Recognizing Multimodal Entailment

Cesar Ilharco[γ] Afsaneh Shirazi[γ] Arjun Gopalan[ρ] Arsha Nagrani[ρ] Blaž Bratanič[γ]
Chris Bregler[ρ] Christina Liu[ρ] Felipe Ferreira[γ] Gabriek Barcik[ρ] Gabriel Ilharco[ω]
Georg Osang[α] Jannis Bulian[ρ] Jared Frank[γ] Lucas Smaira[δ] Qin Cao[ρ]
Ricardo Marino[γ] Roma Patel[β] Thomas Leung[ρ] Vaiva Imbrasaite[ρ]

[γ]Google [ρ]Google Research [β]Brown University [ω]University of Washington [α]IST Austria [δ]DeepMind

[γρδ]{ ilharco, afsaneh, arjung, anagrani, blazb, bregler, christinafunk, felipeg, gbarcik, jbulian, jaredfrank, lsmaira, qincao, ricm, leungt, vimbrasaite } @google.com
[ω]gamaga@cs.washington.edu, [α]georg.osang@ist.ac.at, [β]roma_patel1@brown.edu

Abstract

How information is created, shared and consumed has changed rapidly in recent decades, in part thanks to new social platforms and technologies on the web. With ever-larger amounts of unstructured and limited labels, organizing and reconciling information from different sources and modalities is a central challenge in machine learning.

This cutting-edge tutorial aims to introduce the multimodal entailment task, which can be useful for detecting semantic alignments when a single modality alone does not suffice for a whole content understanding. Starting with a brief overview of natural language processing, computer vision, structured data and neural graph learning, we lay the foundations for the multimodal sections to follow. We then discuss recent multimodal learning literature covering visual, audio and language streams, and explore case studies focusing on tasks which require fine-grained understanding of visual and linguistic semantics question answering, veracity and hatred classification. Finally, we introduce a new dataset for recognizing multimodal entailment, exploring it in a hands-on collaborative section.

Overall, this tutorial gives an overview of multimodal learning, introduces a multimodal entailment dataset, and encourages future research in the topic.

1 Website

multimodal-entailment.github.io

2 Type of the tutorial

Cutting edge.

3 Diversity considerations

- Instructors affiliated in 6 different countries.

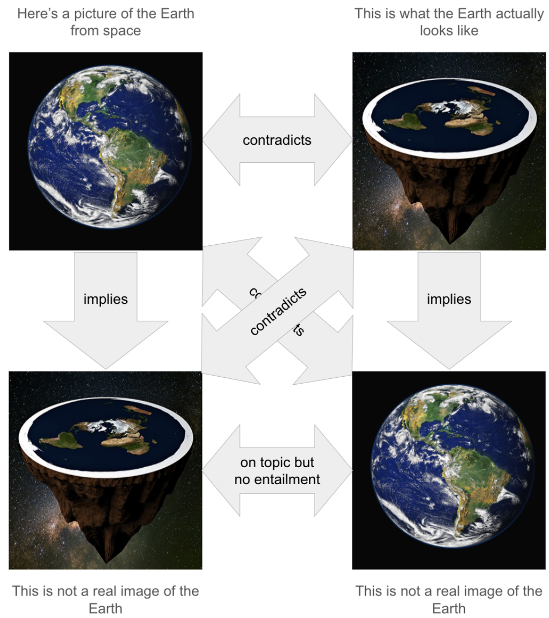

Figure 1: Example of multimodal entailment where texts or images alone would not suffice for semantic understanding or pairwise classifications.

- 3 academia and 3 industry affiliations.
- 6 female organizers.
- 5 female instructors.
- Participation of senior (up to Research Director) and junior (PhD candidate) instructors.
- Recognizing Multimodal Entailment can help with automated fact-checking, prompting for (re)focusing on traditionally underserved audiences (Scheufele and Krause, 2019).

4 Prerequisites

- Programming or other tools: Familiarity with Python and a high level machine learning framework.
- Machine Learning: Basic understanding of deep learning for Natural Language Processing and Computer Vision is desired, but not

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 29–30, August 1st, 2021.

critical for a successful completion of the tutorial.

5 Reading list

Bui et al. (2017); Vaswani et al. (2017); Peters et al. (2018); Devlin et al. (2018); Lan et al. (2019); Raffel et al. (2019); Ngiam et al. (2011); Lu et al. (2019a,b); Tan and Bansal (2019); Su et al. (2019); Sun et al. (2019b,a); Alayrac et al. (2020).

6 Tutorial presenters

Afsaneh Shirazi, Arjun Gopalan, Arsha Nagrani, Cesar Ilharco, Christina Liu, Gabriel Barcik, Jannis Bulian, Jared Frank, Lucas Smaira, Qin Cao, Ricardo Marino and Roma Patel.

7 Open access

We agree to allow the publication of slides and video recording of the tutorial in the ACL Anthology. Teaching materials will be openly available.

8 Acknowledgements

We would like to thank Abby Schantz, Abe Ittycheriah, Aliaksei Severyn, Allan Heydon, Aly Grealish, Andrey Vlasov, Arkaitz Zubiaga, Ashwin Kakarla, Chen Sun, Clayton Williams, Cong Yu, Cordelia Schmid, Da-Cheng Juan, Dan Finnie, Dani Valevski, Daniel Rocha, David Price, David Sklar, Devi Krishna, Elena Kochkina, Enrique Alfonseca, Françoise Beaufays, Isabelle Augenstein, Jialu Liu, John Cantwell, John Palowitch, Jordan Boyd-Graber, Lei Shi, Luís Valente, Maria Voitovich, Mehmet Aktuna, Mogan Brown, Mor Naaman, Natalia P, Nidhi Hebbar, Pete Aykroyd, Rahul Sukthankar, Richa Dixit, Steve Pucci, Tania Bedrax-Weiss, Tobias Kaufmann, Tom Boulos, Tu Tsao, Vladimir Chtchetkine, Yair Kurzion, Yifan Xu and Zach Hynes.

References

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks.

Thang D. Bui, Sujith Ravi, and Vivek Ramavajjala. 2017. Neural graph machines: Learning neural networks using graphs.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019a. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2019b. 12-in-1: Multi-task vision and language representation learning.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Dietram A. Scheufele and Nicole M. Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Learning video representations using contrastive bidirectional transformer.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. Videobert: A joint model for video and language representation learning.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.