

# Joint Processing of Language and Visual Data for Better Automated Understanding

Edited by

Marie-Francine Moens<sup>1</sup>, Lucia Specia<sup>2</sup>, and Tinne Tuytelaars<sup>3</sup>

1 KU Leuven, BE, [sien.moens@cs.kuleuven.be](mailto:sien.moens@cs.kuleuven.be)

2 Imperial College London, GB, [lspecia@gmail.com](mailto:lspecia@gmail.com)

3 KU Leuven, BE, [tinne.tuytelaars@esat.kuleuven.be](mailto:tinne.tuytelaars@esat.kuleuven.be)

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19021 “Joint Processing of Language and Visual Data for Better Automated Understanding”. It includes a discussion of the motivation and overall organization, the abstracts of the talks, and a report of each working group.

**Seminar** January 6–11, 2019 – <http://www.dagstuhl.de/19021>

**2012 ACM Subject Classification** Computing methodologies → Natural language processing, Computing methodologies → Natural language generation, Computing methodologies → Computer vision tasks, Computing methodologies → Scene understanding, Computing methodologies → Image representations

**Keywords and phrases** multimodal learning, representation learning, computer vision, natural language processing, machine learning

**Digital Object Identifier** 10.4230/DagRep.9.1.1


**Edited in cooperation with** Guillem Collell

## 1 Executive Summary

*Marie-Francine Moens*

*Lucia Specia*

*Tinne Tuytelaars*

**License**  Creative Commons BY 3.0 Unported license  
© Marie-Francine Moens, Lucia Specia, and Tinne Tuytelaars

The joint processing of language and visual data has recently received a lot of attention. This emerging research field is stimulated by the active development of deep learning algorithms. For instance, deep neural networks (DNNs) offer numerous opportunities to learn mappings between the visual and language media and to learn multimodal representations of content. Furthermore, deep learning recently has become a standard approach for automated image and video captioning and for visual question answering, the former referring to the automated description of images or video with descriptions in natural language sentences, the latter to the automated formulation of an answer in natural language to a question in natural language about an image.

Apart from aiding image understanding and the indexing and search of image and video data through the natural language descriptions, the field of jointly processing language and visual data builds algorithms for grounded language processing where the meaning of



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license.  
Joint Processing of Language and Visual Data for Better Automated Understanding,  
*Dagstuhl Reports*, Vol. 9, Issue 1, pp. 1–27

Editors: Marie-Francine Moens, Lucia Specia, and Tinne Tuytelaars



Dagstuhl Reports  
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

natural language is based on perception and/or actions in the world. Grounded language processing contributes to automated language understanding and machine translation of language. Recently, it has been shown that visual data provide world and common-sense knowledge that is needed in automated language understanding.

Joint processing of language and visual data is also interesting from a theoretical point of view for developing theories on the complementarity of such data in human(-machine) communication, for developing suitable algorithms for learning statistical knowledge representations informed by visual and language data, and for inferencing with these representations.

Given the current trend and results of multimodal (language and vision) research, it can be safely assumed that the joint processing of language and visual data will only gain in importance in the future. During the seminar we have discussed theories, methodologies and real-world technologies for joint processing of language and vision, particularly in the following research areas:

- Theories of integrated modelling and representation learning of language and vision for computer vision and natural language processing tasks;
- Explainability and interpretability of the learned representations;
- Fusion and inference based on visual, language and multimodal representations;
- Understanding human language and visual content;
- Generation of language and visual content;
- Relation to human learning;
- Datasets and tasks.

The discussions have attempted to give an answer to the following research questions (a non-exhaustive list):

- Which machine learning architectures will be best suited for the above tasks?
- How to learn multimodal representations that are relational and structured in nature to allow a structured understanding?
- How to generalize to allow recognitions that have few or zero examples in training?
- How to learn from limited paired data but exploiting monomodal models trained on visual or language data?
- How to explain the neural networks when they are trained for image or language understanding?
- How to disentangle the representations: factorization to separate the different factors of variation and discovering of their meaning?
- How to learn continuous representations that describe semantics and that integrate world and common-sense knowledge?
- How to reason with the continuous representations?
- How to translate to another modality?
- What would be effective novel evaluation metrics?

This Dagstuhl Seminar has brought together an interdisciplinary group of researchers from computer vision, natural language processing, machine learning and artificial intelligence to discuss the latest scientific realizations and to develop a roadmap and research agenda.

## 2 Table of Contents

### Executive Summary

<i>Marie-Francine Moens, Lucia Specia, and Tinne Tuytelaars</i> . . . . .	1
---	---

### Overview Talks

Science in Computer Science	
<i>Andrei Barbu</i> . . . . .	5
Grounded Language Learning in Virtual Environments	
<i>Stephen Clark</i> . . . . .	5
RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes	
<i>Aykut Erdem</i> . . . . .	6
Language Based Image Manipulation	
<i>Erkut Erdem</i> . . . . .	6
Learning from Multilingual Multimodal Data	
<i>Desmond Elliott</i> . . . . .	7
Women also Snowboard: Overcoming Bias in Image Captioning	
<i>Lisa Anne Hendricks</i> . . . . .	7
Visual Context for Verb Sense Disambiguation	
<i>Frank Keller</i> . . . . .	8
Countering Language Drift through Grounding / Reinterpreting Wittgenstein	
<i>Douwe Kiela</i> . . . . .	8
Mining for Meaning: From Vision to Language through Multiple Networks Consensus	
<i>Marius Leordeanu</i> . . . . .	9
Describing Similarities and Differences in Related Videos	
<i>Florian Metze</i> . . . . .	9
Visual Dialogue without Vision or Dialogue	
<i>Siddharth Narayanaswamy</i> . . . . .	10
Conversational Mobile Robots: Integrating Vision, Language, and Planning	
<i>Jean Oh</i> . . . . .	10
Grounding Semantic Roles in Images	
<i>Carina Silberer</i> . . . . .	11
Pervasive Attention: 2D CNNs for Sequence-to-Sequence Prediction	
<i>Jakob Verbeek</i> . . . . .	11

### Working Groups

Representations	
<i>Tinne Tuytelaars, Vera Demberg, David Hogg, Dietrich Klakow, Chiraag Lala, Jindrich Libovický, Pranava Madhyastha, Siddharth Narayanaswamy, Bernt Schiele</i>	12
Visual and Language Understanding	
<i>Marie-Francine Moens, Luisa Coheur, Stephen Clark, Erkut Erdem, Anette Frank, Jean Oh</i> . . . . .	14

Challenges in Language Generation Applications	
<i>Lucia Specia, Loïc Barrault, Thales Bertaglia, Erkut Erdem, Lisa Anne Hendricks, Pecina Pavel, Florian Metze, Jean Oh . . . . .</i>	17
Modeling Human Learning	
<i>Raffaella Bernardi, Zeynep Akata, Andrei Barbu, Ozan Caglayan, Stephen Clark, Guillem Collell, Desmond Elliott, Raquel Fernandez, Orhan Firat, Stella Frank, Frank Keller, Douwe Kiela, Pecina Pavel, David Vernon, Josiah Wang . . . . .</i>	21
Explainability	
<i>Tinne Tuytelaars, Luisa Coheur, Vera Demberg, Lisa Anne Hendricks, Dietrich Klakow, Jindrich Libovický, Pranava Madhyastha, Marie-Francine Moens, Siddharth Narayanaswamy, Bernt Schiele . . . . .</i>	23
Tasks: Creating Simulated Worlds from Existing Media	
<i>David Hogg, Raffaella Bernardi, Desmond Elliott, Raquel Fernandez, Stella Frank, Marius Leordeanu, Jean Oh, Pavel Pecina, Lucia Specia, Jakob Verbeek, David Vernon . . . . .</i>	24
Tasks and Datasets for Vision and Language	
<i>Stephen Clark, Zeynep Akata, Andrei Barbu, Loïc Barrault, Raffaella Bernardi, Ozan Caglayan, Aykut Erdem, Erkut Erdem, Orhan Firat, Anette Frank, Stella Frank, David Hogg, Frank Keller, Douwe Kiela, Chiraag Lala, Marius Leordeanu, Florian Metze, Lucia Specia . . . . .</i>	25
<b>Participants . . . . .</b>	<b>27</b>

## 3 Overview Talks

### 3.1 Science in Computer Science

*Andrei Barbu (MIT – Cambridge, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Andrei Barbu

**Joint work of** Andrei Barbu, Yevgeni Berzak, David Mayo, Boris Katz

The methodology used to run experiments in computer science, for determining the performance of systems and humans, has been the same for several decades despite the huge advances in machine learning techniques. In particular, the ability of deep networks to effectively exploit unforeseen correlations in datasets means that many datasets are no longer good predictors of real-world performance. We demonstrate this with a new methodology to collect large-scale image datasets on Mechanical Turk while controlling for biases. De-biasing data is standard in other sciences and we believe computer science should follow: put simply, collecting images at random from some source does not guarantee those images are not highly biased. In another domain, we demonstrate how human performance has been significantly overstated for syntactic parsing problems and demonstrate that existing systems are overfitting to the particular biased methodology used to annotate parsing datasets. As datasets grow in importance in machine learning, and computer science in general, it is time that we adopt the lessons from other sciences: stringent controls, independently collecting test and training sets, and characterizing human performance at scale to create baselines for machines and to discover new biases.

### 3.2 Grounded Language Learning in Virtual Environments

*Stephen Clark (Google DeepMind – London, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Stephen Clark

**Joint work of** Stephen Clark, Felix Hill, Karl Moritz Hermann, Phil Blunsom

**Main reference** Felix Hill, Karl Moritz Hermann, Phil Blunsom, Stephen Clark: “Understanding Grounded Language Learning Agents”, CoRR, Vol. abs/1710.09867, 2017.

**URL** <https://arxiv.org/abs/1710.09867>

Images provide a means of grounding natural language expressions in another modality. However, there are limits to what images can provide in this regard, including a lack of state change and a lack of agent interaction. In this talk we describe work in grounding language in the actions of an embodied agent, where the environment of the agent is a simulated virtual world. We focus on some limitations of the current work, and discuss how far such an approach could take us in the goal of developing intelligent agents in both a virtual and the real world.

### 3.3 RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes

*Aykut Erdem (Hacettepe University – Ankara, TR)*

**License** © Creative Commons BY 3.0 Unported license  
© Aykut Erdem

**Joint work of** Semih Yagcioglu, Aykut Erdem, Ekut Erdem, Nazli Ikizler-Cinbis  
**Main reference** Semih Yagcioglu, Aykut Erdem, Ekut Erdem, Nazli Ikizler-Cinbis: “RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes”, In Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, Brussels, Belgium, Oct 31-Nov 4, 2018, pp. 1358–1368, ACL, 2018.  
**URL** <https://aclweb.org/anthology/D18-1166>

Understanding and reasoning about cooking recipes is a fruitful research direction towards enabling machines to interpret procedural text. In this talk, we present RecipeQA, a dataset for multimodal comprehension of cooking recipes. It comprises of approximately 20K instructional recipes with multiple modalities such as titles, descriptions and aligned set of images. With over 36K automatically generated question-answer pairs, we design a set of comprehension and reasoning tasks that require joint understanding of images and text, capturing the temporal flow of events and making sense of procedural knowledge. Our preliminary results indicate that RecipeQA serves as a challenging test bed and an ideal benchmark for evaluating machine comprehension systems. The data and leaderboard are available at <https://hucvl.github.io/recipeqa/>

### 3.4 Language Based Image Manipulation

*Erkut Erdem (Hacettepe University – Ankara, TR)*

**License** © Creative Commons BY 3.0 Unported license  
© Erkut Erdem

**Joint work of** Levent Karacan, Zeynep Akata, Aykut Erdem, Erkut Erdem  
**Main reference** Levent Karacan, Zeynep Akata, Aykut Erdem, Erkut Erdem: “Manipulating Attributes of Natural Scenes via Hallucination”, CoRR, Vol. abs/1808.07413, 2018.  
**URL** <http://arxiv.org/abs/1808.07413>

Recently, much progress has been made towards realistic image synthesis. In particularly, different flavors and improved versions of Generative Adversarial Networks (GANs) have achieved impressive results along this direction. Using GANs as a backbone, we present our efforts on language based image manipulation. In our first study, we explore building a two-stage framework for enabling users to directly manipulate high-level attributes of a natural scene. The key to our approach is a deep generative network which can hallucinate images of a scene as if they were taken at a different season (e.g., during winter), weather condition (e.g., in a cloudy day) or time of the day (e.g., at sunset). In our second work, we present a novel approach for language conditioned editing of fashion images. Our approach employs a GAN-based architecture which allows the users to edit an outfit image by feeding in different descriptions to generate new outfits.

### 3.5 Learning from Multilingual Multimodal Data

*Desmond Elliott (University of Copenhagen – DK)*

**License** © Creative Commons BY 3.0 Unported license  
 © Desmond Elliott  
**Joint work of** Elliott Desmond, Ákos Kádár  
**Main reference** Desmond Elliott, Ákos Kádár: “Imagination Improves Multimodal Translation.” In Proc. of the Eighth International Joint Conference on Natural Language Processing, EMNLP 2017, Taipei, Taiwan, November 27-December 1, 2017, (Volume 1: Long Papers) Vol. 1, pp. 130–141, ACL 2017.  
**URL** <https://aclweb.org/anthology/I17-1014>

We speak about two perspectives on learning from multilingual and multimodal data. In the language generation setting of multimodal machine translation, we discuss whether we should use visual representations as an input variable, or as a variable that the model learns to predict. In the image–sentence retrieval setting, we discuss experiments on when it is useful to train with multilingual annotations, as opposed to monolingual annotations.

### 3.6 Women also Snowboard: Overcoming Bias in Image Captioning


*Lisa Anne Hendricks (University of California – Berkeley, US)*

**License** © Creative Commons BY 3.0 Unported license  
 © Lisa Anne Hendricks  
**Joint work of** Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, Anna Rohrbach  
**Main reference** Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, Anna Rohrbach: “Women Also Snowboard: Overcoming Bias in Captioning Models”, in Proc. of the Computer Vision – ECCV 2018 – 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III, Lecture Notes in Computer Science, Vol. 11207, pp. 793–811, Springer, 2018.  
**URL** [https://doi.org/10.1007/978-3-030-01219-9\\_47](https://doi.org/10.1007/978-3-030-01219-9_47)

Most machine learning methods are known to capture and exploit biases of the training data. While some biases are beneficial for learning, others are harmful. Specifically, image captioning models tend to exaggerate biases present in training data (e.g., if a word is present in 60% of training sentences, it might be predicted in 70% of sentences at test time). This can lead to incorrect captions in domains where unbiased captions are desired, or required, due to over-reliance on the learned prior and image context. In this work we investigate generation of gender-specific caption words (e.g., man, woman) based on the person’s appearance or the image context. We introduce a new Equalizer model that ensures equal gender probability when gender evidence is occluded in a scene and confident predictions when gender evidence is present. The resulting model is forced to look at a person rather than use contextual cues to make a gender-specific predictions. The losses that comprise our model, the Appearance Confusion Loss and the Confident Loss, are general, and can be added to any description model in order to mitigate impacts of unwanted bias in a description dataset. Our proposed model has lower error than prior work when describing images with people and mentioning their gender and more closely matches the ground truth ratio of sentences including women to sentences including men. We also show that unlike other approaches, our model is indeed more often looking at people when predicting their gender.

### 3.7 Visual Context for Verb Sense Disambiguation

*Frank Keller (University of Edinburgh, GB)*

**License**  Creative Commons BY 3.0 Unported license  
© Frank Keller

**Joint work of** Frank Keller, Gella Spandana, Elliott Desmond


**Main reference** Spandana Gella, Frank Keller, Mirella Lapata: “Disambiguating Visual Verbs”, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 41(2), pp. 311–322, 2019.

**URL** <https://doi.org/10.1109/TPAMI.2017.2786699>

Every day millions of images are uploaded on the web. To process images at such a large scale it is important to build automatic image understanding systems. A step towards understanding the content of an image is to be able to recognize actions (or verbs) depicted in the image. This type of image understanding can then be integrated with natural language processing to build systems that interact with humans for tasks such as image retrieval. In this talk, we present models for integrating visual and textual contexts for: (i) Verb Classification: automatically identifying verbs that denote actions depicted in images; (ii) Visual Sense Disambiguation: fine-grained analysis of how visual context can help disambiguate different meanings of verbs; (iii) Multilingual Sense Disambiguation: using visual sense disambiguation across languages to benefit tasks such as machine translation.

### 3.8 Countering Language Drift through Grounding / Reinterpreting Wittgenstein

*Douwe Kiela (Facebook – New York, US)*

**License**  Creative Commons BY 3.0 Unported license  
© Douwe Kiela

The distributional hypothesis (Harris, 1952; Firth, 1957) has come to be one of the corner stones of modern natural language processing (NLP). It is the foundation upon which word embeddings are built, for instance. Oftentimes, the distribution hypothesis is seen as an incarnation of Wittgenstein’s famous “meaning is use” paradigm. In this talk, we argue that that conception of Wittgenstein as “the godfather of the distributional hypothesis” is misguided, and that Wittgenstein has rather different lessons to teach us.

The distributional hypothesis has two big issues. First, it forces us to define symbols only in terms of other symbols, which leads to the well-studied grounding problem. Second, and this is less well-studied, it presumes a passive observational stance towards language, where we just observe “the company that words keep”. We argue that Wittgenstein’s actual theory of meaning and language does not suffer from either of these issues, and that we should take this account of language more seriously. The resultant interpretation of Wittgenstein leads to a new research program for true natural language understanding, centering around active language usage in “multi-agent grounded language games”. We finish with some quick examples of research we have done at FAIR that goes in that direction.

### 3.9 Mining for Meaning: From Vision to Language through Multiple Networks Consensus

*Marius Leordeanu (University Politehnica of Bucharest, RO)*

**License** © Creative Commons BY 3.0 Unported license

© Marius Leordeanu

**Joint work of** Leordeanu, Marius; Iulia, Duta; Andrei Liviu, Nicolicioiu, Simion-Vlad, Bogolin

**Main reference** Iulia Duta, Andrei Liviu Nicolicioiu, Simion-Vlad Bogolin, Marius Leordeanu: “Mining for meaning: from vision to language through multiple networks consensus”, In Proc. of the 29th British Machine Vision Conference, BMVC 2018, Newcastle, GB, Sep 3-6, 2018.

**URL** <http://bmvc2018.org/contents/papers/1031.pdf>

Describing visual data into natural language is a very challenging task, at the intersection of computer vision, natural language processing and machine learning. Language goes well beyond the description of physical objects and their interactions and can convey the same abstract idea in many ways. It is both about content at the highest semantic level as well as about fluent form. Here we propose an approach to describe videos in natural language by reaching a consensus among multiple encoder-decoder networks. Finding such a consensual linguistic description, which shares common properties with a larger group, has a better chance to convey the correct meaning. We propose and train several network architectures and use different types of image, audio and video features. Each model produces its own description of the input video and the best one is chosen through an efficient, two-phase consensus process. We demonstrate the strength of our approach by obtaining state of the art results on the challenging MSR-VTT dataset.

### 3.10 Describing Similarities and Differences in Related Videos

*Florian Metze (Carnegie Mellon University – Pittsburgh, US)*

**License** © Creative Commons BY 3.0 Unported license

© Florian Metze

**Joint work of** Jindrich Libovický, Shruti Palaskar, Spandana Gella, Florian Metze

**Main reference** Jindrich Libovický, Shruti Palaskar, Spandana Gella, Florian Metze: “Multimodal abstractive summarization of opendomain videos”, In NeurIPS Workshop on Visually Grounded Interaction and Language (ViGIL), 32nd Conference on Neural Information Processing Systems, NIPS 2018, Montréal, Canada. Dec 3-8, NIPS, 2018.

**URL** <https://nips2018vigil.github.io/static/papers/accepted/8.pdf>

Multimodal and abstractive summarization of open-domain videos requires summarizing the contents of an entire video in a few short sentences, while fusing information from multiple modalities, in our case video and audio (or text). Different from traditional news summarization, the goal is less to “compress” text information only, but to provide a fluent textual summary of information that has been collected and fused from different source modalities. In this talk, we introduce the task of abstractive summarization for open-domain videos, we show how a sequence-to-sequence model with hierarchical attention can integrate information from different modalities into a coherent output, and present pilot experiments on the How2 corpus of instructional videos. We also present a new evaluation metric for this task called Content F1 that measures semantic adequacy rather than fluency of the summaries, which is covered by ROUGE and BLEU like metrics.

### 3.11 Visual Dialogue without Vision or Dialogue

*Siddharth Narayanaswamy (University of Oxford, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Siddharth Narayanaswamy

**Joint work of** Siddharth Narayanaswamy, Daniela Massiceti, Puneet Dokania, Philip Torr

**Main reference** Daniela Massiceti, Puneet K. Dokania, N. Siddharth, Philip H. S. Torr: “Visual Dialogue without Vision or Dialogue”, CoRR, Vol. abs/1812.06417, 2018.

**URL** <https://arxiv.org/abs/1812.06417>

We characterize some of the quirks and shortcomings in the exploration of Visual Dialogue (VD) – a sequential question-answering task where the questions and corresponding answers are related through given visual stimuli. Using an embarrassingly simple method based on Canonical Correlation Analysis (CCA) on the standard dataset gets near state-of-the-art performance for some standard metric. In contrast to current complex and over-parametrized architectures that are both compute and time intensive, this method ignores the visual stimuli, ignores the sequencing of dialogue, does not need gradients, uses off-the-shelf feature extractors, has at least an order of magnitude fewer parameters, and learns in practically no time. These results are indicative of issues in current approaches to Visual Dialogue relating particularly to implicit dataset biases, under-constrained task objectives, and over-constrained evaluation metrics.

### 3.12 Conversational Mobile Robots: Integrating Vision, Language, and Planning

*Jean Oh (Carnegie Mellon University – Pittsburgh, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Jean Oh

**Joint work of** Jean H. Oh, Arne Suppé, Felix Duvallet, Abdeslam Boularias, Luis E. Navarro-Serment, Martial Hebert, Anthony Stentz, Jerry Vinokurov, Oscar J. Romero, Christian Lebiere, Robert Dean

**Main reference** Jean H. Oh, Arne Suppé, Felix Duvallet, Abdeslam Boularias, Luis E. Navarro-Serment, Martial Hebert, Anthony Stentz, Jerry Vinokurov, Oscar J. Romero, Christian Lebiere, Robert Dean: “Toward Mobile Robots Reasoning Like Humans”, in Proc. of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA., pp. 1371–1379, AAAI Press, 2015.

**URL** <http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9766>

As robots are envisioned to become ubiquitous in our personal and work environments, there have been growing interests in developing intuitive ways for lay users to interact with such autonomous systems, raising several challenging questions. For instance, can we command robots in natural language? Can robots describe what they observe or explain what they have done or plan to do? How can we train our robots to understand rich semantic context, utilizing a vast amount of sensor data that is available in multiple modalities? In this talk, we discuss various hurdles in addressing these challenges in several robotics application domains including social navigation, autonomous driving, disaster response, and military robots. We will also discuss general limitations of datasets and evaluation metrics in interdisciplinary research and propose alternative directions.

### 3.13 Grounding Semantic Roles in Images

*Carina Silberer (UPF – Barcelona, ES)*

**License** © Creative Commons BY 3.0 Unported license  
© Carina Silberer

**Joint work of** Carina Silberer, Manfred Pinkal

**Main reference** Carina Silberer, Manfred Pinkal: “Grounding Semantic Roles in Images”, in Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 – November 4, 2018, pp. 2616–2626, Association for Computational Linguistics, 2018.

**URL** <https://aclanthology.info/papers/D18-1282/d18-1282>

Images of everyday scenes can be interpreted and described in many ways, depending on the perceiver and the context in which the image is presented, where the context may be natural language data or a visual sequence. The interpretation of a (visual) scene is related to the determination of who did what to whom, etc. This may require a joint processing or reasoning with possibly multiple (extra-)linguistic information sources (e.g., text, images).

To facilitate the joint processing over multiple sources, it is desirable to induce representations of texts and visual scenes which do encode this kind of information, and in, essentially, a congruent and generic way. In this talk we present our approach towards this goal: We address the task of visual semantic role labeling (vSRL), and learn frame-semantic representations of images. Our model renders candidate participants as image regions of objects, and is trained towards grounding roles in the regions which depict the corresponding participant. We present experimental results which demonstrate that we can train a vSRL model without reliance on prohibitive image-based role annotations, by utilizing noisy data which we extract automatically from image captions using a linguistic SRL system. Furthermore, the frame-semantic visual representations which our model induces yield overall better results on supervised visual verb sense disambiguation compared to previous work.

### 3.14 Pervasive Attention: 2D CNNs for Sequence-to-Sequence Prediction

*Jakob Verbeek (INRIA – Grenoble, FR)*

**License** © Creative Commons BY 3.0 Unported license  
© Jakob Verbeek

**Joint work of** Jakob Verbeek, Maha Elbayad, Laurent Besacier

**Main reference** Maha Elbayad, Laurent Besacier, Jakob Verbeek: “Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction”, in Proc. of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 – November 1, 2018, pp. 97-107, Association for Computational Linguistics, 2018.

**URL** <https://aclweb.org/anthology/K18-1010>

Current state-of-the-art machine translation systems are based on encoder-decoder architectures, that first encode the input sequence, and then generate an output sequence based on the input encoding. Both are interfaced with an attention mechanism that recombines a fixed encoding of the source tokens based on the decoder state. We propose an alternative approach which instead relies on a single 2D convolutional neural network across both sequences. Each layer of our network re-codes source tokens on the basis of the output sequence produced so far. Attention-like properties are therefore pervasive throughout the network. Our model yields excellent results, outperforming state-of-the-art encoder-decoder systems, while being conceptually simpler and having fewer parameters.

## 4 Working Groups

### 4.1 Representations

*Tinne Tuytelaars (KU Leuven, BE), Vera Demberg (Universität des Saarlandes, DE), David Hogg (University of Leeds, GB), Dietrich Klakow (Universität des Saarlandes, DE), Chiraag Lala (University of Sheffield, GB), Jindrich Libovický (Charles University – Prague, CZ), Pranava Madhyastha (Imperial College London – GB), Siddharth Narayanaswamy (University of Oxford, GB), Bernt Schiele (MPI für Informatik – Saarbrücken, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Tinne Tuytelaars, Vera Demberg, David Hogg, Dietrich Klakow, Chiraag Lala, Jindrich Libovický, Pranava Madhyastha, Siddharth Narayanaswamy, Bernt Schiele

As a group, we had a discussion on *representations*. The way data in general, and images, video and text in particular, are represented defines, to a large extent, what information can (easily) be extracted from the data. Representations are also key when it comes to estimating the similarity between two or more items. With the popularity of deep learning methods, many of the representations used today are learned in a data-driven manner, making it harder to assess what they really capture.

We started the discussion with a definition of the concept of representation, as vision and language communities seemed to have somewhat different views on this. A first important aspect is the *representation scheme* or *format* (e.g., do we represent the data as a vector, matrix, tensor, graph, set or something else?). For most of our discussion, we focused on the case of a vectorial representation. Especially for vision people, this seemed a natural choice, while text people use a larger variety of representations. The main reason to opt for other, more complex formats is to make some of the structure within the data more explicit (e.g., the spatial dimensions of an image captured by a matrix or tensor, or dependencies made explicit in a graph representation). Even within a given format, different representations are possible. Ultimately, these are all the result of some transformations applied to the input data, typically with the aim to remove redundancy, remove noise and highlight relevant information.

Representations can be designed ('hand-crafted') or learned from data (typically with neural networks). In the latter case, the representation obtained depends on i) the network architecture (e.g., receptive field), ii) the loss used to train the model (e.g., reconstruction of the input data, semantic classification, etc.), and iii) the training data.

We discussed properties of the "ideal representation". Desired characteristics (sometimes conflicting) of good representations include:

- *Compactness*;
- *Robustness*: A small change in the input data does not have a big impact on the resulting representation;
- *Disentangled*: Different aspects of the data are stored in different subspaces, e.g., illumination vs. geometry vs. material for objects; ethnicity vs. facial expression vs. age for faces; or intent vs. style vs. language for text messages;
- *Explicit*: Easy to interpret by humans, with different elements being 'name-able';
- *Transferable*: A good representation generalizes well beyond the initial training conditions;
- *Probabilistic*: A good representation incorporates information about uncertainty.

Additionally, depending on the context, additional characteristics include

- *Static vs. dynamic*: In case of streaming input data, a good representation gradually changes over time;
- *Task-specific vs. universal*: In some cases it can be beneficial to have a representation that's tuned towards one specific task; yet ideally one can imagine the existence of a universal representation, from which task-specific ones can be derived by projecting on some lower-dimensional subspace;
- *Granularity*: Representations can be considered at different levels of granularity, e.g., objects or scenes in images; or words, sentences or stories in text. For images, there's the additional distinction between 2D and 3D representations.

We had a further discussion about *implicit vs. explicit representations*, and concluded that this distinction is related to the difference between model-driven vs. data-driven approaches, and closely linked to explainability. One interpretation is to measure the degree of explicitness of a representation as the amount of work that is needed (either by a human or a machine) to derive knowledge from it. Making a representation more explicit than often implies imposing more constraints based on prior knowledge (a model).

In the context of language research, both discrete and continuous representations are used. Understanding of language goes beyond semantics (e.g., intent) and explicit representations. Structured representations are an example of more explicit representations – yet opinions seemed to be mixed as to whether this is something we should strive for or maybe not really needed. There was agreement though that explicit symbolic units are probably insufficient to capture the richness of natural language.


In the context of vision research, a trend towards more explicit representations can be observed in the sense that several works aim at incorporating domain knowledge such as geometric constraints in the neural network models. There's also old work explicitly designing neural networks with weights derived from physics or geometry. The work on Generative Query Networks, on the other hand, is an example of a powerful implicit representation.

We continued with a discussion on the interaction between visual and textual representations. Both modalities are complementary, and visual understanding cannot be reduced to just mapping images to words. Language, e.g., in the form of image captions or textual descriptions, seems especially useful to provide compositionality in human learning, leading to more abstract interpretations and better generalization. It can also help to focus attention. At the same time, the two communities think about representations quite differently: while the language community is mostly aiming to deal with the ambiguity in language and stresses the fact that a lot of information is implicit, vision researchers aim for a precise, absolute description, and mostly ignore implicit aspects.

Open research questions include the development of hybrid models combining data-driven learning with more explicit representations (especially on the vision side), representations of the 3D world (vision), better cross-modal representations, and unsupervised machine translation using statistics or transfer learning (language).

## 4.2 Visual and Language Understanding

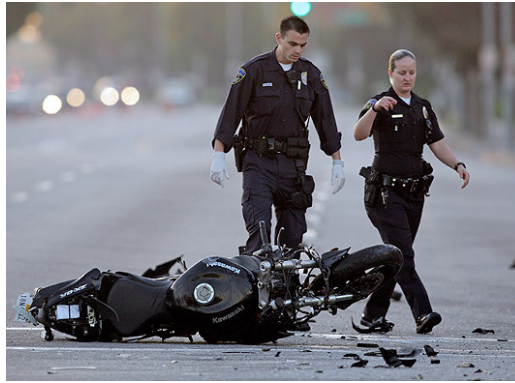
Marie-Francine Moens (KU Leuven, BE), Stephen Clark (Google DeepMind – London, GB), Luisa Coheur (INESC-ID – Lisbon, PT), Erkut Erdem (Hacettepe University – Ankara, TR), Anette Frank (Universität Heidelberg, DE), Jean Oh (Carnegie Mellon University – Pittsburgh, US)

License  Creative Commons BY 3.0 Unported license

© Marie-Francine Moens, Luisa Coheur, Stephen Clark, Erkut Erdem, Anette Frank, Jean Oh

This working group focused on the broad topic of visual and language understanding and went deeper into the need and potential of modeling alignment between multimodal representations and their components.

An image is worth more than a thousand words: it is possible to generate many different descriptions of an image ranging from what is actually seen in the image to what could be inferred from the image. For instance, the picture of an accident (as in Figure 1) could generate the description of the accident setting, but it is more difficult to determine what the causes of the accident are or whether someone is injured.

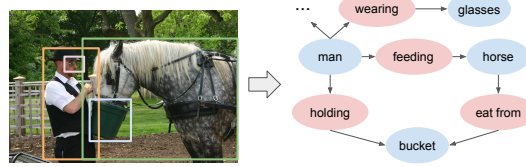


■ **Figure 1** Picture taken from [4].

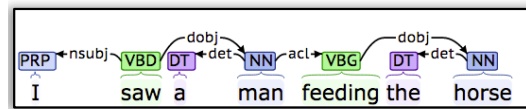
To accomplish the latter, we need some form of *reasoning* both with the *content of the image* and *prior knowledge*. Also, very often a command given to a robot is multimodal, that is, it is composed of the language command and the visual context in which the command is given. In all these cases, *knowing what information in the language is aligned with the visual data and which is not*, is valuable in understanding the command and correct action upon it.

In this working group we also discussed the properties of the visual and language data. Visual data are restricted to what could be visually perceived, but on the other hand expresses and evokes lots of knowledge. Language is often compact, ambiguous and abstract, but offers ways to be very specific, for instance, by being able to communicate negation and modal aspects of what is expressed. *Both modalities often function in a complementary way*. For instance, language leaves certain information implicit which could be learned from other modalities such as vision, acoustics or engineered knowledge (e.g., [12]). So, *complementarity helps for completing missing information*. On the other hand, overlapping or aligned information between visual and language data helps in *disambiguating* polysemous words or ambiguous image patterns.

All the above favors a multimodal representation that explicitly identifies overlaps and complementarity of the two modalities which would allow to disentangle information from the



■ **Figure 2** Scene graph from [9].



■ **Figure 3** Dependency parse (Stanford CoreNLP).

two modalities and encourage the interpretability of the representation. Such a representation would help both the semantic parsing of an image or a text as well as the generation of each modality.

The structure of a scene image or a sentence can be modeled as a graph. Scene graphs (an example shown in Figure 2) are currently very popular to capture the content of an image in terms of objects and their relationships [9, 8].

They are generated to describe an image (e.g., [9]), or images are generated from scene graphs (e.g., [3]). Graphs are also popular to structure language utterances and are the results of a (neural) dependency parse or a (neural) semantic parse (Figure 3).

A dependency parse of a sentence can in its turn easily be translated into a scene graph. As a result, these graph structures offer anchors for finding alignments between language structures and visual structures as well as between their composing components.

Humans are very good at making sense of scenes or utterances composed of objects that they have never seen before in that combination due to their ability and understanding of compositionality. A necessary condition for compositional interpretation is to recognize the components that make up a scene or language utterance, and understanding their relations. It would be interesting to *take advantage of the inherent compositionality of both images and language* and *integrate these properties into the learned representations*. This would entail that we can decompose representations of a whole image or video, of complete language utterances or discourses, and thus complete multimodal inputs into representations of their components.

Consequently, it would be interesting to construct structured multimodal representations. Initial attempts in this direction were made by [6], [5], [10], [7], [11]. This could lead to possible advances in *alignment, attention models, compositionality, and incremental learning*. For instance, multimodal alignment is then seen as finding relationships and correspondences between sub-components of instances from two or more modalities as in [1].

Visual data and language operate sometimes at different levels of abstraction. Possible advances in hierarchical alignment could help in finding corresponding and complementary content across modalities. An example in that direction is a hierarchical multimodal attention-based NN for image captioning as proposed by [2].

To conclude we aim at aligned, (de)composable image and linguistic representations. Explicit alignments allow to identify *overlapping vs. complementary content* and facilitate *system interpretability*. In addition, these would allow for easy and interpretable integration of symbolic (e.g., available from a knowledge resource) and continuous representations. To

reach that goal we will need more research on how to make structure more explicit through joint learning of entities and relationships in both modalities, and to align them – possibly at different levels of granularity. In the long term this could lead to inducing task-specific multimodal semantic grammars.

## References

- 1 T. Baltrušaitis, C. Ahuja, and L. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, Feb 2019.
- 2 Yong Cheng, Fei Huang, Lian Zhou, Cheng Jin, Yuejie Zhang, and Tao Zhang. A hierarchical multimodal attention-based neural network for image captioning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’17, pages 889–892, New York, NY, USA, 2017. ACM.
- 3 Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1219–1228, 2018.
- 4 Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *ACL. The Association for Computer Linguistics*, 2016.
- 5 Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- 6 Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.
- 7 Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *CoRR*, abs/1511.06361, 2016.
- 8 Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. Scene graph parsing as dependency parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 397–407, 2018.
- 9 Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- 10 Hongteng Xu, Licheng Yu, Dixin Luo, Hongyuan Zha, and Yi Xu. Dictionary learning with mutually reinforcing group-graph structures. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI’15, pages 3101–3107. AAAI Press, 2015.
- 11 Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
- 12 Guillem Collell, Luc Van Gool, Marie-Francine Moens. Acquiring common sense spatial knowledge through implicit spatial templates. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI’18, pages 6765–6772. AAAI Press, 2018.

### 4.3 Challenges in Language Generation Applications

*Lucia Specia (Imperial College London, GB), Loïc Barrault (Université du Mans, FR), Thales Bertaglia (University of Sheffield, GB), Erkut Erdem (Hacettepe University – Ankara, TR), Lisa Anne Hendricks (University of California – Berkeley, US), Pecina Pavel (Charles University – Prague, CZ), Florian Metz (Carnegie Mellon University – Pittsburgh, US), Jean Oh (Carnegie Mellon University – Pittsburgh, US)*

**License** © Creative Commons BY 3.0 Unported license

© Lucia Specia, Loïc Barrault, Thales Bertaglia, Erkut Erdem, Lisa Anne Hendricks, Pecina Pavel, Florian Metz, Jean Oh

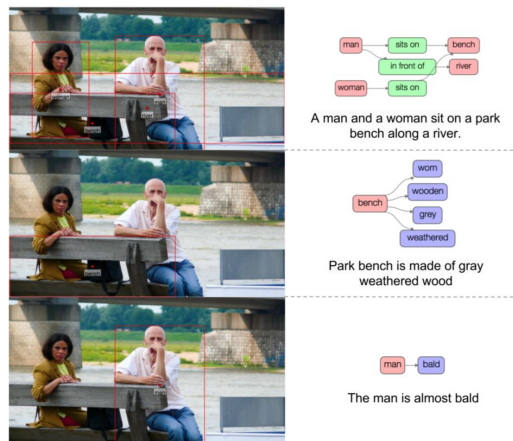
This group discussed limitations in current work in vision and language where the final objective is to generate language, given as input images/video and optionally language (e.g. the text in the source language in machine translation). The core points discussed were the types of visual information, the modality fusion architectures/approaches, how to benefit from the temporal aspect in videos, how to evaluate outputs, and a good task/dataset would be for the area. The discussion items in these topics are summarised below.

**Visual information and representations.** This part of the discussion focused on current work going beyond dense visual features (i.e. CNN layers) to exploit detections or proposals, mostly for objects, for example Neural Baby Talk [10] and the Bottom-Up Top-Down approach [2] for image captioning. It was suggested that using detections for places, actions, “stuff” and potentially others visual elements could be beneficial, but that there certainly are not detectors for everything. Another factor that is disregarded in current work is relevance of content in images for a given task. For example, for image description, the objects detected may not be the most important/interesting parts of the image to describe. Datasets with saliency information such as the Saliency Benchmark<sup>1</sup> could be useful. There is some work in this area which attempts to focus on interesting parts of the image using a loss function to make descriptions across images different from one another (e.g. [11]), or learning from eye-tracking data on how humans do it (e.g. [14]). Further structural information beyond what is in the image and its relevance for a task could also be beneficial, such as relationships between elements in the image. For texts, this can be done via dependency parsing or semantic role labelling. For images, one can consider scene graphs, such as in the Visual Genome dataset [15]. Examples from the Visual Genome dataset (<https://visualgenome.org/>) are given in Figures 4 and 5. However, this type of information is harder to generate automatically and reliability from images. A promising direction is to move beyond structure to external information, e.g. metadata or the use of multiple images for each instance. An example of work in this direction is [3], for personalized textual output. The variety of visual elements that can be detected and realistic differences in distributions between training and test sets is also important. We discussed image domain captioning via **nocaps**, a benchmark for novel object captioning at scale [1]<sup>2</sup> where 166K captions are generated for 15K images from Open Images.

**Fusion Approaches/architectures.** The ways in which visual information is currently fused with textual information is suboptimal. Few attempts to improve on that have been made, for example, as mentioned before, recent image captioning approaches such as Neural Baby Talk [10] and the Bottom-Up Top-Down approach [2] using object information, the use

<sup>1</sup> <http://saliency.mit.edu/home.html>

<sup>2</sup> <https://nocaps.org>



**Figure 4** Examples from visual genome dataset (<https://visualgenome.org/>).

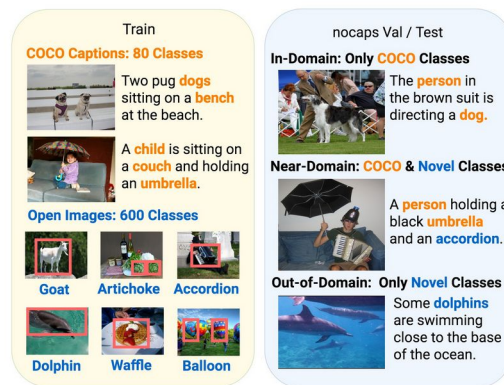


■ **Figure 5** Examples from visual genome dataset (<https://visualgenome.org/>).

of convolutional feature maps at decoding time also for image captioning [4], the FILM architectures for visual question answering [12], or deep context end-to-end contextual speech recognition [13].

**Storytelling.** In this case we are interested in exploring the sequential and temporal aspects of visual information, generally videos. The challenge we face is to move away from static images or sets of frames from actual video data, potentially along with acoustics and language. This should help capture more information (e.g. about actions). One dataset that can help in this direction is the How2 dataset [16] of instructional videos. Examples are approaches that attempt to align multiple asynchronous sub-sequences, e.g. Localizing Moments in Video with Temporal Language [7]. Additionally, an interesting direction is that of methods that do not assume multi-way parallelism between modalities, in other words, when some modalities may be missing for some instances. Ultimately, one should be able to generate textual stories from videos, but the field is far from that stage. A starting point could be approaches to sorting multimodal image-caption pairs.

**Evaluation.** Better metrics are needed to evaluate vision and language tasks where language is generated. Instead of looking at the text information only (string similarity between system output and reference text), metrics should look at object hallucination and missing objects. For the former, [8] provides interesting insights. Another improvement direction is to move away from string matching (e.g. BLEU) or synonymy-level matching (e.g. Meteor) into



■ **Figure 6** The nocaps benchmark for novel object captioning (at scale) (<https://nocaps.org>).

matching at a more semantic level. For image captioning, Word Mover’s Distance, which uses word embedding as the unit to match, works better than other existing metrics [5]. For visual dialogue and storytelling, evaluation tends to be done for utterances independently. This is suboptimal as there many alternative good stories that can be generated. Text coherence models could be helpful but have not yet been investigated for this purpose.

**Task/dataset.** The discussion on what could be an interesting new task and dataset for language generation from images was not conclusive but some thoughts included the idea of a task that expresses common sense/creativity, and that would leverage multiple videos/images, such as multi-video textual summarization, where intermediate tasks involve measuring differences and similarities between videos. Visual dialogue with a plan/task/end-goal in mind is also an interesting direction. One could base the data collection on existing text-only dialogue datasets, e.g. [9] and [6].

To conclude, for language generation tasks, the field seems to be moving towards representing images in more structured ways, but there are a number of open questions, for example: do we need the same structure for all vision and language tasks? Does it matter if it is a tree or graph? Should it be hierarchical? Do we care about all elements and relationships? Or only salient unusual ones? Each of these open questions should lead to interesting research in the area.

## References

- 1 Harsh Agrawal, Karan Desai, Xinlei Chen, Rishabh Jain, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. *arXiv preprint arXiv:1812.08658*, 2018.
- 2 Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.
- 3 Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 895–903, 2017.
- 4 Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander Schwing, and David A Forsyth. Diverse and controllable image captioning with part-of-speech guidance. *arXiv preprint arXiv:1805.12589*, 2018.

- 5 Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis and Erkut Erdem. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 199–209, 2017.
- 6 He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1766–1776, 2017.
- 7 Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1380–1390, 2018.
- 8 Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.
- 9 Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, 2017.
- 10 Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, 2018.
- 11 Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2212, 2017.
- 12 Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- 13 Golan Pundak, Tara N Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. Deep context: end-to-end contextual speech recognition. *arXiv preprint arXiv:1808.02480*, 2018.
- 14 Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. *arXiv preprint arXiv:1902.03751*, 2019.
- 15 Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 684–699, 2018.
- 16 Shou-I Yu, Lu Jiang, and Alexander Hauptmann. Instructional videos for unsupervised harvesting and learning of action examples. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 825–828. ACM, 2014.

## 4.4 Modeling Human Learning

*Raffaella Bernardi (University of Trento, IT), Zeynep Akata (University of Amsterdam, NL), Andrei Barbu (MIT – Cambridge, US), Ozan Caglayan (Université du Mans, FR), Stephen Clark (Google DeepMind – London, GB), Guillem Collell (KU Leuven, BE), Desmond Elliott (University of Copenhagen, DK), Raquel Fernandez (University of Amsterdam, NL), Orhan Firat (Google Inc. – Mountain View, US), Stella Frank (University of Edinburgh, GB), Frank Keller (University of Edinburgh, GB), Douwe Kiela (Facebook – New York, US), Pecina Pavel (Charles University – Prague, CZ), David Vernon (CMU Africa – Kigali, RW), Josiah Wang (University of Sheffield, GB)*

**License** © Creative Commons BY 3.0 Unported license

© Raffaella Bernardi, Zeynep Akata, Andrei Barbu, Ozan Caglayan, Stephen Clark, Guillem Collell, Desmond Elliott, Raquel Fernandez, Orhan Firat, Stella Frank, Frank Keller, Douwe Kiela, Pecina Pavel, David Vernon, Josiah Wang

The group has focused on the connection between what is known from Cognitive Neuroscience about human learning and how such findings can be of inspiration when developing machine learning systems. We first shared our knowledge on the topic and then highlighted those research questions that would be worth to address in order to model human learning.

**Human development.** From Cognitive Neurosciences it has been highlighted the distinction between “developmental” and “learning” phases children go through. During the developmental phase, progresses are divided into clear not overlapping steps which are experienced by all children in the very same order (e.g., understanding before speaking). In neonates, development is manifested by the emergence of new forms of action and the acquisition of predictive control of these actions. It has been highlighted that mastery of action relies critically on prospection, i.e., the perception and knowledge of upcoming events. Repetitive practice of new actions is not focused on establishing fixed patterns of movement but on establishing the possibilities for prospective control in the context of these actions [12]. Neonats go through a “perceptual narrowing” mechanism which has been proved to be involved in language as well as in face recognition [4]. After age 2 till 10 a “synaptic pruning” process starts: 50% of the synapses are eliminated thereby increasing the efficiency of the neural network [7]. Furthermore, it has been noted that in human development a major role is played by “core abilities” which enable infants to acquire core knowledge which act as building blocks for scaffolding new cognitive abilities and more complex cognitive tasks. This core knowledge relates to perception of objects, numerosity and people.

**The role of language.** Besides the general overview above we zoom into the question of which is the role of language in human development. We mention the work by [8] showing that language may accelerate learning of other skills, in particular it could serve as effective priors, facilitating perception and integration of sensory information. The importance of internal talk (a.k.a., inner speech or sub-vocalization), and the findings about tongue anesthesia disrupts performance in mental tasks found in children but not in adults (after inner speech appeared) [10]. Finally, it has been pointed out the important role of language in structuring our memories has claimed in well established theories (e.g., [11]).

**Continual machine learning.** The importance of bringing into machine learning the developmental approach has been strongly advocated in [9]. We tried to get a picture of the advancement in ML which could be connected with the discussion summarized above on human development. In particular, the following works have been mentioned: continuation methods, viz., start with simple objectives and a simple model, then increase the task

complexity and use more neurons [3]; few-shot meta-learning, viz., start with a subset of classes, then introduce new classes, and at end all the classes [6]; knowledge distillation, viz., pre-train the model on core skills and then plug it to transfer the knowledge for learning other skills [2] or start with a big model, then compress it and specialize it on a fine-grained distinction by fine-tuning; reinforcement learning and curiosity-driven learning [5, 1].

**The role of language in learning multisensory skills.** What is the role of language in learning multisensory skills? Does language accelerate or amplify learning? Is language necessary to learn some specific concepts or is vision enough? Does language bring advantages to development/learning?

**Relation among tasks in continual learning.** Can we capture the formal relation holding between tasks useful in multimodal and multi-task settings? What should be the formal relationship between the tasks? What is a good multimodal multi-task setting? Combine pre-training with multi-task learning, how does this relate to continuation methods? Does Bayesian learning/inference have anything to offer in this context?

## References

- 1 Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *ICLR*, 2019.
- 2 Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS*, 2015.
- 3 Caglar Gulcehre, Marcin Moczulski, Francesco Visin, and Yoshua Bengio. Mollifying networks. In *ICLR*, 2017.
- 4 Olivier Pascalis, Michelle de Haan, and Charles A. Nelson. Is face processing species-specific during the first year of life? *Science*, 296, 2002.
- 5 Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*, volume 2017, 2017.
- 6 Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018.
- 7 Edalmarys Santos and Chad A. Noggle. *Encyclopedia of Child Behavior and Development*, chapter Synaptic Pruning, pages 1464–1465. Springer US, 2011.
- 8 Irina Simanova, Jolien C. Francken, Floris P. de Lange, and Harold Bekkering. Linguistic priors shape categorical perception. In *Language, Cognition and Neuroscience*, volume 31, pages 159–165, 2016.
- 9 L. B. Smith and L.K. Slone. A developmental approach to machine learning. *Frontiers in Psychology*, 2017.
- 10 A. Sokolov. *Inner speech and thought*, chapter V. Springer Science and Business Media, 2012.
- 11 E. Tulving and D.M. Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5):352–373, 1973.
- 12 David Vernon, Claes von Hofsten, and Luciano Fadiga. *A Roadmap for Cognitive Development in Humanoid Robots*. Springer, 2010.

## 4.5 Explainability

*Tinne Tuytelaars (KU Leuven, BE), Luisa Coheur (INESC-ID – Lisbon, PT), Vera Demberg (Universität des Saarlandes, DE), Lisa Anne Hendricks (University of California – Berkeley, US), Dietrich Klakow (Universität des Saarlandes, DE), Jindrich Libovický (Charles University – Prague, CZ), Pranava Madhyastha (Imperial College London – GB), Marie-Francine Moens (KU Leuven, BE), Siddharth Narayanaswamy (University of Oxford, GB), Bernt Schiele (MPI für Informatik – Saarbrücken, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Tinne Tuytelaars, Luisa Coheur, Vera Demberg, Lisa Anne Hendricks, Dietrich Klakow, Jindrich Libovický, Pranava Madhyastha, Marie-Francine Moens, Siddharth Narayanaswamy, Bernt Schiele

The demand for and research into explainable AI is increasing. Talking about open challenges in vision and language, generating explanations of decisions and models was one that we all scored high on the research agenda. On the one hand, explanations can help us to better understand the models we are using, which may lead to more insight as well as better models per se. On the other hand, explainability is considered a crucial step towards building trust, and as such necessary to make the technology accepted in real world applications. Explanations can take many forms. Typical examples are causal graphs, graphical models, textual explanations, visual heatmaps, visualizations of CNN filters, or diagrams. They should be geared towards humans, thus intuitive and easy to interpret.

Since this is a relatively young research topic, there is some discussion about terminology, and in particular the difference between interpretation and explanation. Some consider interpretation to refer to introspection, explaining a model as a whole, whereas explanation relates to predictions or decisions for a specific input. Others consider interpretation to be more related to internal representations and understanding, whereas explanations are the formal way to communicate about this with others. For instance, a large decision tree may be interpretable, but if it's too big it's not explainable (i.e., explanation should be given at a higher level).

Explainable AI can, in general, be achieved in two ways: i) as a posthoc analysis, given a model – e.g., by discovering emerging structure, or using a surrogate model to explain the original one; or ii) by designing the models with interpretability in mind from the start – e.g., by adding an objective during training for interpretability, or by designing novel machine learning architectures that are better explainable. Some models lend themselves better for explainability than others by nature, e.g., graphical models. But explainability probably is also influenced by the data used to train a model.

A critical aspect when it comes to studying explainability is the evaluation of different methods. One criterion could be the time it takes a user to understand the interpretation. Another one may be related to how comprehensive it is, i.e., how much relevant information is retained. Since explanations are geared towards humans, it makes sense to say that humans need to be in the loop for evaluations. However, can humans really evaluate the quality of explanation? For instance, they may not like biases, even though it is a correct explanation of the model. They may also prefer explanations that are in line with their own interpretation. It's unclear how to deal with this subjectivity. The kind of explanations people like most are not necessarily the ones that are objectively the most correct. For instance, it has been shown that people prefer explanations for why a specific advertisement was selected that are not too specific.

An interesting question that popped up was, whether explanations can be used to improve the model or system itself. This is similar to the notion of reflective learning observed in humans. The fact that decisions can be explained in a consistent manner could increase the

confidence of the system in its decisions. Examples in this direction include self-labeling based on clustering, or adaptive gradient descent. When using textual explanations for visual problems or v.v., the addition of an extra modality can also improve robustness and encourage more disentangled representations.

## 4.6 Tasks: Creating Simulated Worlds from Existing Media

*David Hogg (University of Leeds, GB), Raffaella Bernardi (University of Trento, IT), Desmond Elliott (University of Copenhagen, DK), Raquel Fernandez (University of Amsterdam, NL), Stella Frank (University of Edinburgh, GB), Marius Leordeanu (University Politehnica of Bucharest, RO), Jean Oh (Carnegie Mellon University – Pittsburgh, US), Pavel Pecina (Charles University – Prague, CZ), Lucia Specia (Imperial College London, GB), Jakob Verbeek (INRIA – Grenoble, FR), David Vernon (CMU Africa – Kigali, RW)*

**License** © Creative Commons BY 3.0 Unported license

© David Hogg, Raffaella Bernardi, Desmond Elliott, Raquel Fernandez, Stella Frank, Marius Leordeanu, Jean Oh, Pavel Pecina, Lucia Specia, Jakob Verbeek, David Vernon

Our discussion focused on the use of the vast repository of on-line audio-visual media, such as TV shows and movies, to create a generative model for virtual worlds. These worlds would:

- Be realistic in visual and auditory modalities;
- Follow a narrative involving simulated people (agents) behaving in a natural way;
- Be configurable for selected situations, environments, cultural and emotional norms, mirroring the content in the source media;
- Be interactive, enabling active and natural participation of real people using VR equipment.

Other potential media sources include: proceedings of world governments, customer service interactions, video-conference recordings, and AV sensors on domestic robots and autonomous vehicles.

### Societal impact of such technology:

- Entertainment – interactive TV, enabling role-playing in shows; narrative transfer into new contexts (‘Romeo and Juliet’ into the ‘Friends’ genre); researching movie locations.
- Education and training – language learning; skills coaching, including generic skills; learning maths, presentation skills; robot learning.
- General Media – conceptual search/comparison on narratives, scene contexts; more empathetic agents; promotion of cultural understanding; mapping media into new cultural contexts; generated worlds as a novel communication medium between humans.
- Health – therapy for people who have difficulty recognising social cues (e.g., in ASD); culturally sensitive telemedicine.

The technology could be a transformative tool for the behavioural sciences, for example in studying human adaptability to low-fidelity agents (non-human speech patterns, prosody-only sound profiles, language abilities, masked facial appearance and expression, response latency, gender neutral agents); cultural modes of communication (distance apart, eye contact); and use of mental imagery and language in conceptual reasoning (e.g., planning and prediction of future actions).

**Research challenges:**

There are central challenges in achieving photo and audio realism, capturing behavioural characteristics and understanding narrative within a generative model acquired from on-line media. Beyond this, there are many interesting and fundamental research questions, such as: (1) How to create, control and manipulate virtual worlds using language cues, for increasingly complex environments; (2) How to ensure that language use acquired in the virtual world (e.g., in learning a foreign language, training a robot) is deployable in the real world, (3) The extent to which people can learn new skills from observation and non-physical interaction alone (e.g., car maintenance); (4) What is a good meaning representation of a person in learning through interaction with skilled agents; (5) How the same situation can be interpreted differently by different people, and how different views can converge; (6) What are good feedback/interaction strategies; (7) 'Teaching' versus 'doing' as pedagogical strategies; (8) What representations are needed to discover the social roles of agents, and generate dialogue to further an agent's social and long-term collective goals; (9) Is learning in a virtual environment better than learning from people in the real world, perhaps because agents are potentially friendly, better teachers, and always available, or is enabling people to avoid human contact problematic in the long term?

**4.7 Tasks and Datasets for Vision and Language**

*Stephen Clark (Google DeepMind – London, GB), Zeynep Akata (University of Amsterdam, NL), Andrei Barbu (MIT – Cambridge, US), Loïc Barrault (Université du Mans, FR), Raffaella Bernardi (University of Trento, IT), Ozan Caglayan (Université du Mans, FR), Aykut Erdem (Hacettepe University – Ankara, TR), Erkut Erdem (Hacettepe University – Ankara, TR), Orhan Firat (Google Inc. – Mountain View, US), Anette Frank (Universität Heidelberg, DE), Stella Frank (University of Edinburgh, GB), David Hogg (University of Leeds, GB), Frank Keller (University of Edinburgh, GB), Douwe Kiela (Facebook – New York, US), Chiraag Lala (University of Sheffield, GB), Marius Leordeanu (University Politehnica of Bucharest, RO), Florian Metze (Carnegie Mellon University – Pittsburgh, US), Lucia Specia (Imperial College London, GB)*

**License** © Creative Commons BY 3.0 Unported license

© Stephen Clark, Zeynep Akata, Andrei Barbu, Loïc Barrault, Raffaella Bernardi, Ozan Caglayan, Aykut Erdem, Erkut Erdem, Orhan Firat, Anette Frank, Stella Frank, David Hogg, Frank Keller, Douwe Kiela, Chiraag Lala, Marius Leordeanu, Florian Metze, Lucia Specia

A discussion group was led on tasks and datasets located at the intersection of Vision and Language. First of all, everyone seemed to agree that a lack of suitable datasets and tasks was a bottleneck for the development of AI systems operating at this intersection, and that developing better tasks should be a priority.

One discussion centered around the question of whether we could find “One Task to Rule Them All”, much as automatic speech recognition (ASR) seems to have found such a task in the guise of minimizing word error rate (WER). Minimizing WER is only a proxy for many of the actual tasks that we want to do with ASR, but it seems to have been a good enough proxy that it has led to clear advances in the field.

One suggestion for Vision and Language was conditional language modeling, and in particular caption generation. The point was made that we should be clear what we mean by caption generation, since this can refer to at least three distinct tasks: captions for newspaper

photographs, captions for videos, and image descriptions. Typically researchers mean the third option – an image description task – when talking about caption generation, so we focused on that.

The image description task is problematic for a number of reasons, but perhaps most fundamentally because it is often unclear which aspect of the image to focus on when providing a description, and this makes evaluation especially difficult. This relates to a broader problem of there often being a lack of a clear goal, or application, when eliciting the captions (either from a human or a machine). So one suggestion for creating better tasks and datasets is to focus on an application first, for example caption generation for visually-impaired people. It was also suggested that Visual Question Answering may be better in this regard, since it relates to an information need that a viewer of the image may have.

We then moved onto a discussion about whether it would be possible to analyse a dataset and corresponding task along a number of core dimensions, which could then be used to provide a useful summary to potential users of that dataset, and could also act as a useful guide when creating it. The idea was that there would be a relatively small number of such dimensions – perhaps as few as three – that would capture the “essence” of any dataset. However, when attempting to come up with these core dimensions, we quickly discovered that there are many important dimensions along which a dataset can vary. Examples include: number of modalities, temporal complexity, interactivity, number of agents, world complexity, linguistic complexity, existence of biases, dataset size, whether the task requires reasoning, whether the dataset tests for generalization capabilities, whether it exercises the cognitive core (intuitive physics, intuitive psychology, semantic memory).

We also attempted to analyse an existing dataset – the How2 dataset - along these dimensions. How2 is a multimodal collection of instructional videos which a number of the working group were familiar with. Again, it was surprisingly difficult to categorize this dataset along the chosen dimensions, but it was felt that this could still be a useful exercise, and that a final set of dimensions – obtained after a few cycles of use – could be a useful resource. It was also suggested that this classification may make a useful university class exercise.

Finally, we took a broader perspective on vision and language tasks, and considered what the “ultimate application” might be in this space, including the possibility of other modalities. We fixed on an “embodied Alexa”, something like the ultimate robot butler, which could tidy rooms, fix up lunch and dinner, take the children to school and so on. It was felt that computer vision technology was a long way from being in a strong enough state to be usable in such an application, but that interestingly NLP technology – especially as far as semantic parsing is concerned – might be in a better state; however, it was acknowledged that the general problem of knowledge acquisition would still need to be solved, and we’re a long way from that.

Will we have such an application in 25 years? Who knows, but the general consensus seemed to be probably not, at least in the complete form described above, but that there may be a limited, caricature of such a robot butler available (much like Alexa is currently a caricature of a fully-functioning natural language understanding system).

## Participants

- Zeynep Akata  
University of Amsterdam, NL
- Andrei Barbu  
MIT – Cambridge, US
- Loïc Barrault  
Université du Mans, FR
- Raffaella Bernardi  
University of Trento, IT
- Thales Bertaglia  
University of Sheffield, GB
- Ozan Caglayan  
Université du Mans, FR
- Stephen Clark  
Google DeepMind – London, GB
- Luísa Coheur  
INESC-ID – Lisbon, PT
- Guillem Collell  
KU Leuven, BE
- Vera Demberg  
Universität des Saarlandes, DE
- Desmond Elliott  
University of Copenhagen, DK
- Aykut Erdem  
Hacettepe University –  
Ankara, TR
- Erkut Erdem  
Hacettepe University –  
Ankara, TR
- Raquel Fernández  
University of Amsterdam, NL
- Orhan Firat  
Google Inc. –  
Mountain View, US
- Anette Frank  
Universität Heidelberg, DE
- Stella Frank  
University of Edinburgh, GB
- Lisa Anne Hendricks  
University of California –  
Berkeley, US
- David C. Hogg  
University of Leeds, GB
- Frank Keller  
University of Edinburgh, GB
- Douwe Kiela  
Facebook – New York, US
- Dietrich Klakow  
Universität des Saarlandes, DE
- Chiraag Lala  
University of Sheffield, GB
- Marius Leordeanu  
University Politehnica of  
Bucharest, RO
- Jindrich Libovický  
Charles University – Prague, CZ
- Pranava Madhyastha  
Imperial College London – GB
- Florian Metze  
Carnegie Mellon University –  
Pittsburgh, US
- Marie-Francine Moens  
KU Leuven, BE
- Siddharth Narayanaswamy  
University of Oxford, GB
- Jean Oh  
Carnegie Mellon University –  
Pittsburgh, US
- Pavel Pecina  
Charles University – Prague, CZ
- Bernt Schiele  
MPI für Informatik –  
Saarbrücken, DE
- Carina Silberer  
UPF – Barcelona, ES
- Lucia Specia  
Imperial College London, GB
- Tinne Tuytelaars  
KU Leuven, BE
- Jakob Verbeek  
INRIA – Grenoble, FR
- David Vernon  
CMU Africa – Kigali, RW
- Josiah Wang  
University of Sheffield, GB

