Vision-Language Pretraining: Current Trends and the Future

https://vlp-tutorial-acl2022.github.io/

Aishwarya Agrawal
University of Montreal,
Mila, DeepMind
aishwarya.agrawal@mila.quebec

Damien Teney
Idiap Research Institute
contact@damienteney.info

Aida Nematzadeh
DeepMind
nematzadeh@deepmind.com

1 Description

In the last few years, there has been an increased interest in building multimodal (vision-language) models that are pretrained on larger but noisier datasets where the two modalities (*e.g.*, image and text) loosely correspond to each other (*e.g.*, Lu et al., 2019; Radford et al., 2021). Given a task (such as visual question answering), these models are then often fine-tuned on task-specific supervised datasets. (*e.g.*, Lu et al., 2019; Chen et al., 2020; Tan and Bansal, 2019; Li et al., 2020a,b). In addition to the larger pretraining datasets, the transformer architecture (Vaswani et al., 2017) and in particular self-attention applied to two modalities are responsible for the impressive performance of the recent pretrianed models on downstream tasks (Hendricks et al., 2021).

This approach is appealing for a few reasons: first, the pretraining datasets are often automatically curated from the Web, providing huge datasets with negligible collection costs. Second, we can train large models once, and reuse them for various tasks. Finally, these pretraining approach performs better or on par to previous task-specific models. An interesting question is whether these pretrained models – in addition to their good task performance – learn representations that are better at capturing the alignments between the two modalities.

In this tutorial, we focus on recent vision-language pretraining paradigms. Our goal is to first provide the background on image–language datasets, benchmarks, and modeling innovations before the multimodal pretraining area. Next we discuss the different family of models used for vision-language pretraining, highlighting their strengths and shortcomings. Finally, we discuss the limits of vision-language pretraining through statistical learning, and the need for alternative approaches such as causal modeling.

We believe that the computational linguistics (CL) community will benefit from this tutorial in multiple ways. Language grounding research often uses or evaluates the most successful vision-language approaches. Better understanding of the shortcomings and strengths of these approaches – which we hope our tutorial provides – will pave the way for building stronger language grounding agents. Moreover, vision-language pretraining has been inspired by its parallel in pretraining language models. As a result, the CL community has a special role in thinking about the future of vision-language approaches using lessons learned from language pretraining.

2 Type of the Tutorial

This is a cutting-edge tutorial focusing on discussing the new trends in vision-language pretraining: if recent models result in better representations and how they contribute to downstream tasks. We plan to mostly discuss recent papers from 2018 and after but will also include influential papers from before 2018 that have played a crucial role in the current vision-language paradigms.

3 Target Audience

We expect the target audience to be researchers interested in the intersection of vision and language, such as the language grounding or grounded communication researchers. This tutorial is also of interest for junior students who are starting their career. Familiarity with recent architectures such as transformers is a useful but not needed for attending the tutorial.

4 Outline of the Tutorial

- Introduction: the goal of the tutorial (5 minutes)

- Vision-language landscape before the pretraining era (55 minutes)

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 38 - 43
May 22-27, 2022 ©2022 Association for Computational Linguistics

  - Motivation for vision-language research from both application and research point of views.
  - Popular vision-language tasks, datasets and benchmarks (e.g., image-retrieval, referring expressions, image captioning, visual question answering).
  - Task specific modelling approaches and fundamental innovations before the pretraining era (e.g., CNN + LSTM based approaches, language guided image attention, multimodal pooling, compositional networks).

- Vision-language pretraining (VLP) (60 minutes)
  - Inspiration from pretraining successes in NLP (transformers, BERT, GPT).
  - Different families of VLP models (all are transformer based models):
    * Models using task-specific heads for each downstream task (e.g., ViLBERT, LXMERT, UNITER, OSCAR, VinVL).
    * Models treating all downstream tasks as language generation tasks, i.e. no task-specific head (e.g., VL-T5, VLBART, SimVLM).
    * Models using VLP data for improving performance on vision tasks (e.g., CLIP, ALIGN).
    * Models using VLP data for improving performance on language tasks, including multilingual data (e.g., Vokenization, M3P, VL-T5, SimVLM).
  - Different VLP datasets and how they affect the downstream task performance w.r.t their size, degree of noise, and similarity with downstream datasets.

- Beyond statistical learning in vision-language (55 minutes)
  - Challenges yet to be tackled in vision-language research that are inherent limitations of the mainstream machine learning approach. These challenges include shortcut learning, sensibility of distribution shifts, model biases, adversarial vulnerabilities, and generally poor out-of-distribution generalization. We will also briefly cover privacy and fairness concerns when collecting large scale datasets, and the problem of models amplifying biases.
  - Background on causal reasoning necessary to formalize these issues and introduce potential solutions.
  - Existing benchmarks and other possible evaluation procedures that go beyond the traditional i.i.d. setting and allow diagnosing these issues: contrast examples, pairs of counterfactual examples, out-of-distribution test sets, etc.
  - Methods for learning better models by exploiting expert knowledge / inductive biases (Cadène et al., 2019; Ramakrishnan et al., 2018) or by utilizing different training paradigms (e.g., across multiple environments (Arjovsky et al., 2019; Teney et al., 2020b) or from pairs of training examples (Gokhale et al., 2020; Teney et al., 2020a)).

- Conclusion: main takeaways and future research (5 minutes)

5 Breadth of the Tutorial

We will mainly cover other people's work (as outlined in §4 and §7). More specifically, we expect the tutorial to include less than 15% of instructors' work – speakers will spend at most 10 minutes presenting their prior work.

6 Diversity Considerations

We are planning to increase diversity in a few ways: First, the topic of the tutorial is multidisciplinary bringing together researchers from diverse backgrounds (such as language, vision, and representation learning). We also plan to discuss how vision-language pretraining can benefit multilingual applications through grounding multiple languages into vision. Second, the instructors are from diverse backgrounds including their career stage (mid-career / junior), geography, gender, as well as their institution (academia / industry). Third, we will share our reading list, slides, and the recording of the talk publicly for people who cannot attend the conference in person, and also as a resource for junior researchers who are starting their career.

7 Reading List

- Popular vision-language tasks, datasets and benchmarks (Plummer et al., 2015; Kazemzadeh et al., 2014; Mao et al., 2015; Chen et al., 2015; Antol et al., 2015; Krishna et al., 2016; Hudson and Manning, 2019).
- Task specific modelling approaches before the pretraining era (Antol et al., 2015; Yang et al., 2015; Lu et al., 2016; Anderson et al., 2017; Fukui et al., 2016; Andreas et al., 2015).
- *Pretraining models in NLP (Devlin et al., 2018; Brown et al., 2020).
- VLP models with task-specific heads (Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2020; Li et al., 2020b; Zhang et al., 2021).
- VLP models without task-specific heads (Cho et al., 2021; Wang et al., 2021).
- VLP models for improving performance on vision tasks (Radford et al., 2021; Jia et al., 2021).
- VLP models for improving performance on language tasks (Tan and Bansal, 2020; Huang et al., 2020; Cho et al., 2021; Wang et al., 2021).
- Analyzing VLP models (Hendricks et al., 2021; Frank et al., 2021; Hendricks and Nematzadeh, 2021; Bugliarello et al., 2020).
- Shortcomings of vision-language models (Agrawal et al., 2016; Rohrbach et al., 2018; Gan et al., 2020; Ross et al., 2020; van Miltenburg, 2016; Misra et al., 2015; Raji et al., 2020; Zhao et al., 2017a).
- Methods and evaluation benchmarks that go beyond the traditional i.i.d. setting (Agrawal et al., 2017; Cadène et al., 2019; Ramakrishnan et al., 2018; Teney et al., 2020c; Arjovsky et al., 2019; Teney et al., 2020b; Gokhale et al., 2020; Teney et al., 2020a; Ilse et al., 2020; Agarwal et al., 2019).

* It would be great if the audience could read these papers before the tutorial, but it is okay even if they do not get a chance, as we will briefly cover these topics in the tutorial.

8 Instructors

Aishwarya Agrawal [webpage: `https://www.iro.umontreal.ca/~agrawal`] is an Assistant Professor in the Department of Computer Science and Operations Research at the University of Montreal. She is also a Canada CIFAR AI Chair and a core academic member of Mila – Quebec AI Institute. She also spends one day a week at DeepMind as a Research Scientist. Aishwarya's research interests lie at the intersection of computer vision, deep learning and natural language processing. Aishwarya is one of the two lead authors on the VQA paper (Antol et al., 2015) that introduced the task and the VQA v1.0 dataset. She has played an active role in releasing the dataset to the public. She is, in particular, keen about building vision-language models that generalize to out-of-distribution datasets. She used to co-organize the annual VQA challenge and workshop, and has given numerous invited talks (see `https://www.iro.umontreal.ca/~agrawal/index.html#talks`).

Damien Teney [webpage: `https://www.damienteney.info`] is a research scientist heading the machine learning group at the Idiap Research Institute in Switzerland. He is known for his work at the intersection of computer vision, machine learning, and natural language processing. He was part of the team that won the Visual Question Answering Challenge at CVPR 2017, which introduced the bottom-up/top-down attention mechanisms that are now ubiquitous for vision and language. His current research focuses on out-of-distribution generalization and learning methods inspired by causal reasoning. He has given multiple introductory talks on these topics and is a regular invited speaker at workshops and seminars on vision and language (e.g., VQA workshop at CVPR 2021, Vision and Language workshop at ACCV 2018).

Aida Nematzadeh [webpage: `http://www.aidanematzadeh.me`] is a staff research scientist at DeepMind. Her research interests are in the intersection of computational linguistics, cognitive science, and machine learning. Her recent work has focused on multimodal learning and evaluation and analysis of neural representations. She co-instructed a tutorial on "Language Learning and Processing in People and Machines" at NAACL 2019, and has given numerous invited talks (see `http://aidanematzadeh.`

me/talks.html).

9 Ethics Statement

Vision-language systems have many potential applications beneficial for society:

- Aiding visually impaired users in understanding their surroundings (Human: `What is on the shelf above the microwave?` AI: `Canned containers.`),
- Teaching children through interactive demos (AI captioning a picture of Dall Sheep: `That is Dall Sheep. You can find those in Alaska.`),
- Aiding analysts in processing large quantities of visual surveillance data (Analyst: `What kind of car did the man in red shirt leave in?` AI: `Blue Toyota Prius.`),
- Interacting with in-home physical robots (Human: `Is my laptop in my bedroom upstairs?` AI: `Yes.` Human: `Is the charger plugged in?`),
- Making visual social media content more accessible (AI: `Your friend Bob just uploaded a picture from his Hawaii trip.` Human: `Great, is he at the beach?` AI: `No, on a mountain.`).

But like most other technology, such vision-language systems could also be used for potentially harmful applications such as:

- Invasion of individual's privacy by using vision-language systems to query streams of video data being recorded by CCTV cameras at public places.
- Visually impaired users often need assistance with parsing data containing personal information (Ahmed et al., 2015), such as credit cards, personal mails etc. Vision-language systems providing such assistance could be configured to leak / retain such personally identifiable information.

In addition to the above potentially harmful applications of vision-language systems, there exist ethical concerns around fairness and bias. The vision-language models, as other deep learning based models (Zhao et al., 2017b), could potentially amplify the biases present in the data they are trained on. Since the training data (images and language) captures stereotypical biases present in the society (e.g, the activity of cooking is more likely to be performed by a woman than a man), amplification of such stereotypes by vision-language systems is concerning as it has the potential to harm the users in the relevant groups (based on gender, race, religion etc.) by entrenching existing stereotypes and producing demeaning portrayals (Brown et al., 2020).

To raise awareness about such ethical concerns and to promote discussions among researchers, the last part of the tutorial ("Beyond statistical learning in vision-language") will focus on such shortcomings of existing models and we will discuss some methods that aim to tackle some of these challenges.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2019. Towards causal VQA: revealing and reducing spurious correlations by invariant and covariant semantic editing. *CoRR*, abs/1912.07538.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. *CoRR*, abs/1606.07356.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2017. Don't just assume; look and answer: Overcoming priors for visual question answering. *CoRR*, abs/1712.00377.

Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. 2015. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3523–3532.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, abs/1707.07998.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2015. Deep compositional question answering with neural module networks. *CoRR*, abs/1511.02799.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2020. Multimodal pretraining unmasked: Unifying the vision and language berts. *CoRR*, abs/2011.15124.

Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases in visual question answering. *CoRR*, abs/1906.10169.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. *CoRR*, abs/2102.02779.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *CoRR*, abs/1606.01847.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *CoRR*, abs/2006.06195.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. *arXiv preprint arXiv:2009.08566*.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *arXiv preprint arXiv:2102.00529*.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. *CoRR*, abs/2106.09141.

Haoyang Huang, Lin Su, Di Qi, Nan Duan, Edward Cui, Taroon Bharti, Lei Zhang, Lijuan Wang, Jianfeng Gao, Bei Liu, Jianlong Fu, Dongdong Zhang, Xin Liu, and Ming Zhou. 2020. M3P: learning universal representations via multitask multilingual multimodal pre-training. *CoRR*, abs/2006.02635.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: a new dataset for compositional question answering over real-world images. *CoRR*, abs/1902.09506.

Maximilian Ilse, Jakub M. Tomczak, and Patrick Forré. 2020. Selecting data augmentation for simulating interventions.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, abs/2102.05918.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pretraining. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *CoRR*, abs/1606.00061.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2015. Generation and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283.

Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross B. Girshick. 2015. Learning visual classifiers using human-centric annotations. *CoRR*, abs/1512.06974.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. *CoRR*, abs/2001.00964.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. *CoRR*, abs/1810.03649.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *CoRR*, abs/1809.02156.

Candace Ross, Boris Katz, and Andrei Barbu. 2020. Measuring social biases in grounded vision and language embeddings. *CoRR*, abs/2002.08911.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Empirical Methods in Natural Language Processing*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *CoRR*, abs/2010.06775.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020a. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020b. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*.

Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. 2020c. On the value of out-of-distribution testing: An example of goodhart's law. *CoRR*, abs/2005.09241.

Emiel van Miltenburg. 2016. Stereotyping and bias in the flickr30k dataset. *CoRR*, abs/1605.06083.

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.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2015. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CoRR*, abs/2101.00529.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017a. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *CoRR*, abs/1707.09457.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017b. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.