Supplementary Material:
Language-Agnostic Visual-Semantic Embeddings

1. Hyper-Parameters and Training Details

We choose hyper-parameters based on the results over validation data. We employ Adam for optimization, with an initial learning rate of 6×10^{-4}, which is further decreased $10\times$ at the 15^{th} epoch for Flickr and M30K datasets, and at the 10^{th} epoch for COCO and YJ Captions datasets. We use a batch size of 128, leading us to select the hard-contrastives among the 127 remaining pairs. We set the margin $\alpha = 0.2$, and the exponential increase of the hard-contrastive weight $\epsilon = 0.991$. The number of filters or neurons f for each layer l is depicted as -version(f_1, f_2, ..., f_l), and thus (384, 448) is an incarnation that generates word embeddings with two fully-connected layers with 384 and 448 neurons, respectively. Note that, unlike several state-of-the-art methods, our models were trained using the same hyper-parameters across all datasets.

2. Additional Experiments

In this section we report additional results that analyze the impact of the number of neurons (Table 1) within LIWE, and its impact on the reduction of the number of parameters when compared to CLMR.

Table 1. Impact of the number of parameters in LIWE.

Method	Image to text		Text to image		#Params(Reduct.)
	R@1	R@10	R@1	R@10	
BERT+GRU (12-layers)	61.4	92.5	43.2	81.4	110.0M (0.08×)
CMLR-GRU	65.8	93.1	47.3	84.2	9.0M (1×)
LIWE-GRU(charGRU)	62.0	92.3	44.8	82.7	0.3M (30×)
LIWE-GRU(128,256)	64.2	93.0	45.5	83.0	0.2M (46×)
LIWE-GRU(128,384)	64.0	93.0	47.2	84.2	0.2M (36×)
LIWE-GRU(256,256)	65.3	92.5	47.4	84.6	0.3M (29×)
LIWE-GRU(128,512)	67.2	93.5	48.0	84.5	0.3M (29×)
LIWE-GRU(384,384)	66.6	94.8	48.8	85.6	0.5M (18×)
LIWE-GRU(512,512)	65.6	93.9	47.2	85.2	0.7M (12×)
LIWE-GRU(1024,512)	67.2	94.4	48.7	85.8	1.3M (7×)

3. Qualitative Results

In this section we show several randomly selected qualitative results for both image retrieval (text-to-image) and image annotation (image-to-text) tasks. Those examples are depicted in Japanese, English and German languages.

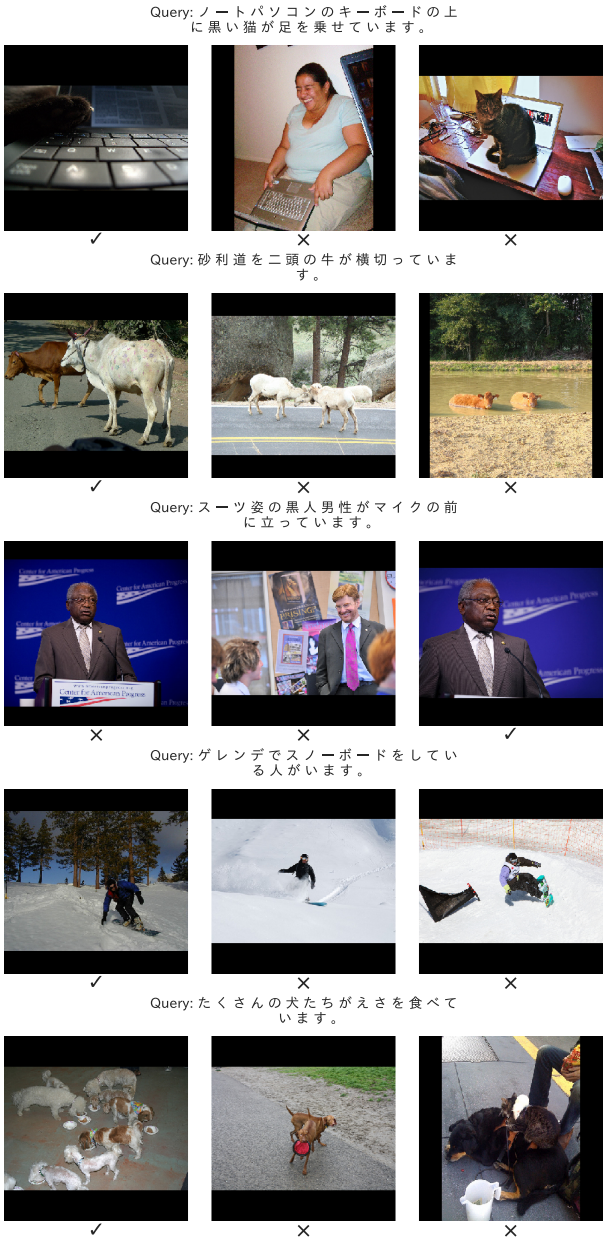
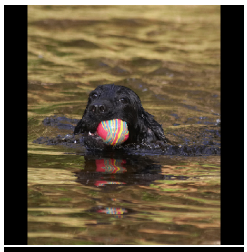


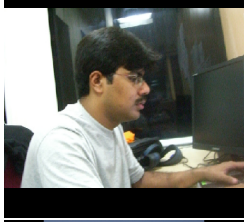
Figure 1. Example of image retrieval in the Japanese language using the language-independent model LIWE.

Figure 2. Example of image retrieval in English and German languages using the language-independent model LIWE.

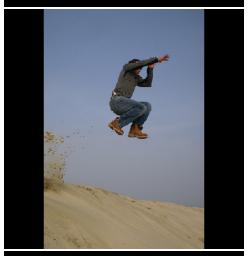
Figure 3. Example of image retrieval in English and German languages using the language-independent model LIWE.



1. Ein schwarzer Hund mit einem bunten Ball im Mund schwimmt im Wasse. ✓
2. Der Hund mit dem Ball im Maul schwimmt im Wasse. ✓
3. Ein Hund schwimmt im Wasser mit einem Ball im Mun. ✓
4. Einen Hund im Wasser der einen Ball hol. ✓
5. schwarzer Hund schwimmen, in den Mund bunte Ku

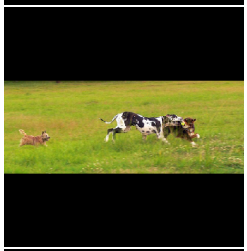


1. Ein asiatisch aussehender Mann mit dunklen Haarer und Brille sitzt vor einem Compute. ✓
2. Der Mann sitzt vor dem P. ✓
3. Ein Mann sitzt vor einem Compute. ×
4. Ein Mann arbeitet am Compute. ✓
5. Ein Mann von oben betrachtet, der an einem Schreibtisch an einem Computer arbeite. ×



1. Ein Mann mit braunen Schuhe und blauer Jeans springt mit Schwung von einer Art Sanddün. ✓
2. Ein Mann in grauen Long Shirt und Blue Jeans springt eine Sanddüne hinab.E. ✓
3. Ein Mann springt eine sandigen Abhang hinunte. ✓
4. Man sieht einen Mann der über einer Düne springt mit ausgestreckten Armen und angezogener
5. Ein Mann springt mit angezogenen Beinen über eine Dün. ✓

1. A woman is standing on the deck of a boat while holding a weight. ✓
2. A woman leans over the bow of a boat while a man walks by. ×
3. A man is standing in a boat holding some netting. ×
4. A woman raising the anchors on a boat. ✓
5. a woman peering over the edge of a boat while a man walks behind her. ×



1. A small black and white dog is followed by a larger dog in a grassy yard. ×
2. Two dogs play together and another runs after in a field. ✓
3. Two black and one white dog interacting in the grass. ×
4. Three dogs play together in the field. ✓
5. Three dogs race towards the viewer over a lawn. ×

1. A young boy is looking at a checkers board with one fist leaning against his head as a partially visible adult sits across the table from him , in a room full of people sitting in chairs around tables. ✓
2. A woman sits at a primary colored children 's table , playing with building blocks , as two girls in pink dresses , and a boy in a red shirt surround her. ×
3. Children are eating food at a table. ×
4. The kids are listening to their teacher. ×
5. two kids sitting at a table eating. ×

1. A young girl swimming in a pool. ✓
2. a young girl swimming in a pool. ✓
3. A dark-skinned girl with goggles and black hair in water. ✓
4. A young girl with goggles strapped to her forehead enjoys the cool pool water. ✓
5. A young girl with goggles is floating in the pool. ✓

Figure 4. Example of image annotation in English and German languages using the language-independent model LIWE.