
</tr>
<tr>
<td><strong>Visual Reasoning</strong></td>
<td></td>
</tr>
<tr>
<td><strong>MaRVL (Liu&amp;Bugliarello+, 2021)</strong></td>
<td>Given two images and a textual description, predict if the description applies to both images (true/false)</td>
</tr>
<tr>
<td>6 Languages: Indonesian, Mandarin, Swahili, Tamil, Turkish</td>
<td></td>
</tr>
<tr>
<td><strong>Question Answering</strong></td>
<td></td>
</tr>
<tr>
<td><strong>QA</strong></td>
<td>Given an image and question about it, predict the answer</td>
</tr>
<tr>
<td>xGQA (Pfeiffer+, 2022)</td>
<td>8 Languages: Bengali, German, Indonesian, Korean, Mandarin, Portuguese, Russian</td>
</tr>
<tr>
<td><strong>Image–Text Retrieval</strong></td>
<td></td>
</tr>
<tr>
<td><strong>IR</strong></td>
<td>Given a caption, retrieve its image</td>
</tr>
<tr>
<td>xFlickr&amp;CO *</td>
<td>8 high-resource languages</td>
</tr>
<tr>
<td><strong>TR</strong></td>
<td>Given an image, retrieve its caption</td>
</tr>
<tr>
<td>WIT (Srinivasan+, 2021)</td>
<td>11 diverse languages</td>
</tr>
</tbody>
</table>

**IMAGE—TEXT RETRIEVAL**

- **Retrieval**
  - ENG: A group of men and women dressed in formal black dresses and suits holding their music books and singing.

**EMANUELE BUGLIARELLO | IGLUE: A BENCHMARK FOR TRANSFER LEARNING ACROSS MODALITIES, TASKS, AND LANGUAGES**

ICML 2022
Experimental Setup

Baselines

👩💻 Implement multilingual V&L Transformers in a single code (VOLTA; Bugliarello+ 2021)
🤖 mUNITER & xUNITER (Liu&Bugliarello+, 2021) M³P (Ni+, 2021) UC² (Zhou+, 2021)
Experimental Setup

**Baselines**

👩💻 Implement multilingual V&L Transformers in a single code ([VOLTA; Bugliarello+ 2021](#))

🤖 mUNITER & xUNITER ([Liu&Bugliarello+, 2021](#)) [M³P (Ni+, 2021)](#) [UC² (Zhou+, 2021)](#)

**Fine-Tuning**

🇺🇸 Train on the English split

💻 On a V100 (16 GB) GPU for less than 12h

**Zero-Shot Transfer**

🌍 Evaluate on multilingual data

**Translate-Test Transfer**

🌍➡🇺🇸 Evaluate on machine translated data
Experimental Setup

**Baselines**

- Implement multilingual V&L Transformers in a single code ([VOLTA; Bugliarello+ 2021](#))
- mUNITER & xUNITER ([Liu&Bugliarello+, 2021](#))
- M³P ([Ni+, 2021](#))
- UC² ([Zhou+, 2021](#))

**Fine-Tuning**

- Train on the English split
- On a V100 (16 GB) GPU for less than 12h

**Zero-Shot Transfer**

- Evaluate on multilingual data

**Translate-Test Transfer**

- Evaluate on machine translated data

**Few-Shot Learning**

- After English fine-tuning, train on few samples in each target language
- Performance as a function of number of shots
- *Max-shot* setup: evaluate with all the few-shot samples (1 run per dataset–language pair)
Experimental Setup

Baselines
- Implement multilingual V&L Transformers in a single code *(VOLTA; Bugliarello+ 2021)*
- mUNITER & xUNITER *(Liu&Bugliarello+, 2021)*
- M3P *(Ni+, 2021)*
- UC² *(Zhou+, 2021)*

Fine-Tuning
- Train on the English split
- On a V100 (16 GB) GPU for less than 12h

Zero-Shot Transfer
- Evaluate on multilingual data

Translate-Test Transfer
- Evaluate on machine translated data

Few-Shot Learning
- After English fine-tuning, train on few samples in each target language
- Performance as a function of number of shots
- *Max-shot* setup: evaluate with all the few-shot samples (1 run per dataset–language pair)

Metric
- Accuracy (XVNLI, xGQA, MaRVL) and Recall@1 (xFlickr&CO, WIT) – *equivalent* in our setup
Results

Zero-Shot Learning
Large zero-shot transfer gap

Zero-Shot Learning
Results

Zero-Shot Learning

Large zero-shot transfer gap

![Bar chart showing results for different tasks and languages](chart.png)

EMANUELE BUGLIARELLO | IGLUE: A BENCHMARK FOR TRANSFER LEARNING ACROSS MODALITIES, TASKS, AND LANGUAGES

ICML 2022
Results

Zero-Shot Learning

Large zero-shot transfer gap

Translate-test transfer » zero-shot transfer

XVNL1

xGQA

MaRVL

xFlickr&CO IR

WIT IR

EMANUELE BUGLIARELLO | IGLUE: A BENCHMARK FOR TRANSFER LEARNING ACROSS MODALITIES, TASKS, AND LANGUAGES

ICML 2022
Results

Zero-Shot Learning

Large zero-shot transfer gap

Translate-test transfer $\gg$ zero-shot transfer

Few-Shot Learning

EMANUELE BUGLIARELLO | IGLUE: A BENCHMARK FOR TRANSFER LEARNING ACROSS MODALITIES, TASKS, AND LANGUAGES  
ICML 2022
**Results**

**Zero-Shot Learning**

- Large zero-shot transfer gap
- Translate-test transfer \(\gg\) zero-shot transfer

**Few-Shot Learning**

- Consistent but moderate gains

---

**Accuracy**

- **XVNLI**
  - Model: 
    - mUNITER
    - xUNITER
    - UC
    - M'P

- **xGQA**
  - Model: 
    - mUNITER
    - xUNITER
    - UC
    - M'P

- **MaRVL**
  - Model: 
    - mUNITER
    - xUNITER
    - UC
    - M'P

- **xFlickr&CO**
  - Model: 
    - mUNITER
    - xUNITER
    - UC
    - M'P

---

**EMANUELE BUGLIARELLO | IGLUE: A BENCHMARK FOR TRANSFER LEARNING ACROSS MODALITIES, TASKS, AND LANGUAGES**

ICML 2022
Conclusion & Outlook

IGLUE: The Image-Grounded Language Understanding Evaluation benchmark

- 5 datasets across 4 tasks in 20 languages
- Zero-shot & few-shot transfer setups show large drops in performance wrt English
- Code, data and pretrained models available online
  - iglue-benchmark.github.io
  - github.com/e-bug/iglue
Conclusion & Outlook

IGLUE: The Image-Grounded Language Understanding Evaluation benchmark
- 5 datasets across 4 tasks in 20 languages
- Zero-shot & few-shot transfer setups show large drops in performance wrt English
- Code, data and pretrained models available online
  - iglue-benchmark.github.io
  - github.com/e-bug/iglue

Next Steps
- Transfer learning across modalities, tasks, and languages
- Single- vs. multi-source transfer
- Beyond image-only tasks (e.g. videos and speech)
Conclusion & Outlook

**IGLUE: The Image-Grounded Language Understanding Evaluation benchmark**
- 5 datasets across 4 tasks in 20 languages
- Zero-shot & few-shot transfer setups show large drops in performance wrt English
- Code, data and pretrained models available online
  - iglue-benchmark.github.io
  - github.com/e-bug/iglue

**Next Steps**
- Transfer learning across modalities, tasks, and languages
- Single- vs. multi-source transfer
- Beyond image-only tasks (e.g. videos and speech)