Learning to Scale Multilingual Representations for Vision-Language Tasks Supplementary

Andrea Burns[1], Donghyun Kim[1], Derry Wijaya[1], Kate Saenko[1,2], and Bryan A. Plummer[1]

[1] Boston University, Boston MA 02215, USA
[2] MIT-IBM Watson AI Lab, Cambridge MA 02142, USA
{aburns4,donhk,wijaya,saenko,bplum}@bu.edu

1 Data Augmentation

We augment the multilingual datasets MSCOCO [5–7] and Multi30K [1–3] with translations from languages with human-generated annotations to other languages using Google Translate. Tables 1 and 2 show what translations were performed for MSCOCO and Multi30K, respectively. The column X refers to all other languages that consist entirely of translations to create the total set of ten languages; *i.e.* for MSCOCO, $X \in$ German, French, Czech, Arabic, Afrikaans, Korean, Russian, and for Multi30K, $X \in$ Chinese, Japanese, Arabic, Afrikaans, Korean, Russian. We compare the effect of using human-generated vs. machine translated sentences at test time in Section 6.

Table 1: Dataset Augmentation for MSCOCO. Arrows signify the use of machine translation, and X refers to all other languages in the total set of ten

Annotation Type	En	Cn	Ja	X
Human Generated	MSCOCO [6]	COCO-CN[5]	YJ Captions [7]	–
Translations	Cn → En Ja → En	En → Cn	En → Ja	En → X

Table 2: Dataset Augmentation for Multi30K. Arrows signify the use of machine translation, and X refers to all other languages in the total set of ten

Annotation Type	En	De	Fr	Cs	X
Human Generated	Flickr30K [9]	Multi30K [3]	Multi30K [2]	Multi30K [1]	–
Translations	De → En Fr → En Cs → En	En → De	En → Fr	En → Cs	En → X

2 Model Parameters

2.1 Exploration Parameters

One component of SMALR is the Hybrid Embedding Model (HEM), which makes use of both language-specific and language-agnostic representations. The Language-Agnostic (LA) baseline refers to only using the shared latent vocabulary, which consists of 40K tokens. We found experimentally that using exploration parameters $p = 0.2$ and $M = 20$ improves downstream performance when using the latent vocabulary. These exploration parameters are used to force the model to randomly select from a set of similar tokens during training rather than always choosing the best matched token in the language-agnostic vocabulary (described in Section 3.1 of the main paper). Tables 3 and 4 demonstrate the difference in mean Recall for image-sentence retrieval with and without our exploration parameters.

Since we find that using the exploration parameters when learning the mapping to the latent vocabulary improves performance, we use them for both the Language-Agnostic and HEM results (and thus is included in the final SMALR training paradigm).

Table 3: MSCOCO Language-Agnostic (LA) Ablation

Model	En	De[1]	Fr[1]	Cs[1]	Cn	Ja	Ar[1]	Af[1]	Ko[1]	Ru[1]	HA	A
LA	64.2	58.8	58.3	52.1	59.0	63.2	61.9	65.3	58.6	58.5	58.3	60.0
LA + Explore	**65.5**	**61.3**	**59.9**	**54.0**	**59.4**	**64.7**	**63.9**	**66.5**	**60.3**	**60.3**	**60.2**	**61.6**

[1]uses translations from English for testing

Table 4: Multi30K Language-Agnostic (LA) Ablation

Model	En	De	Fr	Cs	Cn[1]	Ja[1]	Ar[1]	Af[1]	Ko[1]	Ru[1]	HA	A
LA	73.9	73.0	71.7	72.9	72.0	70.8	72.8	72.0	69.7	72.0	72.2	72.1
LA + Explore	**75.0**	**74.3**	**74.1**	**73.4**	**72.3**	**72.1**	**74.4**	**74.7**	**71.6**	**72.7**	**73.1**	**73.5**

[1]uses translations from English for testing

2.2 Loss Parameters

Training SMALR did not require significant hyperparameter tuning. We found the results were not sensitive to our choice of lambdas used in the SMALR loss,

as defined in Eq. 4 of the main paper. Therefore, λ_1, λ_3 and λ_4 were kept the same as in prior work [16] for consistency. The parameter λ_2 is associated with the MCLM masking loss we introduce, which is determined by grid search over powers of ten on the validation set. On Multi30K, the average mR for SMALR when varying λ_2 has a performance range under one point, see Figure 1 below for exact values.

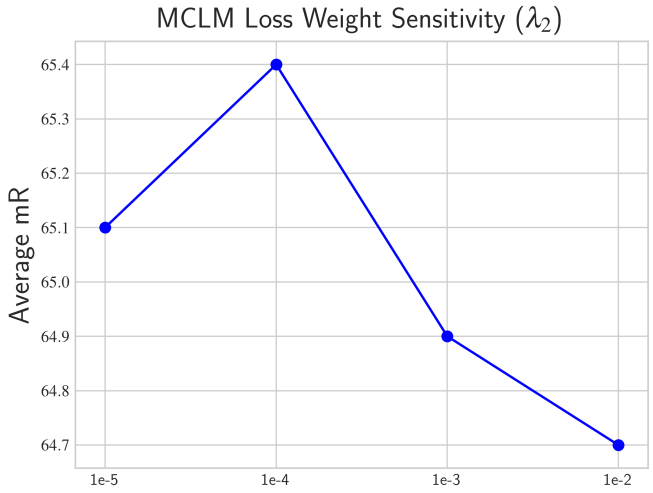

Fig. 1: M30K ablation results for the λ_2 parameter. As shown in Eq. 4 of the main paper, λ_2 determines the contribution of the masking loss to the total SMALR training loss

3 Qualitative Results

We provide two examples for both MSCOCO and Multi30K which show the effect of the Cross-Lingual Consistency (CLC) module used with SMALR. We report results for the CLC-C variant, which makes use of a simple MLP classifier to aggregate scores across language. For a given text query, if it is human generated, we translate it to all other languages and use the predictions from these translations as input to our CLC-C module.

On the left hand side of Figure 2, the original text query is in English and its matching image is incorrectly retrieved, as shown by the red bounding box. However, when CLC-C is used, SMALR is able to correctly retrieve the matching image, as a subset of the translated sentences do correctly retrieve the ground truth image (*e.g.* the German translation). On the right hand side of Figure 2, we also see the same benefit for an original text query in German which is aided by English translations. These two examples demonstrate the benefit of CLC-C for R@1, as CLC-C now correctly retrieves the ground truth image. Additionally, these samples show that every language does not have to make the correct prediction; the CLC-C module can learn to combine predictions to improve performance. As we can see in Figure 2, the images incorrectly retrieved for the original English and German queries "People are walking through a vegetable stall filled market" and "Der mann trägt eine orange wollmütze" contain very similar objects and colors to their respective ground truth images, but these errors are remedied when considering all languages.

In Figure 3, there are two examples for MSCOCO, with original text queries in English and Chinese. Both examples have many translated queries which are able to correctly retrieve the ground truth image, such as French and Russian for English, and English, German, and French (among others) for Chinese. We see again that the original incorrectly retrieved image contains very similar visual semantics (*e.g.* teddy bear for English, baseball field for Chinese) to the ground truth, and the translated sentences help disambiguate subtle details.

Fig. 2: Example of the benefits of using the CLC module on Multi30K

Fig. 3: Example of the benefits of using the CLC module on MSCOCO

4 Masked Cross-Language Modeling Example

SMALR's Masked Cross-Language Model (MCLM) uses two language representations to compute its total loss, namely an average representation, and a sentence-level LSTM representation. The average masked sentence simply removes masked words and then averages each word embedding over the shorter version of the original sentence before predicting the masked token. The masked sentence-level representation retains the same number of words by replacing the masked words with a special [MASK] token; not only does this retain the total word count for a given query, it also maintains grammatical structure by using a LSTM. This representation is passed through a LSTM and fully connected layer before being used to predict the masked token. Figure 4 provides an example of this process; the word boxes represent word embeddings. See Section 3.2 of the main paper for a description of how these representations are used.

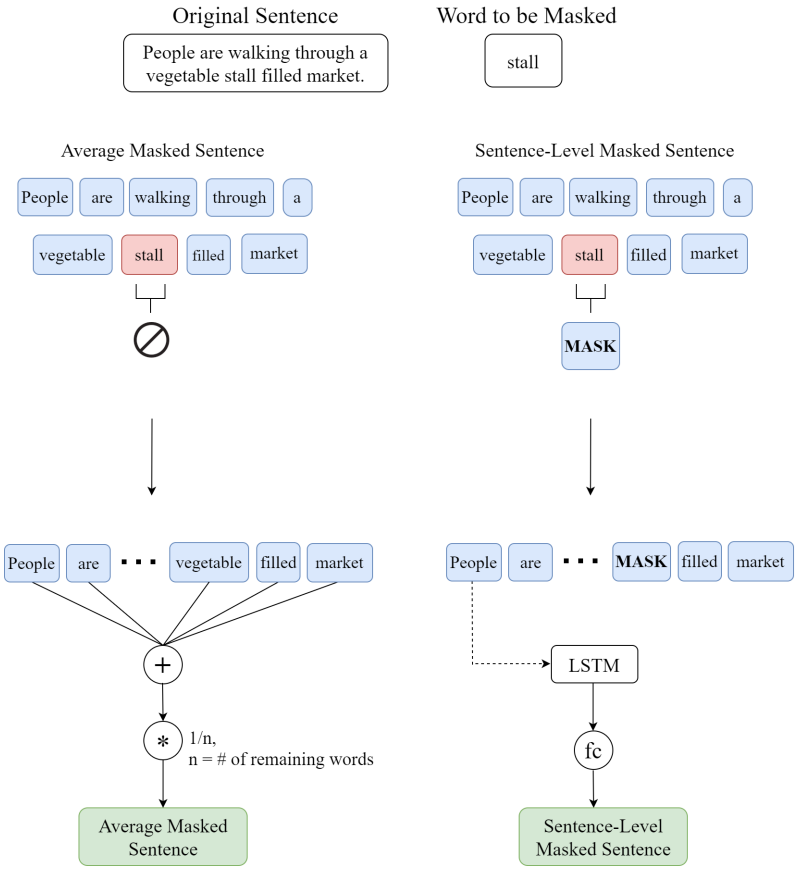

Fig. 4: Variants of masking used in the MCLM module

5 Extended Image-Sentence Retrieval Results

We provide all recall values (Recall@K for K$\in \{1, 5, 10\}$) for all ten languages on image-sentence retrieval with MSCOCO and Multi30K. I-to-S signifies the image to sentence retrieval direction, and S-to-I the sentence to image direction. We shorten "Language-Agnostic" to "LA" and CLC-A, CLC-C to A and C, respectively, due to space constraints. Lastly, "Prior" refers to prior work, "Adapted" refers to prior work that has been adapted to our testing scenario using the author's publicly available code, and "Ours" refers to our SMALR model variants. The number preceding a model refers to the number of languages it was trained on, *e.g.* (3-4) MULE signifies MULE [4] trained on three languages (English, Chinese, Japanese) on MSCOCO, and four on Multi30K (English, German, French, Czech).

Table 5: English bidirectional image-sentence retrieval results using human-generated sentences

Model	MSCOCO							Multi30K						
	I-to-S			S-to-I			mR	I-to-S			S-to-I			mR
	r@1	r@5	r@10	r@1	r@5	r@10		r@1	r@5	r@10	r@1	r@5	r@10	
(a) Prior														
Trans. to En	58.6	86.5	94.1	45.5	79.6	89.5	75.6	58.3	82.9	90.4	41.7	72.0	81.2	71.1
EmbN	61.8	87.6	94.1	47.5	79.8	89.8	76.8	57.9	84.5	90.9	44.3	72.7	84.7	72.0
PAR. EmbN	63.1	89.1	94.1	49.2	82.5	91.5	78.3	52.4	80.1	87.7	41.6	71.5	80.7	69.0
(3-4) MULE	63.9	90.2	95.8	50.9	83.5	92.4	79.5	54.2	82.0	89.9	41.9	72.5	81.1	70.3
(b) Adapted														
(1) S-LIWE	**66.8**	91.2	96.6	52.4	85.1	93.5	80.9	**65.5**	88.9	**95.1**	46.9	77.2	84.5	**76.3**
(2) S-LIWE	62.3	87.3	94.6	48.3	80.7	91.0	77.4	64.5	88.1	94.3	46.4	75.8	84.5	75.6
(10) S-LIWE	61.8	88.2	94.8	47.9	80.3	90.5	77.3	63.8	88.0	93.4	46.4	75.4	84.3	75.2
(10) L-LIWE	63.8	90.2	95.6	50.1	82.9	92.2	79.1	63.9	**89.0**	94.2	46.9	76.8	84.8	75.9
(10) MULE	63.8	88.9	95.5	50.5	83.2	92.0	79.0	55.2	82.1	90.7	42.2	72.2	81.8	70.7
(c) Ours														
LA	56.4	84.9	92.3	46.0	80.5	90.2	75.0	48.1	77.2	86.9	36.5	67.1	77.2	65.5
HEM	61.6	89.0	95.4	50.5	83.3	92.4	78.7	51.3	79.9	88.4	41.8	72.1	81.5	69.2
SMALR	62.9	89.2	95.8	51.1	84.0	92.5	79.3	52.0	81.1	88.4	41.8	72.4	82.1	69.6
SMALR-A	66.6	91.1	97.3	52.8	85.7	93.4	81.2	59.4	83.7	90.2	47.5	77.5	86.1	74.1
SMALR-C	66.5	**91.3**	**97.5**	**53.6**	**86.2**	**94.0**	**81.5**	60.2	83.8	91.0	**47.9**	**77.9**	**86.3**	74.5

Table 6: German bidirectional image-sentence retrieval results using sentences translated from English into German for testing on MSCOCO and human-generated sentences on Multi30K

Model	MSCOCO							Multi30K						
	I-to-S			S-to-I			mR	I-to-S			S-to-I			mR
	r@1	r@5	r@10	r@1	r@5	r@10		r@1	r@5	r@10	r@1	r@5	r@10	
(a) Prior														
Trans. To En	–	–	–	–	–	–	–	34.1	60.4	71.1	19.6	47.4	58.5	48.5
EmbN	–	–	–	–	–	–	–	46.6	73.9	82.2	31.3	59.1	69.0	60.3
PAR. EmbN	–	–	–	–	–	–	–	46.1	76.3	83.2	34.4	62.5	73.0	62.6
(3-4) MULE	–	–	–	–	–	–	–	49.7	77.7	85.7	34.6	63.4	73.5	64.1
(b) Adapted														
(1) S-LIWE	–	–	–	–	–	–	–	**61.1**	**86.6**	**92.7**	42.0	69.9	80.0	**72.1**
(2) S-LIWE	–	–	–	–	–	–	–	51.2	80.2	88.4	35.7	65.7	75.2	66.1
(10) S-LIWE	49.8	79.1	87.3	36.6	69.4	82.4	67.4	50.5	79.0	88.0	34.9	64.3	74.3	65.2
(10) L-LIWE	52.1	84.9	92.6	39.3	73.4	85.0	71.2	51.1	80.5	89.8	35.9	66.6	76.1	66.7
(10) MULE	59.1	88.7	94.9	48.5	81.3	90.6	77.2	45.8	75.8	85.2	35.1	64.6	75.3	63.6
(c) Ours														
LA	54.4	86.2	93.1	44.5	78.6	88.7	74.3	44.0	75.4	85.1	32.2	59.7	71.0	61.3
HEM	59.2	87.2	95.1	49.1	81.8	91.4	77.3	49.2	75.4	83.2	34.5	62.0	72.4	62.8
SMALR	**61.2**	**89.2**	**96.2**	**49.6**	**82.3**	**91.8**	**78.4**	49.9	75.8	85.0	36.9	65.4	75.4	64.7
SMALR-A	–	–	–	–	–	–	–	53.0	77.6	85.8	41.9	72.9	82.3	68.9
SMALR-C	–	–	–	–	–	–	–	52.9	78.8	87.0	**42.6**	**74.2**	**83.1**	69.8

Table 7: French bidirectional image-sentence retrieval results using sentences translated from English into French for testing on MSCOCO and human-generated sentences on Multi30K

Model	MSCOCO							Multi30K						
	I-to-S			S-to-I			mR	I-to-S			S-to-I			mR
	r@1	r@5	r@10	r@1	r@5	r@10		r@1	r@5	r@10	r@1	r@5	r@10	
(a) Prior														
Trans. to En	–	–	–	–	–	–	–	22.5	52.5	63.0	25.1	53.1	63.9	46.7
EmbN	–	–	–	–	–	–	–	31.0	60.4	71.0	35.2	60.3	70.8	54.8
PAR. EmbN	–	–	–	–	–	–	–	37.6	66.0	77.4	37.8	66.4	78.2	60.6
(3-4) MULE	–	–	–	–	–	–	–	38.0	68.4	80.0	38.2	68.9	80.3	62.3
(b) Adapted														
(10) S-LIWE	50.8	79.3	90.4	36.5	70.7	83.2	68.5	39.0	39.0	51.6	37.3	66.7	77.3	51.8
(10) L-LIWE	51.8	81.3	92.2	39.0	73.1	84.7	70.3	40.6	40.7	54.7	37.8	69.3	78.1	53.5
(10) MULE	**60.3**	86.9	94.3	47.8	81.3	90.4	76.8	39.2	70.9	80.7	38.8	70.5	80.2	63.4
(c) Ours														
LA	54.8	83.6	92.6	44.8	79.4	89.7	74.1	35.1	65.8	76.0	39.5	65.6	77.2	59.9
HEM	57.6	87.0	94.0	48.0	80.7	**91.1**	76.4	38.1	70.5	80.6	40.2	69.5	80.6	63.3
SMALR	59.6	**89.7**	**95.9**	**48.7**	**81.9**	91.0	**77.8**	40.6	70.7	81.8	41.1	71.8	80.7	64.5
SMALR-A	–	–	–	–	–	–	–	40.3	**73.4**	80.9	42.2	72.8	81.8	65.2
SMALR-C	–	–	–	–	–	–	–	**41.1**	**73.4**	**82.5**	**42.6**	**73.0**	**82.9**	**65.9**

Table 8: Czech bidirectional image-sentence retrieval results using sentences translated from English into Czech for testing on MSCOCO and human-generated sentences on Multi30K

Model	MSCOCO							Multi30K						
	I-to-S			S-to-I			mR	I-to-S			S-to-I			mR
	r@1	r@5	r@10	r@1	r@5	r@10		r@1	r@5	r@10	r@1	r@5	r@10	
(a) Prior														
Trans. to En	–	–	–	–	–	–	–	23.0	50.9	64.7	25.1	53.4	64.2	46.9
EmbN	–	–	–	–	–	–	–	26.2	51.3	62.5	26.8	50.3	60.8	46.3
PAR. EmbN	–	–	–	–	–	–	–	31.4	58.2	70.1	33.1	60.4	71.6	54.1
(3-4) MULE	–	–	–	–	–	–	–	34.3	63.2	74.2	35.3	63.6	75.5	57.7
(b) Adapted														
(10) S-LIWE	46.8	79.8	90.3	34.6	68.2	82.0	66.9	36.5	36.5	50.0	37.6	64.3	75.2	50.0
(10) L-LIWE	50.7	82.3	92.1	37.6	72.8	84.8	70.1	37.6	37.6	52.9	38.1	66.2	75.2	51.3
(10) MULE	61.6	88.7	94.8	48.8	81.5	91.1	77.8	37.0	66.3	76.4	37.5	64.6	74.8	59.4
(c) Ours														
LA	55.3	84.6	92.4	43.5	76.9	87.8	73.4	31.0	59.6	71.1	32.5	58.5	71.5	54.0
HEM	59.9	88.4	95.4	**49.2**	**82.5**	**91.7**	77.9	35.0	66.9	77.4	36.1	67.4	77.2	60.0
SMALR	**63.2**	**89.6**	**95.7**	**49.2**	82.4	91.6	**78.6**	36.5	69.0	78.0	36.7	68.0	78.2	61.1
SMALR-A	–	–	–	–	–	–	–	41.1	**70.7**	80.4	39.9	**71.8**	**83.0**	64.5
SMALR-C	–	–	–	–	–	–	–	**41.9**	**70.7**	**81.1**	**40.5**	71.7	82.8	**64.8**

Table 9: Chinese bidirectional image-sentence retrieval results using sentences translated from English into Chinese for testing on Multi30K and human-generated sentences on MSCOCO

Model	MSCOCO							Multi30K						
	I-to-S			S-to-I			mR	I-to-S			S-to-I			mR
	r@1	r@5	r@10	r@1	r@5	r@10		r@1	r@5	r@10	r@1	r@5	r@10	
(a) Prior														
Trans. to En	45.9	79.8	89.2	47.8	81.1	89.4	72.2	–	–	–	–	–	–	–
EmbN	49.6	81.6	90.0	47.8	82.1	90.0	73.5	–	–	–	–	–	–	–
PAR. EmbN	47.9	81.4	91.1	47.5	81.6	91.2	73.5	–	–	–	–	–	–	–
(3-4) MULE	51.1	82.6	91.6	49.1	82.4	91.9	74.8	–	–	–	–	–	–	–
(b) Adapted														
(10) S-LIWE	45.1	76.4	88.1	32.7	66.0	79.6	64.5	39.3	68.1	79.2	24.2	51.0	62.5	54.1
(10) L-LIWE	51.4	82.6	91.3	38.1	72.2	84.6	70.0	42.6	72.4	82.4	26.0	53.6	64.7	56.9
(10) MULE	50.8	84.0	92.5	50.3	83.6	92.4	75.6	**47.4**	**77.0**	**85.8**	35.4	64.9	74.4	**64.2**
(c) Ours														
LA	46.0	79.6	90.7	45.9	80.6	91.1	72.3	42.2	72.0	81.6	30.6	59.8	70.0	59.4
HEM	53.2	85.0	93.2	51.3	84.6	93.0	76.7	44.1	74.7	84.4	33.8	63.3	74.4	62.4
SMALR	51.2	86.5	93.8	50.6	84.7	93.3	76.7	45.8	**77.0**	85.0	**35.8**	**65.1**	**75.5**	64.0
SMALR-A	57.5	87.3	94.9	54.8	87.7	95.2	79.6	–	–	–	–	–	–	–
SMALR-C	**58.0**	**87.8**	**95.4**	**55.3**	**88.2**	**95.7**	**80.1**	–	–	–	–	–	–	–

Table 10: Japanese bidirectional image-sentence retrieval results using sentences translated from English into Japanese for testing on Multi30K and human-generated sentences on MSCOCO

Model	MSCOCO							Multi30K						
	I-to-S			S-to-I			mR	I-to-S			S-to-I			mR
	r@1	r@5	r@10	r@1	r@5	r@10		r@1	r@5	r@10	r@1	r@5	r@10	
(a) Prior														
Trans. to En	44.8	74.3	85.4	36.9	71.0	84.7	66.1	–	–	–	–	–	–	–
EmbN	56.0	83.7	90.7	45.5	77.2	87.3	73.2	–	–	–	–	–	–	–
PAR. EmbN	60.1	86.0	92.8	47.7	79.6	89.7	76.0	–	–	–	–	–	–	–
(3-4) MULE	59.6	**86.5**	92.8	47.8	80.8	90.1	76.3	–	–	–	–	–	–	–
(b) Adapted														
(1) S-LIWE	57.2	85.0	93.2	42.2	76.4	87.6	73.6	–	–	–	–	–	–	–
(2) S-LIWE	45.3	78.2	89.5	36.4	68.9	81.2	66.6	–	–	–	–	–	–	–
(10) S-LIWE	45.9	77.9	88.2	34.1	67.5	81.2	65.8	41.8	72.4	82.1	25.3	52.4	63.4	56.2
(10) L-LIWE	51.5	81.4	90.2	39.1	71.4	84.3	69.6	40.5	71.0	82.1	26.2	53.5	64.7	56.3
(10) MULE	59.4	85.2	93.0	47.4	80.1	90.2	75.9	**49.9**	**80.2**	**87.7**	38.1	**69.3**	78.6	**67.3**
(c) Ours														
LA	51.4	83.3	90.3	42.4	76.8	88.1	72.1	48.4	77.2	85.7	35.5	65.1	76.5	64.7
HEM	56.8	86.3	93.8	47.7	81.7	**91.7**	76.3	48.9	78.4	86.0	38.4	68.0	78.3	66.3
SMALR	60.4	86.4	**94.3**	48.5	82.2	91.2	77.2	46.8	79.1	87.6	**38.8**	69.1	**78.8**	66.7
SMALR-A	60.0	84.5	92.9	45.9	78.3	88.6	75.0	–	–	–	–	–	–	–
SMALR-C	**61.9**	86.4	94.0	**49.3**	**81.9**	91.3	**77.5**	–	–	–	–	–	–	–

Table 11: Arabic bidirectional image-sentence retrieval results using sentences translated from English into Arabic for testing

Model	MSCOCO							Multi30K						
	I-to-S			S-to-I			mR	I-to-S			S-to-I			mR
	r@1	r@5	r@10	r@1	r@5	r@10		r@1	r@5	r@10	r@1	r@5	r@10	
(a) Adapted														
(10) S-LIWE	43.4	75.6	86.3	33.1	65.7	78.6	63.8	47.5	76.9	84.7	33.3	61.9	72.0	62.7
(10) L-LIWE	49.1	81.4	90.3	34.4	68.4	81.0	67.5	**48.7**	78.4	**87.9**	34.3	65.3	75.5	65.0
(10) MULE	**60.3**	88.3	94.6	47.9	81.2	90.7	77.2	48.6	78.2	87.4	36.7	66.8	76.9	65.8
(b) Ours														
LA	56.1	85.5	93.6	44.0	78.3	88.7	74.4	44.7	78.1	85.6	34.5	65.2	75.3	63.9
HEM	58.4	87.9	94.9	47.6	81.5	91.4	77.0	45.9	76.8	85.6	36.3	66.2	76.2	64.5
SMALR	60.1	**89.0**	**95.7**	**48.6**	**81.9**	**91.9**	**77.9**	46.2	**78.6**	87.4	**38.3**	**67.7**	**77.9**	**66.0**

Table 12: Afrikaans bidirectional image-sentence retrieval results using sentences translated from English into Afrikaans for testing

Model	MSCOCO							Multi30K						
	I-to-S			S-to-I			mR	I-to-S			S-to-I			mR
	r@1	r@5	r@10	r@1	r@5	r@10		r@1	r@5	r@10	r@1	r@5	r@10	
(a) Adapted														
(10) S-LIWE	46.7	79.1	88.8	35.0	67.4	80.2	66.2	49.8	77.5	85.1	32.8	60.6	70.7	62.8
(10) L-LIWE	49.9	82.2	91.8	36.9	70.6	82.4	68.9	49.5	77.4	86.3	34.0	62.3	72.6	63.7
(10) MULE	62.4	88.1	94.8	48.7	81.5	91.0	77.8	51.3	80.2	**87.7**	39.0	67.7	77.7	67.3
(b) Ours														
LA	55.2	85.1	92.7	45.7	79.9	89.5	74.7	**51.5**	78.9	86.5	37.8	67.2	77.2	66.5
HEM	59.8	86.4	93.9	47.5	81.2	91.2	76.7	47.6	79.3	87.4	38.7	**69.2**	78.9	66.8
SMALR	**62.5**	**88.7**	**95.9**	**48.8**	**82.2**	**91.4**	**78.2**	48.7	**79.7**	87.5	**40.5**	68.8	**79.1**	**67.4**

Table 13: Korean bidirectional image-sentence retrieval results using sentences translated from English into Korean for testing

Model	MSCOCO							Multi30K						
	I-to-S			S-to-I			mR	I-to-S			S-to-I			mR
	r@1	r@5	r@10	r@1	r@5	r@10		r@1	r@5	r@10	r@1	r@5	r@10	
(a) Adapted														
(10) S-LIWE	42.4	76.4	86.9	31.9	63.6	77.4	63.1	38.7	69.1	80.2	24.3	51.8	62.7	54.5
(10) L-LIWE	48.1	79.8	89.5	33.4	66.3	79.9	66.2	41.2	73.5	83.2	26.3	53.4	65.0	57.1
(10) MULE	56.5	85.6	93.5	43.9	78.0	88.5	74.3	**47.1**	76.1	85.3	35.0	63.7	74.6	63.6
(b) Ours														
LA	51.9	85.0	92.2	40.2	73.8	86.4	71.6	43.3	73.4	83.0	31.3	59.4	71.1	60.3
HEM	**57.0**	85.8	94.6	**46.2**	**79.2**	**90.0**	**75.5**	44.4	76.6	85.4	32.9	62.0	72.2	62.3
SMALR	55.7	**86.9**	**94.8**	45.2	78.8	89.4	75.1	45.7	**78.2**	**85.5**	**35.2**	**64.8**	**75.5**	**64.2**

Table 14: Russian bidirectional image-sentence retrieval results using sentences translated from English into Russian for testing

Model	MSCOCO							Multi30K						
	I-to-S			S-to-I			mR	I-to-S			S-to-I			mR
	r@1	r@5	r@10	r@1	r@5	r@10		r@1	r@5	r@10	r@1	r@5	r@10	
(a) Adapted														
(10) S-LIWE	44.7	76.1	86.5	31.4	64.5	78.2	63.6	47.7	77.4	84.6	33.6	62.9	72.6	63.1
(10) L-LIWE	50.7	83.6	91.5	36.7	71.7	83.0	69.6	**49.7**	**79.5**	**86.9**	35.2	65.2	76.0	65.4
(10) MULE	60.8	**89.0**	94.9	48.0	80.4	90.4	77.3	48.3	78.6	86.2	37.1	65.8	76.2	65.4
(b) Ours														
LA	53.7	85.0	92.1	42.4	75.8	87.4	72.7	42.2	72.7	82.8	31.6	60.7	71.7	60.3
HEM	58.3	87.5	94.4	**48.5**	**81.8**	**91.7**	77.0	45.6	75.0	83.6	35.3	63.2	73.1	62.6
SMALR	**62.7**	88.8	**95.0**	48.2	81.7	91.5	**78.0**	48.4	77.3	86.0	**38.0**	**67.2**	**77.5**	**65.7**

6 Testing with Machine Translations

In this section we investigate the effect testing with machine translations rather than human-generated sentences has when comparing methods. For all methods we use models trained on all 10 languages, and test on human-generated and translated sentences for Chinese and Japanese on MSCOCO and German, French, and Czech on Multi30k.

As seen below, there are only minor differences in the performance of each language we tested. Notably, the performance rankings with each dataset are consistent regardless of whether the method is evaluated on human generated test sentences or test sentences translated from English.

Table 15: Comparison of using Human Generated Sentences vs. Translations for testing purposes.

Model	MSCOCO				Multi30k				
	mR		Avg	Rank	mR			Avg	Rank
	Cn	Ja			De	Fr	Cs		
(a) Human generated test sentences									
(10) S-LIWE [8]	64.5	65.8	65.2	6	65.2	51.8	50.0	55.7	6
(10) L-LIWE	70.0	69.6	69.8	5	**66.7**	53.5	51.3	57.2	5
(10) MULE [4]	75.6	75.9	75.8	3	63.6	63.4	59.4	62.1	2
Language-Agnostic	72.3	72.1	72.2	4	61.3	59.9	54.0	58.4	4
HEM	**76.7**	76.3	76.5	2	62.8	63.3	60.0	62.0	3
SMALR	**76.7**	**77.2**	**76.9**	1	64.7	**64.5**	**61.1**	**63.4**	1
(b) Test sentences translated from En									
(10) S-LIWE [8]	64.7	65.8	65.2	6	64.3	49.9	52.1	55.4	6
(10) L-LIWE	70.0	69.6	69.8	5	**65.8**	51.8	54.8	57.5	5
(10) MULE [4]	73.2	75.0	74.1	3	64.4	64.0	64.8	64.4	2
Language-Agnostic	69.7	71.4	70.6	4	61.7	61.1	60.5	61.1	4
HEM	73.5	75.3	74.4	2	63.8	63.5	64.3	63.9	3
SMALR	**74.4**	**75.9**	**75.2**	1	65.1	**65.1**	**65.5**	**65.2**	1

References

1. Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., Frank, S.: Findings of the third shared task on multimodal machine translation. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 304–323 (2018)
2. Elliott, D., Frank, S., Barrault, L., Bougares, F., Specia, L.: Findings of the second shared task on multimodal machine translation and multilingual image description. arXiv:1710.07177 (2017)
3. Elliott, D., Frank, S., Sima'an, K., Specia, L.: Multi30k: Multilingual english-german image descriptions. arXiv:1605.00459 (2016)
4. Kim, D., Saito, K., Saenko, K., Sclaroff, S., Plummer, B.A.: Mule: Multimodal universal language embedding. In: AAAI Conference on Artificial Intelligence (2020)
5. Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G., Xu, J.: Coco-cn for cross-lingual image tagging, captioning and retrieval. IEEE Transactions on Multimedia (2019)
6. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: The European Conference on Computer Vision (ECCV) (2014)
7. Miyazaki, T., Shimizu, N.: Cross-lingual image caption generation. In: Conference of the Association for Computational Linguistics (ACL) (2016)
8. Wehrmann, J., Souza, D.M., Lopes, M.A., Barros, R.C.: Language-agnostic visual-semantic embeddings. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
9. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (TACL) **2**, 67–78 (2014)