Learning to understand and manipulate – recent results in video understanding and robotics manipulation

Abstract: In this talk I will present recent results in video understanding and robotic manipulation. In video understanding self-supervised learning has become increasingly important to leverage the abundance of unlabeled data available on platforms like YouTube. Whereas most existing approaches learn low-level representations, we propose a joint visual-linguistic model to learn high-level features without any explicit supervision. In particular, inspired by its recent success in language modeling, we build upon the BERT model to learn bidirectional joint distributions over sequences of visual and linguistic tokens, derived from vector quantization of video data and off-the-shelf speech recognition outputs, respectively. We use VideoBERT in numerous tasks, including action classification and video captioning. We show that it can be applied directly to open-vocabulary classification, and confirm that large amounts of training data and cross-modal information are critical to performance. Furthermore, we outperform the state-of-the-art on video captioning, and quantitative results verify that the model learns high-level semantic features. Manipulation tasks such as preparing a meal or assembling furniture remain highly challenging for robotics and vision. Traditional task and motion planning methods can solve complex tasks but require full state observability and are not adapted to dynamic scene changes. Recent learning methods can operate directly on visual inputs but typically require many demonstrations and/or task-specific reward engineering. In this work we aim to overcome previous limitations and propose a reinforcement learning approach to task planning that learns to combine primitive skills. First, compared to previous learning methods, our approach requires neither intermediate rewards nor complete task demonstrations during training. Second, we demonstrate the versatility of our vision-based task planning in challenging settings with temporary occlusions and dynamic scene changes. Third, we propose an efficient training of basic skills from few synthetic demonstrations by exploring recent CNN architectures and data augmentation. Notably, while all of our policies are learned on visual inputs in simulated environments, we demonstrate the successful transfer and high success rates when applying such policies to manipulation tasks on a real UR5 robotic arm.

Bio: Cordelia Schmid holds a M.S. degree in Computer Science from the University of Karlsruhe and a Doctorate, also in Computer Science, from the Institut National Polytechnique de Grenoble (INPG). Her doctoral thesis received the best thesis award from INPG in 1996. Dr. Schmid was a post-doctoral research assistant in the Robotics Research Group of Oxford University in 1996–1997. Since 1997 she has held a permanent research position at Inria Grenoble Rhone-Alpes, where she is a research director. Dr. Schmid has been an Associate Editor for IEEE PAMI (2001–2005) and for IJCV (2004–2012), editor-in-chief for IJCV (2013–2018), a program chair of IEEE CVPR 2005 and ECCV 2012 as well as a general chair of IEEE CVPR 2015, ECCV 2020 and ICCV 2023. In 2006, 2014 and 2016, she was awarded the Longuet-Higgins prize for fundamental contributions in computer vision that have withstood the test of time. She is a fellow of IEEE. She was awarded an ERC advanced grant in 2013, the Humboldt research award in 2015 and the Inria & French Academy of Science Grand Prix in 2016. She was elected to the German National Academy of Sciences, Leopoldina, in 2017. In 2018 she received the Koenderink prize for fundamental contributions in computer vision that have withstood the test of time and in 2019 the Royal Society Milner award. Starting 2018 she holds a joint appointment with Google research.

Host: Martial Hebert
Point of Contact: Stephanie Matvey (smatvey@andrew.cmu.edu)