Probing Multimodal Embeddings for Linguistic Properties: the visual-semantic case

COLING 2020

Adam Dahlgren Lindström.
Suna Bensch, Johanna Björklund, Frank Drewes
Department of Computing Science
Umeå University

Ideas to discuss during the conference session

- How can we create inconsistencies in language in a visual context?
- How do we best employ probing techniques?
- How do we bridge the gap between visual context and language?
- What are the effects of focusing on grounding language versus regular language?

Introduction

We extend work on probing tasks to the multimodal domain with the following motivation:

- Semantic embeddings (e.g. `word2vec`) – a success story
- Difficult to interpret what models learn outside of metrics
- Increasing interest in
    - Multimodal machine learning
    - Interpret-/explainability
- Language alone is not enough to resolve semantic uncertainties

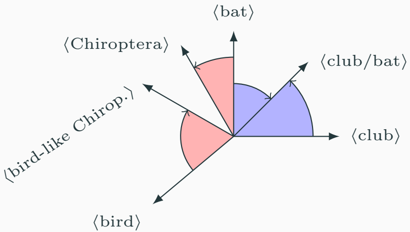
Our work

In this presentation we will see

- extension of unimodal probing tasks to a multimodal setting,
- concrete probing tasks for visual-semantic embeddings,
- how to gain valuable insights from probing (multimodal) embeddings,
- how language and vision clearly complement each other,
- why probing is a delicate process.

Background

- Probing to understand models (Conneau et al., 2018; Hewitt and Liang, 2019; Tenney et al., 2019)
- Multimodal machine learning a lively field (Baltrusaitis et al., 2019; Beinborn et al., 2018)
  - Multimodality adds another dimension of (un)interpretability.



A tiny bat is held by someone with a camera.

A man gently attempts to feed a baby bird.

A man in shorts is swinging a bat.

A man is swinging a club with both hands.

Figure 1: Image-caption pairs (left) and how vectors representing the words ‘bat’, ‘club’, and ‘bird’ may be affected by the image information (right). Source: MS-COCO dataset (Lin et al., 2014), license CC BY-NC-ND 2.0 and CC BY-NC 2.0, respectively. See also slides 5,7-8.

Probing semantic embeddings

A multimodal probing task

1. is a well-defined classification problem on combined (i.e., joint or coordinated) embeddings of two or more modalities,
2. gives insight into whether and how the multimodal embedding integrates the modalities,
3. has a simple and well-defined structure, so that the results are straightforward to interpret,
4. can be evaluated on standard data sets, or on datasets that can be created from such.

We distinguish between direct and inconsistency probes

Probing tasks 1+2: Direct Probing

Our first two tasks directly probe for information provided in the MS-COCO annotations

- Which MS-COCO object categories are in the image
- The number of objects seen in the image

In the example image below, we see 23 individual annotations, and the three categories person, cow, and umbrella.

Probing task 3: Creating Semantic Inconsistencies

We create inconsistencies in captions following these steps:

1. Pick head of caption using Stanford dependency parser
2. Pick most likely Wordnet synonym set using Lesk algorithm
3. Pick replacement word from a synset in the same Wordnet category
4. Inflect replacement word and mimic capitalization
5. Score 10 modified captions using BERT

This differs from e.g. FOIL-COCO (Shekhar et al., 2017) as the replacement words are not restricted to the MS-COCO categories.

1.1 A *child* holding a flowered umbrella and petting a yak.
1.2 A *checker* holding a flowered umbrella and petting a yak.

2.1 A young *man* holding an umbrella next to a herd of cattle.
2.2 A young *mime* holding an umbrella next to a herd of cattle.

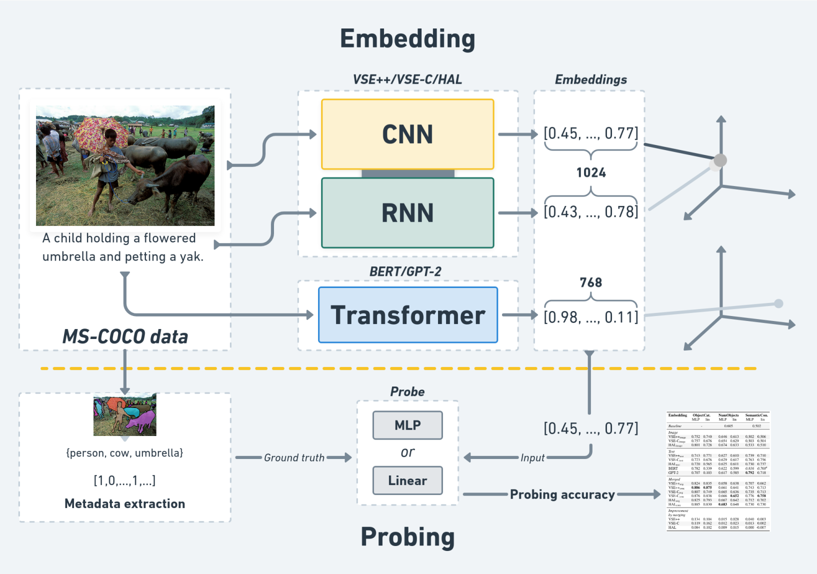
3.1 A young *boy* holding an umbrella touching the horn of a cow.
3.2 A young *wad* holding an umbrella touching the horn of a cow.

4.1 A young *boy* with an umbrella who is touching the horn of a cow.
4.2 A young *bear* with an umbrella who is touching the horn of a cow.

5.1 A *boy* holding an umbrella while standing next to livestock.
5.2 A *fry* holding an umbrella while standing next to livestock.

Figure 2: In task *SemanticCongruence*, the objective is to recognise semantically implausible captions.

Experiments



Results: Object Categories

- Merged embeddings has a significant lead
- HAL seems to rely more on the visual information
- Linear probe shows poor performance on BERT and GPT-2, see (Hewitt and Liang, 2019).
- 3.4%–11.9% improvement from merging

Embedding	ObjectCat.		NumObjects		SemanticCon.	
	MLP	lin	MLP	lin	MLP	lin
Baseline	-		0.605		0.502	
Image						
$VSE{+}{+}_{image}$	0.753	0.768	0.646	0.613	0.502	0.506
$VSE\text{-}C_{image}$	0.754	0.675	0.654	0.629	0.503	0.504
HAL_{image}	0.799	0.730	0.674	0.633	0.533	0.510
Text						
$VSE{+}{+}_{text}$	0.862	0.863	0.627	0.610	0.739	0.710
$VSE\text{-}C_{text}$	0.838	0.805	0.629	0.617	0.763	0.756
HAL_{text}	0.826	0.648	0.625	0.611	0.730	0.737
BERT	0.878	0.365	0.622	0.599	*0.816*	*0.768*[1]
GPT-2	0.811	0.137	0.617	0.585	**0.792**	0.718
Merged						
$VSE{+}{+}_{avg}$	0.862	0.876	0.658	0.638	0.707	0.662
$VSE{+}{+}_{conc}$	**0.911**	**0.901**	0.661	0.641	0.743	0.713
$VSE\text{-}C_{avg}$	0.831	0.783	0.665	0.636	0.735	0.713
$VSE\text{-}C_{conc}$	0.896	0.879	0.666	**0.652**	0.776	**0.758**
HAL_{avg}	0.847	0.820	0.667	0.642	0.712	0.702
HAL_{conc}	0.903	0.849	**0.683**	0.648	0.730	0.730
Improvement by merging						
VSE++	0.049	0.038	0.015	0.028	0.040	0.003
VSE-C	0.058	0.074	0.012	0.023	0.013	0.002
HAL	0.077	0.119	0.009	0.015	0.000	-0.007

Results: Number of objects

- Image embeddings encode the most information
- BERT and GPT-2 consistently outperformed
- $\leq 8\%$ improvement over always choosing largest class
- Few named objects vs. "crowd", or "many cars".
- Note: Most images contain $\leq$10 objects.

Embedding	ObjectCat.		**NumObjects**		SemanticCon.	
	MLP	lin	MLP	lin	MLP	lin
Baseline	-		0.605		0.502	
Image						
VSE++$_{image}$	0.753	0.768	0.646	0.613	0.502	0.506
VSE-C$_{image}$	0.754	0.675	0.654	0.629	0.503	0.504
HAL$_{image}$	0.799	0.730	0.674	0.633	0.533	0.510
Text						
VSE++$_{text}$	0.862	0.863	0.627	0.610	0.739	0.710
VSE-C$_{text}$	0.838	0.805	0.629	0.617	0.763	0.756
HAL$_{text}$	0.826	0.648	0.625	0.611	0.730	0.737
BERT	0.878	0.365	0.622	0.599	*0.816*	*0.768*[2]
GPT-2	0.811	0.137	0.617	0.585	**0.792**	0.718
Merged						
VSE++$_{avg}$	0.862	0.876	0.658	0.638	0.707	0.662
VSE++$_{conc}$	**0.911**	**0.901**	0.661	0.641	0.743	0.713
VSE-C$_{avg}$	0.831	0.783	0.665	0.636	0.735	0.713
VSE-C$_{conc}$	0.896	0.879	0.666	**0.652**	0.776	**0.758**
HAL$_{avg}$	0.847	0.820	0.667	0.642	0.712	0.702
HAL$_{conc}$	0.903	0.849	**0.683**	0.648	0.730	0.730
Improvement by merging						
VSE++	0.049	0.038	0.015	0.028	0.040	0.003
VSE-C	0.058	0.074	0.012	0.023	0.013	0.002
HAL	0.077	0.119	0.009	0.015	0.000	-0.007

Results: Semantic Congruence

- Visual information does not improve accuracy.
- Alternative captions can be identified soley from good language understanding.
    - Better generation would yield better task
- Significant difference in how multimodal embeddings represent language

Embedding	ObjectCat.		NumObjects		SemanticCon.	
	MLP	lin	MLP	lin	MLP	lin
Baseline	-		0.605		0.502	
Image						
VSE++$_{image}$	0.753	0.768	0.646	0.613	0.502	0.506
VSE-C$_{image}$	0.754	0.675	0.654	0.629	0.503	0.504
HAL$_{image}$	0.799	0.730	0.674	0.633	0.533	0.510
Text						
VSE++$_{text}$	0.862	0.863	0.627	0.610	0.739	0.710
VSE-C$_{text}$	0.838	0.805	0.629	0.617	0.763	0.756
HAL$_{text}$	0.826	0.648	0.625	0.611	0.730	0.737
BERT	0.878	0.365	0.622	0.599	*0.816*	*0.768*[3]
GPT-2	0.811	0.137	0.617	0.585	**0.792**	0.718
Merged						
VSE++$_{avg}$	0.862	0.876	0.658	0.638	0.707	0.662
VSE++$_{conc}$	**0.911**	**0.901**	0.661	0.641	0.743	0.713
VSE-C$_{avg}$	0.831	0.783	0.665	0.636	0.735	0.713
VSE-C$_{conc}$	0.896	0.879	0.666	**0.652**	0.776	**0.758**
HAL$_{avg}$	0.847	0.820	0.667	0.642	0.712	0.702
HAL$_{conc}$	0.903	0.849	**0.683**	0.648	0.730	0.730
Improvement by merging						
VSE++	0.049	0.038	0.015	0.028	0.040	0.003
VSE-C	0.058	0.074	0.012	0.023	0.013	0.002
HAL	0.077	0.119	0.009	0.015	0.000	-0.007

Discussion

We have shown that

- Visual and linguistic information complement each other
- Concatenated embeddings give best overall performance
    - Lack of language understanding compensated for by visual information
- Embeddings seem to have slightly different focus
- Linear probe results most likely more reliable
- Difficult to model `NumObjects` (`a dog and a tree` vs. `a crowd`)

Limitations and future work

- Limitations inherited from MS-COCO
  - Image annotations are flawed
  - The language used in grounding datasets differs from general NLP datasets (Tan and Bansal, 2020)
- Other probes
  - Per-class probing
  - Image manipulation
  - Introspective model probing, similar to (Tenney et al., 2019)
- Other datasets

Ideas to discuss during the conference session

- How can we create inconsistencies in language in a visual context?
- How do we best employ probing techniques?
- How do we bridge the gap between visual context and language?
- What are the effects of focusing on grounding language versus regular language?

Learn more

- https://github.com/dali-does/vse-probing

References

Baltrusaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Beinborn, L., Botschen, T., and Gurevych, I. (2018). Multimodal grounding for language processing. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 2325–2339. Association for Computational Linguistics.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*, pages 2126–2136.

Hewitt, J. and Liang, P. (2019). Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743. Association for Computational Linguistics.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Proceedings of the 13th European Conference on Computer Vision, ECCV 2014, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755.

Shekhar, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., and Bernardi, R. (2017). FOIL it! Find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pages 255–265.

Tan, H. and Bansal, M. (2020). Vokenization: Improving language understanding with contextualized, visual-grounded supervision.

Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. In *Association for Computational Linguistics*.