Improving the Cross-Lingual Generalisation in Visual Question Answering

Farhad Nooralahzadeh, Rico Sennrich

Department of Computational Linguistics, University of Zurich

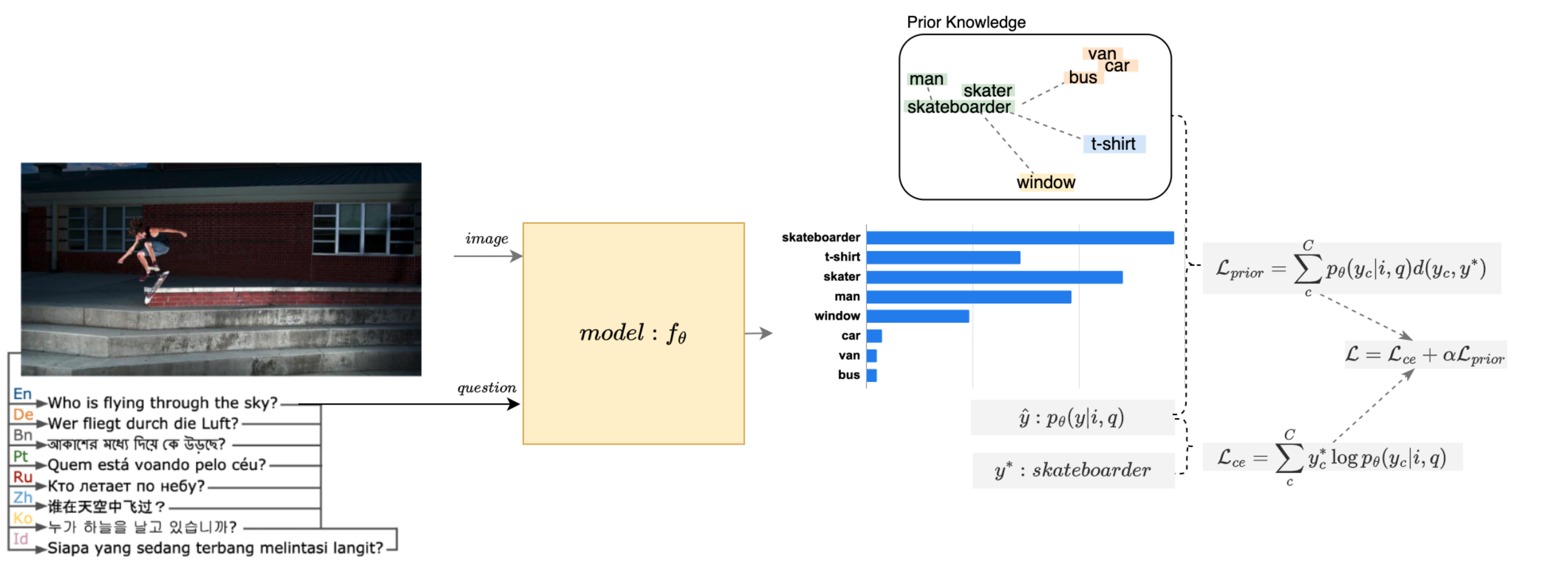
fahrad.nooralahzadeh@uzh.ch, sennrich@cl.uzh.ch

In short

We explore the poor performance of **multilingual vision-language pretrained model** on a zero-shot **cross-lingual visual question answering** (VQA) task, where models are fine-tuned on English visual-question data and evaluated on **7 typologically** diverse languages. We improve cross-lingual transfer with three strategies:

- We introduce a **linguistic prior objective** to augment the cross-entropy loss with a similarity-based loss to guide the model during training.
- We learn a **task-specific subnetwork** that improves cross-lingual generalisation and reduces variance without model modification.
- We augment training examples using **synthetic code-mixing** to promote alignment of embeddings between source and target languages.

1. Incorporating Linguistic Prior



We formalize the **distance score** $d(y_c, y^*)$ between the ground truth label and others in the label space by using two sources of **linguistic knowledge**:

- **WordNet** (prior$_{wn}$):

$$d(y_c, y^*) = \begin{cases} 0 & \text{if } y_c \text{ and } y^* \text{ are } synonyms \\ d_1 & \text{if } y_c \text{ is } hyponym \text{ of } y^* \\ d_2 & \text{if } y_c \text{ is } hypernym \text{ of } y^* \\ 1 & \text{otherwise} \end{cases}$$

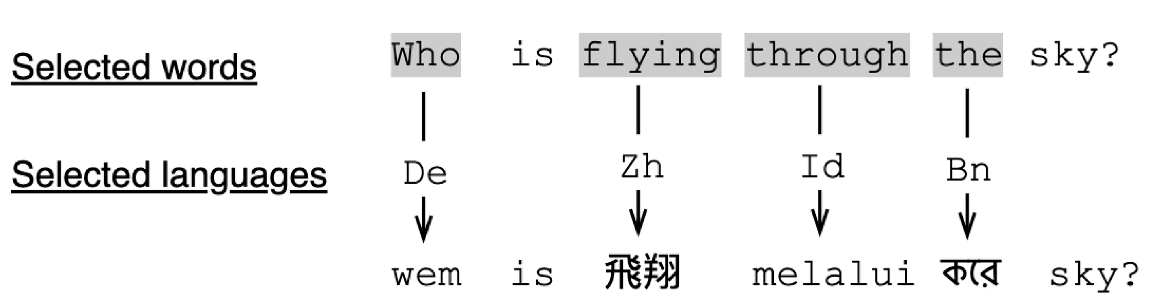
- **Word Embeddings** (prior$_{em}$):

$$d(y_c, y^*) = CosineDistance(emb_{y^*}, emb_{y_c})$$

2. Task-specific Sparse Fine-tuning (SFT)

- **Task-specific** and **language-neutral** components of **multilingual pretrained models**, which capture **commonalities** among languages.
- There exists a **sparse, separated trainable subnetwork** (i.e. a winning ticket) capable to match or even outperform the original neural network.
- $Step_0$**:** To obtain a **subnetwork** $f(.; M \odot \theta)$ where $M \in \{0,1\}^{|\theta|}$ represents a binary mask and $\odot$ is element-wise multiplication using **Iterative Magnitude Pruning** (IMP).
- $Step_1$**:** Having the **pruning mask** M, we perform fine-tunning again. In this step, only the **unmasked parameters** are **trained** while the masked ones are **kept frozen**.

3. Code-Mixing (CDM)



To make full use of the **cross-lingual alignment** information and better fine-tuning, we construct code-mixed data in **target languages**.

Results

	Model	En	Bn	De	Id	Ko	Pt	Ru	Zh	Avg
	Fine-tune model on English training set (Zero-Shot)									
UC2	Our Baseline	54.92	19.99	42.00	28.44	22.40	30.92	28.55	31.19	29.07
	Baseline (Bugliarello et al. 2022)	55.19	19.98	42.85	28.67	21.36	30.41	30.99	31.15	29.35
	Liu et al. (2022)	58.57±0.2	26.23 ±1.5	49.51 ±1.1	38.92 ±1.3	36.48 ±1.3	39.76 ±0.6	41.72 ±0.3	**46.52** ±0.9	39.87
	With prior$_{wn}$	55.77±0.02	23.66 ±0.76	47.93 ±0.19	35.67 ±1.43	34.57 ±1.81	37.46 ±1.35	40.08±0.54	40.08±4.31	37.06
	With prior$_{em}$	56.09 ±0.14	23.97±2.56	48.13 ±0.78	36.87±1.90	34.14 ±3.56	38.18 ±2.55	41.07 ±0.86	41.76±1.89	37.73
	With prior$_{em}$ + SFT	56.56±0.10	23.53 ±1.97	49.54 ±0.27	36.79 ±0.46	34.56 ±0.49	38.95 ±0.19	41.18 ±0.23	43.40 ±0.21	38.28
	With prior$_{em}$ + CDM	54.37±0.01	27.38 ±0.02	46.66 ±1.70	20.88 ±2.33	36.32±1.11	40.81 ±2.06	43.48 ±0.18	30.62 ±1.46	35.16
	With prior$_{em}$ + SFT + CDM	55.21±0.08	**30.96** ±1.33	**50.30** ±0.22	**41.68** ±0.74	**39.57** ±0.65	**43.43** ±0.60	**44.58** ±0.92	44.80 ±0.78	**42.19**
M3P	Our Baseline	54.02	17.24	32.40	23.77	25.57	32.91	32.32	27.39	27.37
	Baseline (Bugliarello et al. 2022)	53.75	18.64	33.42	32.48	25.11	31.40	27.50	28.65	28.17
	Liu et al. (2022)	46.70±0.7	29.75 ±1.4	39.52 ±1.3	36.73 ±1.6	35.67 ±1.1	37.59 ±0.8	37.93 ±0.9	36.15 ±0.9	36.19
	With prior$_{wn}$	55.91±0.20	22.38 ±0.38	39.48 ±1.73	29.31 ±2.27	35.15±0.86	39.00 ±0.17	38.92 ±0.31	35.74 ±0.79	34.28
	With prior$_{em}$	56.33±0.09	22.93 ±3.19	40.10 ±0.54	30.63 ±0.05	35.35 ±2.14	38.85 ±1.09	39.95 ±0.03	36.97 ±0.07	34.97
	With prior$_{em}$ + SFT	56.18 ±0.00	22.07 ±0.46	40.29 ±0.17	27.04 ±0.11	34.62 ±0.06	38.39 ±0.24	39.44 ±0.00	36.32 ±0.49	34.02
	With prior$_{em}$ + CDM	54.35±0.64	28.71 ±1.73	43.57 ±0.33	**38.89** ±2.07	38.06 ±0.44	41.93 ±0.59	41.64 ±1.11	38.80 ±1.22	38.80
	With prior$_{em}$ + SFT + CDM	55.58±0.12	**31.53** ±1.47	**46.19** ±0.54	34.60 ±0.49	**40.21** ±0.91	**42.87** ±0.58	**42.32** ±1.19	**42.25** ±0.60	**40.00**
	Translate everything to English and use the English-only model (Translate-Test)									
UC2	(Bugliarello et al. 2022)	55.19	49.31	52.61	50.34	48.62	52.17	49.95	48.32	50.19
M3P	(Bugliarello et al. 2022)	53.75	47.79	51.01	49.35	47.64	51.21	47.76	47.04	48.83

Error Analysis

We investigate the effect of **synonymy** relations among the **target labels** on xGQA **evaluation results**.

	Model	Avg. w/o Syn.	Avg. w Syn.	Diff.
UC2	Our Baseline	29.07	29.96	+0.89
	With prior$_{wn}$	37.06	38.91	+1.85
	With prior$_{em}$	37.73	39.06	+1.33
	With prior$_{em}$+ SFT	38.28	39.67	+1.39
	With prior$_{em}$ + SFT + CDM	**42.19**	**43.90**	+1.71
M3P	Our Baseline	27.37	31.83	+4.56
	With prior$_{wn}$	34.28	37.70	+3.42
	With prior$_{em}$	34.97	38.85	+3.88
	With prior$_{em}$ + SFT	34.02	38.25	+4.23
	With prior$_{em}$ + SFT + CDM	**40.00**	**43.52**	+3.52

We show the **5 most-confused labels** for each language, specifically where the UC2 model predicts a **synonym, hypernym, or hyponym** of the target label.

Model	Lang.	5 most-confused labels label:prediction (rel.)									
Our Baseline	En	girl:woman (hyp)	27	material:color (hpo)	23	lady:woman (hyp)	18	coffee table:table (hyp)	17	zebras:zebra (syn)	16
	Bn	sailboats:sailboat (syn)	3	skater:skateboarder (hpo)	3	plain:field (syn)	2	trees:tree (syn)	2	tank top:shirt (hyp)	1
	De	girl:woman (hyp)	33	material:color (hpo)	21	lady:woman (hyp)	16	woman:girl (hpo)	13	street sign:sign (hyp)	13
	Id	girl:woman (hyp)	28	lady:woman (hyp)	18	skater:skateboarder (hpo)	15	woman:girl (hpo)	14	zebras:zebra (syn)	12
	Ko	girl:woman (hyp)	7	skater:skateboarder (hpo)	7	boy:man (hyp)	2	fire truck:truck (hyp)	2	gown:dress (hyp)	1
	Pt	girl:woman (hyp)	22	skater:skateboarder (hpo)	17	lady:woman (hyp)	13	zebras:zebra (syn)	12	woman:girl (hpo)	11
	Ru	girl:woman (hyp)	32	skater:skateboarder (hpo)	17	lady:woman (hyp)	17	woman:girl (hpo)	14	cabinets:cabinet (syn)	12
	Zh	girl:woman (hyp)	26	chairs:chair (syn)	15	cabinets:cabinet (syn)	15	skater:skateboarder (hpo)	15	lady:woman (hyp)	15
Our Best Strategy	En	girl:woman (hyp)	28	material:color (hpo)	24	cabinets:cabinet (syn)	20	woman:girl (hpo)	18	zebras:zebra (syn)	16
	Bn	cabinets:cabinet (syn)	29	girl:woman (hyp)	19	skater:skateboarder (hpo)	15	woman:girl (hpo)	12	lady:woman (hyp)	12
	De	girl:woman (hyp)	32	material:color (hpo)	23	lady:woman (hyp)	18	cabinets:cabinet (syn)	17	woman:girl (hpo)	16
	Id	girl:woman (hyp)	27	cabinets:cabinet (syn)	24	woman:girl (hpo)	17	chairs:chair (syn)	17	lady:woman (hyp)	17
	Ko	cabinets:cabinet (syn)	39	girl:woman (hyp)	34	elephants:elephant (syn)	20	woman:girl (hpo)	17	chairs:chair (syn)	17
	Pt	material:color (hpo)	25	girl:woman (hyp)	24	woman:girl (hpo)	20	zebras:zebra (syn)	15	lady:woman (hyp)	15
	Ru	girl:woman (hyp)	33	cabinets:cabinet (syn)	25	material:color (hpo)	19	woman:girl (hpo)	18	lady:woman (hyp)	16
	Zh	cabinets:cabinet (syn)	32	girl:woman (hyp)	27	chairs:chair (syn)	26	zebras:zebra (syn)	25	elephants:elephant (syn)	24

Conclusion

- We present **a series of strategies** to **fine-tune** multilingual vision-language pretrained models for **better cross-lingual generalisation** in the **VQA** task.
- The results indicate **substantial improvements** across target languages. The improvement is **+13.12** and **+12.63** in av- erage accuracy over **all 7 languages** in xGQA compared to **UC2** and **M3P** baselines, respectively.
- We perform an **analysis** of closely related target labels in xGQA:
  - propose **a new metric** that rewards synonymous predictions and further demonstrates the success of the proposed strategies.
  - **highlight** the need for future research on the **label space** and **evaluation metrics** for cross-lingual VQA.