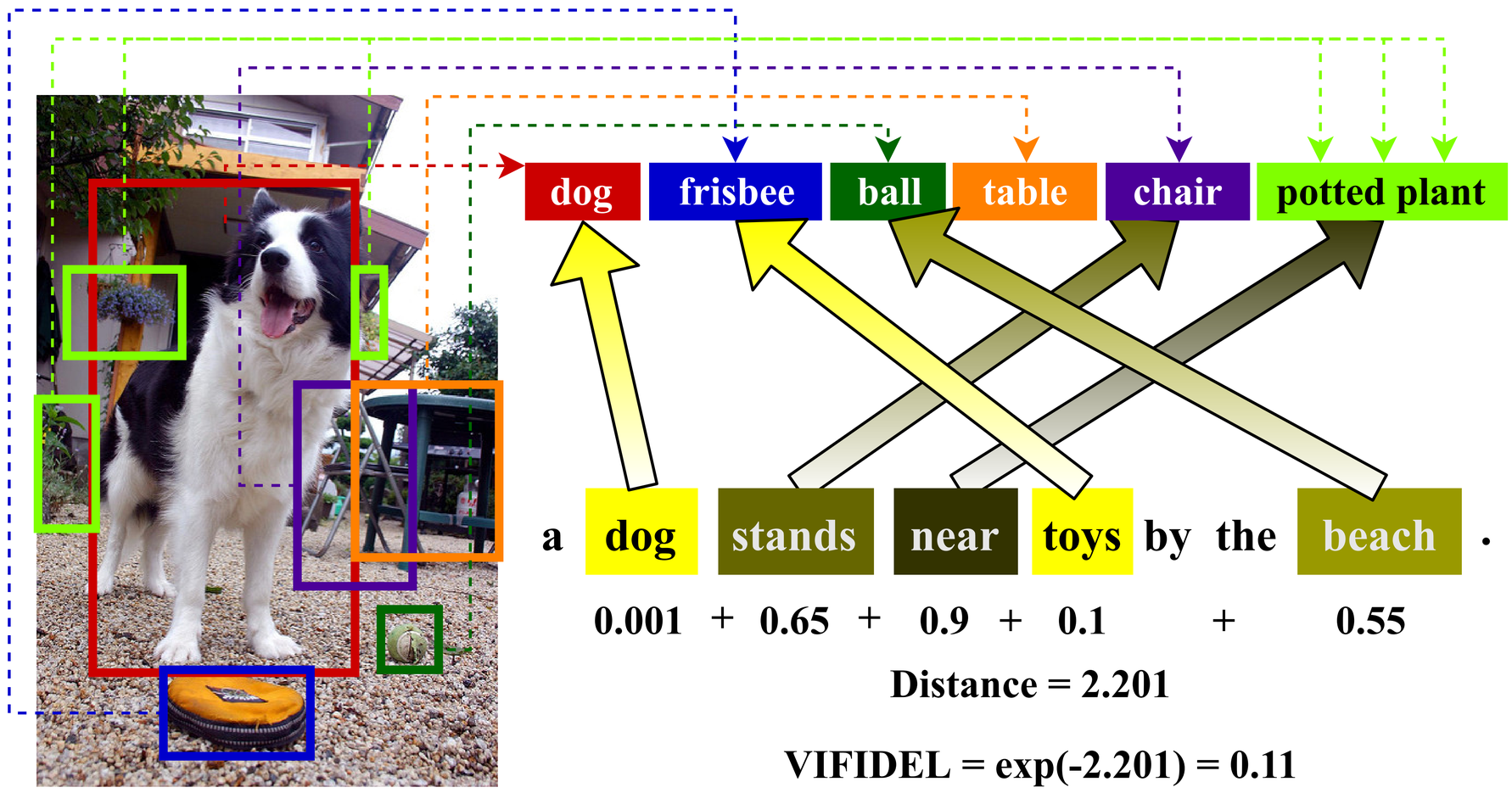
VIFIDEL: Evaluating the Visual Fidelity of Image Descriptions

Pranava Madhyastha, Josiah Wang, Lucia Specia

Imperial College London



Context: Evaluating Image Descriptions

- Existing automatic metrics conflate various criteria implicitly into a single score
- Our contribution: An *image-aware* metric named VIFIDEL
  1. Measures faithfulness of description w.r.t. the image
  2. Explicitly takes image content into account
  3. Also works in the absence of reference descriptions!

Claim: VIFIDEL is useful for fine-grained measurement of descriptions

VIFIDEL in Brief

- Extension of **Wasserstein distance** with weighted Euclidean distance
- Uses information from the images in the form of detected objects
- Consensus-based scores for multiple references

Formally

- For image I and description S:

$$\texttt{VIFIDEL}(I, S) = \exp\left(-\min_{T \geq 0} \sum_{i,j=1}^{n} T_{ij} \text{cost}(i, j)\right)$$

where transport matrix T_{ij} contains information about the proportion of semantic content from the image to the description; *cost* = weighted Euclidean distance.

- To weight importance of word k with a penalty according to human references:

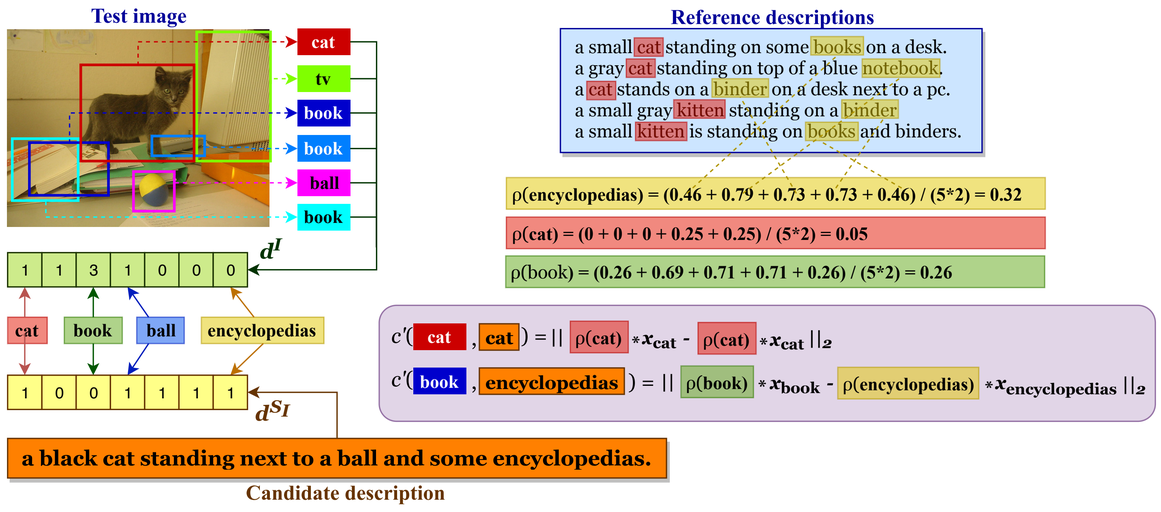
$$\rho_k^I = \frac{1}{M} \sum_{r=1}^{M} \left(\frac{1 - \max_{t \in \{R_r^I\}} \cos(x_k, x_t)}{2}\right)$$

where $\{R_r^I\}$ is the set of content words in the rth reference for image I;
x_t is the word embedding for word t;
M is the number of human references for the image.

- The cost (weighted according to human references) is:

$$\text{cost}(i, j | R^I) = \|\rho_i^I x_i - \rho_j^I x_j\|_2^p$$

At a Glance

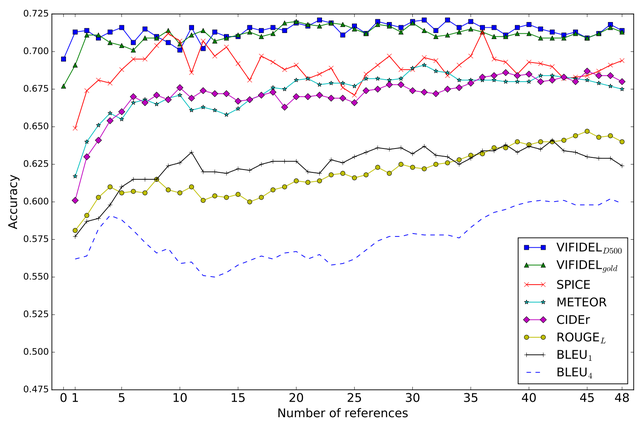


- Weights are computed for *encyclopedias*, *cat* and *books*
- The word *cat* has a low penalty score
- The penalty scores are then used as weights to compute the cost.

Properties

- Semantic matching instead of string matching
- Scores even in the absence of references
- Highly scalable compared to SPICE (dependent on linguistic resources)
- Complements fluency-based metrics
- References are only used to weigh the importance of objects and words
- Extendable with other attributes including the environment
- Language agnostic
- Implementation is open source (QR code below)

Evaluation



Accuracy as a function of # References

- VIFIDEL is more stable and consistently outperforms other metrics for all numbers of references.

Comparison

	References			
	0	1	5	48
BLEU$_1$	-	0.58	0.61	0.62
BLEU$_4$	-	0.56	0.59	0.60
ROUGE$_L$	-	0.58	0.61	0.64
METEOR	-	0.62	0.66	0.68
CIDEr	-	0.60	0.66	0.68
SPICE	-	0.65	0.69	0.69
WMD$_{\text{best}}$	-	0.66	0.70	0.70
WMD$_{\text{worst}}$	-	0.66	0.66	0.66
LM	0.54	0.54	0.54	0.54
VIFIDEL$_{gold}$	0.68	0.69	0.70	**0.71**
VIFIDEL$_{D500}$	**0.69**	**0.71**	**0.72**	**0.71**
VIFIDEL+**LM**	**0.69**	0.70	0.71	0.71
VIFIDEL+**CIDEr**	**0.69**	**0.71**	**0.72**	**0.72**