Learning Semantic Sentence Representations from Visually Grounded Language

Danny Merkx

d.merkx@let.ru.nl

Radboud University, Centre for Language Studies

Supervisors: Dr. Stefan Frank &
Prof. Dr. Mirjam Ernestus

Radboud University

About me:

- M.Sc. in Artificial Intelligence
- Second year PhD at the Centre for language Studies at the Radboud University in Nijmegen, the Netherlands.
- Part of the Language in Interaction Consortium. Our team is working on the nature of the mental lexicon: How to bridge neurobiology and psycholinguistic theory by computational modelling?

LANGUAGE in INTERACTION

Radboud University

Word embeddings

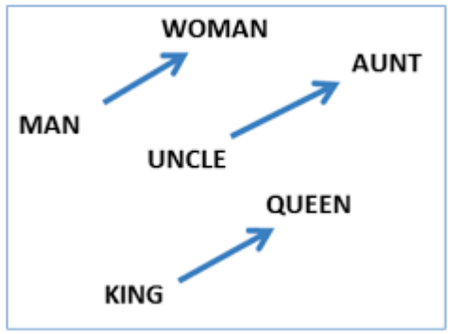
- The idea behind word embeddings has been around for a while:

“You shall know a word by the company it keeps”: John Firth (1957)

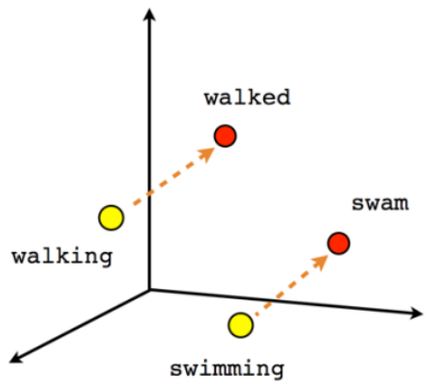
- The idea behind this statement is that words that occur in similar context are similar in meaning.

- We can capture the co-occurrence statistics of our vocabulary using neural networks.
- Simply put, an embedding is just a string of numbers which in this case encodes similarity; we want similar words to have a similar embedding and dissimilar words to have very different embeddings

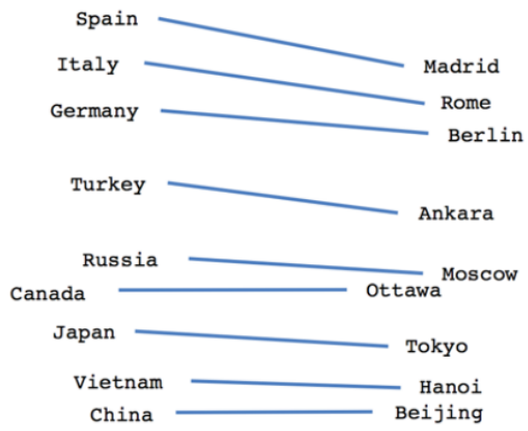
- Such embeddings have shown to be useful in tasks like machine translation and sentiment analysis



Male-Female



Verb tense



Country-Capital

Learning semantic representations of sentences from visually grounded language

+ apple

- Word embeddings can benefit from visual information [1]
- RQ: Can we also incorporate visual information into sentence representations?

Round
Red
Small

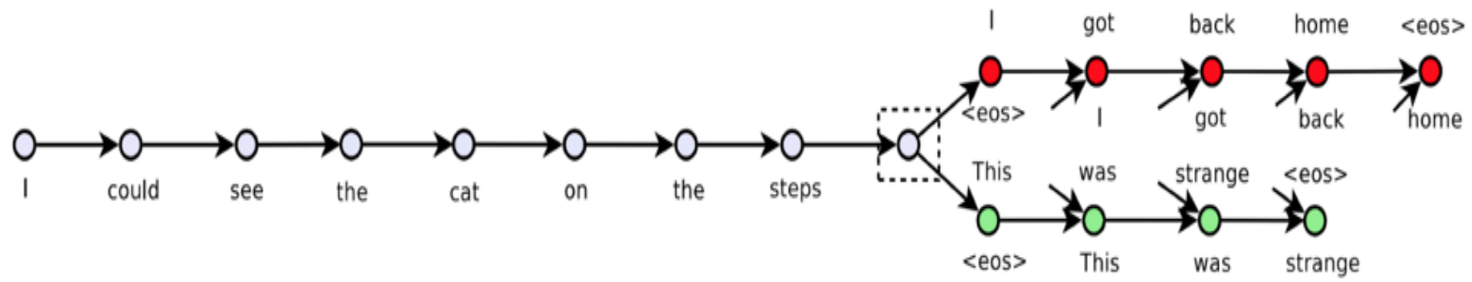
Visual information can be helpful because not many texts would explicitly mention an apple is round, red and small.

[1] Hasegawa, M. Kobayashi, T. & Hiyashi, Y. (2017).Incorporating visual features into word embeddings:
A bimodal autoencoder-based approach. *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Sentence Embeddings

- Surprisingly powerful yet simple method: Bag of Words (BOW)

- More advanced methods try to create learned sentence embeddings, e.g. given a sentence, skip-thought tries to predict the previous and next sentence [2]



[2] Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R. & Fidler S. (2015). Skip-Thought Vectors. arXiv: 1506.06726

Existing methods require pre-trained word embeddings (e.g. GloVe 840B [3])

- We aim to learn sentence representations without the use of word embeddings
- We use character level input instead
- But use image data to enrich the sentence representations
- RQ: can we learn sentence representations without prior assumptions about linguistic units?

Why would we not want to use word embeddings?

- There simply are no such things as word embeddings for speech
- It is implausible that we learn language by first reading billions of tokens before learning how to use words in a sentence

[3] Pennington, J., Socher, R. & Manning, C. (2014) "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

RQ 1 & 2 combined: Can we learn sentence representations from more natural input?

- Tomasello [4] argues that people learn many relatively fixed expressions (e.g., `how-are-you-doing', ‘don’t’) as single linguistic units
- Furthermore, he argues children’s linguistic units early in language acquisition are entire utterances, before their language use becomes more adult-like
- Children learn patterns such as `Where is X' and `Want more X', as they learn to identify ‘slots’ in utterances and their linguistic units become more adult like [5]
- We hope to find similar behaviour in our model, if we do not presuppose any linguistic units but let the system learn whether it is useful to use sub-word or supra-word level information

[4] Tomasello, M. 2000. First steps toward a usage-based theory of language acquisition. Cognitive Linguistics 11 (1/2), 61–82.

[5] Braine, M. D. S. and M. Bowerman 1976. Children’s first word combinations.Monographs of the Society for Research in Child Development 41 (1), 1–104.

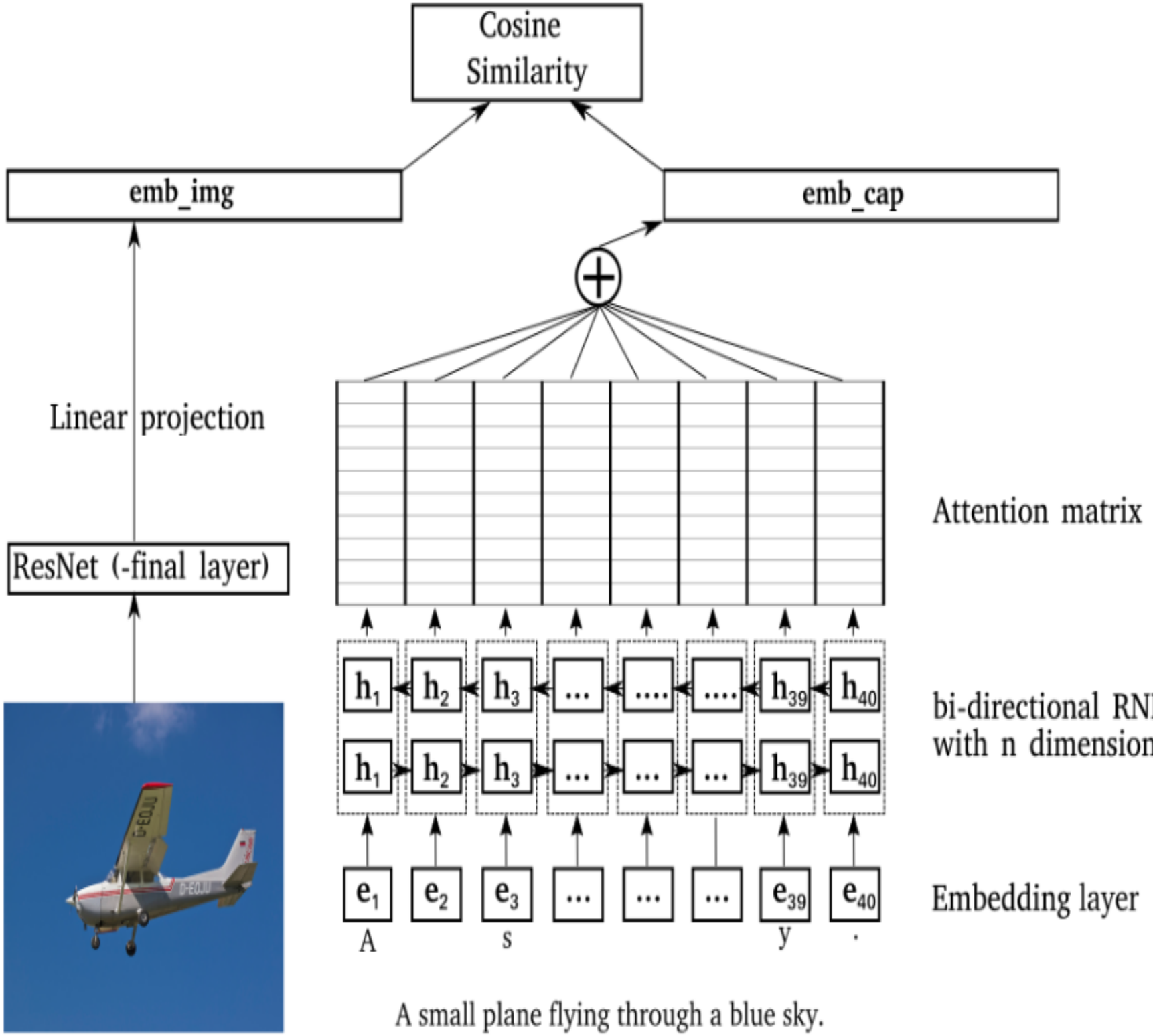
Our multi-modal sentence encoder

Training task: Image-Caption retrieval

- Given an image retrieve the correct caption and vice versa
- This works with text but has also been done with spoken captions
- The resulting embedding space has been called a semantic embedding space but this claim has thus far not been thoroughly investigated [6]
- Kiela et al. have shown that the resulting embeddings are useful in tasks such as sentiment analysis (product, movie reviews), opinion polarity and question-type classification. For those familiar with the toolbox, they tested all SentEval tasks except STS. [7] (n.b. Kiela et al. used GloVe embeddings)

[6] Harwath, D., Torralba, A., & Glass, J. (2016). Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems* (pp. 1858-1866).

[7] Kiela, D., A. Conneau, A. Jabri, and M. Nickel 2018. Learning visually grounded sentencerepresentations.arXiv preprint arXiv: 1809.0253.


Cosine Similarity
emb_img
emb_cap
Linear projection
ResNet (-final layer)
Attention matrix
bi-directional RNN layer with n dimensions
Embedding layer
A
s
y
.
A small plane flying through a blue sky.

Batch hinge loss function

$$L(\theta) = \sum_{(cap,img),(cap',img')\in B} \Big(\max(0, \cos(cap, img') - \cos(cap, img) + \alpha) + \max(0, \cos(img, cap') - \cos(img, cap) + \alpha)\Big)$$

where $(cap, img) \neq (cap', img')$.

We trained models on two datasets of images and corresponding captions

- Flickr8k (8k images, 5 captions per image)
- MSCOCO (123k images, 5 captions per image)

- During training we try to minimise the distance between the image embedding and the embeddings of its captions. We also mismatch image-caption pairs and try to maximise this distance

- a bike parked under a red metal object.
- a bicycle is under a large red structure.
- vintage train station with buildings, bike, and motorcycle.
- a person's bike sitting at a train station.
- a bike parked under a red structure near a building

Semantic Textual Similarity

We evaluate the semantic content on Semantic Textual Similarity (STS), a yearly SemEval task with over 12k sentence pairs and human annotated similarity judgments divided into 24 subsets

Dataset	Similarity	Example pair
SMTeuroparl	3.5	We often pontificate here about being the representatives of the citizens of Europe. We are proud often here to represent the citizens of Europe.
MSRpar	4.6	Myanmar's pro-democracy leader Aung San Suu Kyi will return home late Friday but will remain in detention after recovering from surgery at a Yangon hospital, her personal physician said. Myanmar's pro-democracy leader Aung San Suu Kyi will be kept under house arrest following her release from a hospital where she underwent surgery, her personal physician said Friday.
FNWN	2.0	An agent has attempted to achieve a goal, and the actual outcome of the agent's action has been resolved, so that it either specifically matches the agent's intent (e.g. success) or does not match it (e.g. failure). Having succeeded or being marked by a favorable outcome.
Question-Question	4.0	Should I drink water during my workout? How can I get my toddler to drink more water?

Training task performance

Flickr8k →

MSCOCO →

Model	Caption to Image				Image to Caption			
	R@1	R@5	R@10	med r	R@1	R@5	R@10	med r
Wehrmann et al. (2018)	26.9±2.7	-	69.6±2.9	4.0	32.4±2.9	-	73.6±2.7	3.0
Dong et al. (2018)	-	-	-	-	36.3±3.0	66.4±2.9	78.2±2.6	-
char-GRU	27.5±2.8	58.2±3.1	70.5±2.8	4.0	38.5±3.0	68.9±2.9	79.3±2.5	2.0
Faghri et al. (2017)	30.3±1.3	59.4±1.4	72.4±1.2	4.0	41.3±1.4	71.1±1.3	81.2±1.1	2.0
Kiela et al. (2018)	17.1±1.0	43.0±1.4	57.3±1.4	8.0	27.1±1.2	55.6±1.4	70.0±1.3	4.0
char-GRU	20.2±1.1	46.9±1.4	60.9±1.4	6.0	25.7±1.2	54.3±1.4	68.8±1.3	4.0

Semantic evaluation

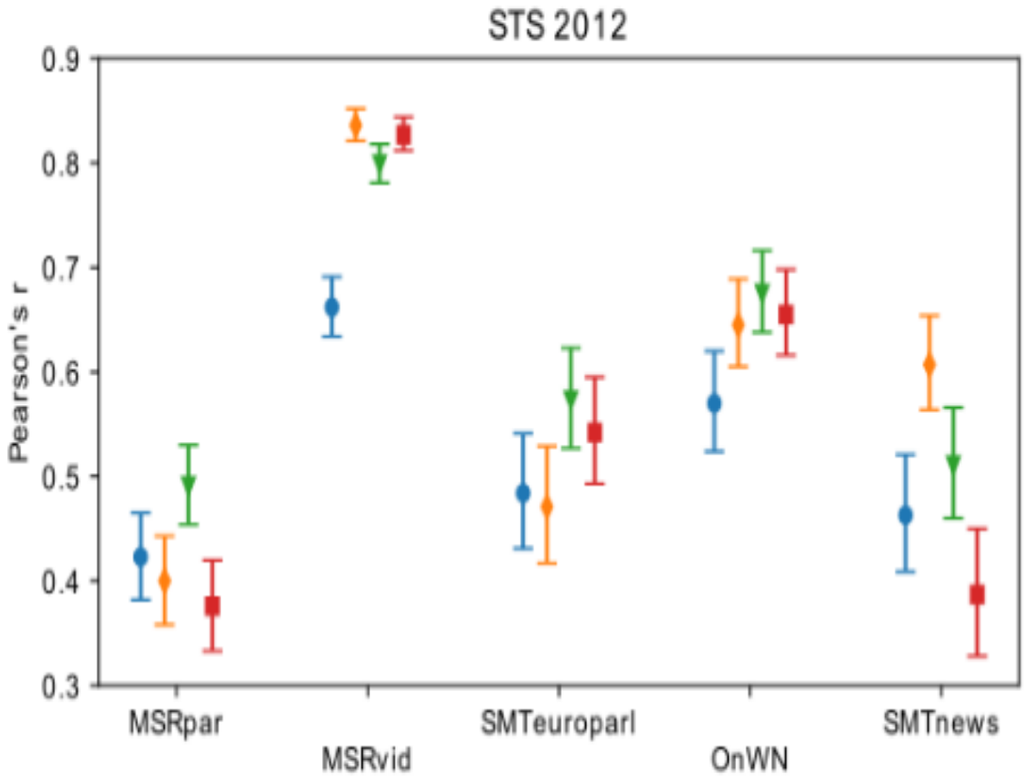
- We compare our model with InferSent [8]
    - trained on natural language inference (SNLI)
    - 500k+ sentence pairs
    - GloVe embeddings
    - But, text only

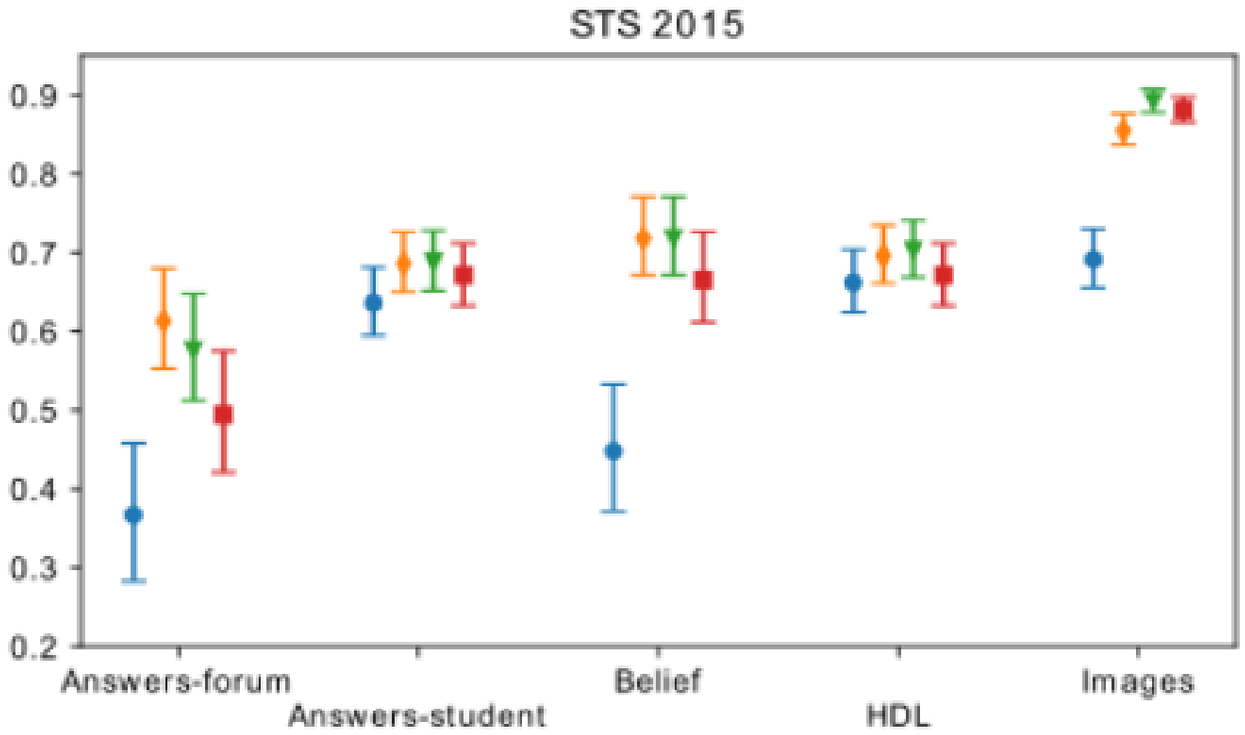
We also compare both models with a bag of words baseline

- How much semantic information can be gained simply from word level knowledge without knowing anything about sentences?
- This shows us what InferSent learns above and beyond what can be gained from the word embeddings

[8] Conneau, A., D. Kiela, H. Schwenk, L. Barrault, and A. Bordes 2017, September. Super-vised Learning of Universal Sentence Representations from Natural Language InferenceData. InProceedings of the 2017 Conference on Empirical Methods in Natural LanguageProcessing, pp. 670–680. Association for Computational Linguistics.

Semantic evaluation





Semantic evaluation: summary

- InferSent outperforms the BOW on 20 out of 24 subtasks
- Our model outperforms the BOW on 18 out of 24 subtasks
- Our model is on par with InferSent on 16 tasks, outperforms InferSent on 3 tasks and is outperformed on 5
- Bag of words model shows reasonable performance based on word level knowledge only
- InferSent learns above and beyond that word level knowledge
- The fact that our model performs on par with InferSent is remarkable given the considerable advantage gained with word embeddings.

Conclusions:

- Our system learns meaningful sentence representations comparable to state-of-the-art models
- Our model is able to extract meaning from multi-modal input*
- We did not need any prior lexical knowledge in the form of word embeddings

* To do: investigate if this is actually due to the visual grounding, or if the same information could be gained from just the Flickr8k/MSCOCO captions.

Spoken sentence embeddings using spoken caption-image retrieval

- We have shown that caption-image retrieval can be used to create meaningful sentence embeddings without word embeddings/prior semantic knowledge
- Our next challenge: Data
- We have the training data (Flickr8k only*) available in spoken format and previous research has shown that image-speech retrieval is possible (although expectedly, R@N takes a huge hit)[9]
- STS is not available in spoken format, 12K+ sentence pairs is to much to collect

* Thanks to Harwath and Glass [10]. Available at: https://groups.csail.mit.edu/sls/downloads/flickraudio/ (google Spoken Flickr8k)

[9] Chrupa la, G., L. Gelderloos, and A. Alishahi 2017, July. Representations of language in a model of visually grounded speech signal. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 613–622. Association for Computational Linguistics.

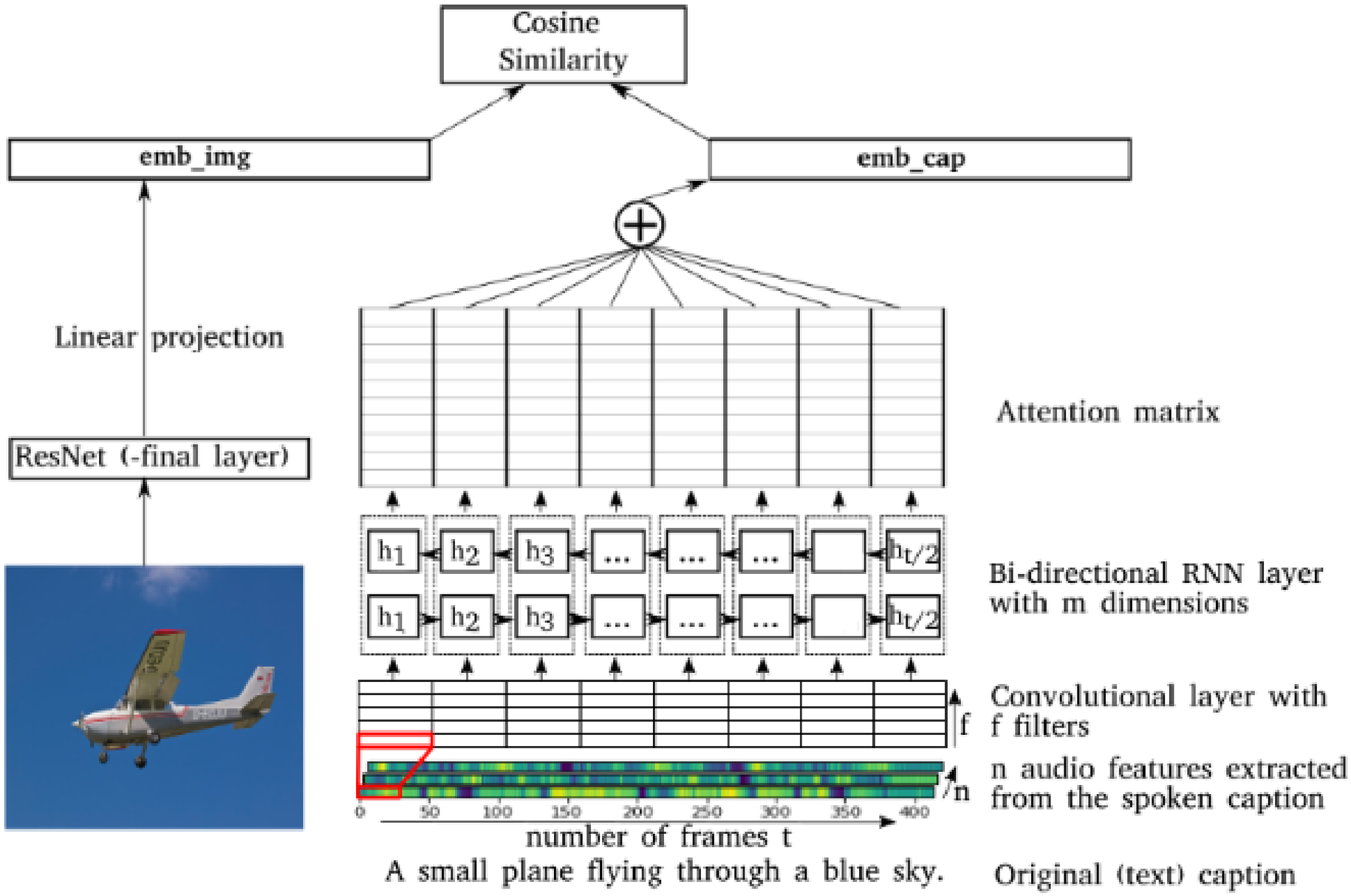
[10] D. Harwath and J. Glass, "Deep Multimodal Semantic Embeddings for Speech and Images," 2015 IEEE Automatic Speech Recognition and Understanding Workshop, pp. 237-244, Scottsdale, Arizona, USA, December 2015

Synthetic Speech

- Synthetic speech is our only viable option
- We use Google Wavenet Text to Speech
- 6 US accent voices, 3 male, 3 female
- Sounds (a lot) better than we expected (almost no warbling, correct intonation for questions, resolves things like St. → Saint, Aug. 10 → August tenth)

- Downsides:
  - probably a lot cleaner and less variable than real speech (i.e. the speech might just be ‘messy text’ to the model).
  - Mismatch between training and test data

We also plan to collect real spoken sentences for at least 1 (preferably 2) subsets of STS



We extract audio features from the speech signal: Mel Frequency Cepstral Coefficients (MFCCs) and Multilingual Bottle Neck features (MBNs)

A 1dimensional convolution is used to subsample the signal in the temporal dimension.

Preliminary results

Our training data was already available in spoken format so we have image-caption retrieval results:

Model	Caption to Image			
	R@1	R@5	R@10	med r
[2]	-	-	17.9±2.4	-
[8]	5.5±1.4	16.3±2.3	25.3±2.7	48
MFCC-GRU	8.0±1.7	24.5±2.7	35.5±3.0	24
MBN-GRU	12.4±2.0	32.4±2.9	44.4±3.1	15
Char-GRU	27.5±2.8	58.2±3.1	70.5±2.8	4
Model	Image to Caption			
	R@1	R@5	R@10	med r
[2]	-	-	24.3±2.7	-
MFCC-GRU	10.6±1.9	30.1±2.8	44.2±3.1	14
MBN-GRU	19.8±2.5	46.7±3.1	59.0±3.0	7
Char-GRU	38.5±3.0	68.9±2.9	79.3±2.5	2

All measures take a huge hit compared to the text model, but especially the multi-lingual bottleneck features perform remarkably well considering previous state-of-the-art results.

[2] Harwath, D. and J. Glass 2015. Deep multimodal semantic embeddings for speech andimages. In2015 IEEE Workshop on Automatic Speech Recognition and Understanding(ASRU), pp. 237–244. IEEE.

[8] Chrupa la, G., L. Gelderloos, and A. Alishahi 2017, July. Representations of language ina model of visually grounded speech signal. InProceedings of the 55th Annual Meetingof the Association for Computational Linguistics, pp. 613–622. Association for Computational Linguistics.

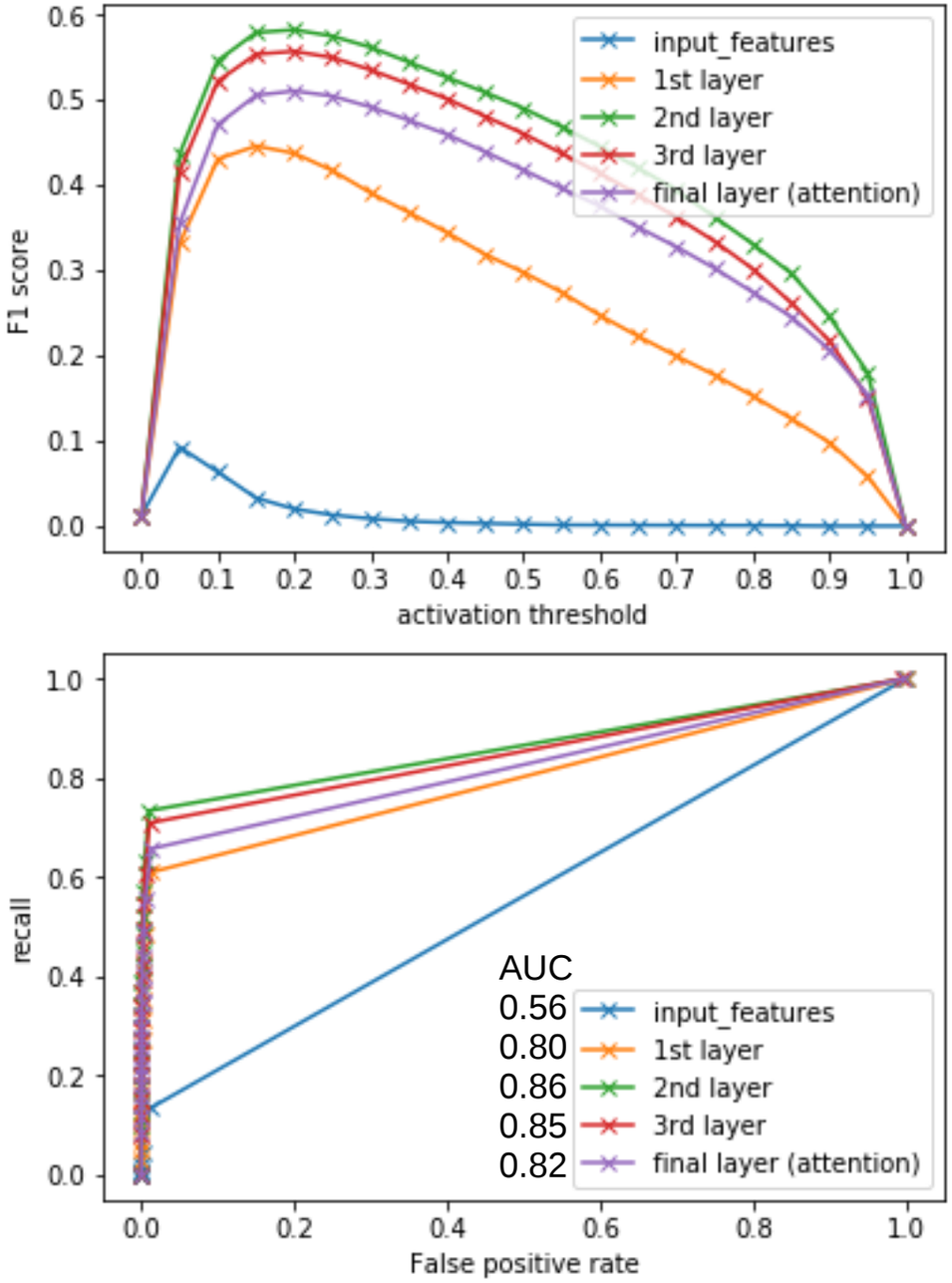
Preliminary results: does our model learn to recognise words?

- Using the written captions we select all words that occur between 50 and 1000 times (to ensure the model has seen it enough to have a chance at learning, but exclude extremely common words such as ‘the’)
- We end up with 460 words, mostly verbs and nouns.
- We train a classifier* on top of our audio caption embeddings, and on top of every intermediate output, which is trained to predict which of the 460 words are present in the caption.

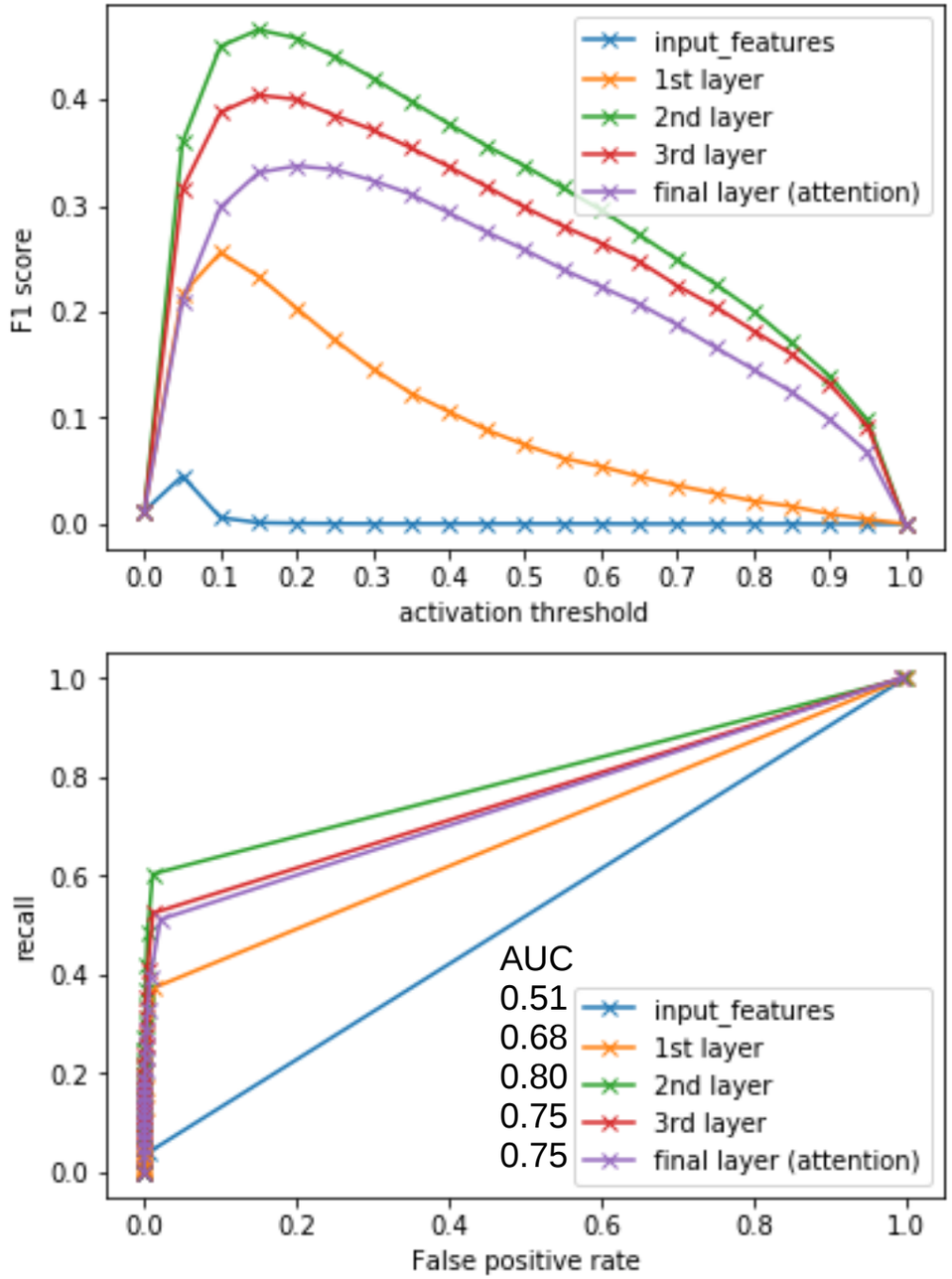
*A simple linear layer of size 460

Word presence detection: Results

MBN model



MFCC model



Future Work

Evaluate the spoken sentence semantics using our synthetic STS data.

We know our model learns to recognise words from the input but our ultimate goal was to see whether the model also learns linguistic units that do not correspond with word boundaries.

We also want to know how these units develop not just over layers but also during training.

Possible collaboration: Visualising the attention layer, and developing methods to investigate what the attention layer is learning.

Future Future work

Incorporate neurophysiological data. Language in Interaction has a huge fMRI database (24 hours of 1 person watching Dr. Who).

People have shown that neural networks learn sensible visual features if you try to predict the fMRI data directly from the video data. What does a model learn if we try to predict the brain response from the audio stream?