Connecting Language and Vision to Actions

ACL 2018 Tutorial

Peter Anderson[⋆]
Australian National University
panderson.me
peter.anderson@anu.edu.au

Abhishek Das[†]
Georgia Tech
abhishekdas.com
abhshkdz@gatech.edu

Qi Wu[⋆]
University of Adelaide
qi-wu.me
qi.wu01@adelaide.edu.au

[⋆]Australian Centre for Robotic Vision [†]Machine Learning and Perception Lab

Abstract

A long-term goal of AI research is to build intelligent agents that can *see* the rich visual environment around us, *communicate* this understanding in natural language to humans and other agents, and *act* in a physical or embodied environment. To this end, recent advances at the intersection of language and vision have made incredible progress – from being able to generate natural language descriptions of images/videos, to answering questions about them, to even holding free-form conversations about visual content! However, while these agents can passively describe images or answer (a sequence of) questions about them, they cannot act in the world (what if I cannot answer a question from my current view, or I am asked to move or manipulate something?). Thus, the challenge now is to extend this progress in language and vision to embodied agents that take actions and actively interact with their visual environments.

1 Tutorial Overview

This tutorial will provide an overview of the growing number of multimodal tasks and datasets that combine textual and visual understanding. We will comprehensively review existing state-of-the-art approaches to selected tasks such as image captioning (Chen et al., 2015), visual question answering (VQA) (Antol et al., 2015; Goyal et al., 2017) and visual dialog (Das et al., 2017a,b), presenting the key architectural building blocks (such as co-attention (Lu et al., 2016)) and novel algorithms (such as cooperative/adversarial games (Das et al., 2017b)) used to train models for these tasks. We will then discuss some of the current and upcoming challenges of combining language, vision and actions, and introduce some recently-released interactive 3D simulation environments designed for this purpose (Anderson et al., 2018b; Wu et al., 2018b; Das et al., 2018). The goal of this tutorial is to provide a comprehensive yet accessible overview of existing work and to reduce the entry barrier for new researchers.

In detail, we will first review the building blocks of the neural network architectures used for these tasks, starting from variants of recurrent sequence-to-sequence language models (Ilya Sutskever, 2014), applied to image captioning (Vinyals et al., 2015), optionally with visual attentional mechanisms (Bahdanau et al., 2015; Xu et al., 2015; You et al., 2016; Anderson et al., 2018a). We will then look at evaluation metrics for image captioning (Vedantam et al., 2015; Anderson et al., 2016), before reviewing how these metrics can be optimized directly using reinforcement learning (RL) (Williams, 1992; Rennie et al., 2017).

Next, on the topic of visual question answering, we will look at more sophisticated multimodal attention mechanisms, wherein the network simultaneously attends to visual and textual features (Fukui et al., 2016; Lu et al., 2016). We will see how to combine factual and commonsense reasoning from learnt memory representations (Sukhbaatar et al., 2015) and external knowledge bases (Wang et al., 2016; Wu et al., 2016), and approaches that use the question to dynamically compose the answering neural network from specialized modules (Andreas et al., 2016a,b; Johnson et al., 2017a,b; Hu et al., 2017).

Following the success of adversarial learning in visual recognition (Goodfellow et al., 2014), it has recently been gaining momentum in language modeling (Yu et al., 2016) and in multimodal tasks such as captioning (Dai et al., 2017) and dialog (Wu et al., 2018a). Within visual dia-

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-Tutorial Abstracts, pages 10–14
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

log, we will look at recent work that uses cooperative multi-agent tasks as a proxy for training effective visual conversational models via RL (Kottur et al., 2017; Das et al., 2017b).

Finally, as a move away from static datasets, we will cover recent work on building active RL environments for language-vision tasks. Although models that link vision, language and actions have a rich history (Tellex et al., 2011; Paul et al., 2016; Misra et al., 2017), we will focus primarily on embodied 3D environments (Anderson et al., 2018b; Wu et al., 2018b), considering tasks such as visual navigation from natural language instructions (Anderson et al., 2018b), and question answering (Das et al., 2018; Gordon et al., 2018). We will position this work in context of related simulators that also offer significant potential for grounded language learning (Beattie et al., 2016; Zhu et al., 2017). To finish, we will discuss some of the challenges in developing agents for these tasks, as they need to be able to combine active perception, language grounding, commonsense reasoning and appropriate long-term credit assignment to succeed.

2 Structure

The following structure is based on an approximately 3 hour timeframe with a break.

1. Introduction (20 min)
   (a) Language, vision and actions
   (b) Overview of relevant tasks and datasets
      i. Historical progression:
         see → communicate → act
2. Image Captioning (30 min)
   (a) Encoder-decoder for image captioning
   (b) Visual attention mechanisms
      i. Soft and hard visual attention
      ii. Semantic attention
      iii. Bottom-up and top-down attention
   (c) Evaluation
      i. CIDEr metric
      ii. SPICE metric
   (d) Reinforcement learning
      i. Policy gradient optimization
      ii. Self-critical sequence training
3. Visual Question Answering (VQA) (30 min)
   (a) Basic VQA architecture
   (b) Multimodal pooling
      i. Hierarchical co-attention
      ii. Compact bilinear pooling (MCB)
   (c) Dynamic network composition
      i. Neural module networks
      ii. Dynamic memory networks
   (d) Incorporating external knowledge
      i. FVQA
      ii. Ask me anything

——————— BREAK ———————

4. Visual Dialog (20 min)
   (a) Task, datasets and evaluation metrics
   (b) Architectures
      i. Hierarchical RNNs
   (c) Cooperative self-talk
   (d) Adversarial learning
5. Static datasets → Active environments (50 min)
   (a) Interactive 3D datasets and simulators
      i. DeepMind Lab
      ii. AI2-THOR
      iii. SUNCG (House3D / MINOS / HoME)
      iv. Matterport3D (Matterport3D Simulator / MINOS)
   (b) Embodied vision-and-language tasks
      i. Interactive Question Answering
      ii. Embodied Question Answering
      iii. Vision-and-Language Navigation
6. Future directions & conclusion (10 min)

3 Presenters

3.1 Peter Anderson

Peter Anderson is a final year PhD candidate in Computer Science at the Australian National University, supervised by Dr Stephen Gould, and a researcher within the Australian Centre for Robotic Vision (ACRV). His PhD focuses on deep learning for visual understanding in natural language. He was an integral member of the team that won first place in the 2017 Visual Question Answering (VQA) challenge at CVPR, and he has made several contributions in image captioning, including achieving first place on the COCO leaderboard in July 2017. He has published at CVPR, ECCV, EMNLP and ICRA, and spent time at numerous universities and research labs including Adelaide University, Macquarie University, and Microsoft Research. His research is currently focused on vision-and-language understanding in complex 3D environments.

3.2 Abhishek Das

Abhishek Das is a Computer Science PhD student at Georgia Institute of Technology, advised by Dhruv Batra, and working closely with Devi Parikh. He is interested in deep learning and its applications in building agents that can see (computer vision), think (reasoning and interpretability), talk (language modeling) and act (reinforcement learning). He is a recipient of an Adobe Research Fellowship and a Snap Research Fellowship. He has published at CVPR, ICCV, EMNLP, HCOMP and CVIU, co-organized the NIPS 2017 workshop on Visually-Grounded Interaction and Language, and has held visiting positions at Virginia Tech, Queensland Brain Institute and Facebook AI Research. He graduated from Indian Institute of Technology Roorkee in 2015 with a Bachelor's in Electrical Engineering.

3.3 Qi Wu

Dr. Qi Wu, is a research fellow in the Australia Centre for Robotic Vision (ACRV) in the University of Adelaide. Before that, he was a postdoc researcher in the Australia Centre for Visual Technologies (ACVT) in the University of Adelaide. He obtained his PhD degree in 2015 and MSc degree in 2011, in Computer Science from University of Bath, United Kingdom. His research interests are mainly in Computer Vision and Machine Learning. Currently, he is working on the vision-to-language problem and he is especially an expert in the area of Image Captioning and Visual Question Answering (VQA). His attributes-based image captioning model got first place on the COCO Image Captioning Challenge Leader Board in the October of 2015. He has published several papers in prestigious conferences and journals, such as TPAMI, CVPR, ICCV, ECCV, IJCAI and AAAI.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018a. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016a. Learning to Compose Neural Networks for Question Answering. In *NAACL-HLT*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016b. Neural Module Networks. In *CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. 2016. Deepmind lab. *arXiv preprint arXiv:1612.03801* .

Xinlei Chen, Tsung-Yi Lin Hao Fang, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325* .

Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. 2017. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied Question Answering. In *CVPR*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. In *CVPR*.

Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. In *ICCV*.

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. In *EMNLP*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. IQA: Visual question answering in interactive environments. In *CVPR*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*.

Oriol Vinyals Quoc V. Le Ilya Sutskever. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017a. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017b. Inferring and executing programs for visual reasoning. In *ICCV*.

Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. In *EMNLP*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*.

Dipendra K Misra, John Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. In *EMNLP*.

Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas M Howard. 2016. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. In *RSS*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end Memory Networks. In *NIPS*.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *CVPR*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2016. FVQA: Fact-based visual question answering. *TPAMI* .

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4):229–256.

Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2016. Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources. In *CVPR*.

Qi Wu, Peng Wang, Chunhua Shen, Ian D. Reid, and Anton van den Hengel. 2018a. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*.

Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. 2018b. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209* .

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2016. SeqGAN: Sequence generative adversarial nets with policy gradient. In *AAAI*.

Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*.