Multilingual Multimodal Learning with Machine Translated Text

Chen Qiu[α] Dan Oneață[β] Emanuele Bugliarello[ς] Stella Frank[ς,ð] Desmond Elliott[ς,ð]

[α]Wuhan University of Science and Technology
[β]University Politehnica of Bucharest
[ς]University of Copenhagen
[ð]Pioneer Centre for AI

Success of Vision-and-language Pretraining

Q1:What color is the plane? A1: White

Q2: How many spots are on this animal? A2: 100

Visual Question Answering

[Rethinking Evaluation Practices in Visual Question Answering: A Case Study on Out-of-Distribution Generalization, Agrawal et al. 2022]

Both images contains a lot of masala vadas.

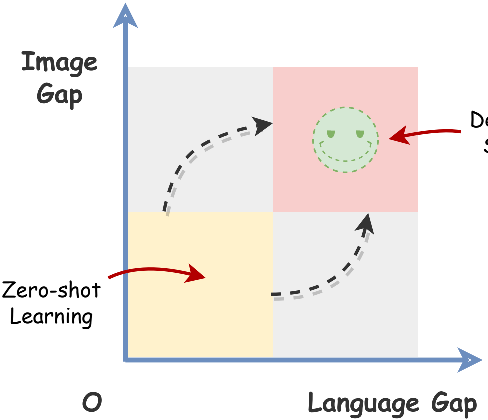
Label: False

Reasoning

[Visual Grounded Reasoning across Languages and Cultures, Liu el al. 2021, EMNLP]

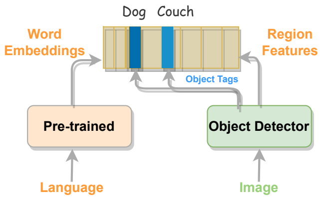
Issues of V&L Pretraining

How to define pretraining strategies that induce high-quality multilingual multimodal representations?

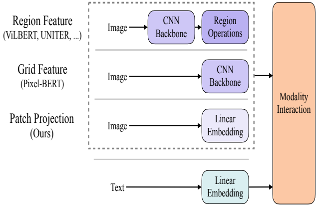


Related Work

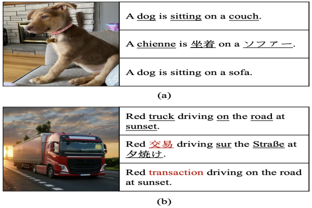
BERT+ENG image-text | Vision Transformer | Multilingual+Image



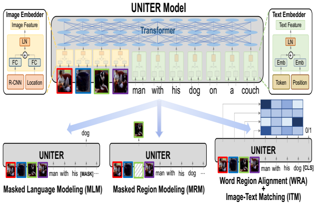
Oscar: Li et al., ECCV2020



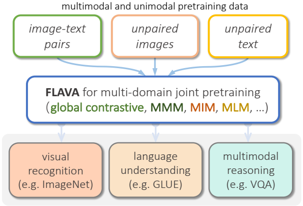
ViLT: Kim et al., ICML2021



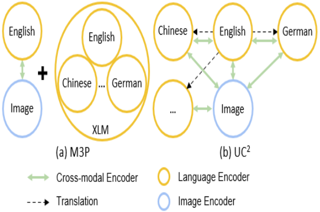
M3P: Ni et al., CVPR2021



UNITER: Chen et al., ECCV2020



FLAVA: Singh et al., CVPR2022



UC2: Zhou et al., CVPR2021

Motivation

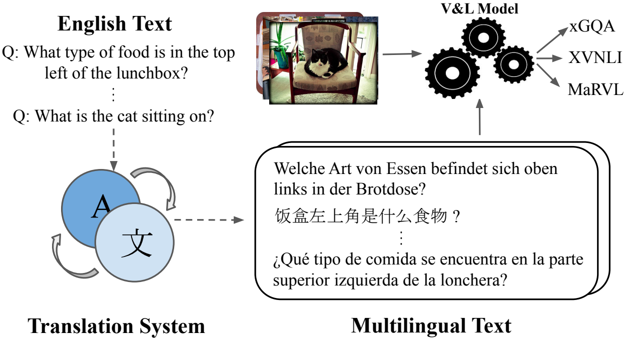
Research Questions:

1. Is translated data useful for fine-tuning and pretraining?
2. How to pretrain on many millions of multilingual translated examples?

Our Contributions:

1. narrow the gap between English and non-English languages on the IGLUE Benchmark [Bugliarello et al. 2022];
2. a reliable approach to filtering out bad translations;
3. provide inexpensive and impressive improvements when evaluating in zero-shot and machine translated scenarios.

TD-MML: Translated Data for Multilingual Multimodal Learning



1. Initial Experiment: Fine-Tuning with Translated Data
2. Translation and Data Preparation
3. Pretraining with Translated Data
4. TD Pretraining and English Fine-tuning *vs* MT Fine-tuning TD-MML

Multilingual Fine-tuning only: Initial Experiments on MaRVL and xGQA

Fine-tuning on multilingual machine-translated data: *inexpensive and viable!*

Approach	**ENG**	**IND**	**SWA**	**TAM**	**TUR**	**CMN**	**avg**
English	**71.6**	55.1	55.5	53.1	56.2	53.1	54.6
MT	67.9	**59.6**	**61.4**	**60.4**	**64.3**	**59.4**	**61.0**

MaRVL accuracy

Approach	**ENG**	**BEN**	**DEU**	**IND**	**KOR**	**POR**	**RUS**	**CMN**	**avg**
English	**54.8**	10.8	34.8	33.7	12.1	22.1	18.8	19.6	21.7
MT	48.1	**41.8**	**46.5**	**45.7**	**44.8**	**46.8**	**46.2**	**45.7**	**45.3**

xGQA accuracy

Pretraining: Translation and Data Preparation

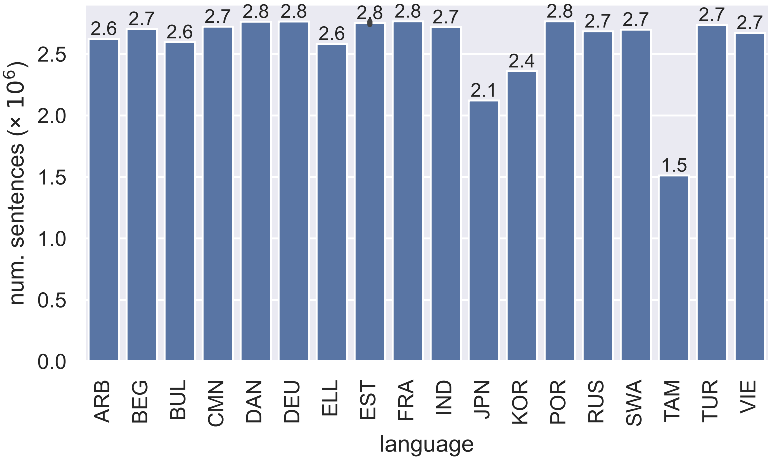
- Translate 2.77M English sentences from the Conceptual Captions datasets into the twenty languages in IGLUE using M2M-100$_{\text{LARGE}}$ [Fan et al. 2021]
- Filter out potential bad data by *Complement of the token-to-type ratio* and *BLEU*

Examples:

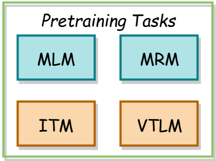
✗	*funny animals of the week, funny animal photo, cute animal pictures*	→	*Animaux drôles, Animaux drôles, Animaux drôles, Animaux drôles* (FRA)
✗	*damask seamless floral pattern, ornament*	→	*Mifano ya Mifano ya Mifano ya Mifano ya Mifano ya Mifano ya Mifano ya Mifano ya Mifano ya Mifano* (SWA)
✗	*plaid, over garment , outfit idea cute fall outfit idea*	→	方格*, over garment, cute fall* (CMN)

Pretraining: Translation and Data Preparation

Number of sentences for pretraining on nineteen non-English IGLUE languages.

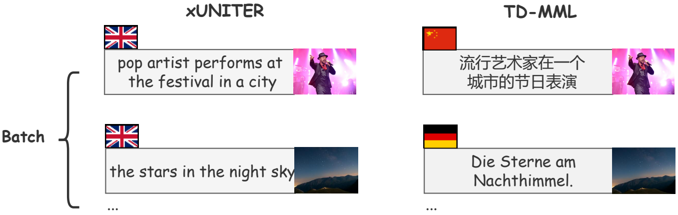


Model: Pretraining with Translated Data



$$\mathcal{L}_{\mathrm{ITM}}(\theta) = -\mathbb{E}_{(\mathbf{x}^l,\mathbf{v})\sim D}\left[c \log s_\theta(\mathbf{x}^l,\mathbf{v}) + (1-c)\log\left(1 - s_\theta(\mathbf{x}^l,\mathbf{v})\right)\right] \quad (1)$$

$$\mathcal{L}_{\mathrm{VTLM}}(\theta) = -\mathbb{E}_{(\mathbf{x}^{\mathrm{ENG}},\mathbf{x}^l,\mathbf{v})\sim D} \log P_\theta\left(\mathbf{x}_a^{\mathrm{ENG}}, \mathbf{x}_b^l \mid \mathbf{x}_{\setminus a}^{\mathrm{ENG}}, \mathbf{x}_{\setminus b}^l, \mathbf{v}\right) \quad (2)$$



xUNITER is trained over 2.77M image–English caption pairs, while TD-MML is pretrained on 52M image–multilingual caption pairs.

Results: Translated Data Pretraining & English Fine-tuning

We evaluate the zero-shot language understanding abilities of the TD-MML model.

Model	NLI	QA	Reasoning	Retrieval			
	XVNLI	xGQA	MaRVL	xFlickr&CO		WIT	
				IR	TR	IR	TR
mUNITER	53.69	9.97	53.72	8.06	8.86	9.16	10.48
xUNITER	58.48	21.72	54.59	14.04	13.51	8.72	9.81
UC^2	62.05	29.35	57.28	20.31	17.89	7.83	9.09
M^3P	58.25	28.17	56.00	12.91	11.90	8.12	9.98
TD-MML	64.84	**35.95**	**59.67**	**21.30**	**26.35**	**9.76**	10.35
- w/o VTLM	**66.28**	33.01	58.14	20.90	24.61	9.14	**10.61**

A substantial improvement for TD-MML across all tasks!

Results: MT Fine-tuning TD-MML

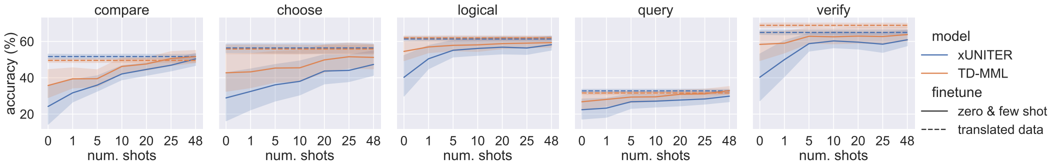
We can combine the machine translated pretraining strategy of TD-MML with additional machine translated fine-tuning.

Type	Method	ENG	IND	SWA	TAM	TUR	CMN	avg
Fine-tune with English-only data (zero-shot)								
—	xUNITER	**71.55**	55.14	55.51	53.06	56.19	53.06	54.59
	TD-MML	69.00	59.04	61.01	56.44	61.95	59.88	59.67
Fine-tune with machine translated data								
Full	xUNITER	67.92	59.57	61.37	60.39	64.32	59.39	61.01
	TD-MML	67.52	59.40	**62.18**	60.55	**66.27**	59.59	**61.60**
Filtered	xUNITER	67.52	**60.82**	61.55	60.63	63.48	59.88	61.27
	TD-MML	67.09	57.62	61.91	**61.35**	64.58	**60.28**	61.15

MaRVL accuracy results for zero-shot cross-lingual evaluation

Few-Shot vs Machine Translated Data

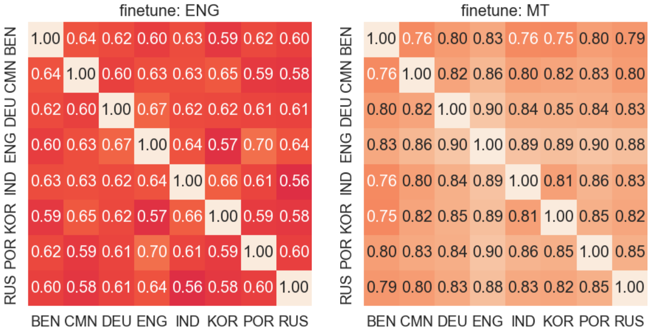
xGQA average accuracy across the languages on the five question types.



- The performance improves with the quantity of training data.
- The machine translated fine-tuning upper bounds the performance of the few-shot approach.

Cross-Language Correlation Analysis

Are the same questions easy or difficult across languages?[1]



◎ MT fine-tuned results show much higher agreement across languages.

[1]We use Cohen's kappa coefficient κ to measure agreement between languages on the xGQA

Qualitative Examples

Qualitative results on the xGQA dataset.

Q1

Q2

Q3

Q4

	Q1		Q2		Q3		Q4	
Q	What is the sign behind the young person?		Is the black and white cat unhappy or happy?		What is covering the man that is wearing jeans?		How is this cooking utensil called?	
A	traffic sign		unhappy		jacket		baking pan	
	fine-tune: ENG	fine-tune: MT	fine-tune: ENG	fine-tune: MT	fine-tune: ENG	fine-tune: MT	fine-tune: ENG	fine-tune: MT
BEN	car	pole	no	lush	bag	umbrella	paper	pretzel
CMN	pole	stop sign	white	happy	coat	umbrella	book	stove
DEU	fire hydrant	stop sign	unhappy	happy	jacket	umbrella	mirror	tea kettle
ENG	stop sign	stop sign	unhappy	happy	towel	umbrella	pan	tea kettle
IND	street sign	stop sign	unhappy	happy	blanket	umbrella	yes	yes
KOR	pole	stop sign	gray	happy	backpack	umbrella	drum	drum
POR	car	car	happy	happy	suitcase	umbrella	table	tea kettle
RUS	stop sign	stop sign	happy	happy	umbrella	umbrella	shelf	pan

Summary

- Translated data improves multilingual multimodal representation learning
- A simple and effective strategy for filtering low-quality translated data
- Results shed light in the importance of explicitly grounding multilingual text

Research Labs

武汉科技大学大数据科学与工程研究院
INSTITUTE OF BIG DATA SCIENCE AND ENGINEERING , WUST
国家新闻出版署富媒体数字出版内容组织与知识服务重点实验室
KEY LABORATORY OF RICH-MEDIA KNOWLEDGE ORGANIZATION AND SERVICE OF DIGITAL PUBLISHING CONTENT, NATIONAL PRESS AND PUBLICATION ADMINISTRATION OF THE PEOPLE'S REPUBLIC OF CHINA

CoAStaL

Chen Qiu
chen@wust.edu.cn
ONTOWEB

Dan Oneață
dan.oneata@gmail.com
CoAStaL

Emanuele Bugliarello
emanuele@di.ku.dk
CoAStaL

Desmond Elliott
de@di.ku.dk
CoAStaL

https://arxiv.org/abs/2210.13134