university of groningen / faculty of mathematics and natural sciences / artificial intelligence

Sim2Real Transfer of Visiolinguistic Representations for Human-Robot Interaction

Graduation Project Proposal
(Computational Intelligence and Robotics)

Georgios Tziafas (s3913171)

January 25, 2021

Internal Supervisors: Dr. S.H. Mohades Kasaei (Artificial Intelligence, University of Groningen)
Dr. L.R.B. Schomaker (Artificial Intelligence, University of Groningen)

Artificial Intelligence / Human-Machine Communication
University of Groningen, The Netherlands

1 Introduction

Humans have the cognitive capacity to process multi-modal data (e.g. vision, language) and make cross-references between parts of the two modalities in real-time effortlessly. They are also very capable of identifying such cross-references in degenerate cases where one modality suffers from noise. However, this is not the case in robotics, where most commonly the visual perception and *Human-Robot Interaction (HRI)* modules are treated separately. In this regime, when verbal input is given to the robot (e.g. in the context of grasping: ”Grasp the *Mug*”) the referring phrases that correspond to objects to be segmented (aka *grounded*, e.g. ”*Mug*”) must be predefined explicitly and hard-coded in the agents behaviour. As a result, the agent is unable to comprehend variants of the predefined object category from the verbal input as it is often the case in real-world scenarios, where objects might be referred by their visual attributes or spatial relation to another object (e.g. ”Grasp the *red mug*” or ”Grasp the *mug* that is *next to* the *laptop*” in the case of multiple mug objects within view). A bridging of the two modules is viable through end-to-end deep learning models that are able to jointly process visual and text data and construct cross-modal representations for image-text pairs. Such representations are then used for the task of **Visual Grounding (VG)** (i.e localizing image regions to phrases in natural language) as well as other downstream vision-and-language tasks.

Learning visiolinguistic representations has been a central theme in recent deep learning literature, with an array of methods originating from different communities and for different applications. However, associated benchmark datasets struggle to cover generic, domain-agnostic content and thus rely on scale to be effectively transferable. The alternative would be to design a domain-specific dataset directed towards HRI applicability, something which comes at the great cost of having to manually collect data and provide linguistic annotations depending on the environment/scenarios that the agent is operating. On the other hand, robotic manipulation and planning have been drastically benefited by the **Sim2Real Transfer** paradigm, in which the agent is fine-tuned in a small-scale supervised dataset of real data after being pre-trained in synthetic environments with automatically generated labels/rewards. This method offers a platform for data-efficient, domain-specific solutions that omits the need for manual annotation by utilizing modern computer graphics and animation tools to simulate the real world.

Inspired by the success of Sim2Real Transfer in machine learning for other robotic domains, as well as by recent works that attempt to study the ability of neural networks to do language-based reasoning on images by training in synthetic data, we propose to apply this methodology in visiolinguistic data for HRI applications. To that end, we assume a scenario of a robot facing a surface with multiple objects in arbitrary arrangements and a human supervisor providing verbal commands for grasping/placing objects. We propose to simulate such an environment with 3D animation and create a synthetic scene dataset, for which we also generate linguistic annotations. An annotated vision-and-language dataset of real objects in similar setup will be collected from subsets of available robotics RGB-D data for evaluation. An array of neural network models will be implemented in order to establish benchmarks for our dataset as well as study the potential of Sim2Real serving as a representation learning method that can be generalized to other popular vision-and-language domains. Finally, we will develop a visual grounding agent based on our model in *Robot Operating*

System (ROS) and perform experiments with a simulated/real robot.

2 Theoretical Framework

Synthetic data generation As mentioned above, pre-training in synthetic datasets/environments is a broadly adapted practise for robotic manipulation, planning and 3D vision domains. For visiolinguistic data, *CLEVR* (Johnson et al., 2016) is a synthetic dataset of abstract object shapes in arbitrary spatial arrangements, equipped with question-answer pairs, used mainly to diagnose the visual reasoning abilities of established models. The *Abstract-Scenes* dataset (Zitnick, Vedantam, & Parikh, 2016) contains manually generated clipart animation scenes used for semantic scene understanding. *Text2Scene* (Tan et al., 2018) generates similar clipart scenes from given textual descriptions. In this work we wish to combine generative methods and 3D simulation to generate synthetic visiolinguistic data for HRI applications.

Vision-and-language benchmarks The core objective of any visiolinguistic model is to align image regions with referring (parts of) phrases. Other objectives can be then treated as downstream tasks learned by combining different classification layers and loss functions on top of the grounding backbone. In the general case, the task of VG addresses multi-query grounding, meaning all noun phrases within the input phrase are localized in different objects in the scene. Benchmarks for this version of the task include *Flickr30k Entities* and *ReferIt* (Plummer et al., 2015; Kazemzadeh, Ordonez, Matten, & Berg, 2014). A closely related task is that of **Referring Expressions Comprehension (REC)**, also referred to as *phrase grounding* or *phrase localization*, in which the entire input phrase is grounded in a single image region. Expressions can generally refer to abstract image regions and not just objects. Benchmarks include the *RefCOCO* suite (X. Chen et al., 2015). The task of **Visual Question Answering (VQA)** (Antol et al., 2015) addresses answering questions about an image, either as multiple choice selection or from a fixed vocabulary of answers. Visual reasoning is often studied as a diagnosing tool for VQA models through the tasks of **Natural Language for Visual Reasoning (NLVR)** and **Visual Entailment (VE)**. In these tasks the model has to judge the validity of a propositional statement or its entailment with the scene semantics. Benchmarks include *NLVR*^{2} and *SNLI-VE* respectively (Suhr, Zhou, Zhang, Bai, & Artzi, 2018; Xie, Lai, Doran, & Kadav, 2019). Most of these datasets also include labels for language-based **Image Retrieval (IR)**, which can be also tackled as a downstream task by grounding architectures. In the scope of this project, we will first perform our Sim2Real experiments in the VG/REC tasks with a baseline grounding model and then seek to augment the model / implement other ones to address more tasks.

Visual grounding Visual grounding architectures typically use LSTM encoders to contextualize phrases locally and then employ a matching objective between image region proposals and textual features (Sadhu, Chen, & Nevatia, 2019). Refined version use cross-modal self-attention mechanisms to capture long range dependencies in the two modalities (Ye, Rochan, Liu, & Wang, 2019). J. Liu and Hockenmaier (2019) formulate phrase grounding as a sequence labeling task where they treat candidate regions as potential labels, and use neural chain Conditional Random Fields (CRFs) to model dependencies among regions for adjacent mentions. More recent approaches extract scene graphs from images-text and tackle grounding as a structure prediction task where visual and textual **Graph Neural Networks (GNNs)** are used to contextualize each modality's representations and a graph similarity metric is introduced to prune the two graphs appropriately (Y. Liu, Wan, Zhu, & He, 2019).

Representation Learning This family of models is aimed to be used as pre-trained representation learners, inspired by the huge success of BERT-like architectures (Devlin, Chang, Lee, & Toutanova, 2019) in the field of NLP. Most popular architectures, including *VisualBERT* (L. H. Li, Yatskar, Yin, Hsieh, & Chang, 2019) and *ViLBERT* (Lu, Batra, Parikh, & Lee, 2019), are typically very parameter-heavy and are trained on massive-scale datasets with self-supervised objectives, functioning as visiolinguistic autoencoders. In such methods, the image is treated as a sequence of image region vectors and multi-modal representations are built with co-attention transformer encoder layers between the visual and textual features. Self-supervised objectives include masked language modeling, masked region modeling, word-region alignment and image-text matching. Newer works (Y. Chen et al., 2019; Lu, Goswami, Rohrbach, Parikh, & Lee, 2019) proposed a **multi-task setting**, in which a curriculum learning approach is followed in order to inject supervision from multiple tasks-domains during unsupervised pre-training. Single-task performance is then benefited from attempting to solve all tasks at once, even after minimal fine-tuning. On the downside, the enormous scale of these models grant them potentially inefficient for real-time application.

3 Research Questions

The proposed work attempts to answer the following research questions:

- Can we successfully apply Sim2Real transfer for visual grounding of natural language in a robotics domain?
- Does synthetic pre-training also improve the performance of state-of-the-art models in other vision-and-language domains?
- Can the resulting models serve as efficient visual grounding agents for online Human-Robot Interaction applications?

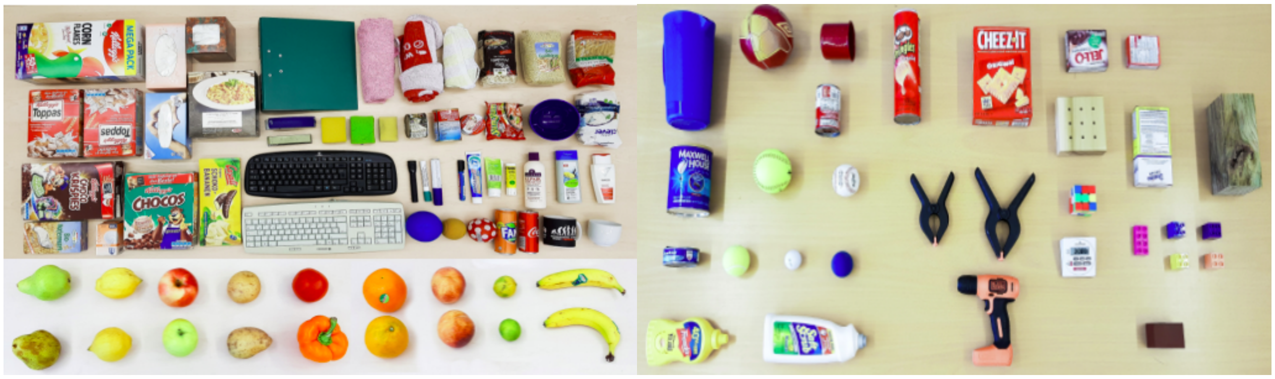
4 Methods

We address the above research questions by proposing the following methodology: First, we establish an HRI dataset of vision-and-language tasks by generating annotations from collected RGB-D data used for learning in robotics. We then apply Sim2Real Transfer by creating simulated scenes of objects similar to the dataset and pre-training our model for vision and language in the synthetic scenes. After model searching and hyper-parameter optimization, we establish benchmarks for our domain and perform domain adaptation experiments in benchmark datasets. Last, we develop software for a visual grounding agent and experiment with simulated/real robots in online HRI scenarios.

4.1 Dataset

In order to build a multi-task vision-and-language dataset for HRI, we will generate linguistic and structural annotations from already existing RGB-D datasets, taken from depth sensors in real scenes, most commonly used for 3D vision. An interesting candidate is the **RGB-D Scene Understanding Suite (SUN RGB-D)** (Song, Lichtenberg, & Xiao, 2015), containing $10,335$ raw RGB-D as well as mixed 2D/3D annotation samples taken by four different depth sensors in house-like environments (kitchen, bedroom etc.). Annotations in the object-level for this dataset contain typical household items - entities (chair, sofa, table, toilet, person etc.).

However, in this work we aspire to use the visiolinguistic model as an interface between a human and a robotic actuator in an online scenario. This suggests that we should strive for data representing simpler scenes picturing groups of graspable objects in different spatial arrangements. Such a dataset most commonly employed for manipulation tasks is the **Object Clutter Indoor Dataset (OCID)** (Suchi, Patten, Fischinger, & Vincze, 2019). OCID is a curated selection of the ARID and YCB (Calli et al., 2015) datasets, containing 96 scenes of 81 different objects placed in incremental number and cluttering overlap either on the floor or in a table. Data are taken from an ASUS Xtion camera placed in two viewpoints of the scenes (direct and top) and enumerate a total of 2346 labeled RGB-D samples. All object categories participating in different scenes are demonstrated in Figure 1. The object catalogue included in OCID scenes is particularly suitable for our experiments, as same objects might be appearing multiple times with different visual attributes (texture, shape, color), granting them very suitable targets for referring expression comprehension learning. The small size and simple controlled setup of the scenes makes it also a relatively easy startup point for simulated reconstruction.



(a) ARID object set

(b) YCB object set

Figure 1: Different object categories included in cluttering scenes in the OCID dataset.

4.2 Generating Annotations

In order to learn joint visual and linguistic representations for our scenes, we will need to generate natural language annotations describing the scenes and each object in relation with it's environment. Most specifically, in order to cover the full range of vision-and-language tasks we would want each sample in our dataset to have the following fields:

- **RGB image**. The RGB-D image data come packed with our dataset.

- **Bounding boxes**. A pixel-wise labeled mask describing the object segmentation of the image scene comes also packed with our dataset. We will need to perform some image processing steps to convert the pixel-wise annotations to bounding box frame coordinates, which will serve as the supervision signal for our grounding tasks.

- **Semantic captions**. Phrases that describe the semantics of the depicted scenes by referring to all objects and regions in full detail. This textual input is used for multi-query VG training. For generating captions we will utilize an image captioning model pre-trained in COCO Captions (X. Chen et al., 2015), with major candidates being Oscar (X. Li et al., 2020) and VLP (Zhou et al., 2019).

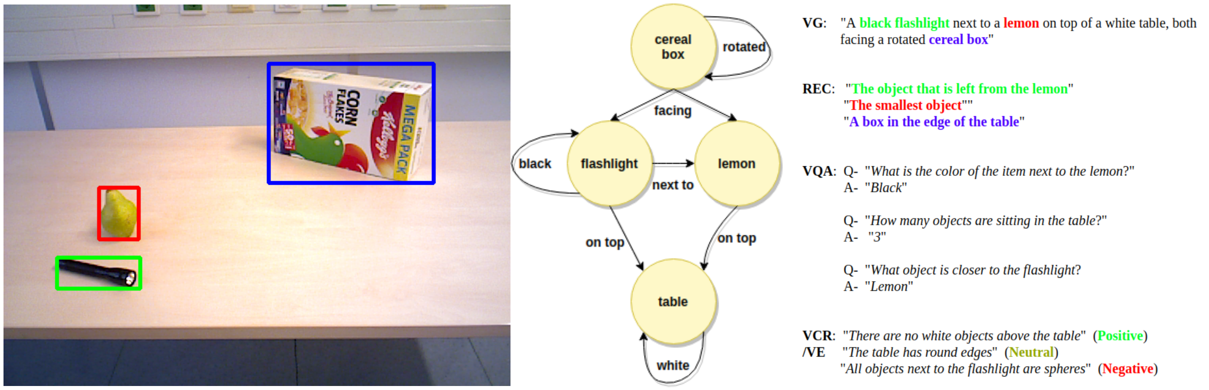
- **Referring expressions**. Phrases that can be grounded to a specific object in the scene by referring to its visual attributes or spatial relations to other objects. These data serve as input for REC training. We will generate multiple captions for each object in every scene and exhaust all possible variations of reference by employing a dense relational captioning model (Kim, Oh, Choi, & Kweon, 2020).

- **Scene graphs**. Scene graphs represent the structure of the depicted scene serving as a form of visual syntax. In such graphs usually objects are represented as nodes, relations between objects (relative location, action etc.) as edges and visual attributes as self-loops (modifiers). GNNs can then be employed to contextualize visual-textual features according to the graphs structure. Extracting scene graphs from images can provide dense graph and sub-graph representations with complex attribute/relation connections (Yang, Lu, Lee, Batra, & Parikh, 2018). However, since some of the graph-based VG methods also require language-driven scene graphs we will utilize a `Python` toolkit based on the *Stanford Scene Graph Parser* (Wang, Liu, Zeng, & Yuille, 2018) that generates scene graphs based on dependency parsing of the corresponding captions.

- **Question-answer pairs**. A question about the depicted scene paired with its answer, used for VQA training. There are no established generative methods for extracting QA samples from images. However, since we are only concerned with visual attributes and spatial relations, we will investigate an analytical method for traversing the extracted scene graphs into multiple QA pairs (e.g. for two connected nodes $\{\mathrm{mug}\}:^{behind} \rightarrow \{laptop\}$ we can generate the pair: [Q:"What is behind the laptop?" - A:"Mug"]).

- **Propositional statements**. A propositional statement about the depicted scene. These data can be generated as a special case of QA samples, where the question is replaced by a propositional statement and the pair answer is either True/False (used for NLVR) or Positive/Neutral/Negative (used for VE).

An example desired sample from annotated OCID (referred to as OCID*) is demonstrated in Figure 2. After annotating OCID and evaluating our annotations usage in building visiolinguistic models, we can repeat the process to generate annotations for bigger scale robotics datasets such as SUN RGB-D and ARID.



VG: "A black flashlight next to a lemon on top of a white table, both facing a rotated cereal box"

REC: "The object that is left from the lemon"
"The smallest object""
"A box in the edge of the table"

VQA: Q- "*What is the color of the item next to the lemon?*"
A- "*Black*"

Q- "*How many objects are sitting in the table?*"
A- "*3*"

Q- "*What object is closer to the flashlight?*
A- "*Lemon*"

VCR /VE: "*There are no white objects above the table*" (**Positive**)
"*The table has round edges*" (**Neutral**)
"*All objects next to the flashlight are spheres*" (**Negative**)

(a) RGB with groundings (b) Scene Graph (c) Gererated Text Samples

Figure 2: Hand-crafted example of structural and linguistic data we opt to generate from an RGB scene of OCID (Suchi et al., 2019)

4.3 Synthetic Data Generation

For generating our synthetic dataset we will make use of the `PyBullet` (Coumans & Bai, 2016–2019) module, which is an easy and fast framework for robotics simulation, especially flavoured towards Sim2Real transfer. We will recreate the setup of the OCID dataset, placing objects either on a rendered floor or a table. Unlike OCID and in order to be able to create robust visual representations for our data, we will include images taken from multiple viewpoints in a 360 range around the objects. Following OCID structure, we will also incrementally increase the number and cluttering overlap of different objects in the scene. We will either manually or by utilizing the provided RGB pictures of OCID objects recreate them in 3D and make sure to include visual variations (object in different color, shape, size, texture) in a systematic manner, serving as a form of data augmentation. An example of our `PyBullet` setup along with the reconstructed image data from the simulated cameras are demonstrated in Figure 3.

We will then repeat the process of the previous section to generate linguistic and structural annotations for our simulated scenes. For each crop in each frame, we will generate multiple referring expressions as well as question-answer pairs and propositions about the target object. The entire

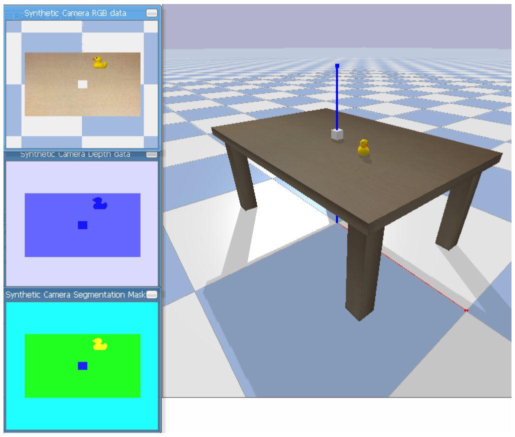

Synthetic Camera RGB data
Synthetic Camera Depth data
Synthetic Camera Segmentation Mask

Figure 3: Example of simulated setup.

synthetic data generation process can be parameterized and automated so with minimal human supervision a new set of scenes - annotations can be generated from a given set of 3D objects.

4.4 Models

We will perform our Sim2Real experiments with the baseline learning architecture demonstrated in Figure 4, adapted from a zero-shot grounding model (Sadhu et al., 2019). Each modality is first encoded implicitly and the encoded features are aligned and concatenated to serve as multimodal representations. The visual module utilizes a Faster R-CNN (Ren, He, Girshick, & Sun, 2015) to extract visual features and make image region proposals. We replace the pre-trained *Glove* embeddings (Pennington, Socher, & Manning, 2014) with a pre-trained BERT encoder for English (Devlin et al., 2019) and the Bi-LSTM phrase contextualization layer with a Transformer encoder layer (Vaswani et al., 2017). By appropriately modifying the classification layer and the objective functions at the output of the fusion module we can adapt the network to tackle other vision-and-language tasks. The zero-shot nature of this model allows for further inspection of the quality of our pre-trained representations by evaluating in the OCID* tasks without fine-tuning on the actual dataset of real objects.

In order to establish benchmarks in OCID* as well as explore the domain adaptation potential of our synthetic data pre-training we will compare the performance of state-of-the-art models in visiolinguistic learning (see Section 2). We will start by implementing one benchmark model per family of methods (attention-based alignment, context graph networks, pre-trained transformer encoders) and employ the multi-task setting when applicable.

4.5 Evaluation

The evaluation of our methodology comes in three different stages. First, we wish to evaluate the proposed Sim2Real Transfer method in a HRI domain. We will do so by comparing the performance of our baseline model in OCID* tasks with and without pre-training in the synthetic scenes. We will

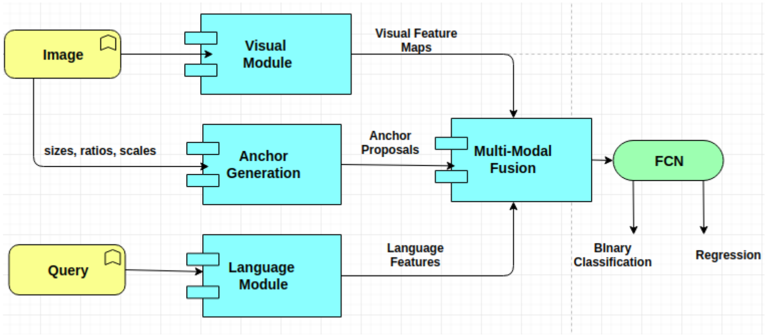


Figure 4: Visual grounding architecture used as baseline model.

perform ablation studies for different orders of fine-tuning needed (zero-shot, few-shot, some % the labels etc.) in order to further evaluate the generalization capacity of our pre-trained visiolingistic representations.

Second, we wish to establish benchmarks for our designed HRI domain. We will implement (a subset of) the models mentioned in the previous section and evaluate them with and without synthetic pre-training. In cases of representation learning architectures (such as VILBERT) that come pre-trained in large annotated corpora we will use synthetic data as an intermediate training step before fine-tuning to OCID*. After we establish the best performing models, we will evaluate the domain adaptation potential of our Sim2Real method by fine-tuning and reporting test scores on a full suite of domains, as shown in Table 1. If time allows it (after finishing the next step) and given the (semi-) automation of the annotation generation process, we can also establish benchmarks in vision-and-language tasks for other RGB-D robotics datasets (SUN RGB-D, ARID).

Last, we wish to asses the applicability of our methods in an online human supervision scenario. We will integrate our models in ROS and perform online experiments with the implemented grounding agent, during which the supervisor manually provides input captions with variations in ambiguity (object name, refer by shape, refer by color, plural etc.) and inspects the model's groundings to qualitatively evaluate its generalization capacity. Different models will be evaluated in terms of speed performance for real-time application. We will record a demo of a human supervisor interacting with a a robotic actuator for grasping and placing objects, displaying the agent's performance on grounding referring expressions and resolving ambiguities through question-answering and visual entailment.

Task	Datasets
VG/REC	RefCOCO, RefCOCO+, RefCOCOg Visual 7W, Guess What, VG-Flickr30k
VQA	VQA, VG QA, GQA
NLVR/VE	NLVR^2, SNLI-VE
IR	IR COCO, IR-Flickr30k

Table 1: Vision-and-language tasks and associated datasets available for evaluation

5 Scientific Relevance for Artificial Intelligence

The main body of this work focuses on the application of the Sim2Real paradigm in vision and language learning for HRI applications, making it scientifically relevant to the intersection of deep learning and robotics. The suggested experiments will showcase if such an approach can be realized in a HRI domain, providing an alternative to the often data-hungry and domain-agnostic established methods. Most specifically, this work aims at making the following contributions:

- Establish a new benchmark dataset of vision-and-language tasks for HRI, equipped with linguistic and structural annotations generated from already existing RGB-D datasets.
- Create a synthetic data generation pipeline for visiolinguistic Sim2Real pretraining. The implemented `PyBullet` code can be re-parametrized by a user to generate new synthetic samples for refined applications (e.g different set of objects), therefore eliminating the need for manual annotation of large datasets.
- Explore the effectiveness of synthetic pre-training in state-of-the-art methods for vision-and-language domains outside robotics
- Build a visual grounding agent available for use in online HRI scenarios.

6 Planning

Our methodology is divided into eight work parts:

- WP1 Bridge gaps in efficient annotation generation and vision-and-language learning methods by studying literature.
- WP2 Build OCID* by annotating the OCID dataset.
- WP3 Generate OCID-like synthetic data in `PyBullet`.
- WP4 Evaluate Sim2Real Transfer of baseline model in OCID*
- WP5 Establish benchmarks for OCID*
- WP6 Evaluate implemented models in other domains
- WP7 Develop a visual grounding agent in ROS and perform experiments in a simulated/real HRI scenario
- WP8 Write thesis and prepare presentation

The timeline of this research, showing the proposed tasks execution plan and interaction between different tasks, is presented in Figure 5

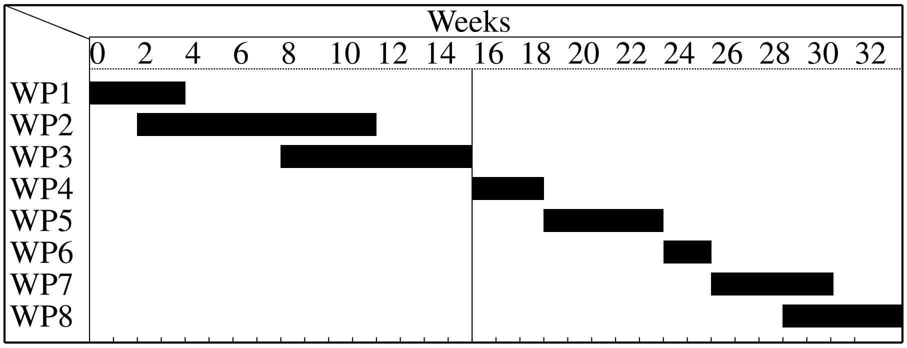


Figure 5: Duration of Work Parts

7 Resources and Support

For creating the synthetic data, extracting the scene graphs, performing all necessary data preprocessing steps as well as implementing the network architectures, I will use my personal computer from home. For all model pre-training, fine-tuning and validation steps, especially in the multi-task setting, I will utilize the virtual Nvidia Tesla v-100 nodes of the Peregrine HPC Cluster for GPU-acceleration. For integrating the trained models in an online HRI demo I will work with my supervisor on a simulated robotic manipulator operated remotely from a university desktop through AnyDesk and/or potentially with the real robot in RUG (if not prohibited due to COVID pandemic).

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: visual question answering. *CoRR*, *abs/1505.00468*. Retrieved from http://arxiv.org/abs/1505.00468

Calli, B., Walsman, A., Singh, A., Srinivasa, S., Abbeel, P., & Dollar, A. M. (2015). Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics Automation Magazine*, *22*(3), 36-52. doi: 10.1109/MRA.2015.2448951

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *CoRR*, *abs/1504.00325*. Retrieved from http://dblp.uni-trier.de/db/journals/corr/corr1504.html#ChenFLVGDZ15

Chen, Y., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., . . . Liu, J. (2019). UNITER: learning universal image-text representations. *CoRR*, *abs/1909.11740*. Retrieved from http://arxiv.org/abs/1909.11740

Coumans, E., & Bai, Y. (2016–2019). *Pybullet, a python module for physics simulation for games, robotics and machine learning.* http://pybullet.org.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N19-1423 doi: 10.18653/v1/N19-1423

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. B. (2016). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, *abs/1612.06890*. Retrieved from http://arxiv.org/abs/1612.06890

Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. L. (2014). Referit game: Referring to objects in photographs of natural scenes. In *Emnlp*.

Kim, D.-J., Oh, T.-H., Choi, J., & Kweon, I. S. (2020). *Dense relational image captioning via multi-task triple-stream networks.*

Li, L. H., Yatskar, M., Yin, D., Hsieh, C., & Chang, K. (2019). Visualbert: A simple and performant baseline for vision and language. *CoRR*, *abs/1908.03557*. Retrieved from http://arxiv.org/abs/1908.03557

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... Gao, J. (2020). *Oscar: Object-semantics aligned pre-training for vision-language tasks.*

Liu, J., & Hockenmaier, J. (2019). Phrase grounding by soft-label chain conditional random field. *CoRR*, *abs/1909.00301*. Retrieved from http://arxiv.org/abs/1909.00301

Liu, Y., Wan, B., Zhu, X., & He, X. (2019). Learning cross-modal context graph for visual grounding. *CoRR*, *abs/1911.09042*. Retrieved from http://arxiv.org/abs/1911.09042

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, *abs/1908.02265*. Retrieved from http://arxiv.org/abs/1908.02265

Lu, J., Goswami, V., Rohrbach, M., Parikh, D., & Lee, S. (2019). 12-in-1: Multi-task vision and language representation learning. *CoRR*, *abs/1912.02315*. Retrieved from http://arxiv.org/abs/1912.02315

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *In emnlp*.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the 2015 ieee international conference on computer vision (iccv)* (p. 2641–2649). USA: IEEE Computer Society. Retrieved from https://doi.org/10.1109/ICCV.2015.303 doi: 10.1109/ICCV.2015.303

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, *abs/1506.01497*. Retrieved from http://arxiv.org/abs/1506.01497

Sadhu, A., Chen, K., & Nevatia, R. (2019). Zero-shot grounding of objects from natural language queries. In *Proceedings of the ieee international conference on computer vision* (pp. 4694–4703).

Song, S., Lichtenberg, S. P., & Xiao, J. (2015). Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Cvpr* (p. 567-576). IEEE Computer Society. Retrieved from http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#SongLX15

Suchi, M., Patten, T., Fischinger, D., & Vincze, M. (2019). Easylabel: A semi-automatic pixel-wise object annotation tool for creating robotic RGB-D datasets. In *International conference on robotics and automation, ICRA 2019, montreal, qc, canada, may 20-24, 2019* (pp. 6678–6684). Retrieved from https://doi.org/10.1109/ICRA.2019.8793917 doi: 10.1109/ICRA.2019.8793917

Suhr, A., Zhou, S., Zhang, I., Bai, H., & Artzi, Y. (2018). A corpus for reasoning about natural language grounded in photographs. *CoRR*, *abs/1811.00491*. Retrieved from http://arxiv.org/abs/1811.00491

Tan, Fuwen, Feng, Song, Ordonez, & Vicente. (2018, 09). Text2scene: Generating abstract scenes from textual descriptions.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, Y., Liu, C., Zeng, X., & Yuille, A. L. (2018). Scene graph parsing as dependency parsing. *CoRR*, *abs/1803.09189*. Retrieved from http://arxiv.org/abs/1803.09189

Xie, N., Lai, F., Doran, D., & Kadav, A. (2019). Visual entailment: A novel task for fine-grained image understanding. *CoRR*, *abs/1901.06706*. Retrieved from http://arxiv.org/abs/1901.06706

Yang, J., Lu, J., Lee, S., Batra, D., & Parikh, D. (2018). Graph R-CNN for scene graph generation. *CoRR*, *abs/1808.00191*. Retrieved from http://arxiv.org/abs/1808.00191

Ye, L., Rochan, M., Liu, Z., & Wang, Y. (2019). Cross-modal self-attention network for referring image segmentation. *CoRR*, *abs/1904.04745*. Retrieved from http://arxiv.org/abs/1904.04745

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. J., & Gao, J. (2019). Unified vision-language pre-training for image captioning and VQA. *CoRR*, *abs/1909.11059*. Retrieved from http://arxiv.org/abs/1909.11059

Zitnick, C., Vedantam, R., & Parikh, D. (2016, 09).
IEEE Trans Pattern Anal Mach Intell. 2016;38(4):627-638.