Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics

Douwe Kiela, Anita L. Verő and Stephen Clark

Computer Laboratory, University of Cambridge

1. Resources for Multi-Modal Semantics

Distributional models suffer from the *grounding problem*:

Grounding problem: the fact that the meaning of a word is represented as a distribution over other words does not account for the fact that human semantic knowledge is *grounded in physical reality and sensorimotor experience.* (Harnad, 1990)

Multi-modal semantics addresses this by *enhancing* linguistic representations with extra-linguistic perceptual input, usually using **images**.

Open questions about representation learning techniques and data sources:

Does the improved performance over bag-of-visual-words extend to **different convolutional network architectures**?

How important is the **source** of images? Is there a difference between **search engines** and **manually annotated** data sources? Does the **number of images** obtained for each word matter?

Do these findings extend to **different languages** beyond English?

2. CNN Architectures

We use the **MMFeat toolkit** (`https://github.com/douwekiela/mmfeat`) to obtain image representations for three different convolutional network architectures:

- **AlexNet** (Krizhevsky et al., 2012)
- **GoogLeNet** (Szegedy et al., 2015)
- **VGGNet** (Simonyan and Zisserman, 2014)

The models are trained on the **ImageNet** classification task to maximize the multinomial logistic regression objective:

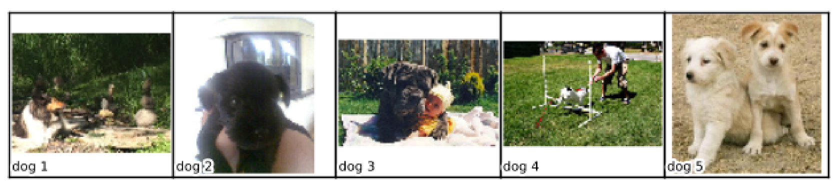
$$-\sum_{i=1}^{D}\sum_{k=1}^{K}\mathbf{1}\{y^{(i)}=k\}\log\frac{\exp(\theta^{(k)\top}x^{(i)})}{\sum_{j=1}^{K}\exp(\theta^{(j)\top}x^{(i)})}$$

3. Image Data Sources

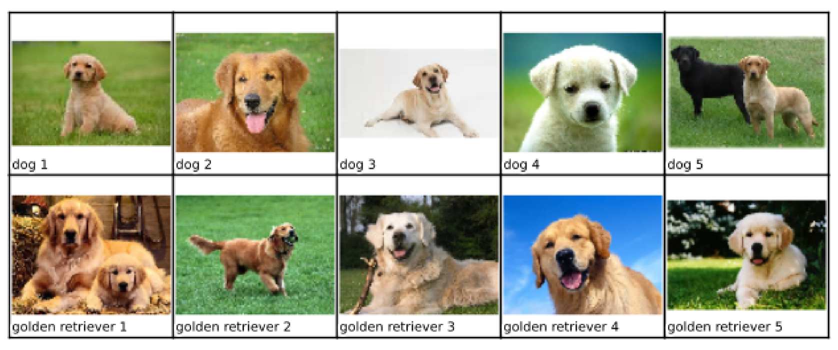
Trade-off of coverage and (human) annotation quality (see paper for a detailed comparison):

ImageNet

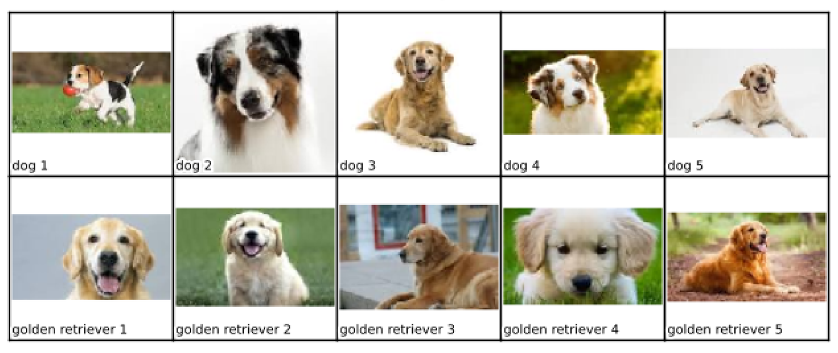
ESP Game dataset



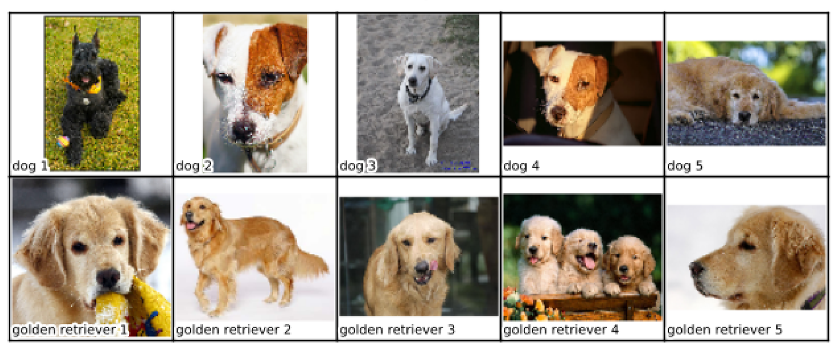
Bing



Google



Flickr



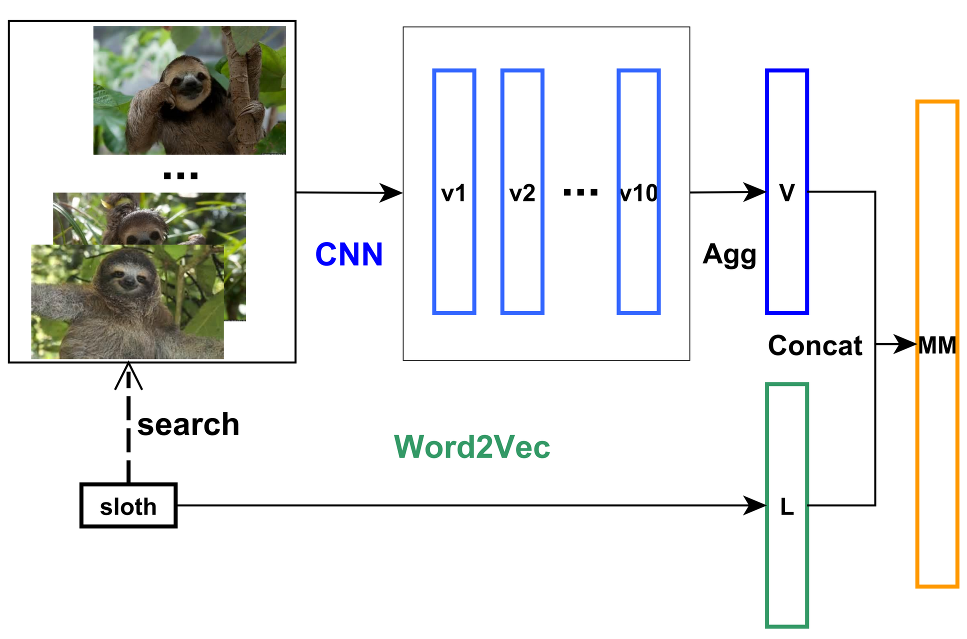
4. Evaluations

Visual representations

- Transfer **convolutional network features**.
- Pre-softmax fully-connected layer from each network.

Aggregating image vectors for one word

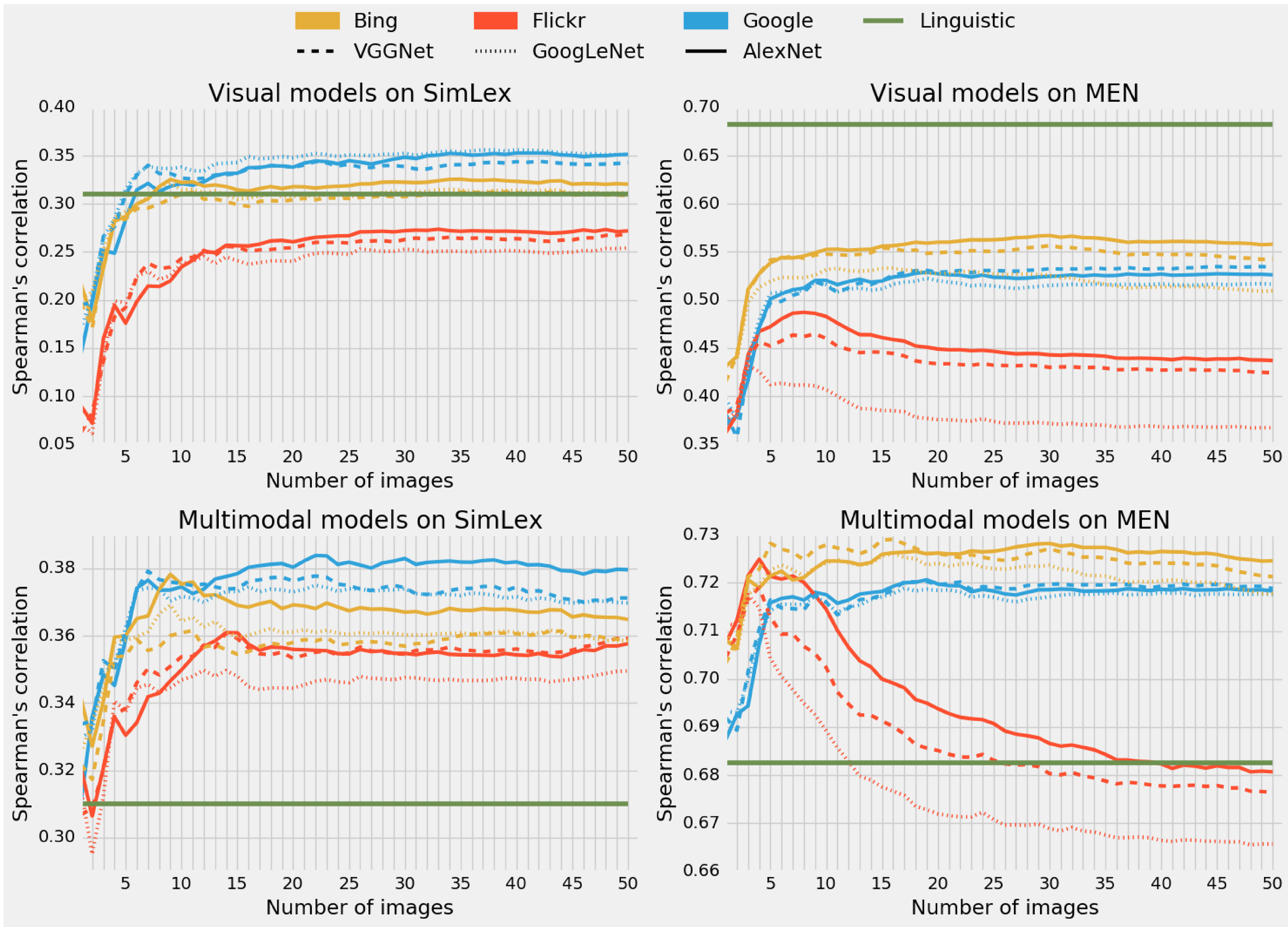
- Element-wise **Mean** or **Maximum**.



Multi-modal representation: concatenating visual and textual vectors.

Standard multi-modal evaluations: **MEN** and **SimLex-999**.

5. Number of images and representation quality



6. Semantic Similarity and Relatedness

Source	Arch. / Agg. / Type/Eval	AlexNet Mean SL	AlexNet Mean MEN	AlexNet Max SL	AlexNet Max MEN	GoogLeNet Mean SL	GoogLeNet Mean MEN	GoogLeNet Max SL	GoogLeNet Max MEN	VGGNet Mean SL	VGGNet Mean MEN	VGGNet Max SL	VGGNet Max MEN
Wikipedia	Text	.248	.654	.248	.654	.248	.654	.248	.654	.248	.654	.248	.654
Google	Visual	.406	.549	.402	.552	.420	.570	.434	.579	.430	.576	.406	.560
	MM	.366	.691	.344	.693	.366	.701	.342	.699	.378	.701	.341	.693
Bing	Visual	.431	.613	.425	.601	.410	.612	.414	.603	.400	.611	.398	.569
	MM	.384	.715	.355	.708	.374	.725	.343	.712	.363	.720	.340	.705
Flickr	Visual	.382	.577	.371	.544	.378	.547	.354	.518	.378	.567	.340	.511
	MM	.372	.725	.344	.712	.367	.728	.336	.716	.370	.726	.330	.711
ImageNet	Visual	.316	.560	.316	.560	.347	.538	.423	.600	.412	.581	.413	.574
	MM	.348	.711	.348	.711	.364	.717	.394	.729	.418	.724	.405	.721
ESPGame	Visual	.037	.431	.039	.347	.104	.501	.125	.438	.188	.514	.125	.460
	MM	.179	.666	.147	.651	.224	.692	.226	.683	.268	.697	.222	.688

Performance on maximally covered datasets (see paper).

7. Multi- and cross-lingual applicability

		SimLex EN	SimLex IT (M)	SimLex IT (C)
Wikipedia	Linguistic	.310	.179	.179
Google	Visual	.340	.231	.238
	Multi-modal	.380	.231	.227
Bing	Visual	.325	.212	.194
	Multi-modal	.373	.227	.207

8. Conclusion

- **Multi-modal** representations consistently outperform linguistic ones.
- Different CNN architectures perform similarly.
- The **choice of data sources** has a bigger impact: Google, Bing and Flickr have the advantage of providing full coverage image datasets.
- The **number of images** has a significant impact on performance that stabilises around 10-20.
- These findings **extend to other languages**.