Hierarchy-based Image Embeddings for Semantic Image Retrieval

Björn Barz and Joachim Denzler

Computer Vision Group, Friedrich Schiller University Jena, Germany

Computer Vision Group

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

Code available!
github.com/cvjena/semantic-embeddings

WHY DO WE NEED TO INTEGRATE PRIOR SEMANTIC KNOWLEDGE?

The New York Times

GOOGLE

Google Photos Mistakenly Labels Black People 'Gorillas'

BY CONOR DOUGHERTY JULY 1, 2015 7:01 PM Comment

Email Share Tweet Save More

Google continued to apologize Wednesday for a flaw in Google Photos, which was released to great fanfare in May, that led the new application to mistakenly label photos of black people as "gorillas."

The company said it had fixed the problem and was working to figure out exactly how it happened.

"We're appalled and genuinely sorry that this happened," said a Google representative in an emailed statement. "We are taking immediate action to prevent this type of result from appearing."

Semantic similarity does not always correlate with visual similarity.

visually similar

semantically similar

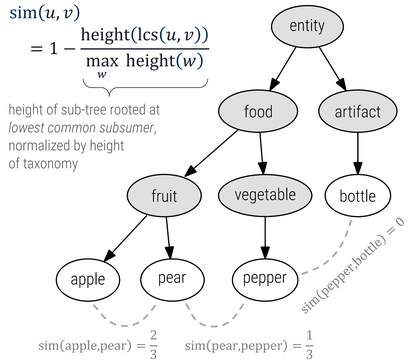
- Simply removing the "gorilla" class is a sub-optimal "fix" and only works for classification, not for content-based image retrieval.
- Better: learning a feature representation that carries semantic information.

MEASURING SEMANTIC SIMILARITY

- Previous works focused on learning the semantic similarity between classes either from text [1] or the images themselves [2, 3].
- However, semantic knowledge is already available in the form of **class taxonomies** for almost all concepts of the world (WordNet [4], Wikispecies, Open Tree of Life, ...).
- We use a **semantic similarity measure** derived from this graph of concepts [5].

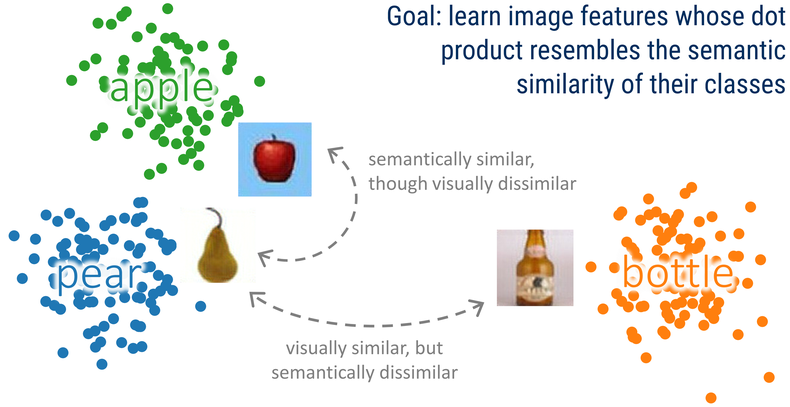
$$\mathrm{sim}(u,v) = 1 - \frac{\mathrm{height}(\mathrm{lcs}(u,v))}{\max_w \mathrm{height}(w)}$$

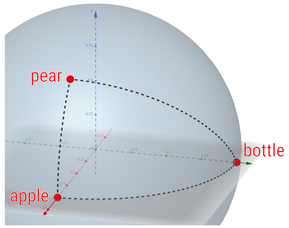
height of sub-tree rooted at *lowest common subsumer*, normalized by height of taxonomy



HIERARCHY-BASED SEMANTIC EMBEDDINGS

Goal: learn image features whose dot product resembles the semantic similarity of their classes





Class embeddings on the unit hyper-sphere do not need to be learned but can be **computed explicitly** by solving:

$$\forall_{1\leq i,j\leq n}: \ \varphi(c_i)^T\varphi(c_j) = \mathrm{sim}(c_i, c_j)$$
$$\forall_{1\leq i\leq n}: \ \|\varphi(c_i)\| = 1$$

class embedding function — known classes

Map images onto their target class embeddings in this semantic space using a simple loss:

Cosine Distance

$$\mathcal{L}_{\mathrm{CORR}}(x,y) = 1 - \psi(x)^T\varphi(c_y)$$

L2-normalized CNN output

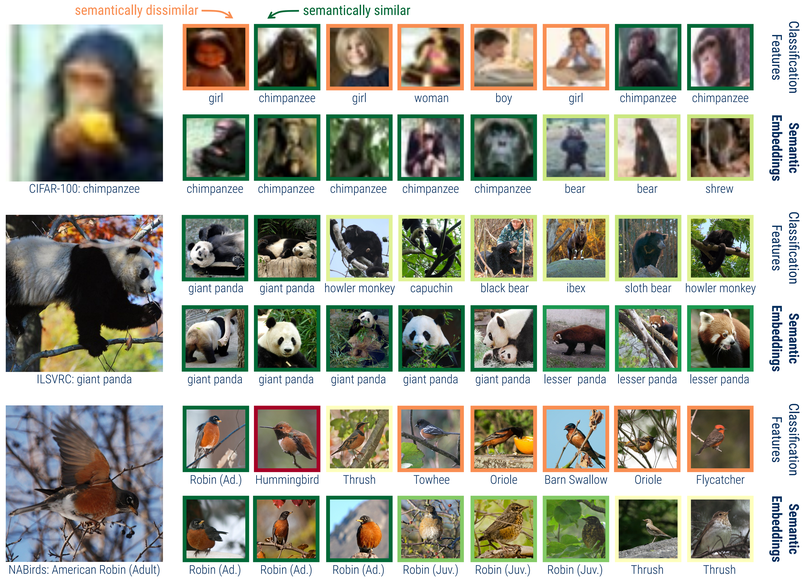
with additional classification objective

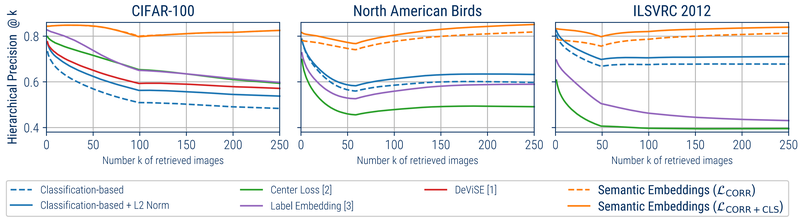
$$\mathcal{L}_{\mathrm{CORR+CLS}}(x,y) = \mathcal{L}_{\mathrm{CORR}}(x,y) + \lambda\cdot\mathcal{L}_{\mathrm{XENT}}(f(\psi(x)),y)$$

cross-entropy loss

additional FC layer with softmax

RESULTS





- Semantically more consistent retrieval results.
- Semantically meaningful feature space based on prior knowledge.
- Straightforward training.

References

[1] Frome, Corrado, Shlens, Bengio, Dean, Ranzato, Mikolov. "DeViSE: A Deep Visual-Semantic Embedding Model." NIPS 2013.

[2] Wen, Zhang, Li, Qiao. "A Discriminative Feature Learning Approach for Deep Face Recognition." ECCV 2016.

[3] Sun, Wei, Ren, Ma. "Label Embedding Network: Learning Label Representation for Soft Training of Deep Networks." arXiv:1710.10393, 2017.

[4] Fellbaum, "WordNet." Wiley Online Library, 1998.

[5] Deng, Berg, Fei-Fei. "Hierarchical Semantic Indexing for Large-Scale Image Retrieval." CVPR 2009.