Computer Vision Center

Imaging Knowledge

Self-supervised Learning and Multimodal embeddings

Marçal Rossinyol

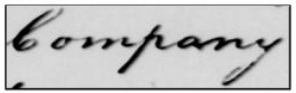
4th Summer School on Machine Learning
IIIT Hyderabad
July 11th 2019

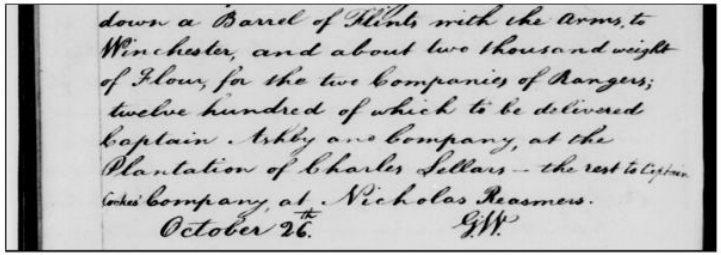
Learning Multimodal Embeddings

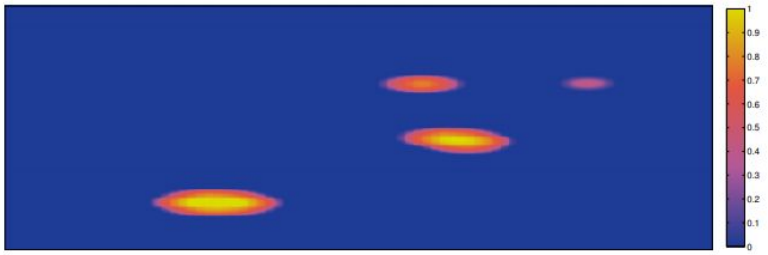
Word Spotting

- Given a large collection of documents and a query (either an exemplar image or a text string) retrieve the locations where it is likely to find such word
- Not transcription and then indexing!

Company

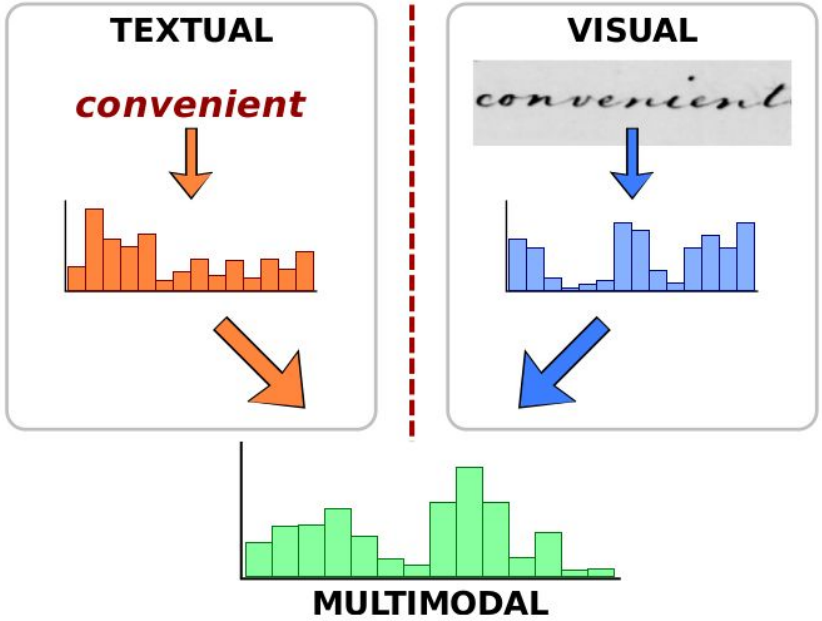






Word Spotting

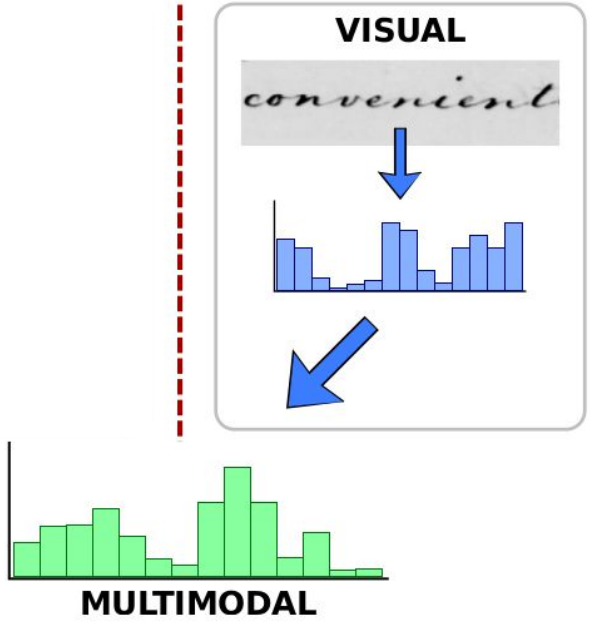
Embed **textual** and **visual** representations in a common space



D. Aldavert et al. "Integrating visual and textual cues for query-by-string word spotting." ICDAR, 2013.
J. Almazan et al. "Word spotting and recognition with embedded attributes." TPAMI, 2014.
S. Sudholt et al. "PHOCNet: A deep convolutional neural network for word spotting in handwritten documents," ICFHR, 2016.

Word Spotting

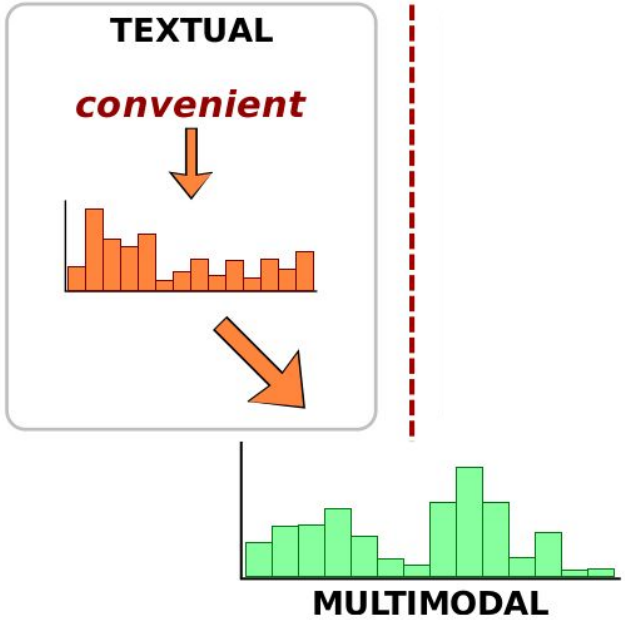
Indexing a new collection, by just having access to the images, no the labels



Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

Word Spotting

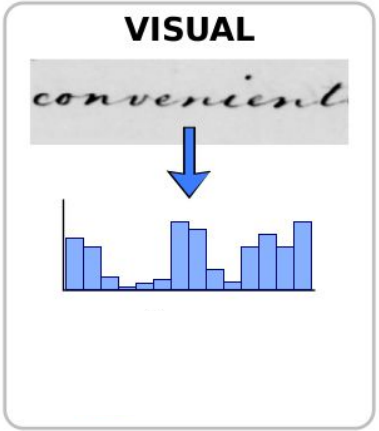
Querying a new word, by just typing it, we do not have any exemplar image



Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

Word Spotting

We do have many discriminative visual features for handwriting words

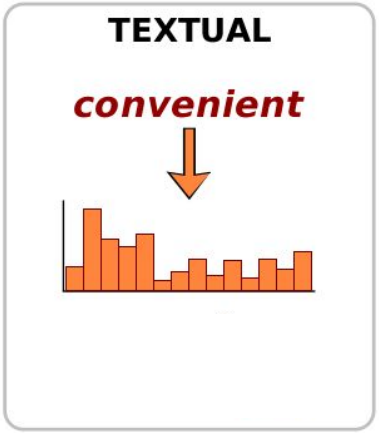


Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

Word Spotting

But what about relevant string features?

- Word2Vec ?
- FastText ?
- Bag of n-grams
- DCToW
- PHOC



Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

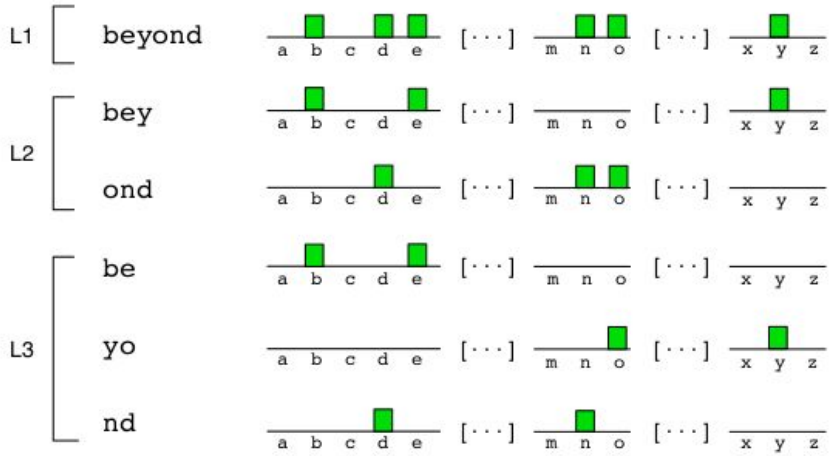
Word Spotting

PHOC

Pyramidal Histogram of Characters

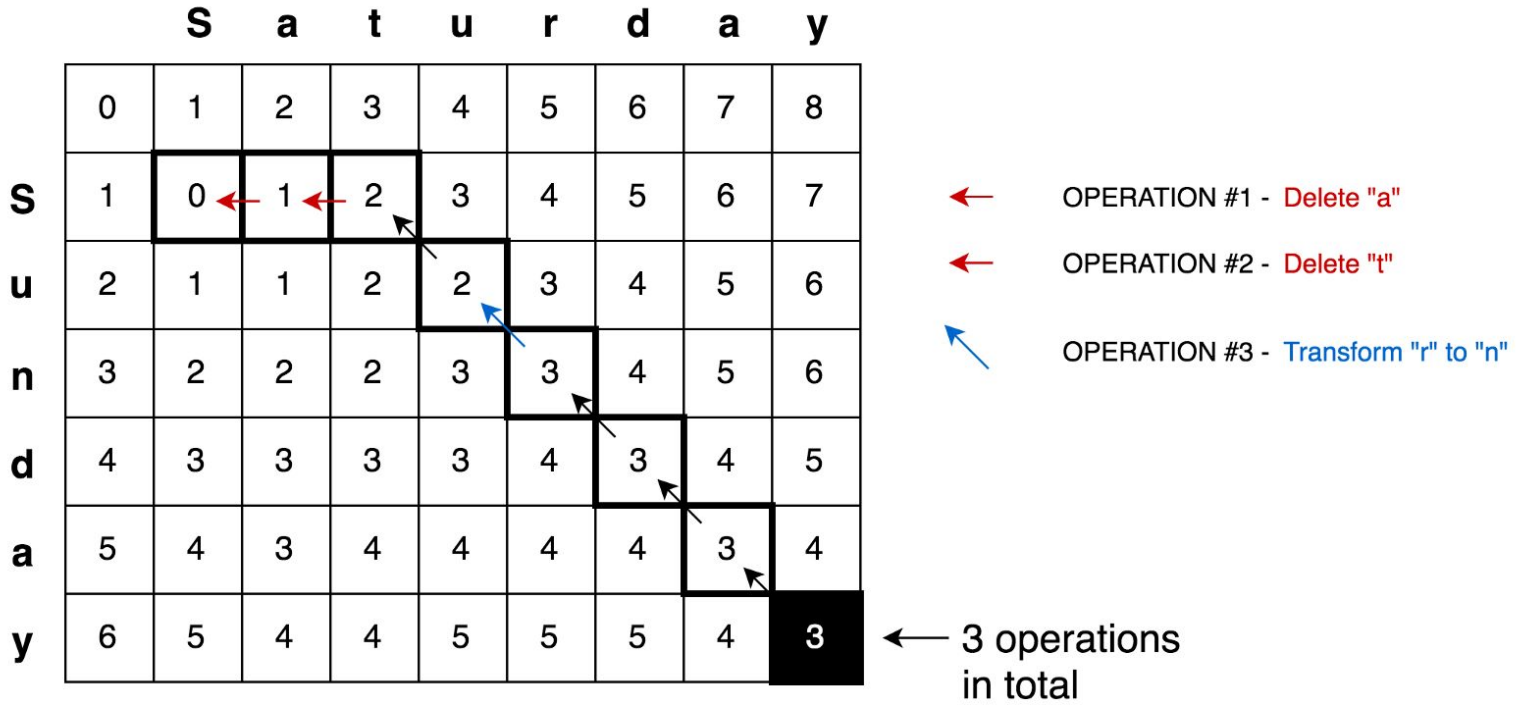
Important drawback:

small changes in words are **huge** in the vectorial space



Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

Levenshtein Distance

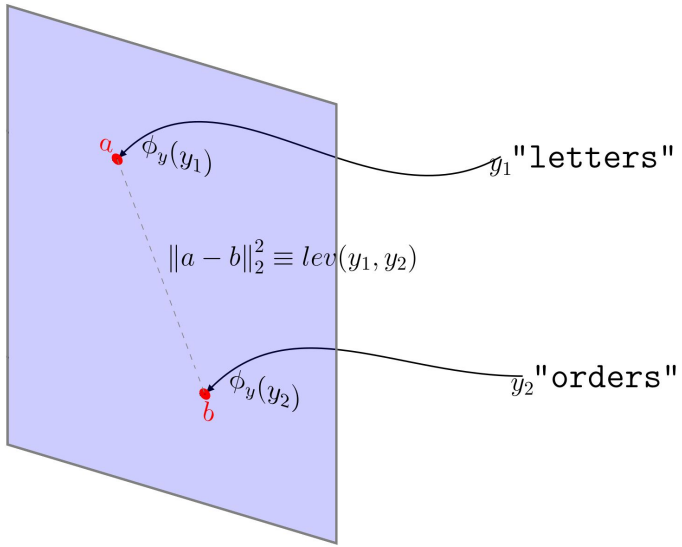


Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

LSDE

So, why not **learn** a new embedding space between strings and images that correlates with the Levenshtein Distance?

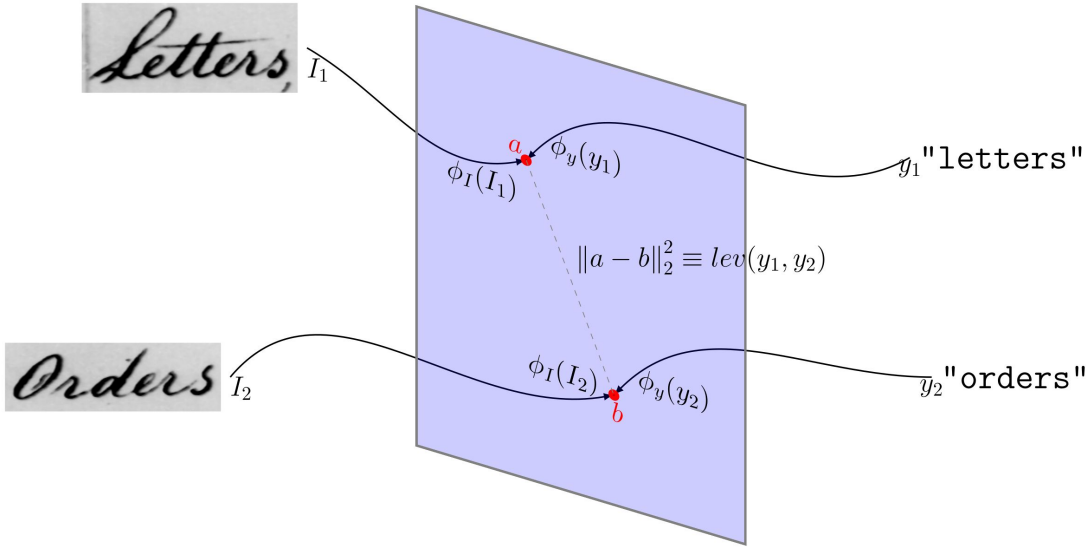
So, why not **learn** a new embedding space between strings and images that correlates with the Levenshtein Distance?



Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

LSDE

So, why not **learn** a new embedding space between strings and images that correlates with the Levenshtein Distance?



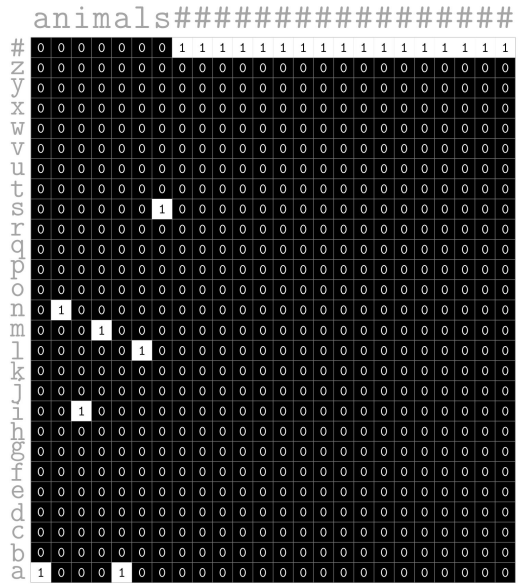
Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

LSDE

Each encoded character corresponds to a one-hot vector in the corresponding matrix column for its particular position on the string.

Fixed max string length 😞

"animals" ⟹



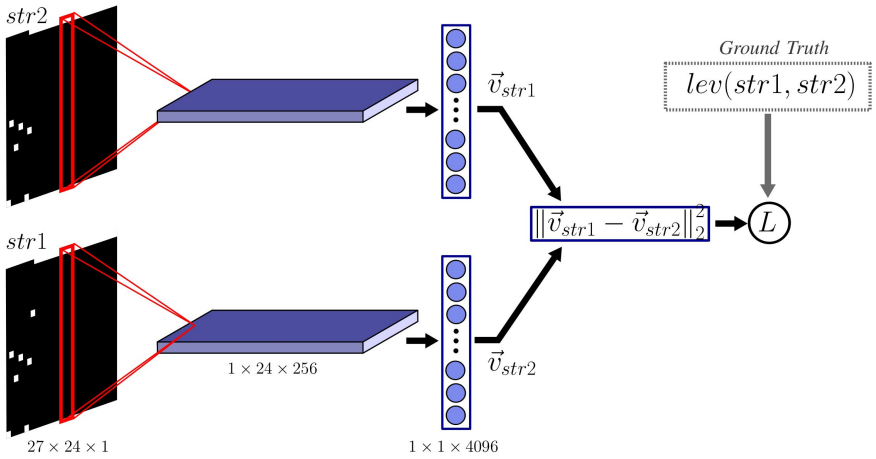
Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

LSDE

Upon this matrix representation of text strings our CNN model applies a convolutional layer with 256 kernels of size 27 × 3 and a fully connected layer with 4096 output neurons.

The network is trained with a siamese setup, presented with arbitrary pairs of text strings, using the following loss function:

$$L = (\sum(\vec{v}_{str1}, \vec{v}_{str2})^2 - lev(str1, str2))^2$$

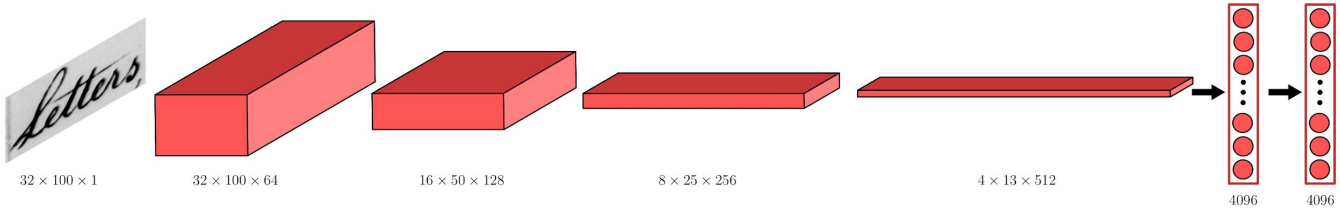


Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

LSDE

Once the string embedding model is trained we can use it to teach an image embedding model so that, given a word image as input, it regresses at its output the LSDE representation of the corresponding string provided by the string embedding model.

At training time either the Cross Entropy or Euclidean loss functions can be used to learn the optimal weights for embedding images into our learned Levenshtein space.

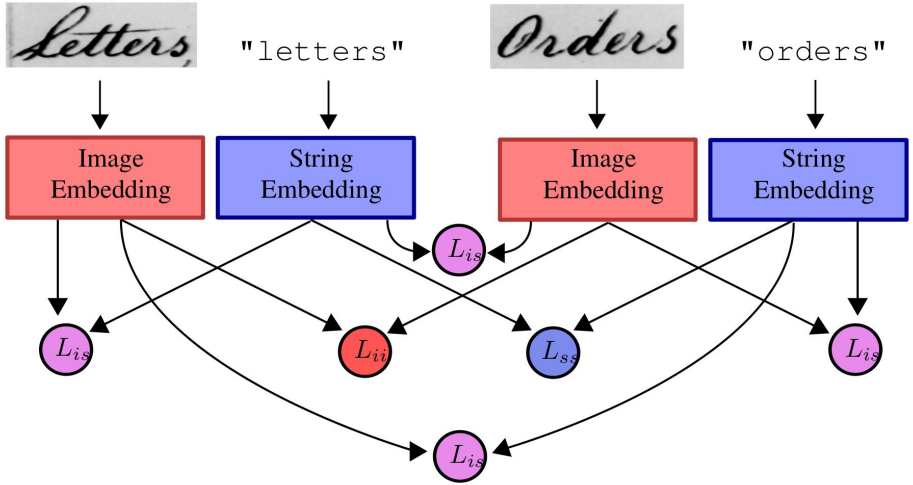


Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

LSDE

The joint image-string embedding is initialized from the two trained models.

A combined loss improves both models by fostering Euclidean-Levenshtein equivalence among all possible pairs of words representations.



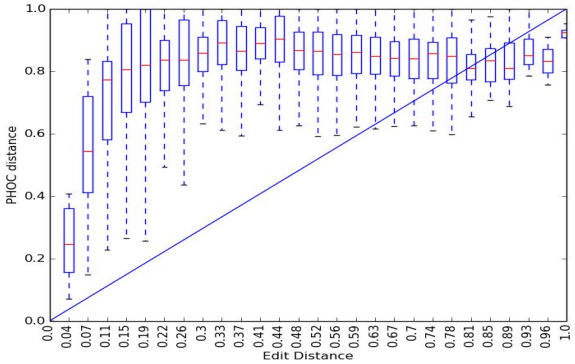
Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

LSDE

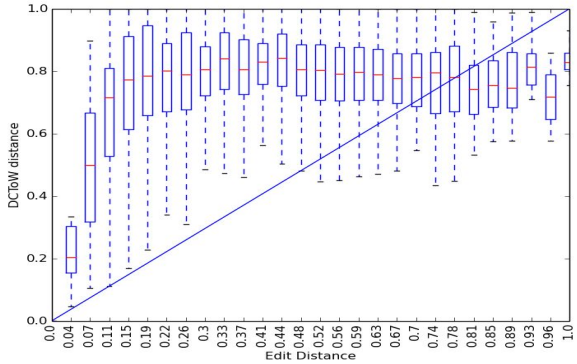
Some results.

Correlation between string representations and edit distances

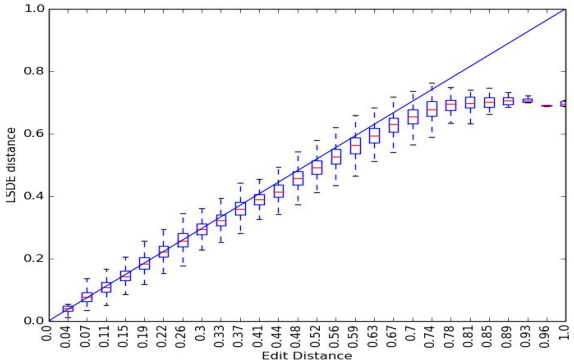
PHOC



DCToW



LSDE

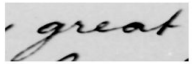
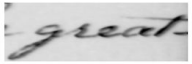
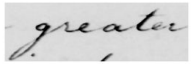
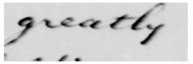
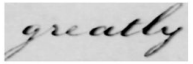
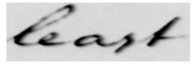
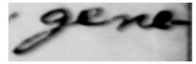
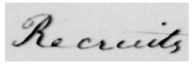
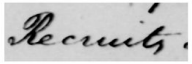
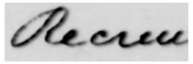
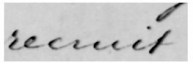
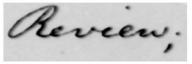
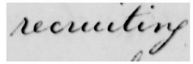
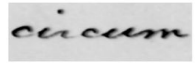
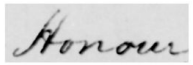
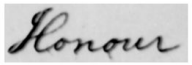
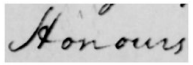
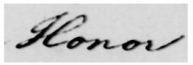
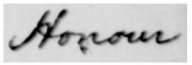
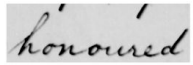
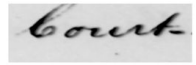
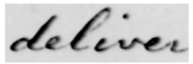
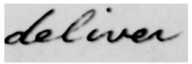
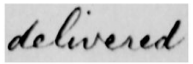
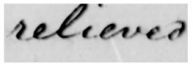
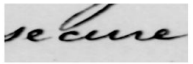
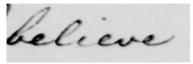


Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

LSDE

Some results:

Qualitative results

great	 great ED=0	 great ED=0	 greater ED=2	 greatly ED=2	 greatly ED=2	 least ED=3	 gene ED=3
recruits	 recruits ED=0	 recruits ED=0	 recru ED=3	 recruit ED=1	 review ED=5	 recruiting ED=3	 circum ED=7
honour	 honour ED=0	 honour ED=0	 honours ED=1	 honor ED=1	 honour ED=0	 honoured ED=2	 court ED=4
deliver	 deliver ED=0	 deliver ED=0	 delivered ED=2	 relieved ED=3	 secure ED=5	 believe ED=3	 else ED=4

Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

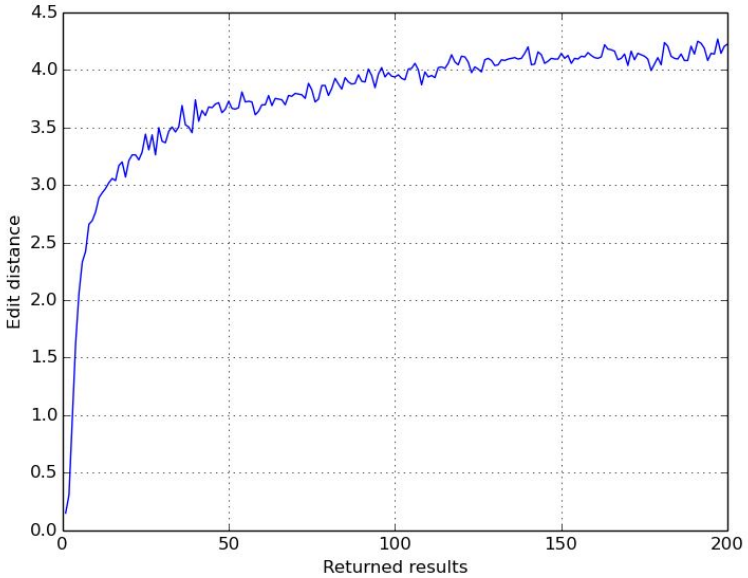
Some results

mAP on the George Washington dataset

Method	**QBS mAP**
Aldavert et al. (Aldavert 2013)	56.54
Frinken et al. (Frinken 2012)	84.00
Almazan et al. (Almazan 2014)	91.29
Sudholt et al. (Sudholt 2016)	92.64
Krishnan et al. (Krishnan 2016)	92.84
Wilkinson et al. (Wilkinson 2016)	**93.69**
LSDE	91.31

Some results

Average Levenshtein distance between query and results

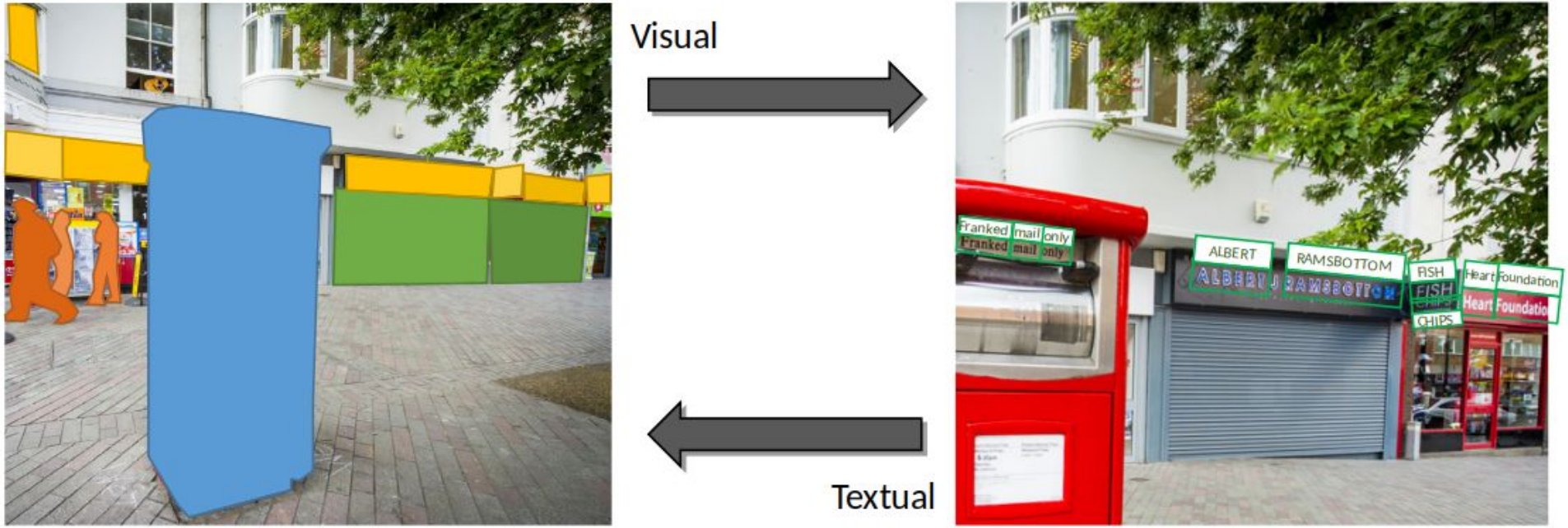


Gomez, Rossinyol, Karatzas. "LSDE: Levenshtein Space Deep Embedding for Query-by-string Word Spotting." ICDAR17

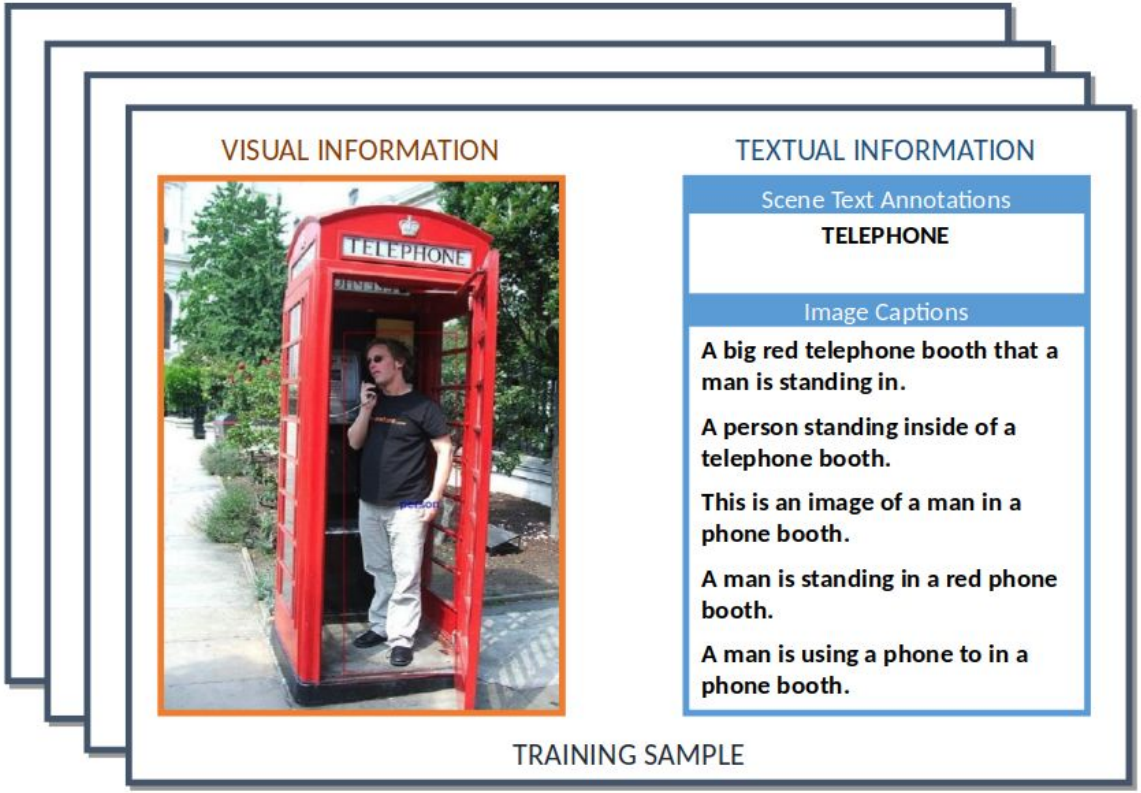
Text and Images

Text and Images

Aim: to give computer vision models the ability to read, while exploiting visual information to improve that reading ability.

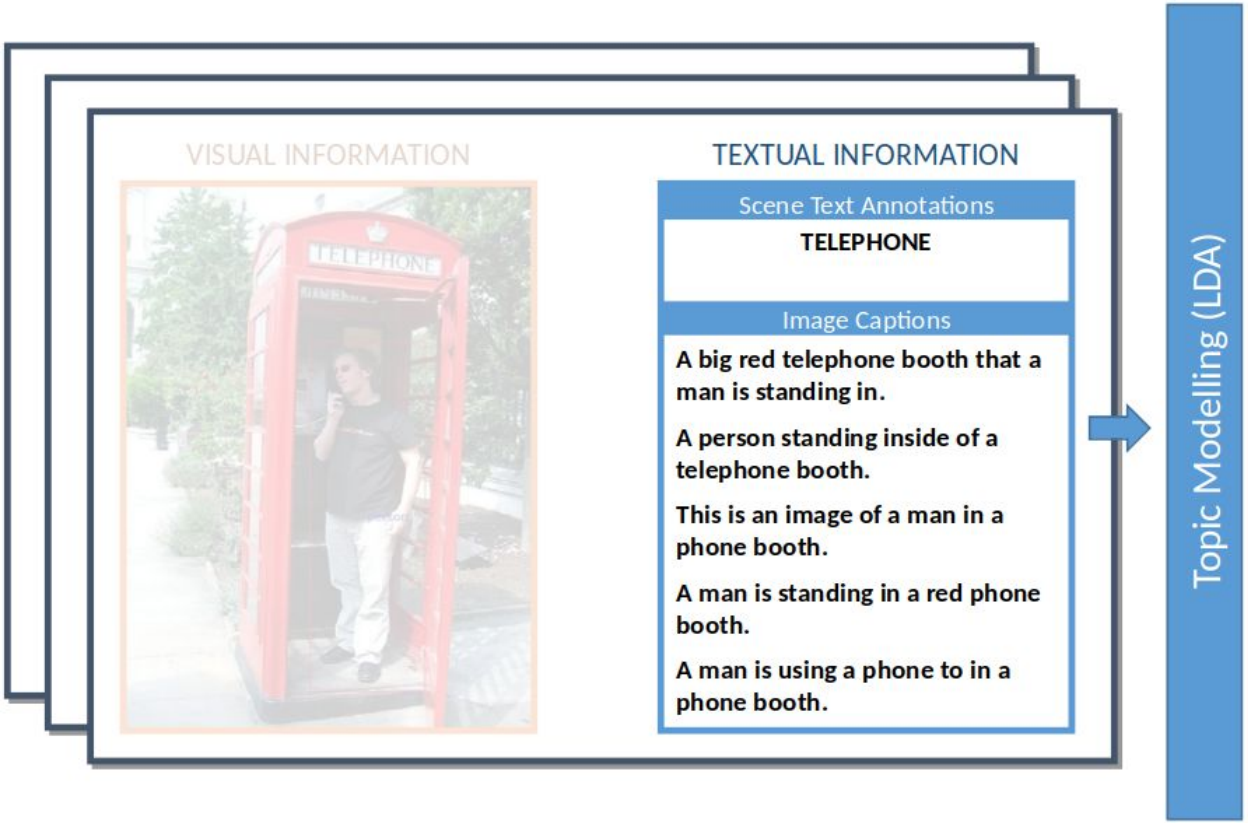


Dynamic lexicon generation



Y. Patel, L. Gomez, M. Rusiñol, D. Karatzas, “Dynamic Lexicon Generation for Natural Scene Images”, IWRR 2016

Dynamic lexicon generation



Y. Patel, L. Gomez, M. Rusiñol, D. Karatzas, “Dynamic Lexicon Generation for Natural Scene Images”, IWRR 2016

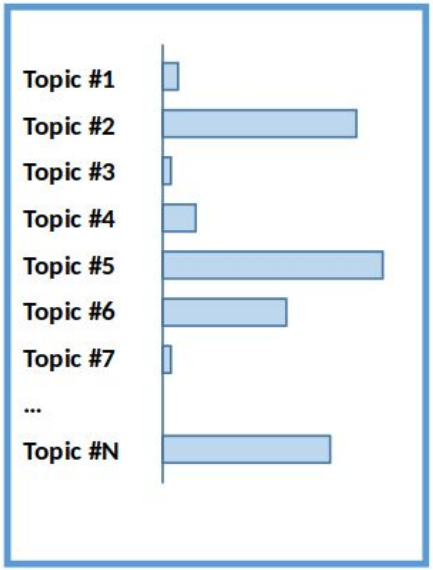
Dynamic lexicon generation

Learn a model to predict the corresponding topic space given an image.

VISUAL INFORMATION

CNN

P (TOPIC | TEXT)



Y. Patel, L. Gomez, M. Rusiñol, D. Karatzas, “Dynamic Lexicon Generation for Natural Scene Images”, IWRR 2016

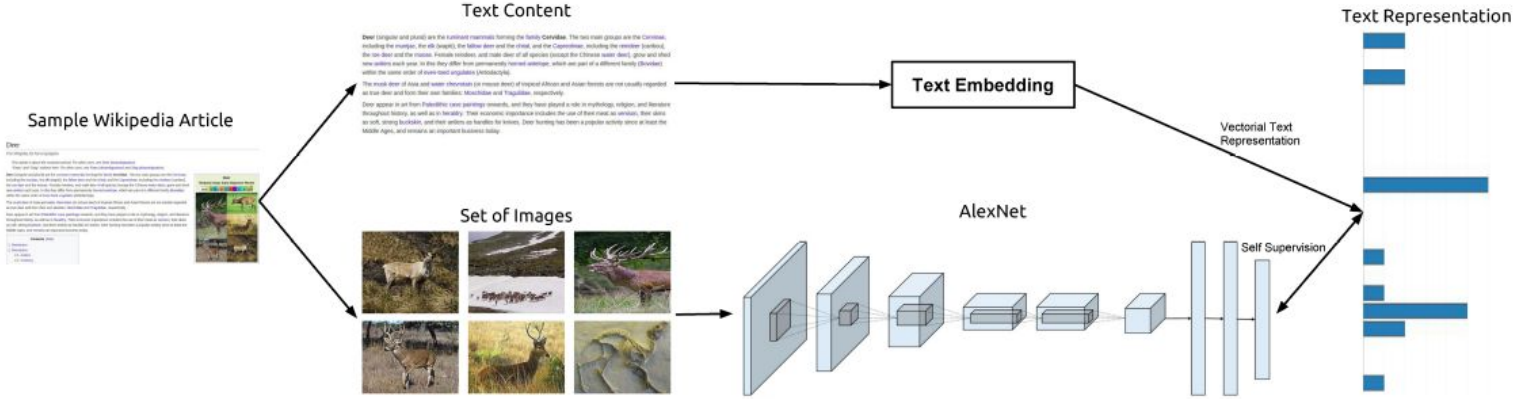
Dynamic lexicon generation

Image	Word Rank	Image	Word Rank
	Fire: 2 Hydrant: 4		Wii: 11
	High: 182 Street: 1		Tennis: 1

Y. Patel, L. Gomez, M. Rusiñol, D. Karatzas, “Dynamic Lexicon Generation for Natural Scene Images”, IWRR 2016

TextTopicNet

Wikipedia articles comprise a **textual description** of a subject as well as **illustrative images** supporting the text.



A text embedding framework generates a global contextual representation of the textual information. This vectorial text representation of the entire text article is used to provide the self-supervision for the training of the CNN.

L. Gomez et al, “Self-supervised learning of visual features through embedding images into text topic spaces”, CVPR 2017

Y. Patel et al, “Self-Supervised Visual Representations for Cross-Modal Retrieval”, ICMR 2019

TextTopicNet

We train our models on a subset of the English Wikipedia articles. **1.7M** unique articles and **4.2M** images.

Top-5 most relevant words and top-6 most relevant images for four of the discovered topics.

L. Gomez et al, “Self-supervised learning of visual features through embedding images into text topic spaces”, CVPR 2017

Y. Patel et al, “Self-Supervised Visual Representations for Cross-Modal Retrieval”, ICMR 2019

TextTopicNet

	Method	aer	bk	brd	bt	btl	bus	car	cat	chr	cow	din	dog	hrs	mbk	prs	pot	shp	sfa	trn	tv
TextTopicNet	TextTopicNet (Wikipedia)	**71**	**52**	**47**	**61**	**26**	**49**	**71**	**46**	**47**	**36**	**44**	**41**	72	**62**	**85**	**31**	**40**	**42**	**72**	**44**
	TextTopicNet (ImageCLEF)	67	44	39	53	20	49	68	42	43	33	41	35	70	57	82	30	31	39	65	41
Self-Supervised Methods	Sound [7]	69	45	38	56	16	47	65	45	41	25	37	28	**74**	61	85	26	39	32	69	38
	Texton-CNN	65	35	28	46	11	31	63	30	41	17	28	23	64	51	74	9	19	33	54	30
	K-means	61	31	27	49	9	27	58	34	36	12	25	21	64	38	70	18	14	25	51	25
	Motion [22]	67	35	41	54	11	35	62	35	39	21	30	26	70	53	78	22	32	37	61	34
	Patches [4]	70	44	43	60	12	44	66	52	44	24	45	31	73	48	78	14	28	39	62	43
	Egomotion [6]	60	24	21	35	10	19	57	24	27	11	22	18	61	40	69	13	12	24	48	28
Fully Supervised Methods	ImageNet [52]	79	**71**	**73**	75	**25**	60	80	**75**	51	**45**	60	**70**	**80**	**72**	**91**	42	**62**	56	82	62
	Places [2]	**83**	60	56	**80**	23	**66**	**84**	54	**57**	40	**74**	41	**80**	68	90	**50**	45	**61**	**88**	**63**

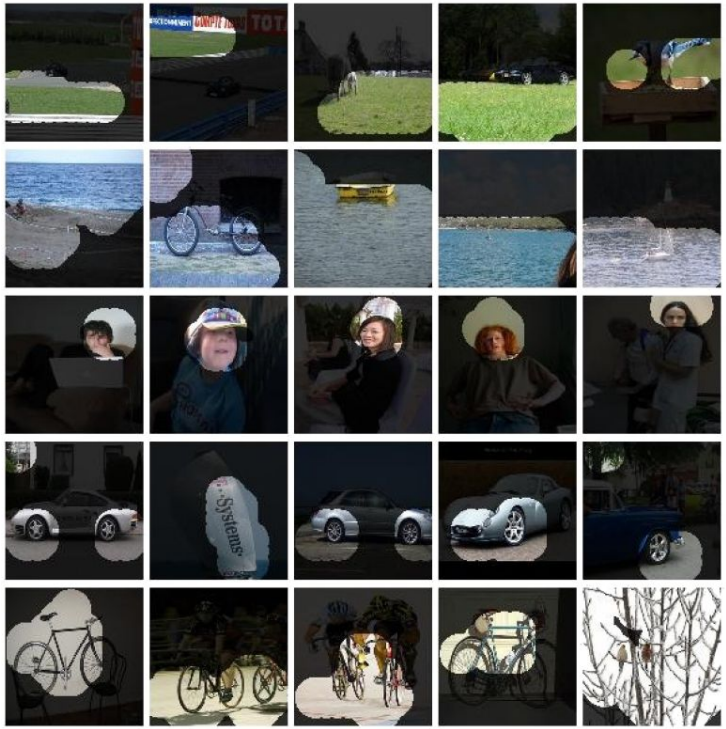
PASCAL VOC2007 per-class average precision (AP) scores for the classification task. One-vs-all Linear SVM classifier with pool5 features

	Method	max5	pool5	fc6	fc7
TextTopicNet	TextTopicNet (Wikipedia)	-	**51.9**	**54.2**	**55.8**
	TextTopicNet (ImageCLEF)	-	47.4	48.1	48.5
Self-Supervised Methods	Sound [7]	39.4	46.7	47.1	47.4
	Texton-CNN	28.9	37.5	35.3	32.5
	K-means [20]	27.5	34.8	33.9	32.1
	Tracking [22]	33.5	42.2	42.4	40.2
	Patch pos. [4]	26.8	46.1	-	-
	Egomotion [6]	22.7	31.1	-	-
Fully Supervised Methods	ImageNet [52]	**63.6**	**65.6**	**69.6**	**73.6**
	Places [2]	59.0	63.2	65.3	66.2

PASCAL VOC2007 %mAP for image classification. One-vs-all Linear SVM classifier, features from different top-layers.

L. Gomez et al, “Self-supervised learning of visual features through embedding images into text topic spaces”, CVPR 2017
Y. Patel et al, “Self-Supervised Visual Representations for Cross-Modal Retrieval”, ICMR 2019

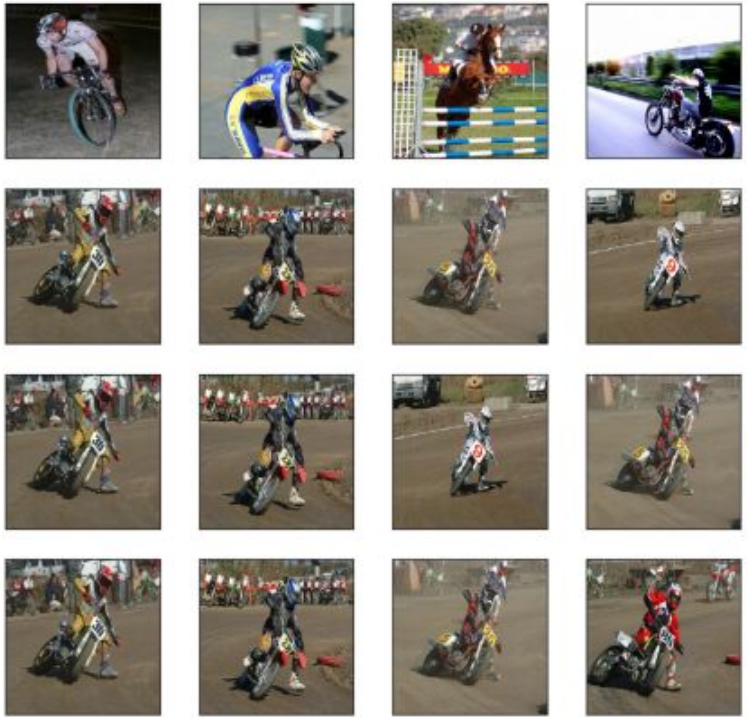
TextTopicNet



L. Gomez et al, “Self-supervised learning of visual features through embedding images into text topic spaces”, CVPR 2017

Y. Patel et al, “Self-Supervised Visual Representations for Cross-Modal Retrieval”, ICMR 2019

TextTopicNet



Last-layer features: **semantic** retrieval

Earlier layers: **visual similarity** retrieval

L. Gomez et al, “Self-supervised learning of visual features through embedding images into text topic spaces”, CVPR 2017
Y. Patel et al, “Self-Supervised Visual Representations for Cross-Modal Retrieval”, ICMR 2019

Semantic Retrieval

Why stop at Wikipedia?

Wikipedia:
1.7M articles in English with 4.2M associated illusrtative images.

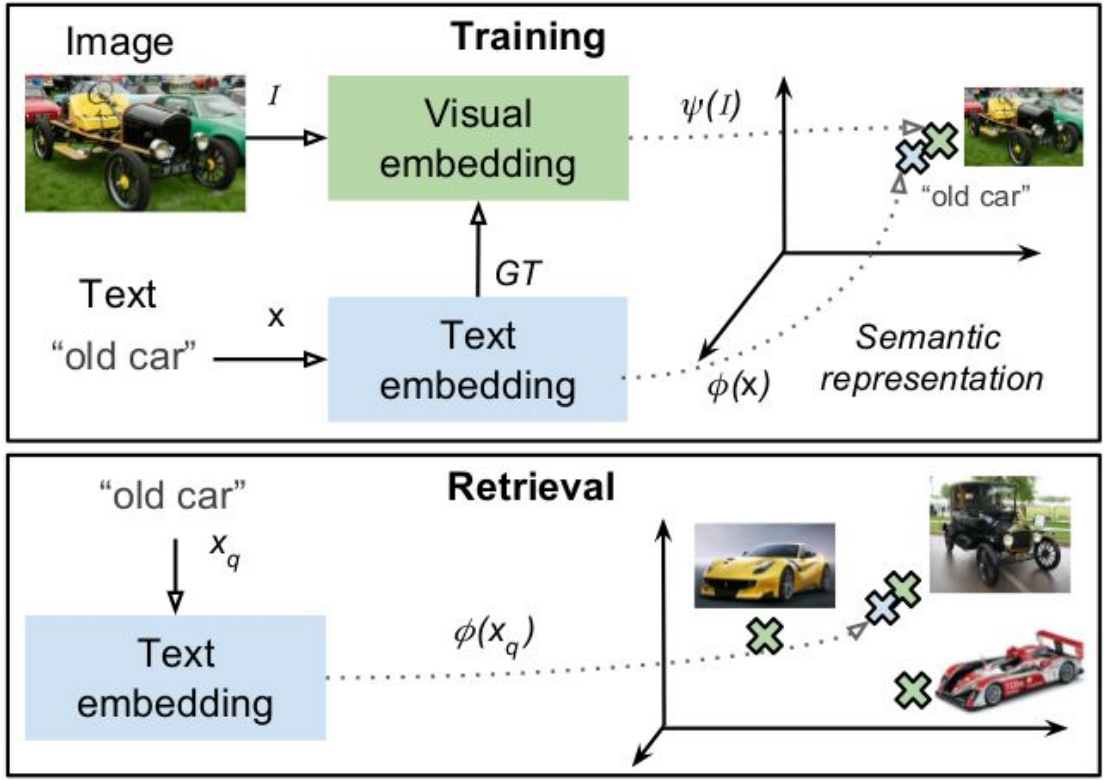
WebVision:
2.4M Flickr and Google images associated to ImageNet classes.

InstaCities1M:
1M Instagram images associated with one of the 10 most populated English speaking cities.

Gomez et al. “Learning to Learn from Web Data through Deep Semantic Embeddings” MULA 2018
Gomez et al. “Self-Supervised Learning from Web Data for Multimodal Retrieval”, arXiv:1901.02004, 2019

Semantic Retrieval

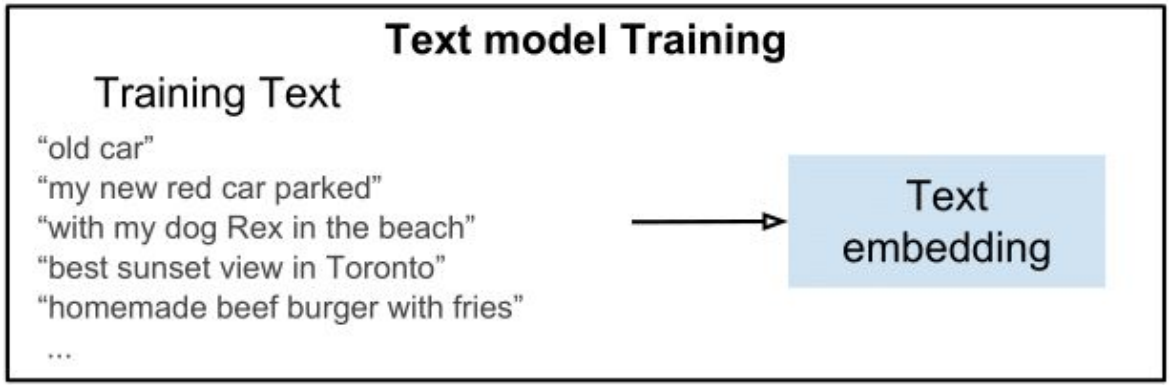
Proposed pipeline



Gomez et al. "Learning to Learn from Web Data through Deep Semantic Embeddings" MULA 2018

Gomez et al. "Self-Supervised Learning from Web Data for Multimodal Retrieval", arXiv:1901.02004, 2019

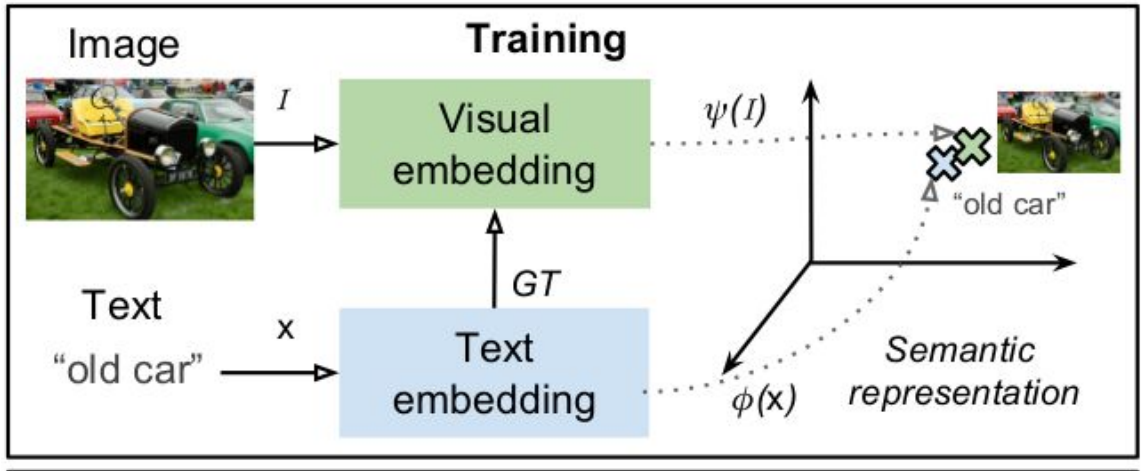
Semantic Retrieval



- **LDA** [1]: Learns latent topics from a collection of text documents and maps words to a vector of probabilities of those topics.
- **Word2Vec** [2]: Learns relationships between words automatically using a feed-forward neural network.
- **Doc2Vec** [3]: Is an extension of Word2Vec to documents.
- **GloVe** [4]: It is a count-based model. It learns the word vectors by essentially doing dimensionality reduction on the co-occurrence counts matrix.
- **FastText** [5]: While Word2Vec and GloVe treat each word in a corpus like an atomic entity, FastText treats each word as composed of character n-grams. So the vector of a word is made of the sum of this character n-grams.

Gomez et al. “Learning to Learn from Web Data through Deep Semantic Embeddings” MULA 2018
Gomez et al. “Self-Supervised Learning from Web Data for Multimodal Retrieval”, arXiv:1901.02004, 2019

Semantic Retrieval

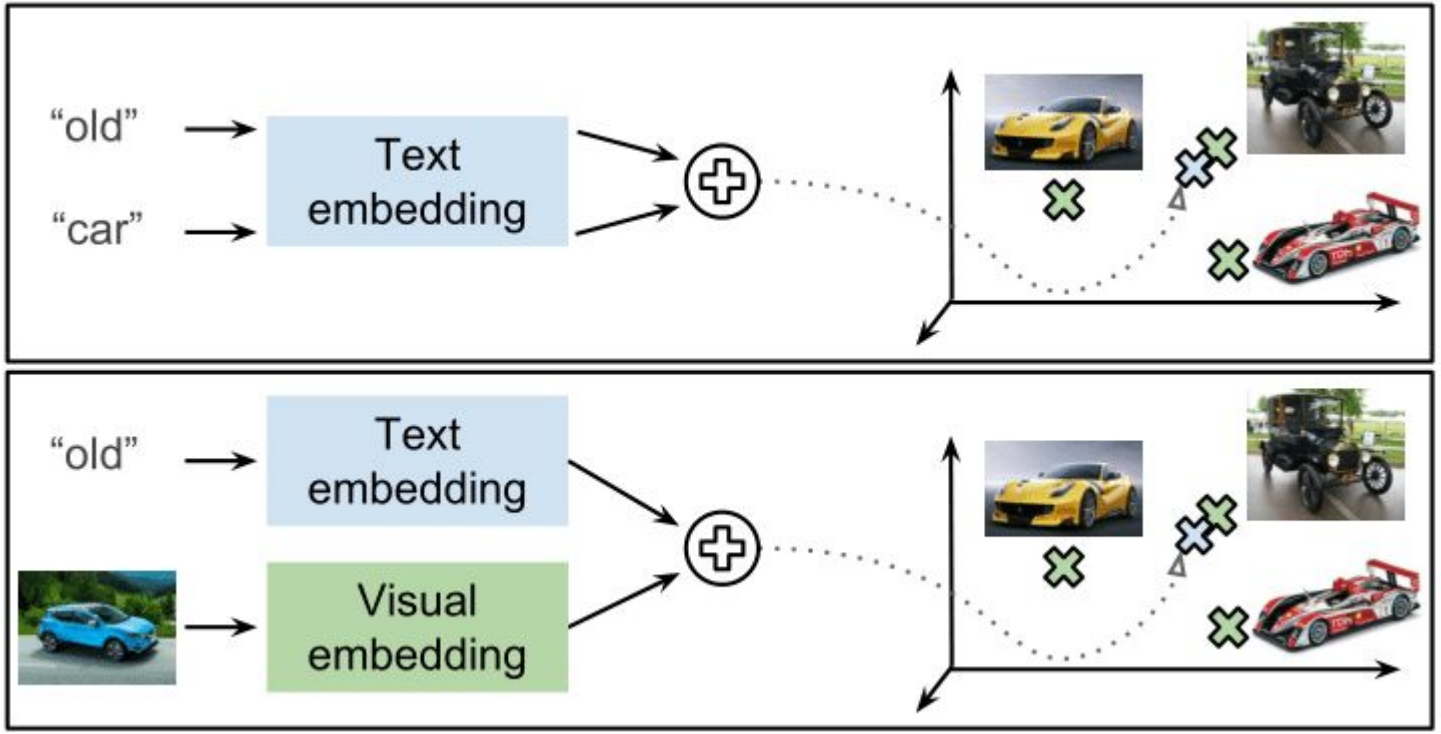


Visual embedding: A GoogleNet CNN is trained to regress text embeddings from the correlated images minimizing a sigmoid cross-entropy loss.

We get a text and a visual model that can embed respectively text and images in a common space with semantic structure.

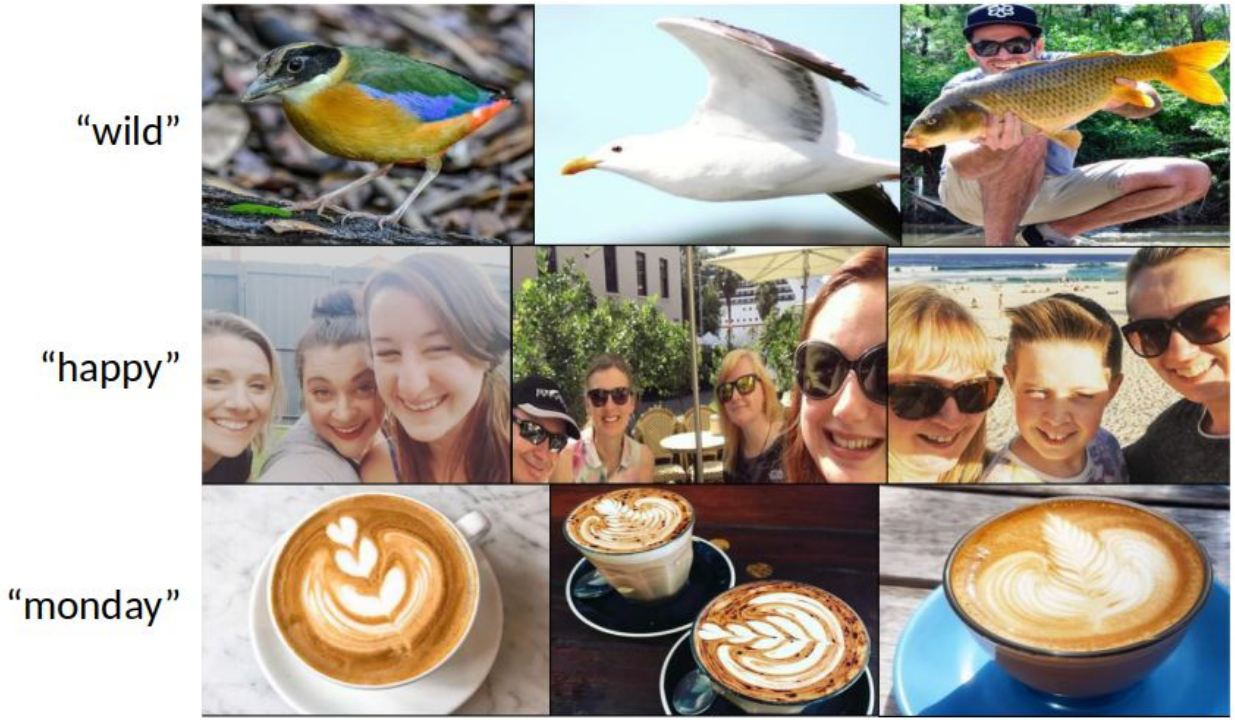
Gomez et al. "Learning to Learn from Web Data through Deep Semantic Embeddings" MULA 2018
Gomez et al. "Self-Supervised Learning from Web Data for Multimodal Retrieval", arXiv:1901.02004, 2019

Semantic Retrieval



Gomez et al. "Learning to Learn from Web Data through Deep Semantic Embeddings" MULA 2018
Gomez et al. "Self-Supervised Learning from Web Data for Multimodal Retrieval", arXiv:1901.02004, 2019

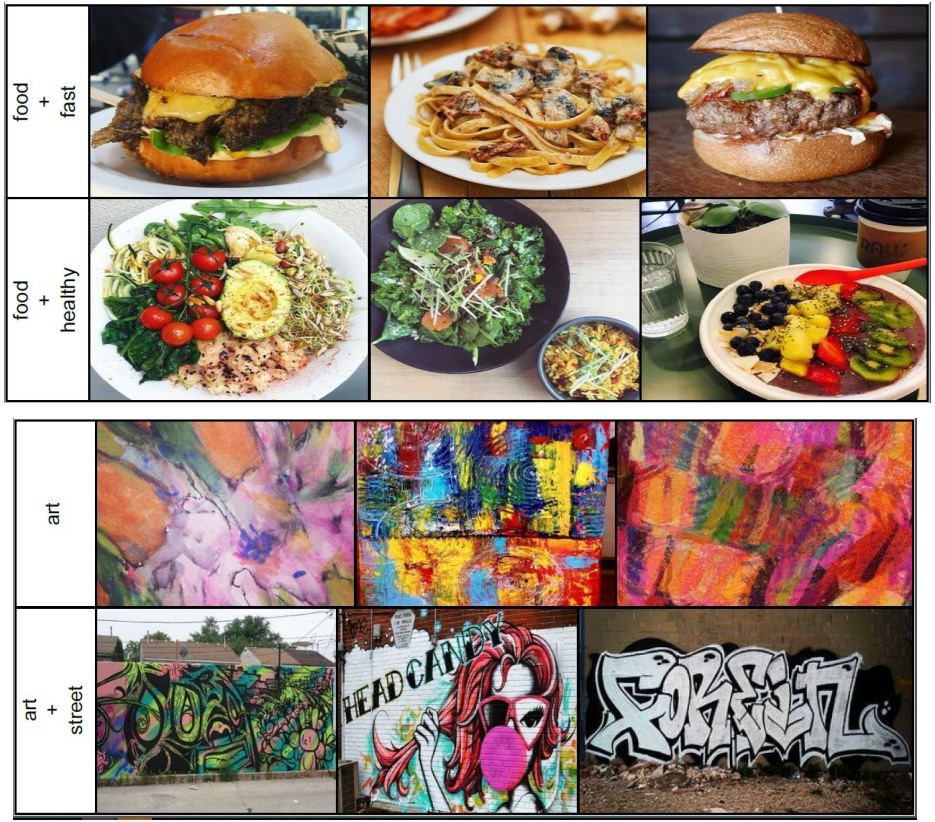
Semantic Retrieval



Model trained with Word2Vec on InstaCites1M

Gomez et al. “Learning to Learn from Web Data through Deep Semantic Embeddings” MULA 2018
Gomez et al. “Self-Supervised Learning from Web Data for Multimodal Retrieval”, arXiv:1901.02004, 2019

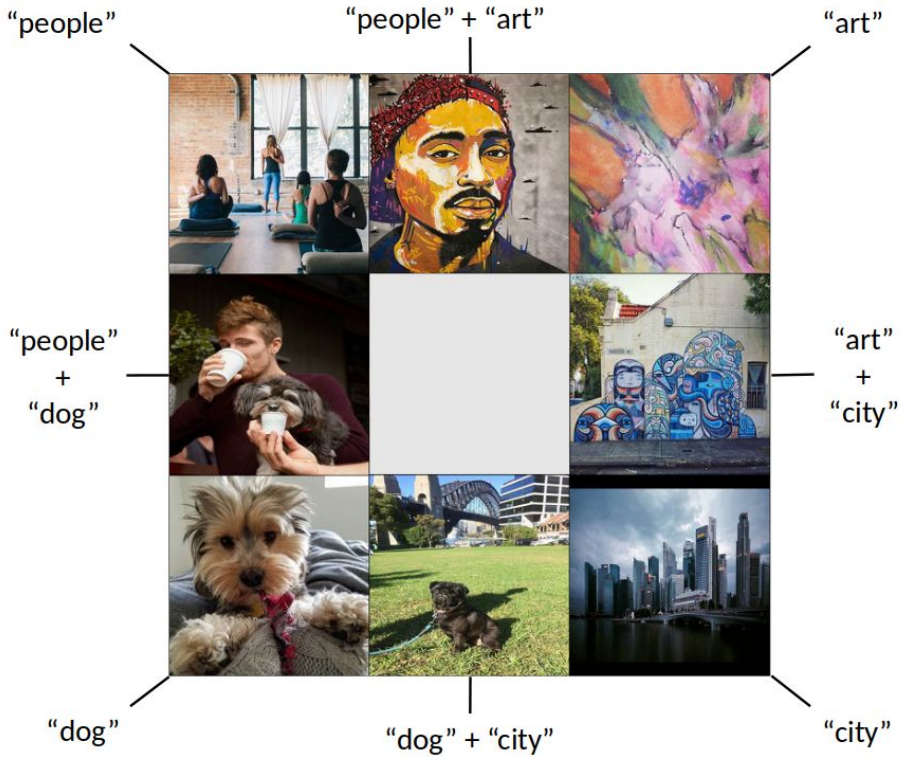
Semantic Retrieval



Gomez et al. “Learning to Learn from Web Data through Deep Semantic Embeddings” MULA 2018

Gomez et al. “Self-Supervised Learning from Web Data for Multimodal Retrieval”, arXiv:1901.02004, 2019

Semantic Retrieval

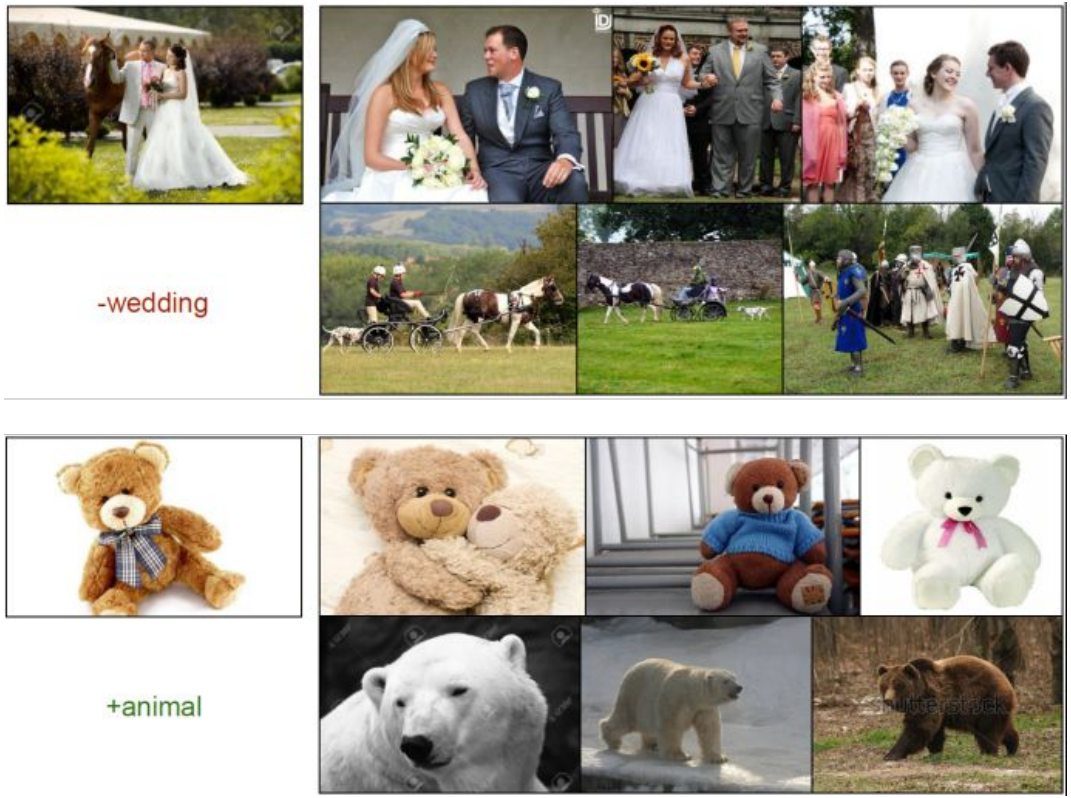


Gomez et al. “Learning to Learn from Web Data through Deep Semantic Embeddings” MULA 2018
Gomez et al. “Self-Supervised Learning from Web Data for Multimodal Retrieval”, arXiv:1901.02004, 2019

Semantic Retrieval

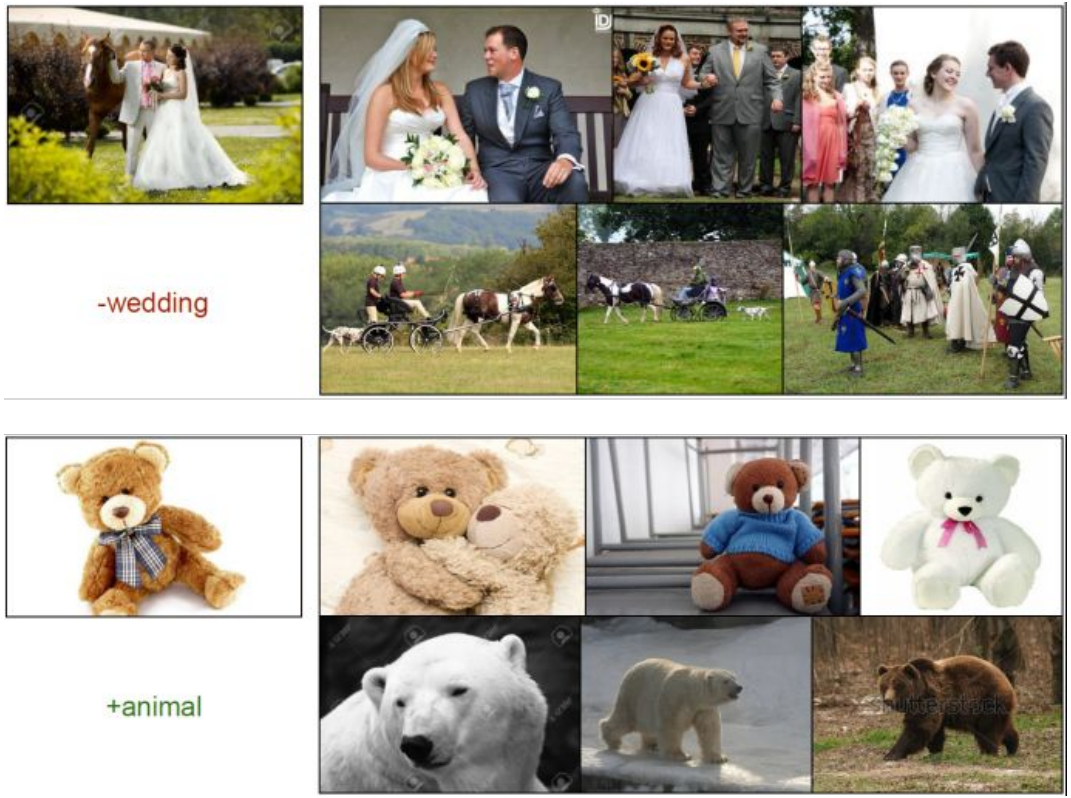
Gomez et al. “Learning to Learn from Web Data through Deep Semantic Embeddings” MULA 2018
Gomez et al. “Self-Supervised Learning from Web Data for Multimodal Retrieval”, arXiv:1901.02004, 2019

Semantic Retrieval



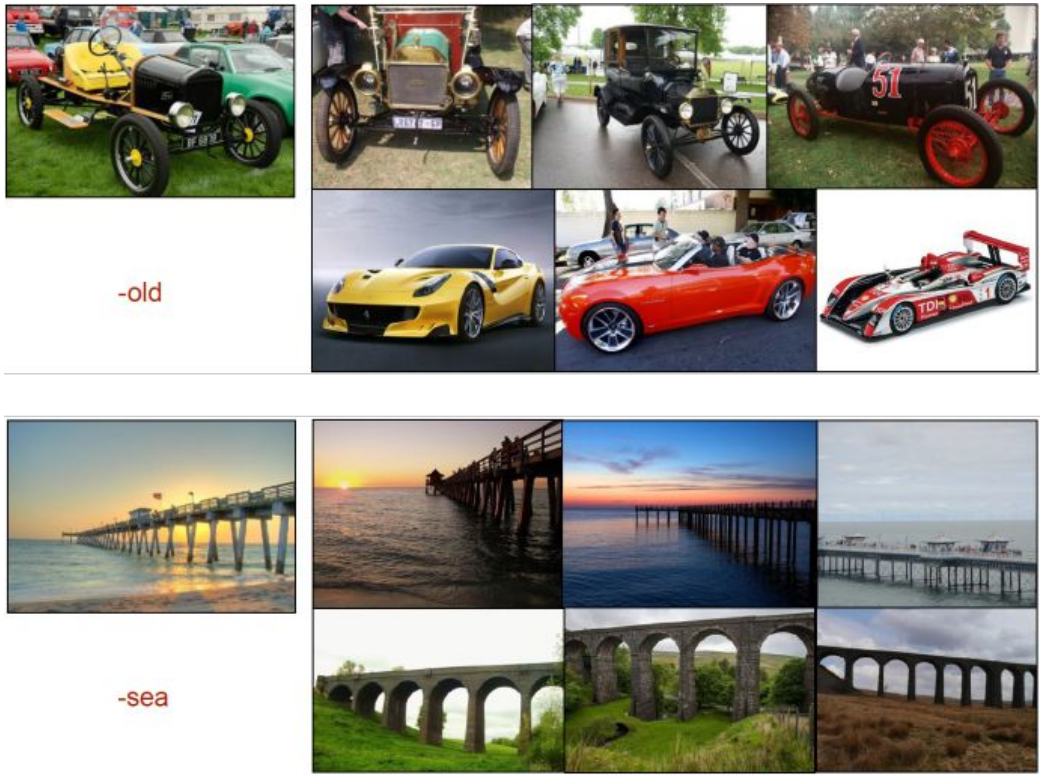
Gomez et al. “Learning to Learn from Web Data through Deep Semantic Embeddings” MULA 2018
Gomez et al. “Self-Supervised Learning from Web Data for Multimodal Retrieval”, arXiv:1901.02004, 2019

Semantic Retrieval



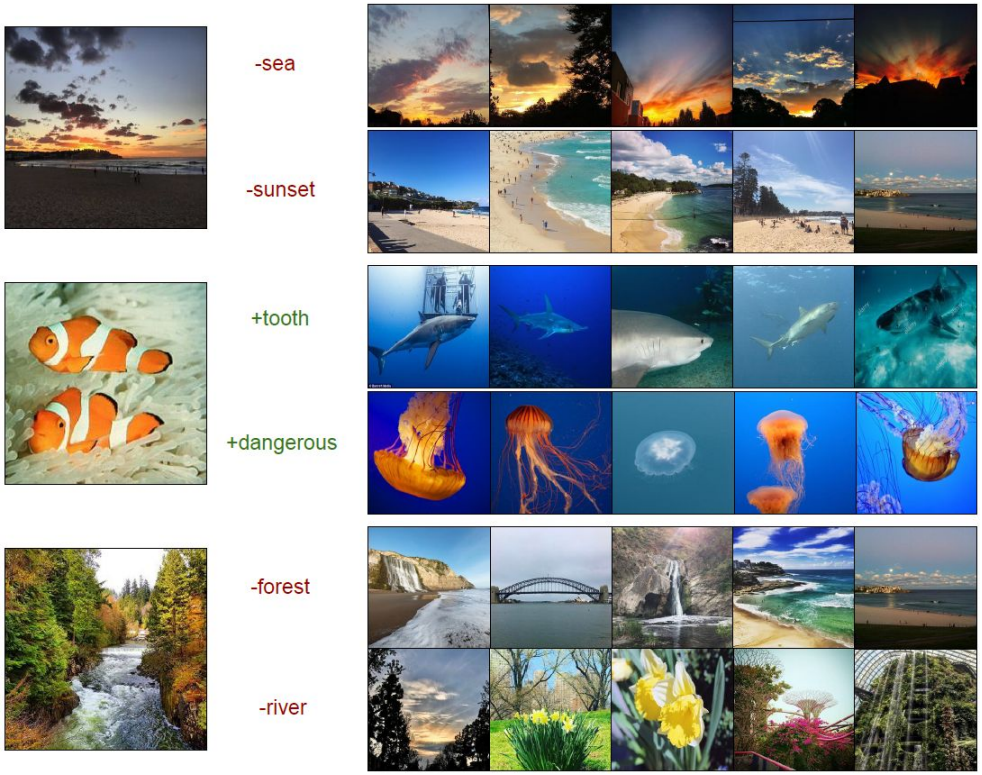
Gomez et al. “Learning to Learn from Web Data through Deep Semantic Embeddings” MULA 2018
Gomez et al. “Self-Supervised Learning from Web Data for Multimodal Retrieval”, arXiv:1901.02004, 2019

Semantic Retrieval



Gomez et al. “Learning to Learn from Web Data through Deep Semantic Embeddings” MULA 2018
Gomez et al. “Self-Supervised Learning from Web Data for Multimodal Retrieval”, arXiv:1901.02004, 2019

Semantic Retrieval



Gomez et al. “Learning to Learn from Web Data through Deep Semantic Embeddings” MULA 2018
Gomez et al. “Self-Supervised Learning from Web Data for Multimodal Retrieval”, arXiv:1901.02004, 2019