Situation-Specific Multimodal Feature Adaptation

Özge Alaçam
Language Technology Group, Dept. of Informatics, Universität Hamburg
Hamburg, Germany
alacam@informatik.uni-hamburg.de

Abstract

Current technological and scientific developments on assistive technologies result in a considerable need for NLP models to successfully grasp the intention of the user in situated settings. Situated language comprehension, where different multimodal cues are inherently present and essential parts of the situations, can not be handled in isolation. In this research proposal, we aim to quantify the influence of each modality including the eye-movements of the speaker as a deictic cue to gain deeper understanding about multimodal interaction. By doing this, we mainly focus on the role of various referential complexities in this interaction. The proposed model encodes the referential complexity of the situated settings in the embedding space during the pre-training phase. This will, in return, implicitly guide the model to adjust to situation-specific properties of an unseen test case.

In this paper, we summarize the challenges of intention extraction and propose a methodological approach to investigate a situation-specific feature adaptation to improve crossmodal mapping and meaning recovery from noisy communication settings.

1 Motivation

In recent years, we have witnessed a considerable increase in the use of assistive technologies that can engage in communication and perform tasks. These can come in different forms like smart speakers and mobile devices that you can command with audio, or more specialized task-oriented robots that can actually realize users' command in 3D environments. The steady increase in the use of collaborative robots (IFR, 2018) in daily life brings along another important Human-Computer Interaction theme: the capability of engaging in a natural and smooth spoken dialog with humans, which is a major scientific and technological challenge. Particularly, being able to follow a communication that conveys thoughts and intentions expressed in a flexible manner without the restrictions of a close-set of commands is a crucial component of assistive robots for the handicapped and elderly people and for the education / entertainment purposes.

Spoken and situated communication is composed of various perceptual (e.g. audio, visual) and representational modalities (e.g. language, deictic eye-movements, gestures). Effectiveness and fluency of human communication capabilities inspire us to develop robust language models that can deal with uncertainties by evaluating all the available information from multiple sources and reach *a good-enough decision*. In order to reach this performance, we need to model our situated language understanding systems to incorporate those modalities and let them interact in a meaningful way. This brings forth some important questions; how to integrate different modalities and how to utilize adaptive processing for effective situation-awareness to be able to deal with cases where some of the modalities are restricted due to noise in the communication channels. This capability for cross-modal integration can be a very important feature in resolving references or executing commands for smart speakers or helper robots that aid people in their daily activities.

2 Situated Language Understanding

In a task-oriented setting (e.g. helper robots completing a given task), the goal of natural language communication is to extract the intention of the speaker. Such communication usually happens in structurally rich visual environments like the one in Figure 1, which contains several glasses of different types (wine and tea glasses), mice (computer mouse and cat toy), windows (open and closed) etc..

Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 81–87
April 20, 2021.

Figure 1: An example image for a living room scenario.

Some of them are even (partially) occluded from the viewer's perspective. The environments usually also include people and their interactions with the objects (actions). Visual information plays a crucial role in determining the referential objects related with the action and to accomplish the task. Thus, computational solutions that incorporate those cues are expected to perform better in grasping the meaning compared to their (text-only) unimodal counterparts (see Alaçam et al. (2020a) for a review).

Determining the correct intention of a user is not always straightforward due to various reasons. Let us take the following sentences as an example:

1. Can you bring me the wine, I want to open it.
2. Can you bring me the wine, I want to drink it.

In both cases, on the syntactic level, the pronoun *it* refers to *the wine*. But in the referential world, the first one clearly refers to *a wine bottle*, while the second (with a lesser degree of certainty) refers to *a glass of wine*.

In many cases, even the object names are verbally omitted in spoken utterances. This kind of implicit commands like *"I prefer red, can you open a bottle and bring it to me?"*, require the hearer to reconstruct the underlying intention *"Open a bottle of red wine, and bring the bottle"* (Gundel et al., 2012). Alternatively, depending on the spatial arrangements of the agents and the objects in the room, the intention of the speaker might be slightly different and more complex. For example, when the empty glasses are closer to the listener than the speaker, the interpretation might be: *"Open a bottle of red wine, pour the wine into one of the empty wine glasses and bring the glass of wine to me"*. Expressing this intention explicitly most often results in unwieldy utterances.

Furthermore, when the environment is noisy, or the communication partner suffers from a motor or cognitive impairment, multimodal integration plays a more critical role. Noise in communication can originate from various sources. It can be linguistic noise (e.g. spelling mistakes, complex attachments), visual ambiguities (e.g. clutter in the environment, occlusions) or an acoustic noise. Instead of waiting for clarification, combining the uncertain information from the linguistic channel with information from the other ones increases the fluency and the effectiveness of the communication (Garay-Vitoria and Abascal, 2004). One of the most well-known examples to this phenomenon is *the cocktail party effect*, that highlights the human ability to focus on one particular source while inhibiting the noisy ones. When the informativeness of one modality is reduced due to environmental conditions, the human language processing system can successfully adjust itself by relying less on the unclear modality and using other cues in the environment. In this specific scenario, other informative cues provide more reliable information compared to the noisy linguistic input. These cues can come from the surrounding environment and from the communicational partners, and include eye-gaze direction or representational gestures combined with their referential link to the entities in the environment.

Eye-tracking is attracting considerable interest in many assistive technologies such as educational VR systems that provide embodied learning environments or driver monitoring systems. The use of eye-tracking in daily technological products such as mobile phones, laptops and virtual reality headsets is increasing day by day (Brousseau et al., 2020; Rogers, 2019; Khamis et al., 2018). Therefore, incorporating eye-movements in our language comprehension models is an inevitable outcome of these latest developments, and this makes the systematic research on the combination of this modality with others very crucial.

3 Dataset Collection

The success of a situation-aware language comprehension model is highly dependent on the representativeness of the modalities under various conditions. This kind of coverage requires a richly annotated multimodal corpus that displays the variety of language expressiveness and flexibility. Developing such a corpus is a very costly process. Thus, a dataset that profoundly incorporates a variety of

modalities and their various aspects addressing language comprehension tasks is currently not available. Therefore, we plan to train the model on a set of available multimodal datasets (as listed below), by using whichever modality constellations they can offer (including datasets from various domains like psycholinguistics, language technology, computer vision, human-robot interaction etc.) in a stepwise manner; namely starting with simple / few relations, then gradually increasing the complexity of the interactions.

There are general-purpose multimodal datasets that can be used for training:

- MS COCO (Lin et al., 2014) : an object detection and captioning dataset with >200 K labeled images and 5 captions in a sentence form for each image
- Flicker30k (Plummer et al., 2015): 31 K images collected from Flickr, together with 5 reference sentences
- ImageNET (Deng et al., 2009): 14 M annotated images, hierarchically organized (w.r.t. WordNet)
- MVSO (Jou et al., 2015): 15 K visual concepts across 12 languages, 7.36 M images

Additionally, there are multimodal datasets that were created for a specific task:

- HuRIC 2.0 (Bastianelli et al., 2014): audio files (656 sentences) paired with their transcriptions referring to commands for a robot
- LAVA (Berzak et al., 2016): 237 sentences, with 2 to 3 interpretations per sentence, and a total of 1679 videos that depict visual variations of each interpretation
- CLEVR-Ref+ (Liu et al., 2019): 100 K synthetic images with several referring expressions
- Eye4Ref (Alaçam et al., 2020b): 86 systematically controlled sentence-–image pairs and 2024 eye-movement recordings from various referentially complex situations

Multimodal embeddings will be created from this pool of datasets. Creating embeddings from various data sources will allow us to cover concepts from various aspects such as linguistic, auditory and visual representations. The variety on the visual modality will also help us to capture different visual depictions in a range from synthetic images to photographs. This will increase the representativeness of the concepts in the training dataset that will in return improve the prediction when it comes to unseen environments either in virtual reality or in a real-world setting.

70 % of this collection will be used to create multimodal concept embeddings. The remaining 30 % of the datasets will be included in the test and development sets after semi-automatic and manual annotation of contextual representations, target words, missing words, etc. However, Eye4REF will be used as main testset since it was systematically created to involve referentially complex situations.

4 Objectives

One of the main objectives of this research proposal is to quantify the effect of each modality and their interactions by conducting systematic empirical research with computational modeling and human subject studies. Another objective lies in creating multimodal and multilayer embedding spaces in which the layers will be sensitive to various situation complexities, an approach that has not been considered yet. Moreover, eye-movements of the speakers, as a substantial but underrepresented component of face-to-face communication, are incorporated to further improve NLP methods on meaning extraction and crossmodal reference resolution.

Model. The proposed method will be able to process several modalities that play a crucial role in communication; (i) Linguistic Information (at syntactic and semantic level), (ii) Situational Information, (iii) Prototypical Knowledge and Relations, and (iv) Speech-accompanying eye-movements of the speaker. The initial base model will focus on the first three capabilities by utilizing data-driven language models such as fasttext (Bojanowski et al., 2017) and commonsense knowledge-bases like ConceptNet (Speer et al., 2017). At the same time, two modules that (i) incorporate eye-movements and (ii) perform situation-specific feature adaptation will be developed from scratch. In brief, vocabulary obtained from the pre-trained embeddings is used as a bridge between the modalities. For each vocabulary item, multimodal embeddings will be created by processing every input channel, see Figure 2. For each modality and their joint training, we will utilize an appropriate encoder, such as Fast-R-CNN (Girshick, 2015) for images and attention-based bi-directional LSTMs (e.g. Song et al. (2019), for text and eye-movement data. A neural network ensemble model will be trained on

the embeddings for the task of intended object or action prediction from situated settings with masked information.

Guided Multi-Modal Data Fusion. Based on the vast support provided (Qi et al., 2020; Akbari et al., 2019; Niu et al., 2017; Aytar et al., 2017; Kiros et al., 2014), *a guided multilayer data-driven approach* will be utilized instead of fusing all datasets together without any guidance. Despite their impressive success to solve specific tasks so far, deep learning methods are hardly interpretable in understanding which properties of inputs contribute to the final decision to which degree. Besides, the abstraction capabilities, which are crucial for dealing with new cases, are still very limited. However, the more we know about the interactions among the modalities, the more we can extract and focus on relevant features, and the more we can guide those effective deep learning methods to perform better in an explainable way. This will pave the ground to advance current methods for cross-modal interaction in situated language processing with a comprehensive approach to process more modalities, thus to deal with new situations even under uncertainty.

We plan to obtain concept representations step-by-step and build the concepts over each other with increasing complexity, similar to the development of the human cognitive system. One of the key elements here is to encode referential complexity of the each situated setting in the training data. The multimodal embedding space for each concept will consist of several embeddings, which are sensitive to various complexities (as illustrated in Figure 2) and this structure forms the backbone of the situation-specific adaptation. By automatically classifying the complexity of multimodal input based on the predefined complexity factors, each entry in the datasets will contribute to the respective embeddings in the embedding space. This presents a new approach for creating embeddings, taking input complexity into account. Additionally, this configuration also provides a testbed to investigate another interesting question: does restricting the model to use only a complexity embedding that corresponds with input complexity improve the crossmodal mapping task performance? For example, when the multimodal input refers to a highly featured concept representation (a dinner accompanied by red wine), using a representation that is created from coarse-grain samples (a clip-art of a wine bottle) may yield misclassification and vice versa.

Dynamics among different information sources. In the second objective, we quantify the contribution of each modality and their aspects given the situation to mimic human heuristic processing capability. Language comprehension involves complex sequential decision making and is affected by both uncertainty about the current input and lack of knowledge about the upcoming material. Thus, people use – to a large extent – fast and frugal heuristics, i.e. choosing a good-enough representation (Ferreira, 2003). The heuristic view provides a valid explanation for scenarios with a conversation inside noisy conditions. Instead of waiting / asking for clarification, the model will reach a good-enough decision based on all information gathered through all available input channels. In order to do that, the set of important features given the situated setting should be chosen automatically.

Structuring the embeddings to have separate slots for each modality and for their combinations will allow us to quantify the contribution of each slot individually given the situation in various complexities. Depending on the communication goal or environmental factors, some modalities would contribute to the solution while others could be simply irrelevant or redundant. Understanding the intention of the user requires understanding of which information provided by the modalities is (more) relevant, complementary, or redundant. The human language processing system does this adjustment quite efficiently. Thus, a model will be designed to pick the most effective perceptual and conceptual cues and to ignore the irrelevant ones depending on the situated context. Then, the attention will be channeled towards the most relevant cue sources.

Integrating eye-movements of the speaker. Many eye-tracking technologies in the market employ a sufficient sampling frequency to enable gaze-contingent applications. With advancements in the eye-tracking technology, incorporating eye movements of a speaker or a listener enables us to predict / resolve which entity is being referred to in a complex visual environment (Klerke and Plank, 2019; Mitev et al., 2018; Mishra et al., 2017; Koleva et al., 2015). However, these studies are limited to relatively simple scenes. Situated language understanding in a referentially complex environment or under noisy situations imposes a different level

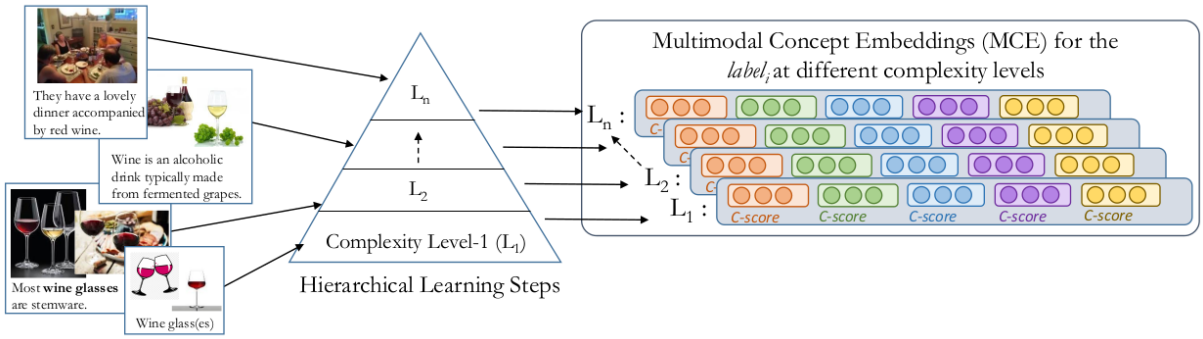

Figure 2: Schematic overview of the hierarchical embeddings.

of challenge that we aim to address. The number of studies that utilize gaze features (Sood et al., 2020; Park et al., 2019; Karessli et al., 2017) is very limited. In this study, we propose to incorporate the eye-movements of the speaker to improve the crossmodal mapping performance. This additional deictic modality may improve the recovery of the intended meaning especially when the communication is noisy (acoustically or visually). The gaze embeddings will be created by using existing eye-movement datasets. However, there are only few big-size eye-movement datasets available (Alaçam et al., 2020b; Wilming et al., 2017; Ehinger et al., 2009). Thus, to enlarge available data, we will conduct a set of experimental studies with increasing referential complexity. There, we will record participants' instructions on a task-oriented scenario and their eye-movements regarding target objects.

Evaluation of the assistive model. After all information sources from various modalities are made available and integrated, the contribution of each modality will be investigated by performing systematic manipulations (e.g. removing a modality one-by-one from the input). The standard accuracy and efficiency metrics will be used for evaluating the models' performance, including the overall runtime, modality-specific accuracy parameters (such as PoS-tag or semantic class accuracy), target mapping accuracy, and accuracy in recovering the missing word.

In addition to evaluating how this model improves the task of reference resolution for acoustically and / or visually noisy settings, its role as assistive technology will be investigated by conducting a user study. The experimental setup will be very similar to the one in the data collection phase. However, this time the participant will interact with a demo model that displays all the above-mentioned capabilities. The model will try to extract user intention by predicting the communicationally relevant objects on the fly. The usability study on the demo model will be evaluated based on the efficiency (how long does it take to reach a decision?), effectiveness (how accurate is the system decision?) and the user satisfaction ratings that will be obtained through the same evaluation metrics and a user survey.

5 Conclusion

In this research proposal, we focus on three factors that can enhance the communication between humans and assistive technologies. The first one is the encoding of the referential complexity of the situated settings while creating multimodal embeddings. As pointed out in (Singh et al., 2020), pre-trained models, that were created by fusing the modalities without constraints, are expected to be an out-of-the-box solution and work well for a variety of simpler tasks. In this research, we propose to encode referential complexity during the training phase to see whether the complexity-sensitive embeddings will improve the tasks of crossmodal mapping and meaning recovery. We believe that this will implicitly direct the model to focus on various textual and visual forms of the same concepts.

The second factor is the inclusion of eye-movements as an additional modality to enhance meaning recovery from noisy settings where some parts of the sentences or visual labels are masked.

At last, this research will also contribute to a better understanding of the contributions of each individual modality, of amodal and modality-specific features and their interactions.

The proposed method will be beneficial for other task-oriented communication scenarios, where the cognitive systems need to understand the intention and to aid the user in the most efficient and effective way, such as educational video-games, training simulations, and assistive navigation systems.

References

Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. 2019. Multi-level multimodal common semantic space for image-phrase grounding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12476–12486, Long Beach, CA, USA.

Özge Alaçam, Xingshan Li, Wolfgang Menzel, and Tobias Staron. 2020a. Crossmodal language comprehension – psycholinguistic insights and computational approaches. Frontiers in Neurorobotics, 14:2.

Özge Alaçam, Eugen Ruppert, Amr R. Salama, Tobias Staron, and Wolfgang Menzel. 2020b. Eye4Ref: A multimodal eye movement dataset of referentially complex situations. In Proceedings of the12th International Conference on Language Resources and Evaluation (LREC), page 2396–2404, Marseille, France.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2017. See, hear, and read: Deep aligned representations. arXiv preprint arXiv:1706.00932.

Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Luca Iocchi, Roberto Basili, and Daniele Nardi. 2014. Huric: a human robot interaction corpus. In LREC, pages 4519–4526.

Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. 2016. Do you see what i mean? visual resolution of linguistic ambiguities. arXiv preprint arXiv:1603.08079.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

Braiden Brousseau, Jonathan Rose, and Moshe Eizenman. 2020. Hybrid eye-tracking on a smartphone with cnn feature extraction and an infrared 3d model. Sensors, 20(2):543.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In IEEE conference on computer vision and pattern recognition, pages 248–255, Miami, Florida, USA.

Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. 2009. Modelling search for people in 900 scenes: A combined source model of eye guidance. Visual cognition, 17(6-7):945–978.

Fernanda Ferreira. 2003. The misinterpretation of noncanonical sentences. Cognitive Psychology, 47(2):164–203.

Nestor Garay-Vitoria and Julio Abascal. 2004. A comparison of prediction techniques to enhance the communication rate. In ERCIM Workshop on User Interfaces for All, pages 400–417, Vienna, Austria. Springer.

Ross Girshick. 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, Santiago, Chile.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 2012. Underspecification of cognitive status in reference production: Some empirical predictions. Topics in Cognitive Science, 4(2):249–268.

IFR. 2018. Executive summary world robotics 2019 service robots.

Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. 2015. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In Proceedings of the 23rd ACM international conference on Multimedia, pages 159–168.

Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. 2017. Gaze embeddings for zero-shot image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4525–4534, Honolulu, USA.

Mohamed Khamis, Florian Alt, and Andreas Bulling. 2018. The past, present, and future of gaze-enabled handheld mobile devices: survey and lessons learned. In Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, pages 1–17, Barcelona, Spain.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539.

Sigrid Klerke and Barbara Plank. 2019. At a glance: The impact of gaze aggregation views on syntactic tagging. In Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), pages 51–61, Hong Kong, China.

Nikolina Koleva, Martín Villalba, Maria Staudte, and Alexander Koller. 2015. The impact of listener gaze on predicting reference resolution. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, volume 2, pages 812–817, Beijing, China.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755, Zurich, Switzerland. Springer.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4185–4194, Long Beach, CA, USA.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 377–387, Vancouver, Canada.

Nikolina Mitev, Patrick Renner, Thies Pfeiffer, and Maria Staudte. 2018. Using listener gaze to refer in installments benefits understanding. In Proceedings of the 40th Annual Meeting of the Cognitive Science Society, pages 2122–2127, Madison, Wisconsin, USA.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2017. Hierarchical multimodal lstm for dense visual-semantic embedding. In The IEEE International Conference on Computer Vision (ICCV), Venice, Italy.

Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. 2019. Few-shot adaptive gaze estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 9368–9377, Seoul, Korea.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision, pages 2641–2649, Santiago, Chile.

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966.

Sol Rogers. 2019. Seven reasons why eye-tracking will fundamentally change vr. https://www.forbes.com/sites/solrogers/2019/02/05/-seven-reasons-why-eye-tracking-will-fundamentally-change-vr/. Accessed: 2021-01-05.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? Digging deeper into visio-linguistic pretraining. arXiv preprint arXiv:2004.08744.

Shengli Song, Haitao Huang, and Tongxiao Ruan. 2019. Abstractive text summarization using lstm-cnn based deep learning. Multimedia Tools and Applications, 78(1):857–875.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. arXiv preprint arXiv:2010.07891.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, San Francisco, CA, USA.

Niklas Wilming, Selim Onat, José P Ossandón, Alper Açık, Tim C Kietzmann, Kai Kaspar, Ricardo R Gameiro, Alexandra Vormberg, and Peter König. 2017. An extensive dataset of eye movements during viewing of complex images. Scientific data, 4(1):1–11.