# Different Modes of Semantic Representation in Image Retrieval

Rory Bennett  Kristina Striegnitz (Advisor)

Computer Science Department, Union College

## Motivation

Image retrieval systems are supposed to only retrieve images that are relevant to a given query. Therefore, they need methods by which to represent the meaning of both the query word and the image, so that these meanings can then be compared. Distributional semantic models typically use semantic vectors to represent words' meanings, based on the extent to which they appear near other words in text. By comparing these semantic vectors, we can compare words' meanings, and thus find words that are simlar or relevant to each other. In this study, I extend this idea, to implement an improved image retrieval system: I build semantic vectors for both words in text and captioned images, and compare these vectors to find, for each query term, the image whose meaning is most relevant to the query's meaning.

## Background

### Key Terms and Techniques

- "Vector space models (VSMs) . . . [embed] words in a continuous vector space where semantically similar words are mapped to nearby points"[1]
- Distributional semantic models (DSMs)[2] find similarities between individual words' VSMs to find similarities between words' meanings (see Table 1)
- Multimodal semantic models integrate information from VSMs across text, images, etc. to find more accurate word similarities[3]
- Propagation: select information from perceptual dataset, add to textual dataset
- Textual data refers to pure text corpus; perceptal data refers to image captions/descriptions

| Word | Semantic Vector |
|---|---|
| cat | $-0.351283, -0.065883, -0.091065, \ldots$ |
| pet | $-0.322014, -0.042600, -0.0781283 \ldots$ |
| tuba | $0.106879, 0.006146, 0.134201 \ldots$ |

Table 1: "Cat" and "pet" have similar vector space representations, because their meanings are similar. "Tuba" is similar to neither word.

- Concrete word examples: *cat*, *chocolate*
- Abstract word examples: *love*, *war*

### Multimodal Distributional Semantics

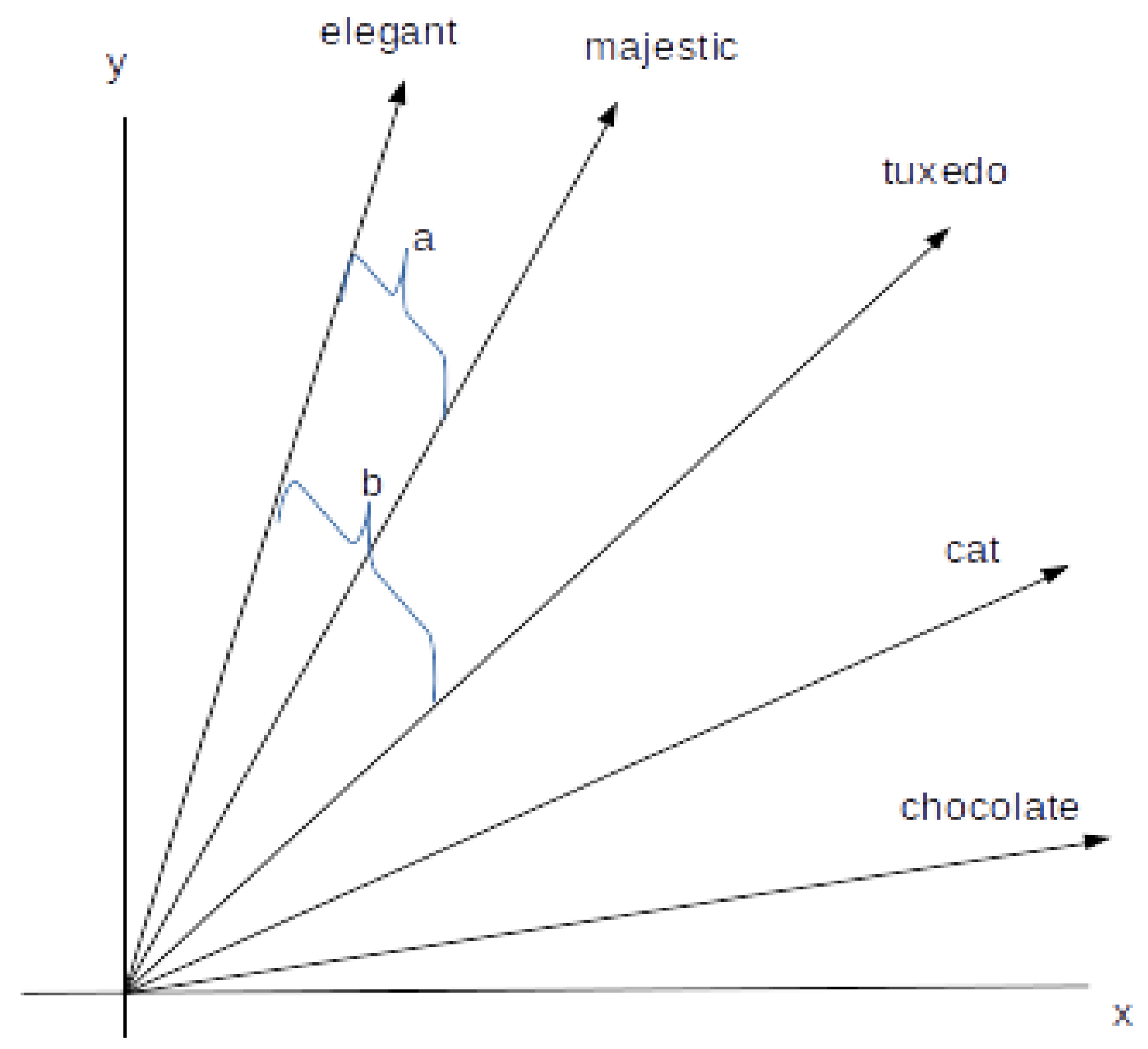
- Hill and Korhonen (H&K): if word occurs in both pure text and captions, then its nearest neighbors in captions should relate to its nearest neighbors in text
- Propagation of words' information to text improved semantic representation (SR) of concrete terms
- For words that occur in ESP-Game or CSLB captions, map word to list of words that co-occur with it → "bag of perceptual features" [4] (BoPF)
- Given a word in text that occurs in both perceptual datasets, H&K inserted "pseudo-sentences" of word and its context words from its BoPF into Text8
- For many abstract terms, H&K's model actually performed worse than when SRs produced by text
- H&K also tried limiting propagation only to concrete words' SRs, whenever they occurred in the text
- Results: improved SRs for abstract and concrete words alike, from purely textual VSMs.
- Suggests that basis for whether word can effectively be represented with images is its concreteness

## Question

When querying an image retrieval system with a word associated with no image, how does propagating perceptual information to the word's SR affect the system's ability to retrieve relevant images?

## Distributional Semantics for Image Retrieval

- Abstract query term "elegant" not likely to be in image captions, so cannot by itself return relevant images
- Even if most similar words to "elegant" are abstract, cannot be used, so must look for most similar that are also in perceptual datasets



*a, b represent cosine distances between semantic vectors

*a < b, so "majestic" is more similar/relevant to "elegant" than "tuxedo" is

Figure 1: High cosine similarity between words' VSMs indicates high similarity between semantic meanings.

## Query-Image Comparison

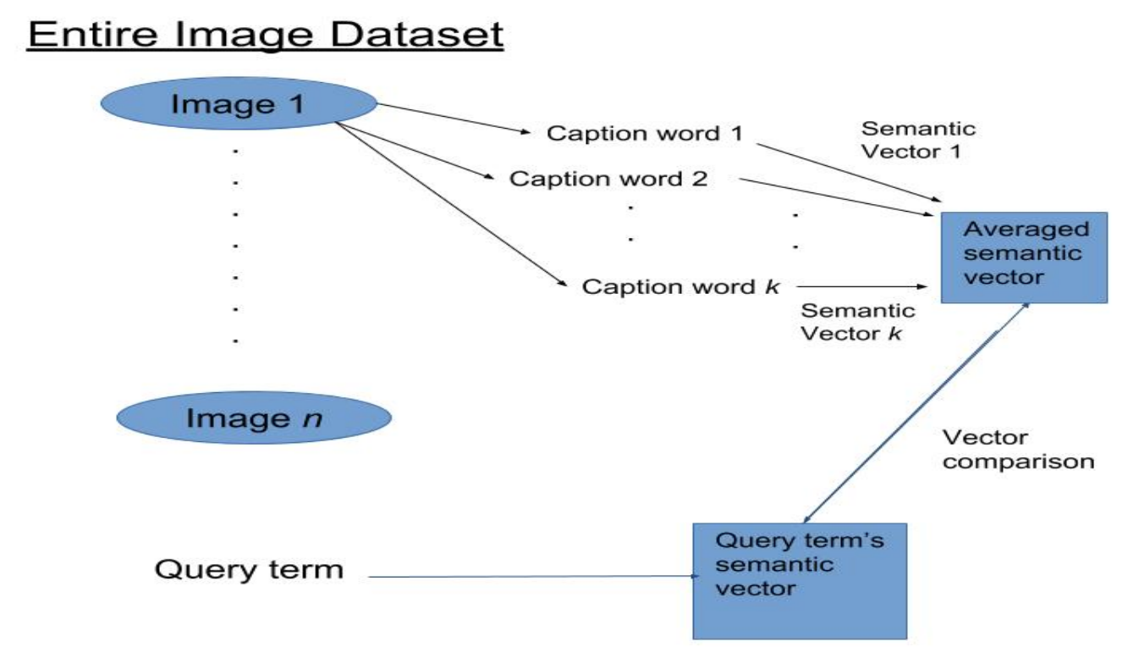


Figure 2: Technique A: for each captioned image, average the vectors of the image's caption words, and compare the average vector with the query term's vector.

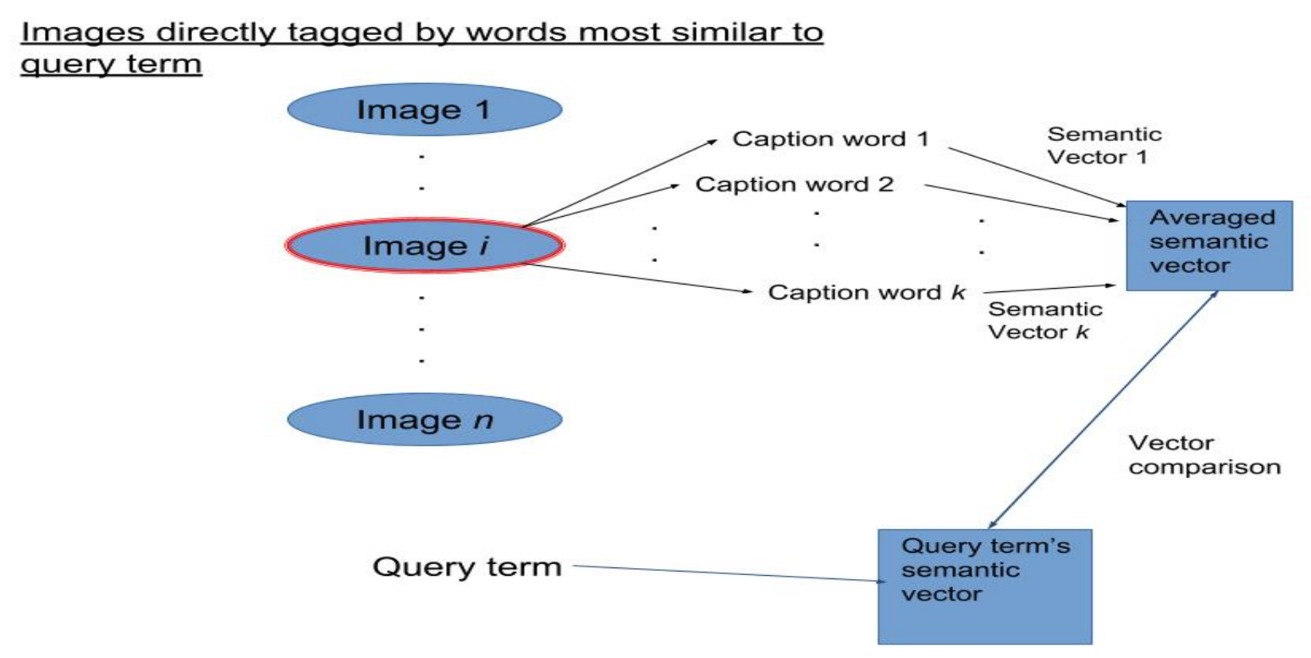


Figure 3: Technique B: only compare query term's vector with images that are directly tagged by words that are relatively similar to the query term in the textual corpus.

## Experiment

- Using H&K's textual and perceptual corpora, apply five different approaches to retrieve images, four of which utilize distributional semantic techniques:
  1. Retrieve images directly tagged by given query
  2. Technique A, using word vectors derived from the plain Text8 corpus,
  3. Technique B, applied to the abovementiond plain Text8 corpus,
  4. Technique A, applied to Text8, augmented by the perceptual information that was propagated to it,
  5. Technique B, applied to the abovementioned augmented Text8 corpus
- Tested techniques over a set of 32 nouns and verbs, of varying concreteness

## Results

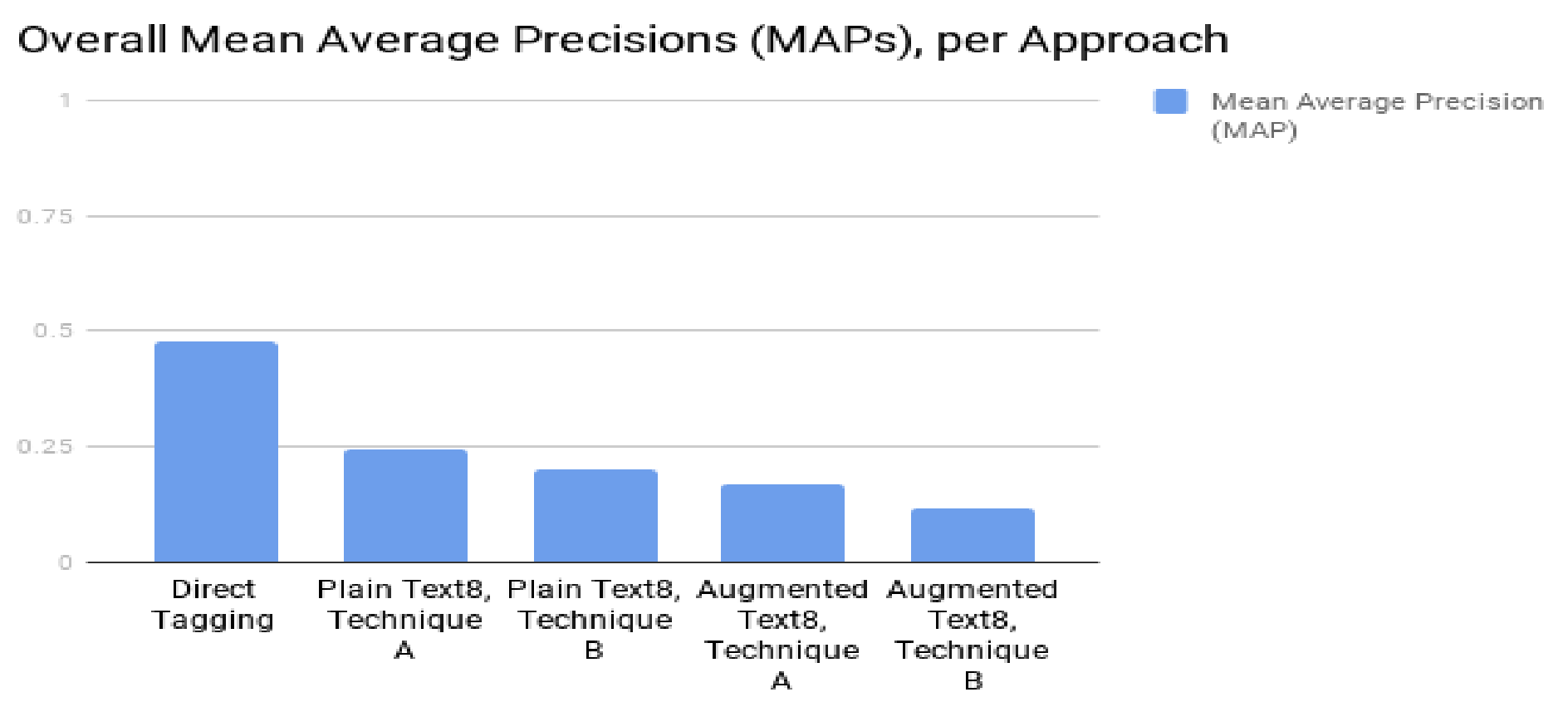


Figure 4: Each approach's mean average precision (MAP), over the average precisions of human relevance ratings of the images returned for each query term.

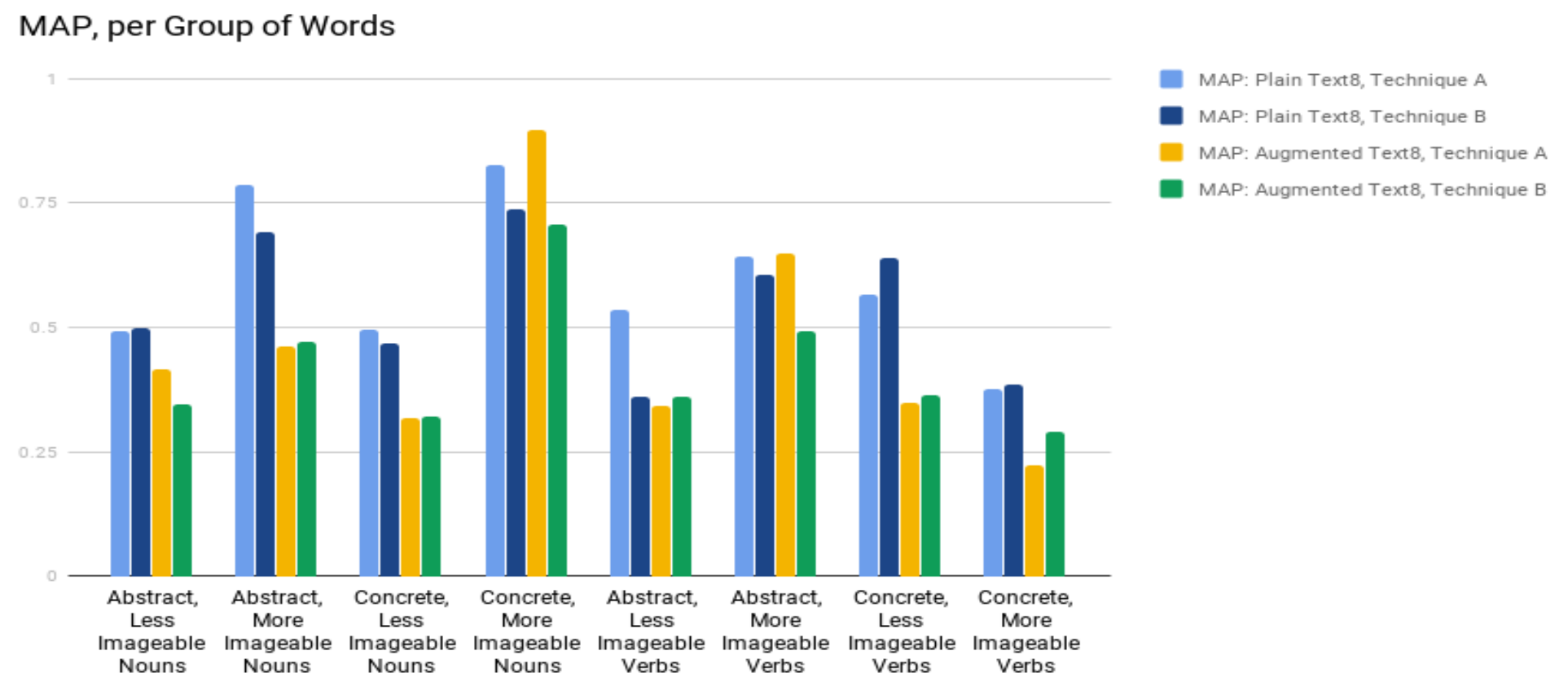


Figure 5: Each approach's mean average precision (MAP), over the average precisions of unique subsets of 4 query terms.

## References

[1] https://www.tensorflow.org/tutorials/word2vec, 2017, Nov. 2.
[2] Turney and Pantel, 2010; Sahlgren, 2006.
[3] Fagarasan, Luana, Eva Maria Vecchi, and Stephen Clark, 2015.
[4] Hill, Felix, and Anna Korhonen, 2014.