Notes on Beyond instance-level retrieval: Leveraging captions to learn a global visual representation for semantic retrieval

Albert Gordo and Diane Larlus

CVPR: 2017

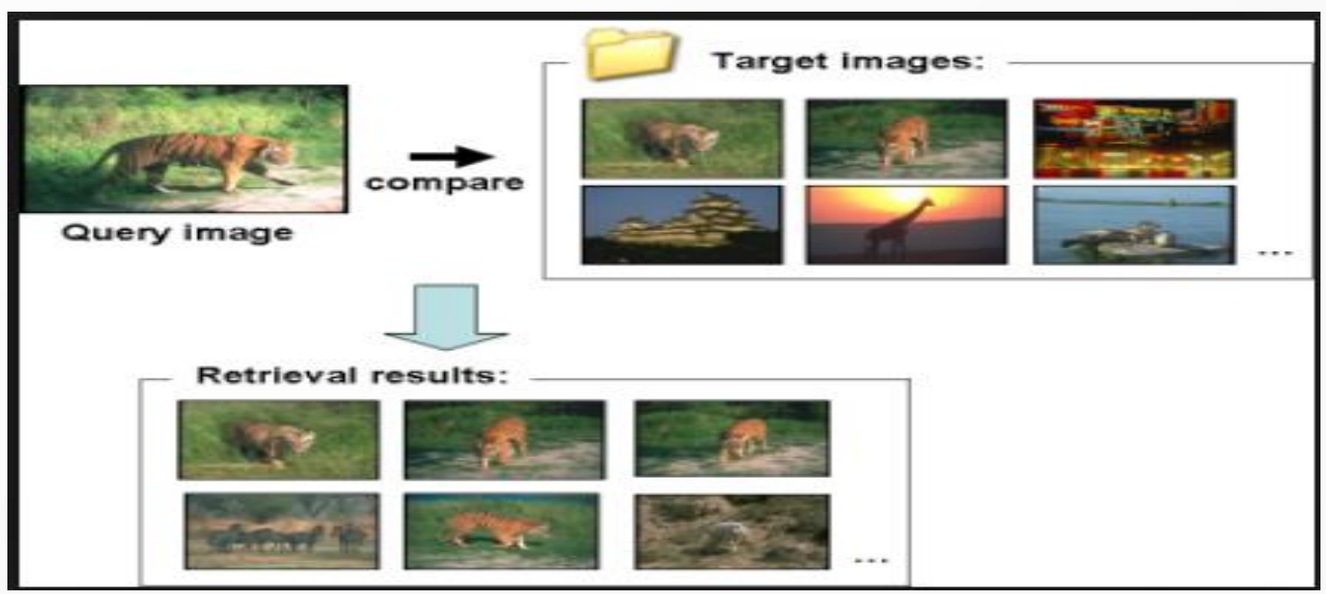
By: Sonit Singh
@Image Analysis Reading Group (IARG)
Macquarie University

Motivation

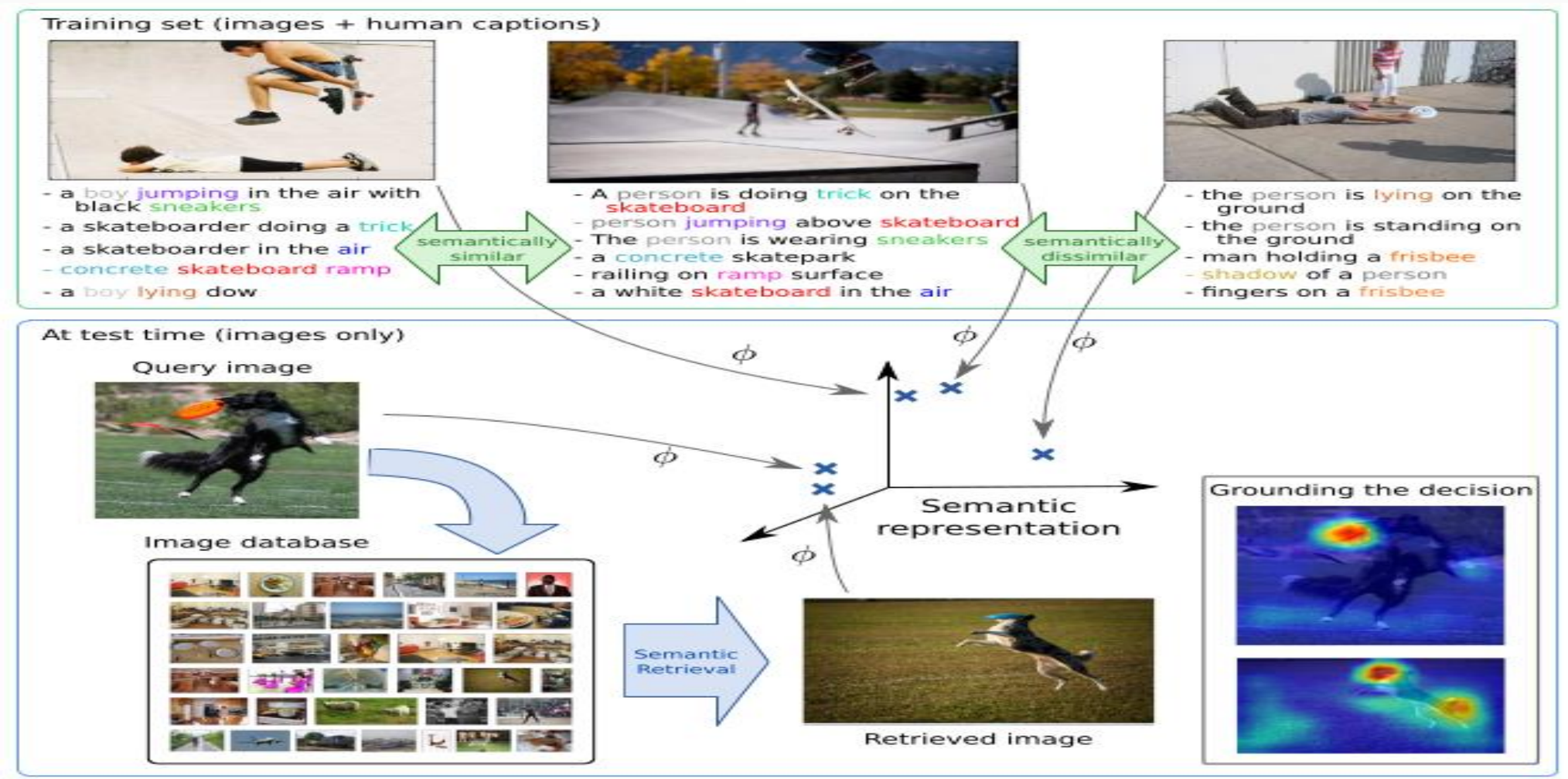
- Existing Systems:
  - Text Based Image Retrieval
  - Content Based Image Retrieval
- Most of the research in image retrieval has focussed on the task of instance-level image retrieval, where the goal is to retrieve images that contain the same object instance as the query image.
- In this paper, authors
  - Move beyond instance-level retrieval and consider the task of semantic image retrieval in complex scenes.

Problem

- CBIR: Given a query image, retrieve all images relevant to that query within a potentially large database of images.
- Existing methods focused on retrieving the exact same instance as in the query image, such as particular object.



Overall Goal: Semantic Retrieval



Contributions

- Validated that the task of semantic image retrieval can be well-defined (because it is also highly subjective).
- Showed that a similarity function based on captions produced by human annotators, available at the training time, constitutes a good computable surrogate of the true semantic similarity.
- Developed a model that leverages the similarity between human-generated captions, to learn how to embed images in a semantic space, where the similarity between embedded images is related to their semantic similarity.
- Developed a model (extending previous one), that leverages the image captions explicitly and learns a joint embedding for the visual and textual representations.

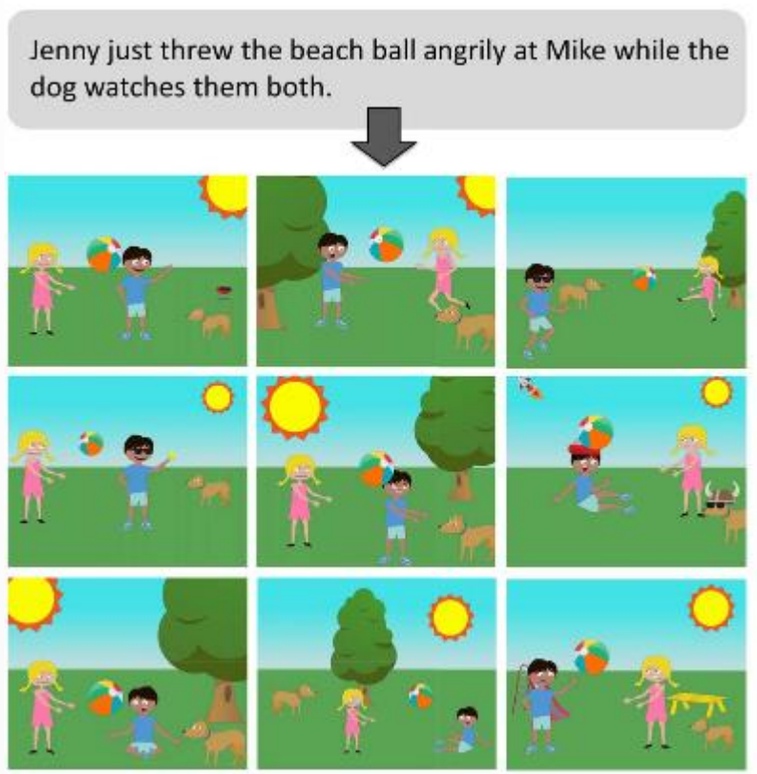
Related Work

- Zitnick and Parikh showed that image retrieval can be greatly improved when detailed semantics is available.

Bringing Semantics Into Focus Using Visual Abstraction

C. Lawrence Zitnick
Microsoft Research, Redmond
larryz@microsoft.com

Devi Parikh
Virginia Tech
parikh@vt.edu



Related Work…

- Image Captioning as a retrieval problem
  - First retrieve similar images, and then transfer caption annotations from the retrieved images to the query images.

Framing Image Description as a Ranking Task:
Data, Models and Evaluation Metrics

Micah Hodosh
Peter Young
Julia Hockenmaier
Department of Computer Science
University of Illinois
Urbana, IL 61801, USA

Our data set of 8,000 Flickr images with 5 crowd-sourced captions

A man is doing tricks on a bicycle on ramps in front of a crowd.
A man on a bike executes a jump as part of a competition while the crowd watches.
A man rides a yellow bike over a ramp while others watch.
Bike rider jumping obstacles.
Bmx biker jumps off of ramp.

ilding.
building.
e brick building.

an old building.

The selected caption …

… describes the image without any errors (score = 4)	… describes the image with minor errors (score = 3)	… is somewhat related to the image (score = 2)	… is unrelated to the image (score = 1)
 *A girl wearing a yellow shirt and sunglasses smiles.*	 *A man climbs up a sheer wall of ice.*	 *A Miami basketball player dribbles by an Arizona State player.*	 *A group of people walking a city street in warm weather.*
 *A boy jumps into the blue pool water.*	 *A dog in a grassy field, looking up.*	 *Basketball players in action.*	
 *A man riding a motor bike kicks up dirt.*	 *Dogs pulling a sled in a sled race.*	 *Two little girls practice martial arts.*	 *A snowboarder in the air over a snowy mountain.*
 *A child jumping on a tennis court.*	 *A boy in a blue life jacket jumps into the water.*	 *A black dog with a purple collar running.*	

Related Work…

- Joint embedding of image and text
  - Many tasks require to jointly leverage images and natural text, such as zero shot learning, language generation, multi-media retrieval, image captioning, and Visual Question Answering.
  - Common Solution: To build a joint embedding for textual and visual cues and to compare the modalities directly in that space.

Related Work: Joint embedding of image and text

- Deep Canonical Correlation Analysis (DCCA)

Deep Correlation for Matching Images and Text

Fei Yan Krystian Mikolajczyk
Centre for Vision, Speech and Signal Processing, University of Surrey
Guildford, Surrey, United Kingdom. GU2 7XH
{f.yan, k.mikolajczyk}@surrey.ac.uk

In contrast to hand-crafted objectives, deep CCA (DCCA) [1] optimises the CCA objective in the deep learning framework. It uses the insight that the total correlation sought in CCA can be maximised by optimising a matrix trace norm, and the gradient of the trace norm with respect to features of the two modalities can be computed. This allows propagating the gradient down in a deep neural network, achieving end-to-end learning.

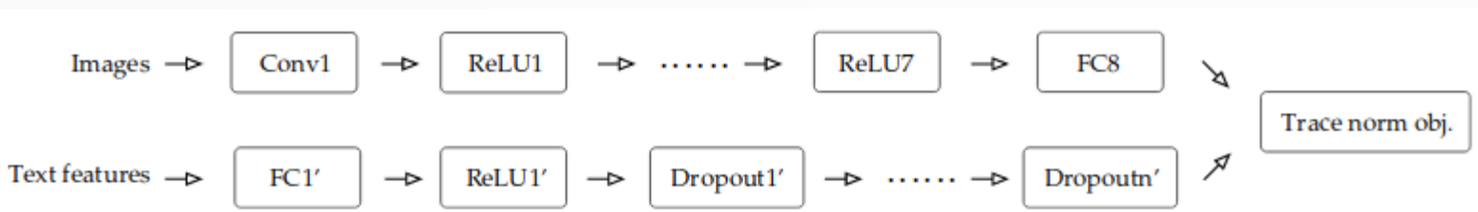


	Image annotation			Image retrieval		
	P@1	P@5	MAP	P@1	P@5	MAP
structured SVM [48]	0.086	0.070	0.050	0.035	0.029	0.035
DCCA	**0.302**	**0.114**	**0.426**	**0.295**	**0.120**	**0.415**

Table 4. Performance on IAPR TC-12.

Related Work: Joint embedding of image and text

- WS-ABIE: Web Scale Annotation By Image Embedding

WSABIE: Scaling Up To Large Vocabulary Image Annotation

Jason Weston[1] and **Samy Bengio**[1] and **Nicolas Usunier**[2]
1 Google, USA
2 Université Paris 6, LIP6, France
{jweston,bengio}@google.com nicolas.usunier@lip6.fr

Image annotation datasets are becoming larger and larger, with tens of millions of images and tens of thousands of possible annotations. We propose a strongly performing method that scales to such datasets by simultaneously learning to optimize precision at the top of the ranked list of annotations for a given image *and* learning a low-dimensional joint embedding space for both images and annotations. Our method, called WSABIE, both outperforms several baseline methods and is faster and consumes less memory.

Table 1: **Summary statistics of the datasets used.**

Statistics	ImageNet	Web-data
Number of Training Images	2,518,604	9,861,293
Number of Test Images	839,310	3,286,450
Number of Validation Images	837,612	3,287,280
Number of Labels	15,952	109,444

Table 3: Nearest annotations in the embedding space learnt by WSABIE on Web-data. Translations (e.g. delphin) and synonyms or misspellings (beckam, mt fuji) have close embeddings.

Annotation	**Neighboring Annotations**
barack obama david beckham santa	barak obama, obama, barack, barrack obama beckham, david beckam, alessandro del piero santa claus, papa noel, pere noel, santa clause
dolphin cows	delphin, dauphin, whale, delfin, delfini, baleine cattle, shire, dairy cows, kuh, horse, cow
rose pine tree	rosen, hibiscus, rose flower, rosa, roze abies alba, abies, araucaria, pine, neem tree
mount fuji eiffel tower	mt fuji, fuji, fujisan, fujiyama, mountain eiffel, tour eiffel, la tour eiffel, big ben, paris
ipod f18	i pod, ipod nano, apple ipod, ipod apple f 18, eurofighter, f14, fighter jet, tomcat, mig 21

Table 8: Examples of the top 10 annotations of three compared approaches: PAMIR^{IA}, One-vs-Rest and WSABIE, on the Web-data dataset. Annotations in **red+bold** are the true labels.

Image	One-vs-Rest	WSABIE
	surf, bora, belize, sea world, balena, wale, tahiti, delfini, surfing, mahi mahi	delfini, orca, **dolphin**, mar, delfin, dauphin, whale, cancun, killer whale, sea world
	**eiffel tower**, tour eiffel, snowboard, blue sky, empire state building, luxor, eiffel, lighthouse, jump, adventure	**eiffel tower**, statue, eiffel, mole antoneliana, la tour eiffel, londra, cctv tower, big ben, calatrava, tokyo tower
	falco, barack, daniel craig, **obama**, barack obama, kanye west, pharrell williams, 50 cent, barrack obama, bono	barrack obama, barack obama, barack hussein obama, barack obama, james marsden, jay z, **obama**, nelly, falco, barack

Related Work: Joint embedding of image and text

- DeViSE: Deep Visual Semantic Embedding Model
  - Learns a linear transformation of visual and textual features with a single-directional ranking loss

DeViSE: A Deep Visual-Semantic Embedding Model

Andrea Frome*, Greg S. Corrado*, Jonathon Shlens*, Samy Bengio
Jeffrey Dean, Marc'Aurelio Ranzato, Tomas Mikolov
* These authors contributed equally.

{afrome, gcorrado, shlens, bengio, jeff, ranzato†, tmikolov}@google
Google, Inc.
Mountain View, CA, USA

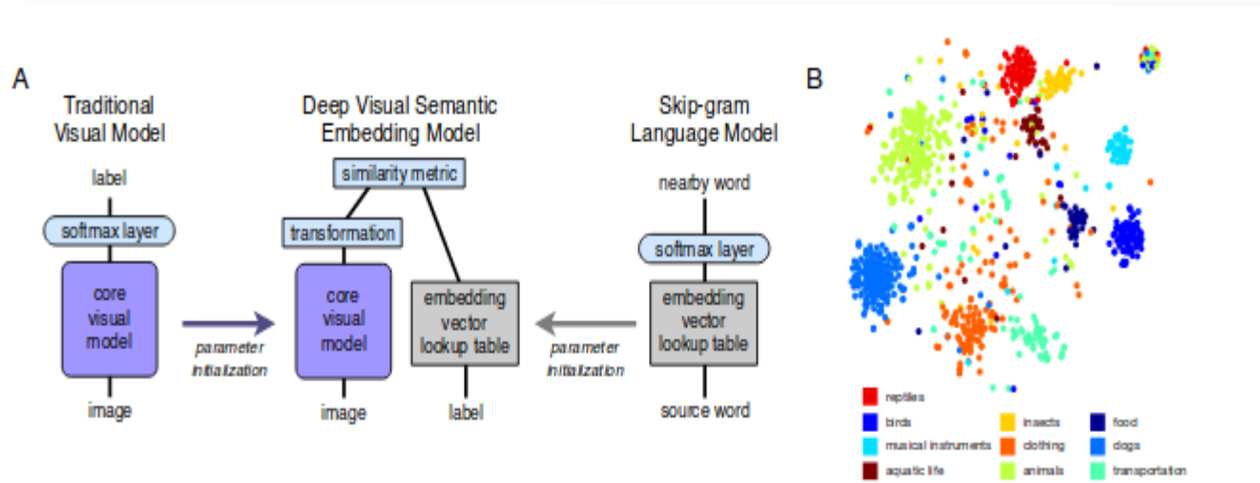


Figure 1: (a) Left: a visual object categorization network with a softmax output layer; Right: a skip-gram language model; Center: our joint model, which is initialized with parameters pre-trained at the lower layers of the other two models. (b) t-SNE visualization [19] of a subset of the ILSVRC 2012 1K label embeddings learned using skip-gram.

Related Work: Joint embedding of i

- Using Bi-directional ranking loss

Deep Visual-Semantic Alignments for Generating Image Descriptions

Andrej Karpathy Li Fei-Fei
Department of Computer Science, Stanford University
{karpathy,feifeili}@cs.stanford.edu

Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models

Ryan Kiros, Ruslan Salakhutdinov, Richard S. Zemel
University of Toronto
Canadian Institute for Advanced Research
{rkiros, rsalakhu, zemel}@cs.toronto.edu

Grounded Compositional Semantics for Finding and Describing Images with Sentences

Richard Socher, Andrej Karpathy, Quoc V. Le*, Christopher D. Manning, Andrew Y. Ng
Stanford University, Computer Science Department, *Google Inc.
richard@socher.org, karpathy@cs.stanford.edu,
qvl@google.com, manning@stanford.edu, ang@cs.stanford.edu

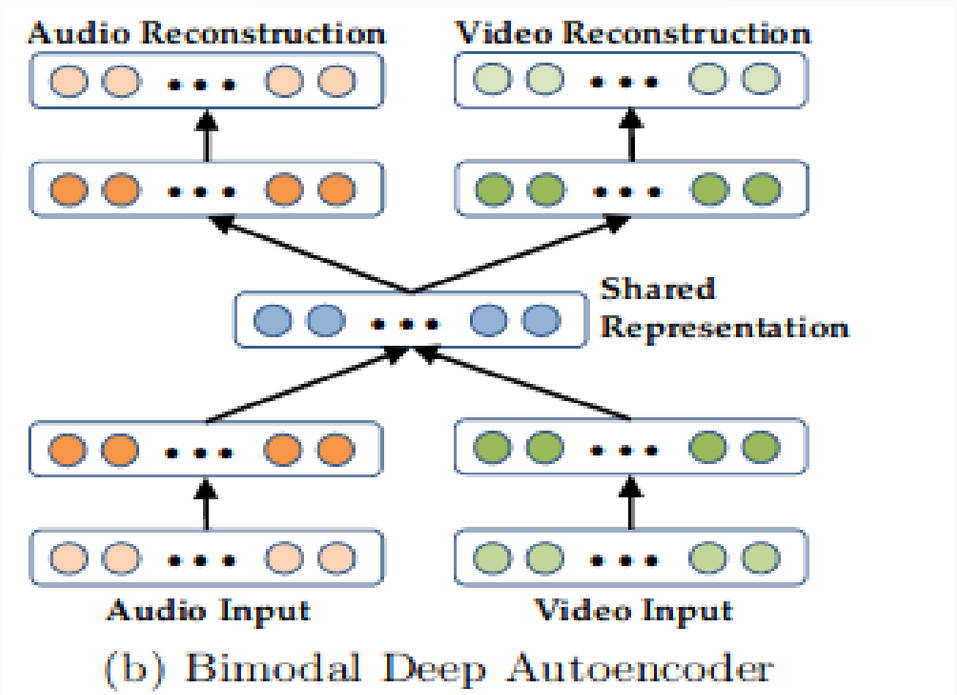
Related Work: Joint embedding of image and text

- Deep methods: Deep Multimodal Auto-Encoders

Multimodal Deep Learning

Jiquan Ngiam[1] JNGIAM@CS.STANFORD.EDU
Aditya Khosla[1] ADITYA86@CS.STANFORD.EDU
Mingyu Kim[1] MINKYU89@CS.STANFORD.EDU
Juhan Nam[1] JUHAN@CCRMA.STANFORD.EDU
Honglak Lee[2] HONGLAK@EECS.UMICH.EDU
Andrew Y. Ng[1] ANG@CS.STANFORD.EDU

[1] Computer Science Department, Stanford University, Stanford, CA 94305, USA
[2] Computer Science and Engineering Division, University of Michigan, Ann Arbor, MI 48109, USA



(b) Bimodal Deep Autoencoder

Related Work: Joint embedding of image and text

- Deep methods: CNN-RNN

Long-term Recurrent Convolutional Networks for Visual Recognition and Description

Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell

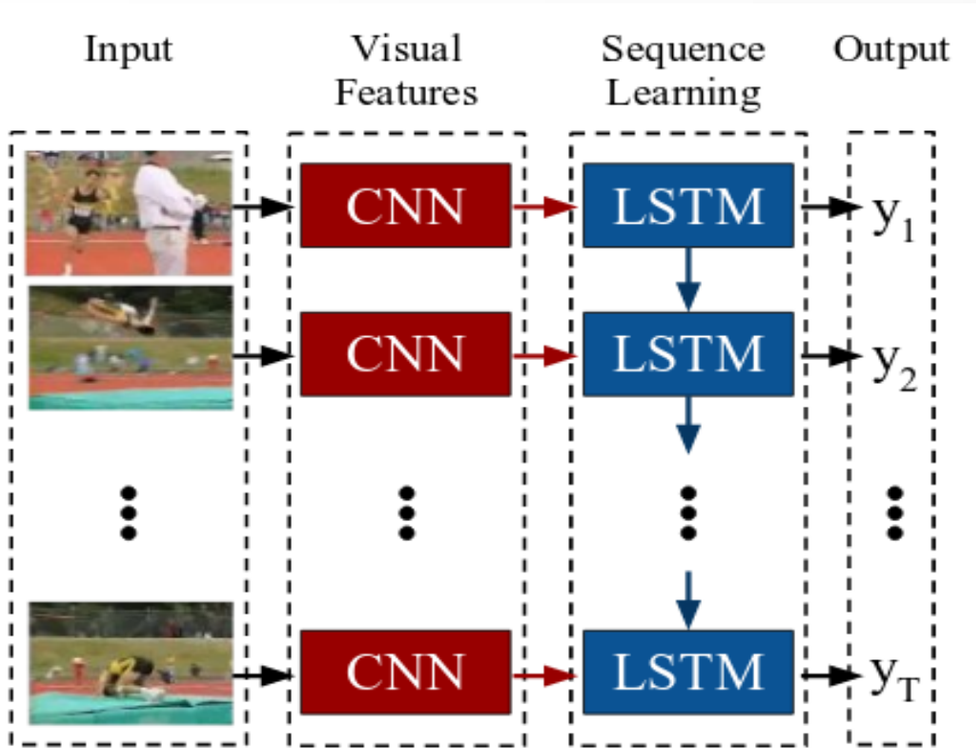


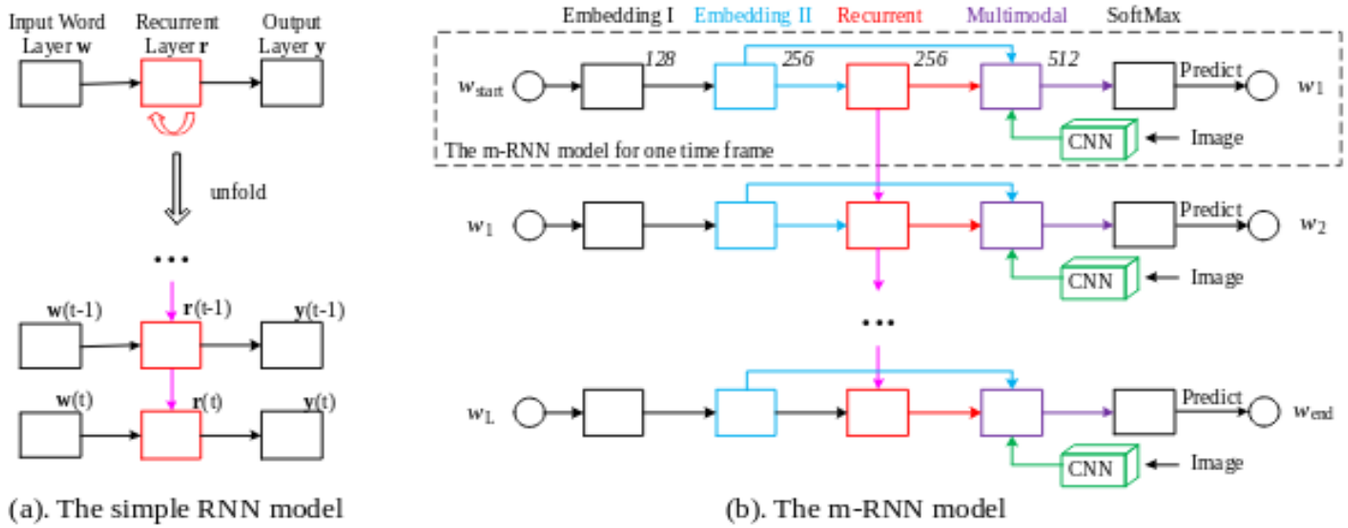
Fig. 1. We propose *Long-term Recurrent Convolutional Networks* (LRCNs), a class of architectures leveraging the strengths of rapid progress in CNNs for visual recognition problems, and the growing desire to apply such models to time-varying inputs and outputs. LRCN processes the (possibly) variable-length visual input (left) with a CNN (middle-left), whose outputs are fed into a stack of recurrent sequence models (*LSTMs*, middle-right), which finally produce a variable-length prediction (right). Both the CNN and LSTM weights are shared across time, resulting in a representation that scales to arbitrarily long sequences.

Related Work: Joint embedding of image and text

- Deep methods: multimodal RNN (mRNN)

DEEP CAPTIONING WITH MULTIMODAL RECURRENT NEURAL NETWORKS (M-RNN)

Junhua Mao
University of California, Los Angeles; Baidu Research
mjhustc@ucla.edu

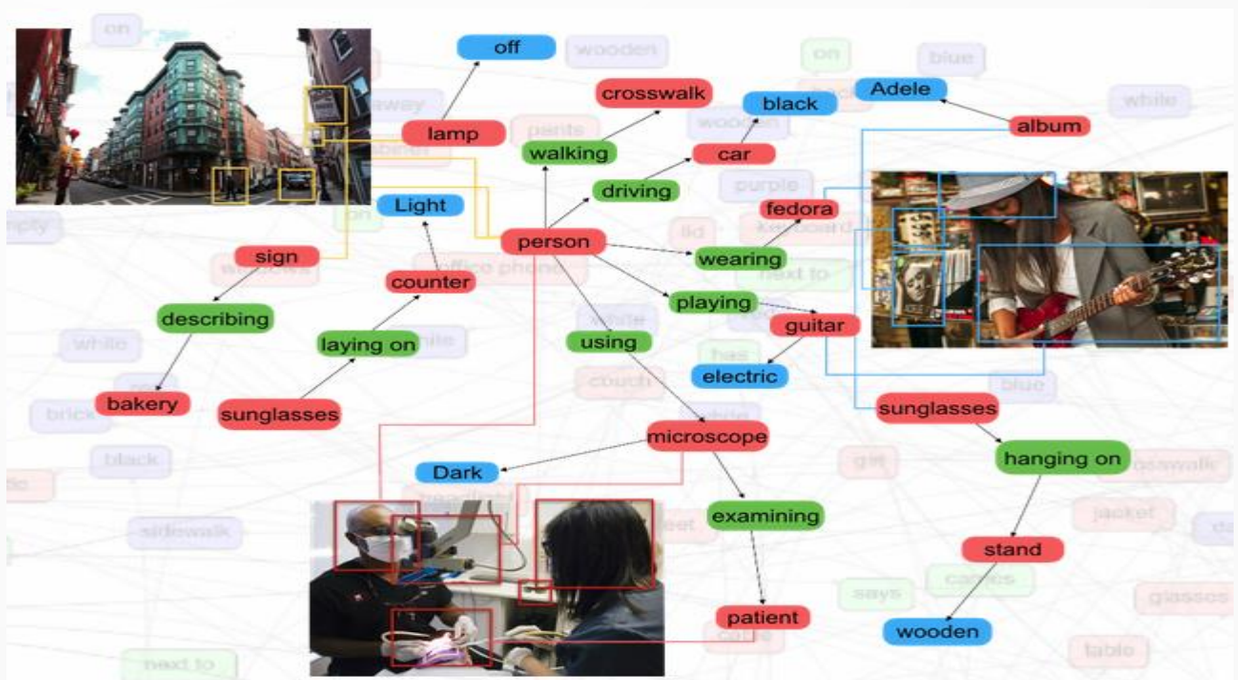


(a). The simple RNN model (b). The m-RNN model

Figure 2: Illustration of the simple Recurrent Neural Network (RNN) and our multimodal Recurrent Neural Network (m-RNN) architecture. (a). The simple RNN. (b). Our m-RNN model. The inputs of our model are an image and its corresponding sentence descriptions. w_1, w_2, ..., w_L represents the words in a sentence. We add a start sign w_{start} and an end sign w_{end} to all the training sentences. The model estimates the probability distribution of the next word given previous words and the image. It consists of five layers (i.e. two word embedding layers, a recurrent layer, a multimodal layer and a softmax layer) and a deep CNN in each time frame. The number above each layer indicates the dimension of the layer. The weights are shared among all the time frames. (Best viewed in color)

User Study: Dataset, Methodology and Inter-user Agreement

- Validating semantic search: Conducted a user study to acquire annotations related to the semantic similarity between images as perceived by users.
- Dataset: Visual Genome composed of 108k images, with a wide range of annotations such as region-level captions, scene graphs, objects, and attributeshttps://visualgenome.org/



User Study: Dataset, Methodology and Inter-user Agreement

- Methodology:
  - Involves 35 annotators (13 women and 22 men)
  - Manually ranking a large set of images according to their semantic relevance to a query image is a very complex, tidious, and time-consuming task.
  - To ease the task to annotators: Triplet ranking problem
- Given a triplet of images, composed of one query image and two other images, annotators were asked to choose the most semantically similar image to the query among the two option.
- To construct the triplets, authors randomly sample query images and then choose two images that are visually similar to the query. This is achieved by extracting image features using ResNet-101, pretrained on ImageNet.
- Two images are sampled from the 50 nearest neighbours to the query in the visual feature space.
  - Inter-user agreement : 87.3

User Study: Dataset, Methodology and Inter-user Agreement

- Agreement with Visual Representations

Method	score
Human annotators	89.1 ± 4.6
Visual baseline: ResNet R-MAC	64.0
Object annotations	63.4
Human captions: METEOR	72.1
Human captions: word2vec + FV	70.1
Human captions: tf-idf	76.3
Generated captions: tf-idf	62.5
Random (x5)	50.0 ± 0.8

Table 1. Top row, inter-human annotation agreement on the image ranking task. Bottom rows: comparison between the semantic ranking provided by human annotators and several visual baselines and methods based on the Visual Genome annotations.

Proposed Methods

5.1. Visual embedding

Our underlying visual representation is the ResNet-101 R-MAC network discussed in Section 3. This network is designed for retrieval [64] and can be trained in an end-to-end manner [25]. Our objective is to learn the optimal weights of the model (the convolutional layers and the projections in the R-MAC pipeline) that preserve the semantic similarity. As a proxy of the true semantic similarity we leverage the tf-idf-based BoW representation over the image captions. Given two images with captions we define their proxy similarity as the dot product between their tf-idf representations.

To train our network we propose to minimize the empirical loss of the visual samples over the training data. If q denotes a query image, d^+ a semantically similar image to q, and d^- a semantically dissimilar image, we define the empirical loss as as $L = \sum_q \sum_{d^+,d^-} L_v(q, d^+, d^-)$, where

$$L_v(q, d^+, d^-) = \frac{1}{2} \max(0, m - \phi_q^T \phi_+ + \phi_q^T \phi_-), \quad (1)$$

m is the margin and $\phi : \mathcal{I} \rightarrow \mathbb{R}^D$ is the function that embeds the image into a vectorial space, *i.e.* the output of our model. We slightly abuse the notation and denote $\phi(q)$, $\phi(d^+)$, and $\phi(d^-)$, as ϕ_q, ϕ_+, and ϕ_-. We optimize this loss with a three-stream network as in [25] with stochastic optimization using ADAM [37].

5.2. A joint visual and textual embedding

In the previous formulation, we only used the textual information (*i.e.* the human captions) as a proxy for the semantic similarity in order to build the triplets of images (query, relevant and irrelevant) used in the loss function. In this section, we propose to leverage the text information in an explicit manner during the training process. This is done by building a joint embedding space for both the visual representation and the textual representation. For this we define two new losses that operate over the text representations associated with the images:

$$L_{t1}(q, d^+, d^-) = \frac{1}{2} \max(0, m - \phi_q^T \theta_+ + \phi_q^T \theta_-), \quad (2)$$

$$L_{t2}(q, d^+, d^-) = \frac{1}{2} \max(0, m - \theta_q^T \phi_+ + \theta_q^T \phi_-). \quad (3)$$

As before, m is the margin, $\phi : \mathcal{I} \rightarrow \mathbb{R}^D$ is the visual embedding of the image, and $\theta : \mathcal{T} \rightarrow \mathbb{R}^D$ is the function that embeds the text associated with the image into a vectorial space of the same dimensionality as the visual features. We define the textual embedding as $\theta(t) = \frac{W^T t}{\|W^T t\|_2}$, where t is the ℓ_2-normalized tf-idf vector and W is a learned matrix that projects t into a space associated with the visual representation.

The goal of these two textual losses is to explicitly guide the visual representation towards the textual one, which we know is more informative. In particular, the loss in Eq. (2) enforces that text representations can be retrieved using the visual representation as a query, implicitly improving the visual representation, while the loss in Eq. (3) ensures that image representations can be retrieved using the textual representation, which is particularly useful if text information is available at query time. All three losses (the visual and

Experiments: Tasks

- To validate the representations produced by proposed semantic embeddings on the semantic retrieval task

  - Evaluated how well the learned embeddings are able to reproduce the similarity surrogate based on the human captions.

  - Evaluated proposed model using the triplet-ranking annotations acquired from users, by comparing how well visual embeddings agree with the human decisions on the triplets.

Experiments: Implementation

- Setup:
  - Visual model: ResNet-101 (pretrained on ImageNet), followed by the R-MAC pooling, projection, aggregation and normalization.
  - Textual features: Encoding the captions using tf-idf, after stemming using Snowball stemmer from NLTK
  - Batch size: 64
  - Optimizer: ADAM
  - LR: 10*e-5
- Metrics: Normalized Discounted Cumulative Gain (NDCG), and Pearson's Correlation Coefficient (PCC)
  - PCC measures the correlation between ground-truth and predicted ranking scores
  - NDCG is the weighted mean average precision

Results and Discussion

Methods and baselines. We evaluate different versions of our embedding. We denote our methods with a tuple of the form ({V, V+T}, {V, V+T}). The first element denotes whether the model was trained using only visual embeddings (V), *cf*. Eq. (1), or joint visual and textual embeddings (V+T), *cf*. Eq. (1)-(3). The second element denotes whether, at test time, one queries only with an image, using its visual embedding (V), or with an image and text, using its joint visual and textual embedding (V+T). In all cases, the database consists only of images represented with visual embeddings, with no textual information.

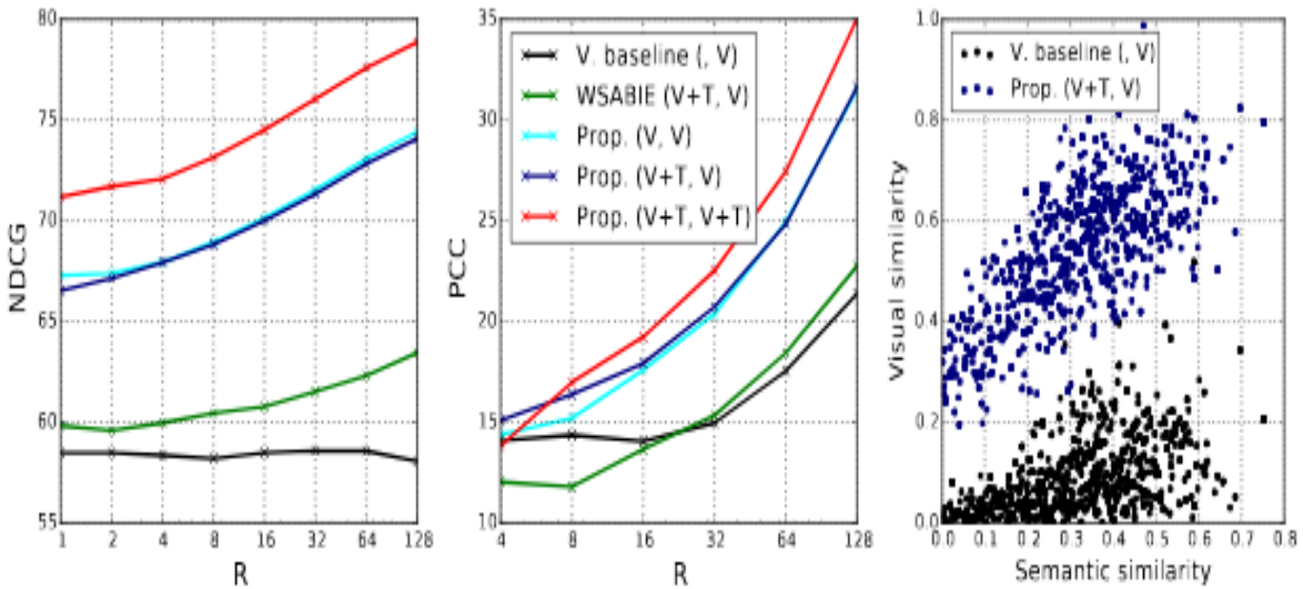


Figure 2. Left and center: NDCG and PCC achieved by the different models as a function of the number of retrieved images R, where the ground truth is determined by the tf-idf similarity. Right: correlation between the ground truth tf-idf similarity and the visual similarity of the baseline and trained models.

Results and Discussion

Methods and baselines. We evaluate different versions of our embedding. We denote our methods with a tuple of the form ({V, V+T}, {V, V+T}). The first element denotes whether the model was trained using only visual embeddings (V), *cf*. Eq. (1), or joint visual and textual embeddings (V+T), *cf*. Eq. (1)-(3). The second element denotes whether, at test time, one queries only with an image, using its visual embedding (V), or with an image and text, using its joint visual and textual embedding (V+T). In all cases, the database consists only of images represented with visual embeddings, with no textual information.

	US	NDCG AUC	PCC AUC
Text oracle			
Caption Tf-idf	76.3	100	100
Query by image			
Random (x5)	50.0 ± 0.8	10.2 ± 0.1	-0.2 ± 0.7
Visual baseline (, V)	64.0	58.4	16.1
WSABIE (V+T, V)	67.8	61.0	15.7
Proposed (V, V)	76.9	70.1	20.7
Proposed (V+T, V)	77.2	68.8	21.1
Query by image + text			
Proposed (V+T, V+T)	78.6	74.4	22.5

Table 2. Comparison of the proposed methods and baselines evaluated according to User-study (US) agreement score, AUC of the NDCG and PCC curves (*i.e.* NDCG AUC and PCC AUC).

Qualitative Results

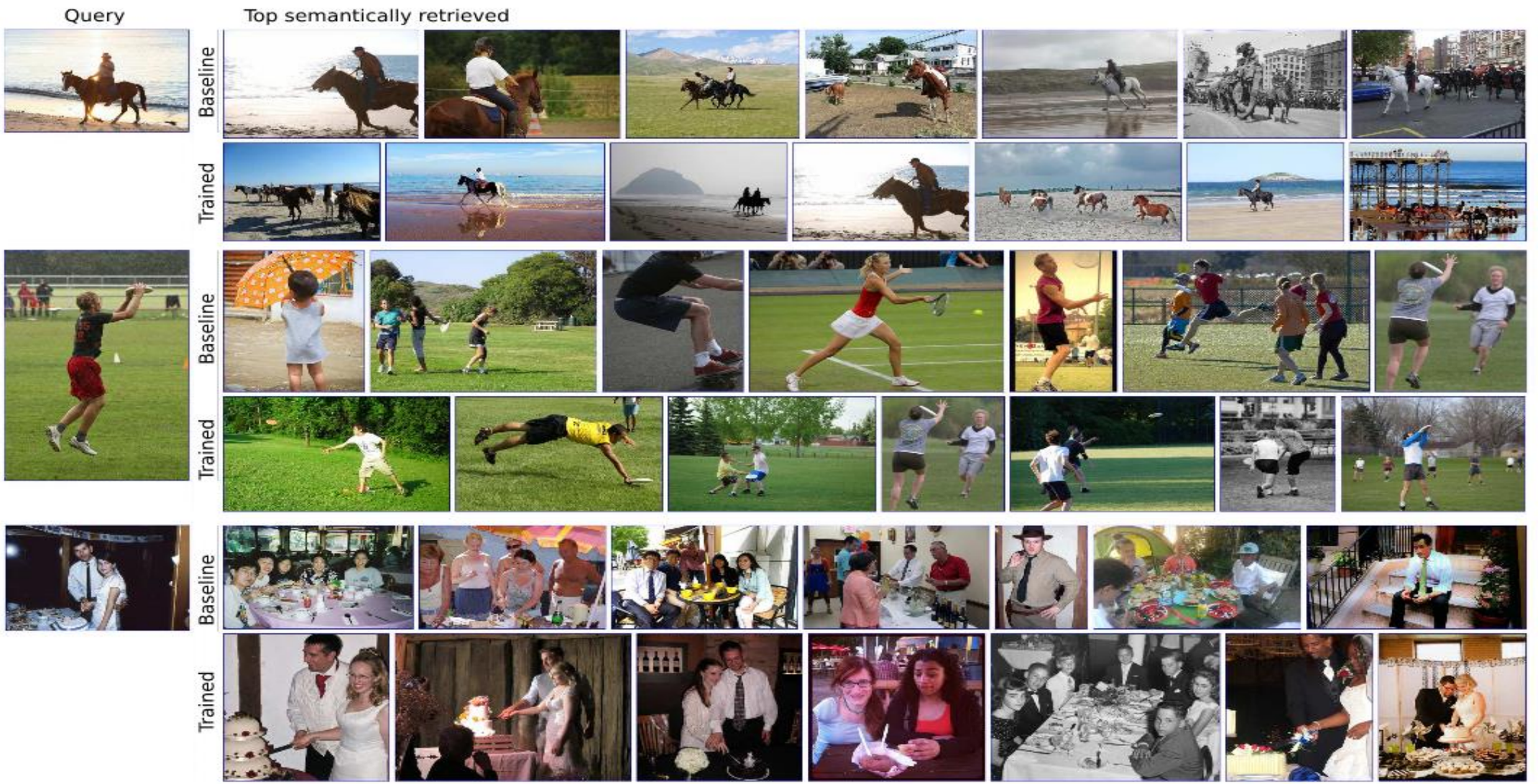


Figure 4. Qualitative results. For every block of images, left: query image. top: top-7 images with the representation pretrained on ImageNet, bottom: top-7 images with our learned representation with the (V+T,V) model.

Qualitative Results

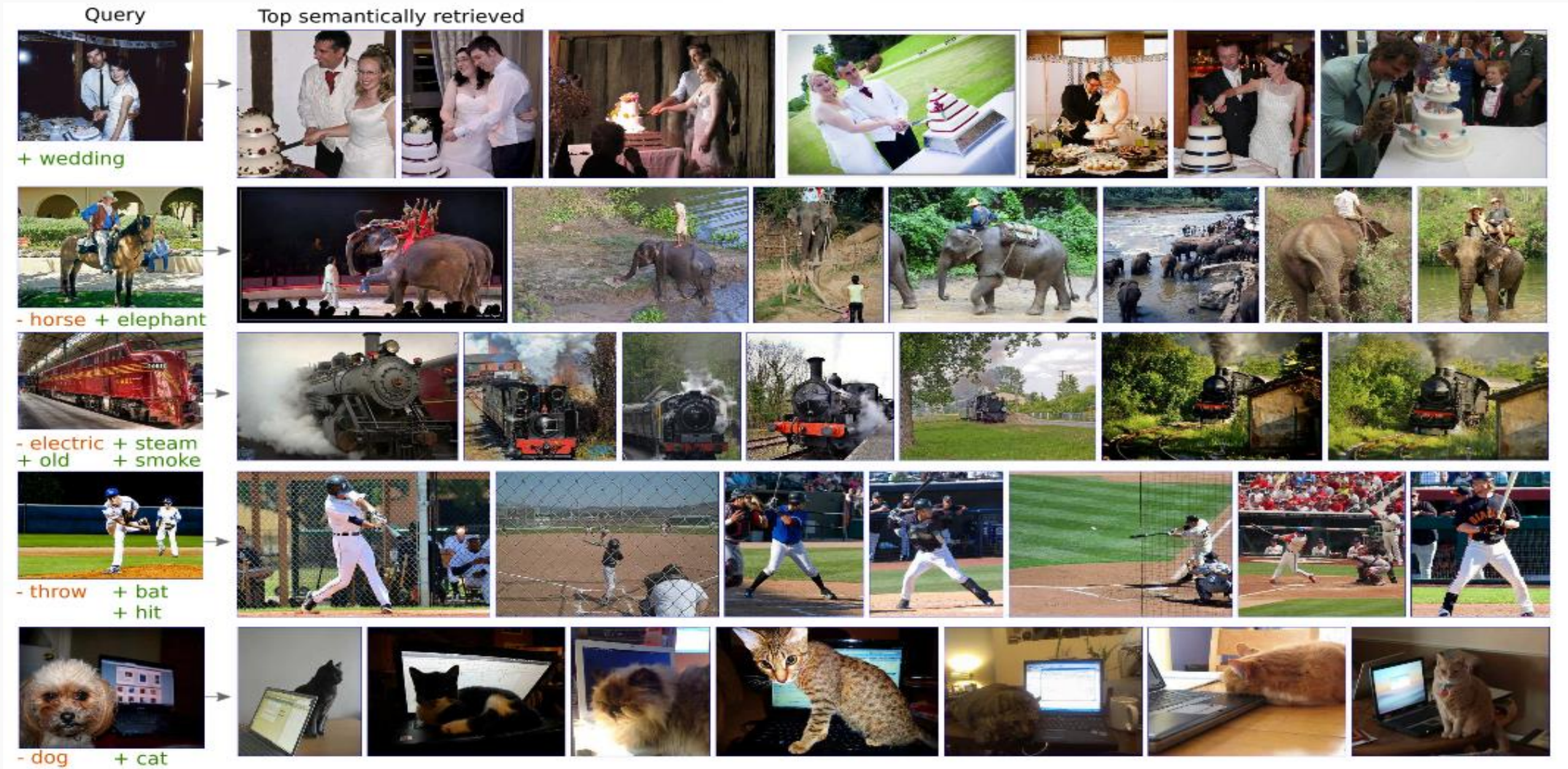


Figure 5. For a set of query images, we use a text modifier as additional query information (concepts are added or removed) to bias the results. Note that the first query is the last one from Figure 4 refined with additional text.

Thanks