Unsupervised Bilingual Lexicon Induction from Mon-lingual Multimodal Data

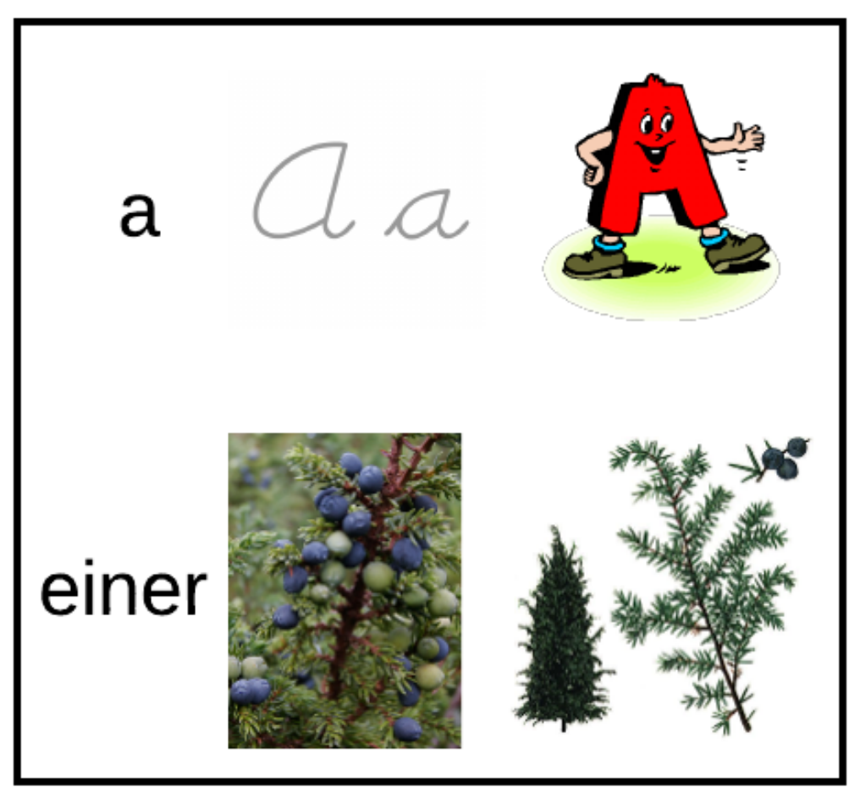
Shizhe Chen[1], Qin Jin[1] and Alexander Hauptmann[2]

{cszhe1, qjin}@ruc.edu.cn, alex@cs.cmu.edu

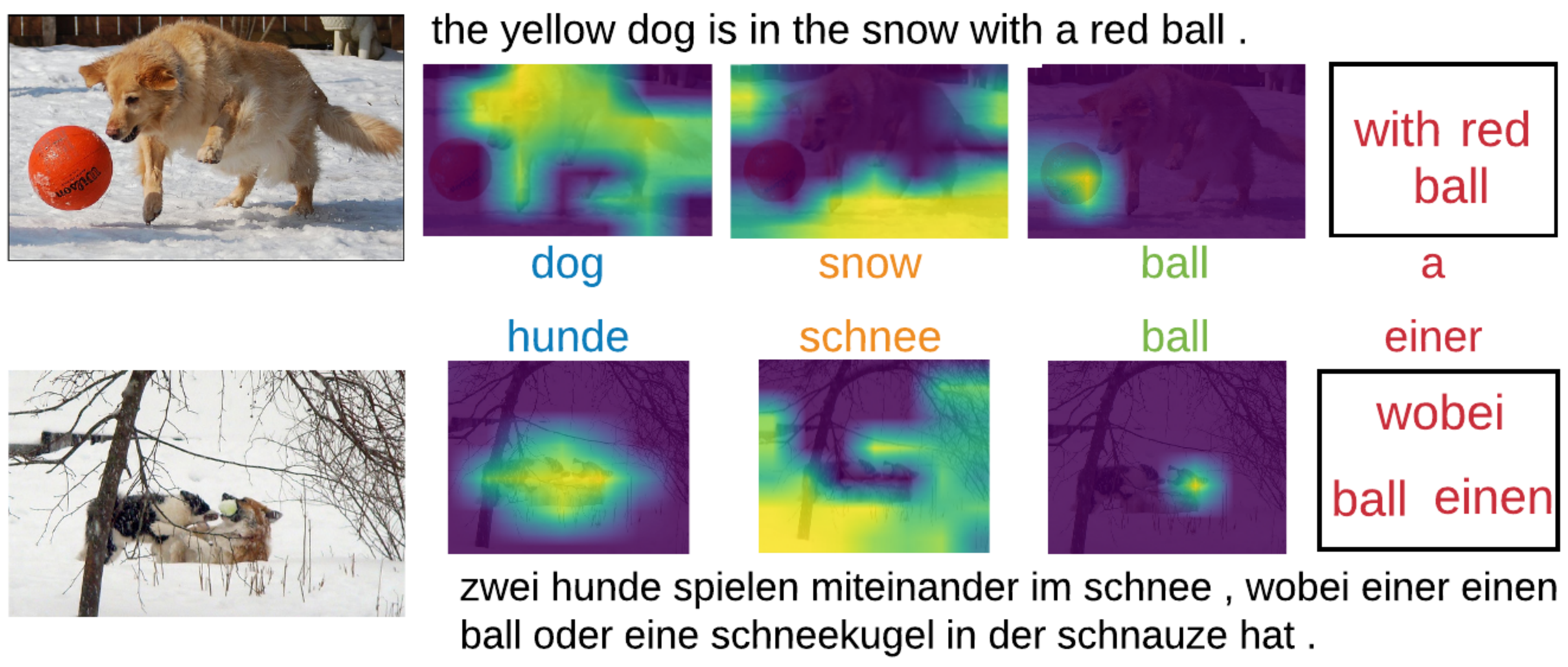
[1]*Renmin University of China*, [2]*Carnegie Mellon University*

Introduction

- **Unsupervised Bilingual Lexicon Induction**
  Aiming to translate words from the source language to the target language without parallel sentences or seed words. It can benefit various natural language processing tasks such as cross-lingual information retrieval, machine translation and so on.
- **Vision-based Unsupervised Approach**
  Utilizing vision as bridge to connect different languages with the assumption that words correlating to similar images should share similar semantic meanings.
- **Prior Vision-based Approach**
  Searching images with multi-lingual words and representing the words as global image features. Drawbacks: 1) requiring object-centered images; 2) unreliable for non-concrete words.

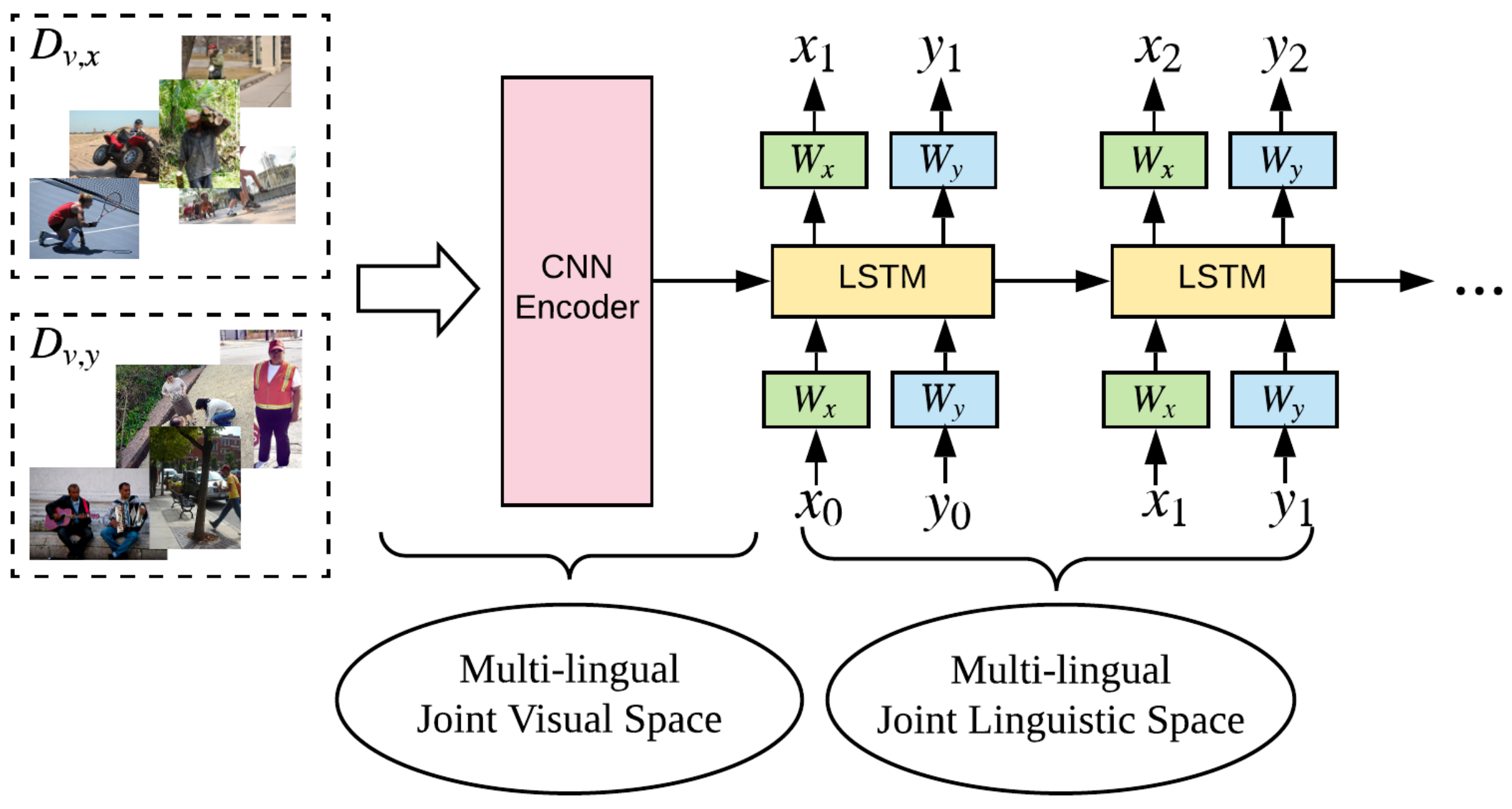


- **Our Contributions**
  1) To the best of knowledge, we are **the first to explore images with sentences for unsupervised bilingual lexicon induction**, which mitigates two main limitations of prior vision-based approaches.
  2) We propose a multi-lingual caption model to induce **linguistic features and grounded visual features** for multilingual words in joint spaces, which are complementary to enhance word translation.
  3) Our approach achieves **significant improvements** over prior vision-based methods in **all part-of-speech classes for different language pairs** such as De-En, Fr-En and Jp-En.

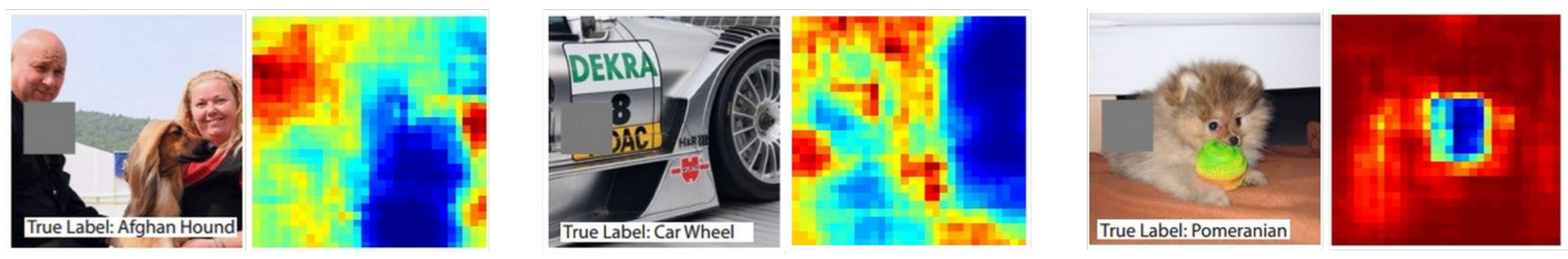

the yellow dog is in the snow with a red ball .
with red ball
dog
snow
ball
a
hunde
schnee
ball
einer
wobei
ball einen
zwei hunde spielen miteinander im schnee , wobei einer einen ball oder eine schneekugel in der schnauze hat .

The Proposed Approach

- **Multilingual Caption Model**
  Sharing the image encoder and language decoder for image captioning in different languages, so it can: 1) explicitly build multi-lingual word embeddings in the joint linguistic space; 2) implicitly ground each word in the joint visual space.



- **Linguistic Features**: embedding words in the same linguistic space.
  Since images in multi-lingual datasets share the same visual space, the sentence features x and y belonging to similar images are bound to be close in the same space. Meanwhile, the language decoder is shared. Therefore, the word embeddings W_x and W_y are enforced into the same space.
- **Grounded Visual Features**: grounding words in the same visual space.
  The visual features are in the same space due to the shared image encoder. In order to ground each word in the image, we occlude over regions of the image to observe the probability change of each word in the sentence, which is learned unsupervisedly and generalizable to different caption models.



Experimental Results

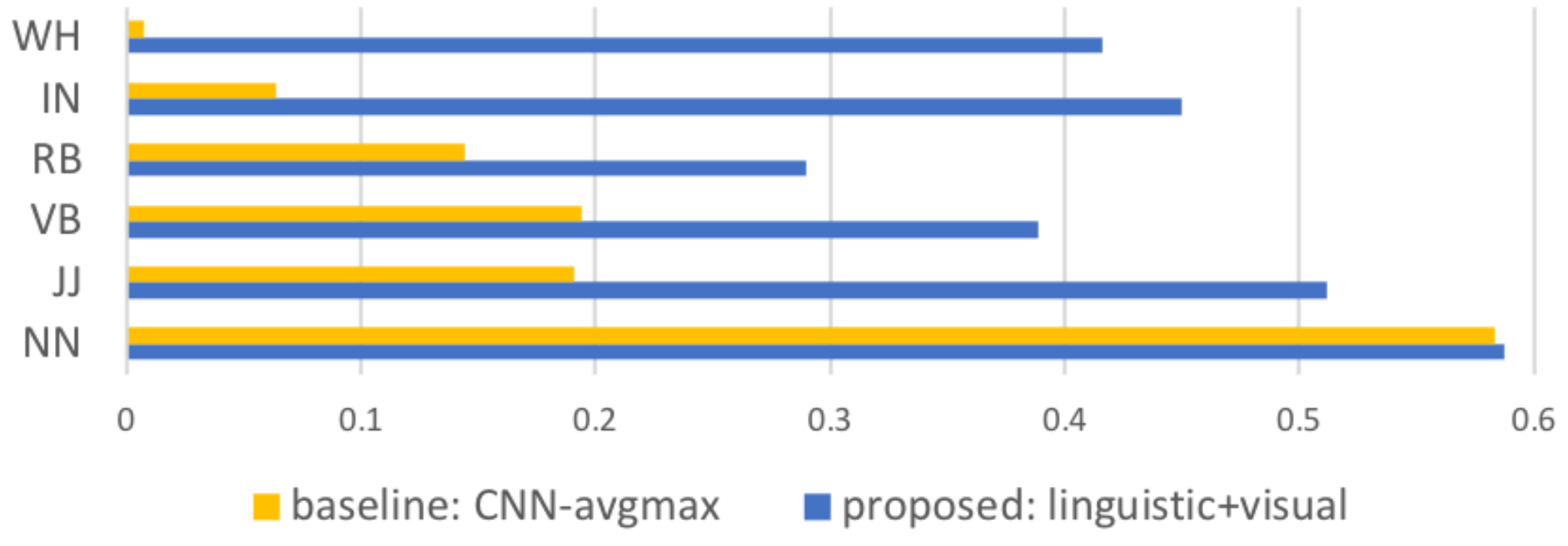
- **Multi-lingual Image Caption Dataset**: 1) Multi30k: 30k images, De-En (5), Fr-En (1); 2) MSCOCO-STAIR: 200k images, Jp-En (5).
- **Multimodal Bilingual Induction Dataset**: 1) De-En (MMID 1311 pairs); 2) Fr-En (MMID 1217 pairs); 3) Jp-En (no multimodal dataset, pure text, 2408 pairs).
- **Baselines**: 1) CNN-mean; 2) CNN-avgmax.
- **Evaluation Metrics**: 1) MRR; 2) P@K.
- **Multi-lingual Image Caption Performance**
  The multi-lingual model utilizes fewer parameters and achieves comparable or better performance than mono-lingual model.

	En	De	Fr
mono-lingual	50.19	44.24	47.45
multi-lingual	51.04	44.75	50.67

- **Bilingual Lexicon Induction Performance**
  1) the linguistic features and grounded visual features are complementary; 2) the grounded visual features is better than the global CNN-mean features; 3) the proposed approach achieves the best performance across different language pairs.

		De-En		Fr-En		Jp-En	
		MRR	P@1	MRR	P@1	MRR	P@1
baselines	CNN-mean	0.262	19.9	0.301	22.8	-	-
	CNN-avgmax	0.430	38.5	0.474	41.9	-	-
proposed model	linguistic	0.467	38.8	0.376	29.3	0.290	22.1
	visual	0.400	31.5	0.387	31.1	0.419	34.2
	linguistic+visual	**0.529**	**45.2**	**0.494**	**42.0**	**0.469**	**38.3**

- **Ablation on Part-of-Speech Labels**
  1) our approach substantially improves the translation performance for all part-of-speech classes; 2) the translation of noun and adj words are easiest based on vision-pivot.



Conclusion and Future Work

In this work, 1) we propose to employ linguistic and visual contexts for unsupervised bilingual lexion induction; 2) we design a multi-lingual caption model to embed the multi-lingual words in the same spaces; 3) the proposed approach outperforms the state-of-the-art vision-based methods on all word types and language pairs. In the future, we will further expand the vision-pivot approaches for zero-resource machine translation.