eman ta zabal zazu

# Evaluating Multimodal Representations on Sentence Similarity: vSTS, Visual Semantic Textual Similarity Dataset

Oier Lopez de Lacalle, Eneko Agirre and Aitor Soroa
IXA research group, University of Basque Country
http://ixa.si.ehu.es/

## Introduction

### Motivation

- **Text understanding**: Success of word representation motivated methods to represent longer sequences of text.
- **Multimodality**: Gained attention on image-caption retrieval, video and text alignment, caption generation, visual question answering, etc.
- **Complementarity**: Visual and text representation for improved language understanding.

### Goal

- Present Visual Semantic Textual Similarity dataset.
- Allow to study if better representation can be built when having access to corresponding images.

### Hypothesis

- **H1**: If image alone are able to predict caption similarity.
- **H2**: If combination of image and text representations allow to improve text only results.

## Semantic Textual Similarity

### Task

- **Assessment** of pairs of sentences according to their degree of similarity.
- **Similarity**: 0 for no meaning overlap - 5 for meaning equivalence.
- **Metric**: Pearson correlation with human judgmets

### Visual Content

- "A man is mid-leap over a stack of wooden steps."
- "A woman is doing gymnastics in a large building"

## The vSTS dataset

### Annotation

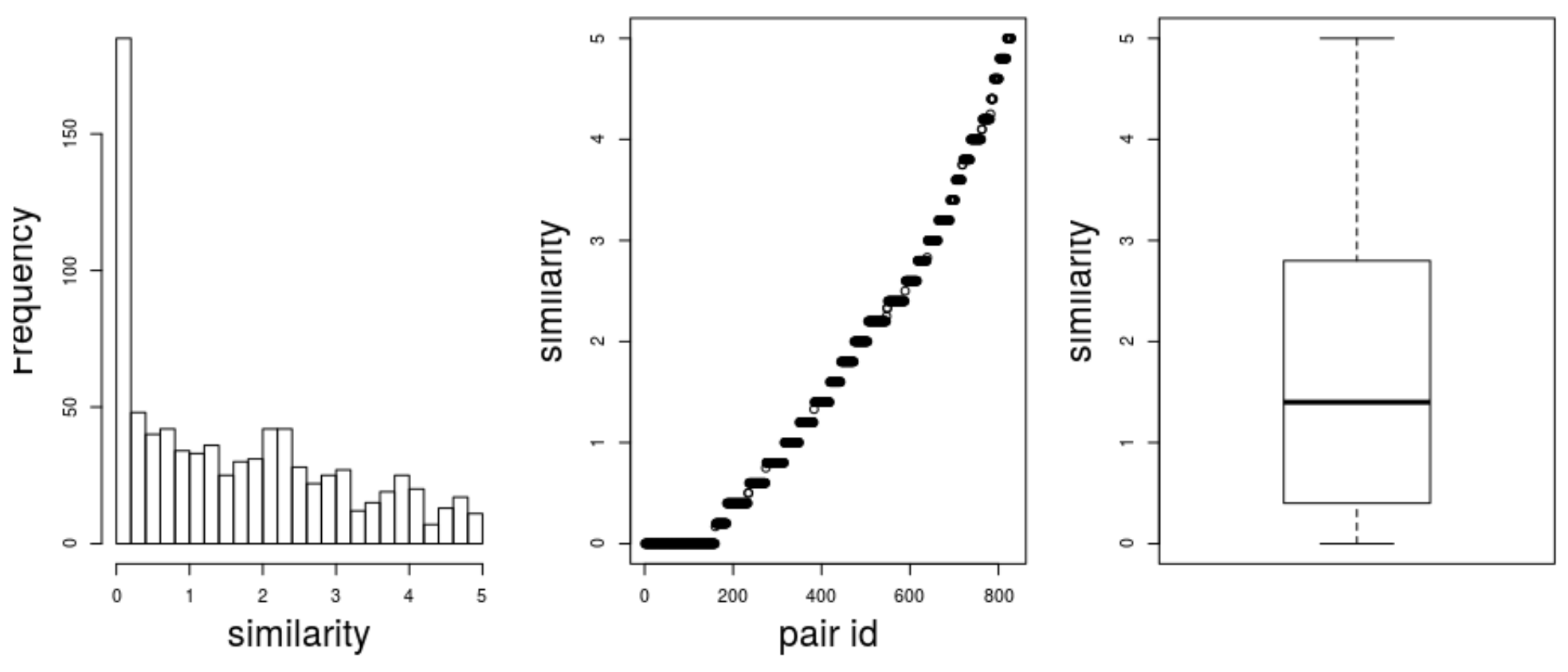
- Caption pairs annotated for the STS benchmark - *Image Description* subset.
- Annotators only had access to text .
- Filter out captions referring to the same image (avoid trivial task).

### Subsets

- **Subset 2014**: Subset of the PASCAL VOC-2008 dataset.
  - Obtained 374 pairs (out of 750 in the original file).
- **Subset 2015**: Subset of Flickr8K benchmark collection for sentence based image descriotion.
  - Obtained 445 pairs (out of 750 in the original file).

### Stats

| subset | #pairs | mean sim | std sim | #zeroes |
|---|---|---|---|---|
| 2014 | 374 | 1.77 | 1.49 | 78 |
| 2015 | 445 | 1.69 | 1.44 | 81 |
| Total | 819 | 1.72 | 1.46 | 159 |



## Experiments

### Settings

- **Dev/Test**: Sample 50% at random preserving the overall similarity distribution.
- **Train**: Part of the text-only STS benchmark dataset as a training set, discarding the examples that overlap with vSTS.
- **Evaluation metric**: Pearson correlation.

### Models

- OVERLAP: Bag-of-words model with cosine similarity.
- CAVERAGE: Glove word embedding based centroid with cosine similarity.
- DAM: Decompositional Attention Model.
- RESNET50: top layer of a pretrained resnet50 model with cosine similarity.

### Combinations

- Combine the predictions of text based models with image based model.
- ⊕: Sum of two outputs.
- ⊗: Multiplication of the output
- LR: Linear regression of two outputs.
  - Parameters estimated with 10fold xval on dev.

## Results

| Modality | Model | Dev set | | | Test set | | |
|---|---|---|---|---|---|---|---|
| TEXT | A - OVERLAP | 0.68 | | | 0.64 | | |
| | B - CAVERAGE | 0.65 | | | 0.67 | | |
| | C - DAM | **0.71** | | | **0.69** | | |
| IMAGE | D - RESNET50 | 0.63 | | | 0.61 | | |
| Combination | | LR | ⊕ | ⊗ | LR | ⊕ | ⊗ |
| TEXT+IMAGE | A+D | 0.77 | 0.77 | 0.77 | 0.76 | 0.75 | 0.75 |
| | B+D | 0.75 | 0.73 | 0.70 | 0.76 | 0.73 | 0.70 |
| | C+D | **0.78** | **0.78** | **0.78** | 0.77 | 0.77 | **0.78** |

### Discussion

**Single models**

- DAM obtains the highest Pearson correlation (expected)
- H1 confirmed: Images alone are valid to predict similarity (0.61)

**Complementarity**

- H2 confirmed: Combination of image and sentence representations obtained the best results (DAM + RESNET50)
- Indications that representation of the real world helps to better understand the text and do better inferences.

## Conclusions & Future Work

### Contributions

- Creation of dataset of caption pairs with human similarity annotations with access to actual images.
- Test the contribution of visual information in STS.
- Experiments confirmed initial hypotheses.

### On going work

- We re-annotated the dataset with scores which are based on both the text and the image.
- First analysis indicate that:
  - Overall similarity values increase when images are present.
  - Similar disagreement on annotators on both settings.
  - High correlation on two annotation frameworks.

### Available at

http://ixa2.si.ehu.eus/~jibloleo/visual_sts.tgz

informatika fakultatea
facultad de informática

Euskal Herriko Unibertsitatea http://www.ehu.eus/eu/web/informatika-fakultatea