Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models

Po-Yao (Bernie) Huang*, Mandela Patrick*, Junjie Hu, Graham Neubig, Florian Metze, Alexander Hauptmann

Carnegie Mellon University
School of Computer Science

University of Oxford

Facebook

Introduction

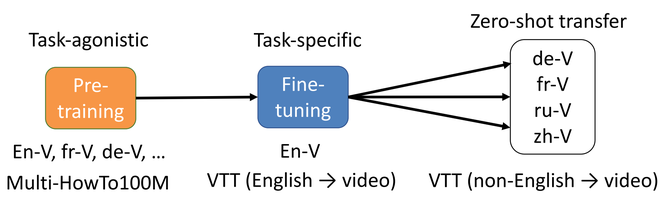
Motivation:
Most vision-language models and V-L tasks are English-centered. It is challenging yet rewarding to generalize V-L models to other 7000 languages.

Our solution:
Zero-shot cross-lingual transfer that utilizes one model to rule all the languages. We propose to learn multilingual multimodal representations with:

- Multilingual Multimodal Transformers
- Multilingual Multimodal Pre-training

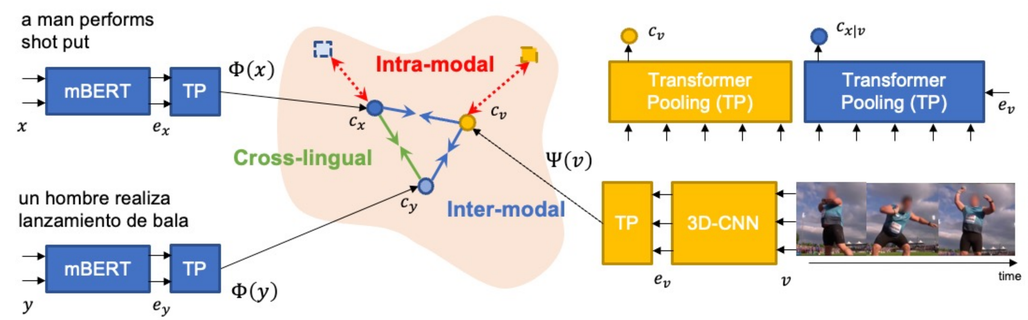
Tasks of our focus in this study:
Multilingual text-video search.

Zero-Shot Cross-Lingual Transfer



	NLP Tasks (e.g., EXTREME)	V-L Tasks (proposed)
Model	Multilingual Transformer	Multilingual multimodal Transformer
Step 1	Task-agonistic multilingual pre-training	Task-agonistic multilingual multimodal pre-training
Step 2	Task-specific English fine-tuning	Task-specific English-vision fine-tuning
Step 3	Zero-shot transfer to non-English tasks	Zero-shot transfer to non-English-vision tasks

Multilingual Multimodal Transformers



Contrastive attraction ⟷ Contrastive repulsion ⟻⟼

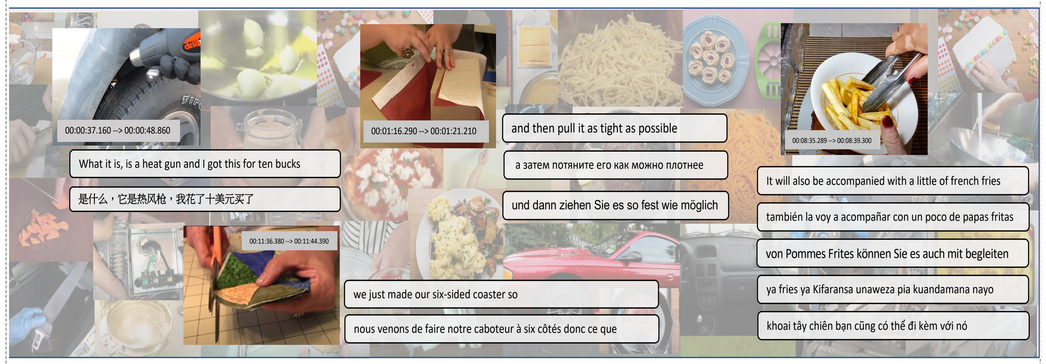
$$\mathcal{L}^{\text{inter}} = \mathcal{L}(\mathcal{X}, \mathcal{V}) + \mathcal{L}(\mathcal{Y}, \mathcal{V})$$
$$\mathcal{L}^{\text{intra}} = \mathcal{L}(\mathcal{X}, \mathcal{X}^m) + \mathcal{L}(\mathcal{Y}, \mathcal{Y}^m) + \mathcal{L}(\mathcal{V}, \mathcal{V}^m)$$
$$\mathcal{L}^{\text{cross}} = \mathcal{L}(\mathcal{X}|\mathcal{V}, \mathcal{Y}|\mathcal{V})$$

$$\mathcal{L}(\mathcal{X}, \mathcal{V}) = -\frac{1}{B}\sum_{i=1}^{B} \log \ell^{\text{NCE}}(\Phi(x_i), \Psi(v_i))$$
$$\ell^{\text{NCE}}(c_x, c_v) = \frac{e^{s(c_x, c_v)}}{e^{s(c_x, c_v)} + \sum_{(x', v') \sim \mathcal{N}} e^{s(c_{x'}, c_{v'})}}$$

Multilingual Multimodal Pre-training

Multi-HowTo100M dataset

- 1.2 million instructional videos. 138 million clips.
- Transcriptions in 9 languages:
  English, German, French, Russian, Spanish, Czech, Swahili, Chinese, Vietnamese.



Experiments

Quantitative Results

(a-1) Multilingual Text-to-Video Search on VTT

Model	*en*	*de*	*fr*	*cs*	*zh*	*ru*	*vi*	*sw*	*es*	Avg↑
mBERT	19.9	11.1	11.6	8.2	6.9	7.9	2.7	1.4	12.0	9.1
mBERT-MP	20.6	11.3	11.9	8.0	7.1	7.7	2.5	1.1	12.5	9.2
mBERT-MMP	21.8	15.0	15.8	11.2	8.4	11.0	3.7	3.4	15.1	11.7
XLM-R	21.0	16.3	17.4	16.0	14.9	15.4	7.7	5.7	17.3	14.7
XLM-R-MP	23.3	17.4	18.5	17.1	16.3	17.0	8.1	6.2	18.5	15.8
XLM-R-MMP	**23.8**	**19.4**	**20.7**	**19.3**	**18.2**	**19.1**	**8.2**	**8.4**	**20.4**	**17.5**
mBERT + translated VTT	19.6	18.2	18.0	16.9	16.2	16.5	8.4	13.0	18.5	16.1
mBERT-MMP + translated VTT	21.5	19.1	19.8	18.3	17.3	18.3	8.9	14.1	20.0	17.4
XLM-R + translated VTT	21.5	19.6	20.1	19.3	18.9	19.1	10.3	12.5	18.9	17.8
XLM-R-MMP + translated VTT	**23.1**	**21.1**	**21.8**	**20.7**	**20.0**	**20.5**	**10.9**	**14.4**	**21.9**	**19.4**

(b) VATEX

Model	English to Video R@1↑	R@5↑	R10↑	Chinese to Video R@1↑	R@5↑	R@10↑
VSE (Kiros et al., 2014)	28.0	64.3	76.9	-	-	-
VSE++ (Faghri et al., 2018)	33.7	70.1	81.0	-	-	-
Dual (Dong et al., 2019)	31.1	67.4	78.9	-	-	-
HGR (Chen et al., 2020a)	35.1	73.5	83.5	-	-	-
Ours (VATEX:*en*-only)	43.5	79.8	88.1	23.9	55.1	67.8
Ours-MMP (VATEX:*en*-only)	**44.4**	80.5	88.7	29.7	63.2	75.5
Ours-MMP (VATEX:*en*, *zh*)	44.3	**80.7**	**88.9**	**40.5**	**76.4**	**85.9**

(a-2) Comparison on VTT

Model	R@1↑	R@5↑	R@10↑
JSFusion (Yu et al., 2018)	10.2	31.2	43.2
JPoSE (Wray et al., 2019)	14.3	38.1	53.0
VidTrans† (Korbar et al., 2020)	14.7	–	52.8
HT100M† (Miech et al., 2019)	14.9	40.2	52.8
Noise† (Amrani et al., 2020)	17.4	41.6	53.6
CE² (Liu et al., 2019)	20.9	48.8	62.4
Ours(VTT:*en*-only)	21.0	50.6	63.6
Ours-MMP (VTT:*en*-only)	**23.8**	**52.6**	**65.0**

(c) Multi30K

Model	M30K # lang.	English to Image R@1↑	R@5↑	R10↑	German to Image R@1↑	R@5↑	R@10↑	Czech to Image R@1↑	R@5↑	R@10↑
OE (Vendrov et al., 2015)	2	21.0	48.5	60.4	25.8	56.5	67.8	-	-	-
VSE++ (Faghri et al., 2018)	2	31.3	62.2	70.9	39.6	69.1	79.8	-	-	-
Pivot (Gella et al., 2017)	2	22.5	49.3	61.7	26.2	56.4	68.4	-	-	-
FB-NMT (Huang et al., 2020a)	2	47.3	75.4	83.5	37.0	64.0	73.1	-	-	-
MULE (Kim et al., 2020)	4	42.2	72.2	81.8	35.1	64.6	75.3	37.5	64.6	74.8
SMALR (Burns et al., 2020)	10	41.8	72.4	82.1	36.9	65.4	75.4	36.7	68.0	78.2
MHA-D (Huang et al., 2019b)	2	50.1	78.1	85.7	40.3	70.1	79.0	-	-	-
Ours (M30K:*en*-only)	1	48.4	78.3	85.9	31.4	61.1	72.6	33.2	65.2	76.1
Ours-MMP (M30K:*en*-only)	1	50.0	79.2	86.8	33.8	63.3	74.7	37.9	68.8	78.2
Ours-MMP (M30K:en, de, cs, fr)	4	**51.6**	**80.1**	**87.3**	**45.1**	**75.6**	**85.0**	**46.6**	**75.9**	**83.4**

Qualitative Results