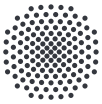


University of Stuttgart
Germany

Institute for Visualization and Interactive Systems
Universitätsstraße 38
70569 Stuttgart

Research Project - InfoTech

Transformer Based Architecture for Belief Prediction in Object-Context Scenarios

John Pravin Arockiasamy

Course of Study:	InfoTech
Examiner:	Prof. Dr. Andreas Bulling
Supervisor:	Matteo Bortoletto M.Sc.
Commenced:	April 3, 2023
Completed:	October 4, 2023

Abstract

In the pursuit of enabling artificial intelligence to understand human beliefs and intentions through nonverbal communication for the purpose of improving human-computer interaction, a recent study by Duan et al has introduced a Benchmark for Human Belief Prediction in Object-context Scenarios (BOSS) dataset [DYT+22b]. The BOSS dataset bridges the gap between the fields of theory of mind and object-Context relations within cognitive science, offering a platform for modeling human beliefs and intentions in object-context scenarios. In this study, we conduct a comprehensive examination of the BOSS dataset, with a particular focus on its input modalities. Earlier attempts by Duan et al. to predict human beliefs using baseline deep learning models revealed limitations in achieving satisfactory accuracy within the BOSS dataset. To address this limitation, we propose the utilization of transformer-based models to enhance belief prediction. Additionally, we investigate the influence of diverse feature extraction techniques when coupled with transformer-based models across different input modalities within the BOSS dataset. Furthermore, we introduce a novel cost function derived from the dataset, which we evaluate alongside the transformer-based models, shedding valuable insights into its effectiveness. Our findings demonstrate that transformer-based models significantly outperform baseline models in predicting human beliefs within the BOSS dataset, though not reaching the highest levels of performance. This underscores the need for further research in this domain while also offering a promising avenue for advancing our understanding of human beliefs in object-context interactions.

Contents

List of Figures

List of Tables

List of Listings

List of Algorithms

1 Introduction

Nonverbal cues like gaze and posture are crucial in human-human communication. As a result, artificial intelligence (AI) should possess the capability to comprehend these nonverbal signals. Nonetheless, a significant difference between humans and AI technology is our ability to infer someone's belief through nonverbal communication, where direct verbal communication is limited [Hin74; Sax06; WCW01]. Remarkably, a high-quality collaboration among individuals can occur even in noisy situations (e.g. a crowded market or a loud office), using nonverbal communication signals such as gaze, gesture, pose, and context. This phenomenon highlights the distinctive human ability to interpret nonverbal communication. Integrating the ability of AI models to infer human beliefs through nonverbal signals could significantly benefit diverse fields such as robot perception [Lee17], computer vision [SPP18], and embodied AI [DYT+22a]), ensuring safe human-robot collaboration and interaction. As a result, substantial research has been dedicated to bridging the gap between AI and humans' ability in non-verbal communication. Nonetheless, accurately modeling human beliefs and intentions remains a challenge in this endeavor.

According to cognitive science research, the ability to infer a person's belief states within a social context can be divided into two fundamental components: *Theory of Mind*, which includes the ability to comprehend and consider another person's beliefs [BLF85; CCT02], and *object-context relations*, which includes the comprehension of objects within specific context [MA11; OJG10]. In preceding years, researchers tackled the challenge of interpreting beliefs and intentions through *Machine Theory of Mind*, often focusing on simplified settings like 2D grid worlds [RPS+18] or specialized games for reinforcement learning [FWCL21; NVNT20]. Consequently, there's a scarcity of real-world data of sufficient quality [DYT+22b]. However, a recent study from the Human Belief Prediction in Object-context Scenarios (BOSS) combined the knowledge of the aforementioned two fundamental components, in the process of modeling human beliefs and intentions [DYT+22b].

Modern AI models have attained a substantial degree of efficacy, enabling them to recognize patterns in a variety of data types including textual, visual, audio, and more. This progress is superficially demonstrated in the work by Duan et al. [DYT+22b], which introduces the Benchmark for Human Belief Prediction in Object-context Scenarios (BOSS) dataset. This dataset serves as a support for training AI models, allowing for the modeling of human beliefs within real-world object context scenarios. In this context, this dataset includes an object-context situation-based video dataset featuring a pair of participants (right and left) collaborating to execute a task involving the selection of objects based on a context through nonverbal communication by inferring and interpreting each other's beliefs. Alongside the video inputs, Duan et al. utilize a variety of deep learning models to extract additional information from the videos, including aspects like participants' pose coordinates, participants' gaze coordinates, and object bounding boxes, with the intention of enhancing the efficiency of the task. Furthermore, Duan et al. used the training samples to create a matrix referred to as the Object-Context Relation (OCR) matrix. This matrix explains the interconnection between the objects placed before the left participants and those placed before the right participants. The ultimate objective is to develop a model capable of accurately predicting

the beliefs of the participants featured in the dataset, utilizing multimodal inputs. To achieve this, Duan et al. deployed certain baseline deep learning models, fine-tuned to capture the beliefs of the participants. However, these models prove to be suboptimal in performance due to their modest design and limited capabilities.

The purpose of this research project is to enhance the accuracy of predicting the beliefs of participants within the BOSS dataset through exploration across the inputs, architecture designs, and cost function domains. Initially, we intend to examine the different input modalities such as participants' pose coordinates, participants' gaze coordinates, object bounding boxes, and OCR matrix derived from the deep learning models and training samples, respectively. Subsequently, we investigated the impact of architectural changes, including transformer-based models and various feature extraction techniques, on the extent of predictive accuracy improvement. Additionally, we introduced a novel object-context-based cost function, designed to capture the relationship between the beliefs of the participants. The intention is to determine whether this approach contributes to an improvement in predictive accuracy, although it was found to be rather ineffective.

2 Related Work

2.1 Transformer Neural Network

The transformer is a deep learning model that, since its introduction in the publication "Attention is All You Need" by [VSP+17], has significantly impacted most natural language processing (NLP) applications and various other areas. Unlike traditional recurrent neural networks (RNNs) or convolutional neural networks (CNNs), transformer models employ a set of mathematical operations called attention or self-attention to capture intricate patterns and dependencies within the input data. These mechanisms allow the model to weigh the importance of each element in the input sequences in relation to all the others, allowing it to successfully capture long-range dependencies and context. The transformer's core innovation is the ability to parallelize computations, enabling it highly efficient for training on large datasets.

One of the core components within the transformer architecture is the self-attention transformer. In the self-attention mechanism, each input element (e.g., a word in a sentence) interacts with all other elements in the sequence. This interaction is utilized to generate weighted representations for each element, emphasizing its relevance to other elements of the sequence. The self-attention mechanism can capture relationships between words in a sentence, making it particularly well-suited for natural language processing tasks, such as translation and sentiment analysis. The Figure 2.1 illustrates schematics of the transformer's architecture.

Another crucial element of the transformer architecture is the cross-attention transformer which allows the transformer to handle multiple sets of input data. It includes both a *query* set and a *key* set, allowing the model to understand how elements in one set relate to elements in the other. This is especially useful in tasks where the model must process information from two different sources such as machine translation. The cross-attention mechanism enables the model to successfully align and transfer information across the two sets of data, resulting in increased translation quality and performance in a variety of tasks involving multiple data sources.

Following the success of transformers in NLP fields, a new transformer model known as Vision Transformer (ViT) was developed by [KDW+21] as an application to the field of computer vision. Since its creation, the vision transformer has substantially enhanced image recognition tasks. The core idea of the vision transformer is to break down the input images as a sequence of patches, flattening the patches into a 1D sequence. Each patch is then linearly embedded and handled as a token, similar to how words are treated in the original transformer.

The development of ViT has provided a crucial foundation for building vision models, leading to some of the most notable applications such as image classification [CFP21; KDW+21], segmentation [YRLW19], object detection [CMS+20; ZSL+20], image generation [CWG+21] and text-image

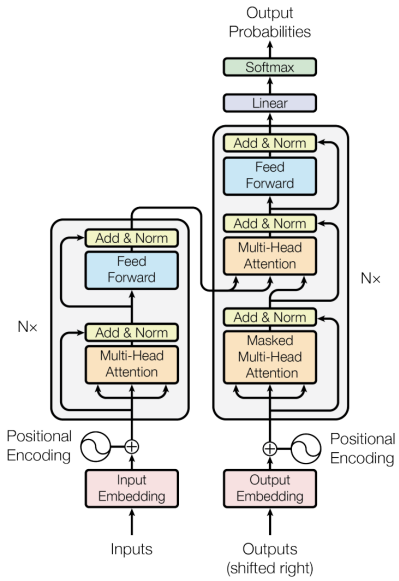


Figure 2.1: The Schematic of the encoder-decoder structure of the transformer architecture from the [VSP+17].

synthesis [RBHP21], among others. Recently, many researchers have explored ViT in new methods to handle high-resolution images [ZDY+21], and in conjunction with the convolution layers to solve image analysis tasks [CMS+20; ZSL+20].

Vision transformers provide a powerful mechanism for capturing global dependencies and contextual understanding in videos, resulting in improved performance in specialized tasks. However, in order to train on a wide scale, ViT requires a huge dataset, which is computationally expensive. Nonetheless, ViT are powerful models for capturing long-term dependencies while having a small number of layers, making them relatively less computationally expensive.

2.2 Theory of Mind and Object-Context Scenarios

In the realm of human cognitive science, researchers break down this social cognitive ability, which involves understanding others' beliefs and deducing their intentions in a social context, into two fundamental ideas: theory of mind (ToM) and object-context scenarios, which are integral to exploring social interaction along with our understanding of the world. ToM is a crucial concept that describes how each individual has distinct intentions, emotions, and feelings that might vary [PW78]. This understanding aids in the insight that individuals' actions are influenced by their beliefs or intentions, which can be crucial for effective communication, empathy, and cooperation. This, in turn, provides substantial advantages in the areas of philosophy of mind, artificial intelligence, and cognitive neuroscience. Additionally, this concept proves beneficial in the area of developmental psychology by aiding in determining the point at which children acquire the ability to comprehend the perspectives of others [Bin05; FBP+15; Sla15]. Object-context scenarios is an important concept that impacts human beliefs about an object [BMA11] [Ho87; OJL10]. For example, human beliefs about an object might change if it was placed in a kitchen or workshop. In this way, this

contextual information serves as a guide for our understanding of the object's purpose, function, and behaviors in a certain context. In a nutshell, object-context scenarios involve perceiving an object within a particular context. This context can offer both direct and indirect information that alters our thoughts and perceptions regarding the object's functionality. The intersection of the theory of mind and object-context scenarios concepts becomes clear when studying how people deduce the thoughts and intentions of others through contextual awareness. An individual with a refined theory of mind can extrapolate another person's thoughts and intents, deducing the motivations behind their actions by observing interactions with an object in a certain situation.

In the field of AI and robotics, the integration of ToM and object-context scenarios holds significant importance, especially when developing tools for human interaction. AI tools that can comprehend not just their physical environment but also anticipate human intentions and beliefs in specific contexts might allow both natural and effective interactions. This idea would significantly improve human-robot interaction, artificial intelligence, and various other fields.

2.3 Machine Theory of Mind

Modeling human beliefs is considered to be one of the difficult challenges in the deep learning era. Inference of others' mental states or beliefs utilizing Bayesian inverse planning [BST09; UBM+09], Partially Observable Markov Decision Processes (POMDP) [BS11; DQGY10], theory-based modeling of social goals [BGT08; KUT+13], and reinforcement learning [HRAD16; WKYL11] focus mostly on a 2D grid-world context. However, our research project aim is to focus on predicting belief states in the real-world object context scenario.

2.4 Inferring Beliefs in Nonverbal Communication

Benchmarks in the field of machine theory of mind mainly rely on artificial toy environments [BS11; UBM+09]. Unfortunately, there is a notable lack of real-world testbeds in this area. Notably, two significant works in this direction have emerged lately. Fan et al. focused on belief dynamics prediction, utilizing a multimodel video dataset capturing social interactions using nonverbal communication between two individuals [FQZ+21]. Fan et al. used a hierarchical energy-based model to predict the dynamics of belief.

In contrast, Duan et al. introduced a multimodal video dataset featuring social interaction with nonverbal communication between a pair of participants, with a focus on predicting belief in collaborative tasks within object-context scenarios [DYT+22b]. This paper, along with the dataset, included some baseline models. However, these models have proved to be ineffective due to their simplistic model architecture designs.

Our research project proposes a new method and evaluates its performance on the BOSS dataset, with a specific interest in belief prediction. Recent developments have demonstrated the success of transformer-based designs in handling multimodal inputs [XZC22]. Consequently, our research project aims to investigate the utilization of transformer-based architectures to capture the spatio-temporal relationships within multimodal inputs and enhance the accuracy of belief prediction.

3 The BOSS Dataset

Duan et al. have presented a dataset known as BOSS, which stands for Benchmark for Human Belief Prediction in Object-context Scenarios [DYT+22b]. The BOSS dataset, which comprises 3D video content, captures the collaboration efforts of two individuals as they work together to accomplish a task by inferring or deducing each other's beliefs in the process. The subsequent sections offer a detailed explanation of the process of data collection and the various multimodal inputs provided in accordance with the paper's content for further understanding.

3.1 Data Collection

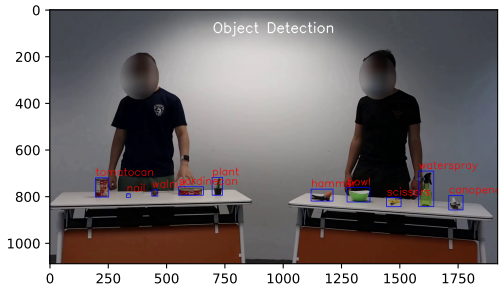
The BOSS dataset (Benchmark for Human Belief Prediction in Object-context Scenarios), as outlined by Duan et al., is a 3D multimodal video [DYT+22b]. This dataset comprises a total of 900 videos (347,490 frames) of ten pairs of participants engaged in collaborative efforts across 15 different contexts to execute tasks aimed at comprehending each other's beliefs through nonverbal communication. However, within the released version, there are only a total of 300 videos or experiments that have been divided into 180 training experiments, 60 test experiments, and 60 validation experiments. Each pair of participants, comprising a left and a right participant, stands before a table featuring a collection of objects. The left participant has a selection of contextual objects, namely Chips, Magazine, Chocolate, Crackers, Sugar, Apple, Wine, Potato, Lemon, Orange, Sardines, TomatoCan, Walnut, Nail, and Plant. Whereas, the right participant has a set of context-related objects, including the WineOpener, Knife, Mug, Peeler, Bowl, Scissors, ChipsCap, Marker, WaterSpray, Hammer, and CanOpener. The selection of these objects is based on their common presence in households, and most of them are part of the Yale-CMU-Berkeley (YCB) dataset [CSW+15]. The YCB dataset includes a wide range of everyday items, featuring diverse characteristics such as shapes, sizes, textures, weight, and rigidity. Additionally, it includes numerous commonly used manipulation tests, allowing for possible replication of this setup in other robotic applications.

In each experimental iteration, an implicit context is provided to the left participant, for instance, "circle the words on the magazine's cover". The left participant then picks the object from the table that corresponds to this implicit context, such as the magazine. Following that, the right participant takes on the task of determining the correct object based on the nonverbal signs provided by the left participant, and related to the object picked earlier, like the marker. This experiment will be repeated until the left participant selects the object relevant to the specific context and the right participant selects the correct object that accurately identifies the context provided. It is important to note that these experimental interactions are conducted without scripted guidance. Duan et al. used an essential process to capture the necessary information for accurately annotating participants' hidden beliefs in relation to the frames. During data collection, participants wore noise-canceling headphones with white noise playing to ensure confidentiality. Subsequently, participants verbally

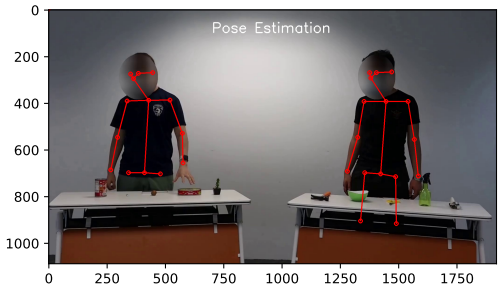
say out the objects on their minds that align with their beliefs, anytime their beliefs are updated. An annotator used the video clips of the experiment as well as audio clips featuring participants voicing out their inner beliefs to obtain a precise ground truth label at every frame. However, it is important to note that even with this approach, achieving 100% accuracy in labeling the belief state for each frame is challenging due to human reaction time. Nevertheless, this approach significantly enhances the accuracy of the ground truth labels, given that they are directly from the participants themselves. Finally, the ground truth or label within this dataset will be the beliefs of both left and right participants. These beliefs will correspond simply to the 26 objects highlighted in the previous paragraph, accompanied by a "None" class. Altogether, there will be a total of 27 distinct classes of participants' beliefs.

3.2 MultiModal Inputs

In addition to the video input, Duan et al. extracted multimodal inputs such as participants' pose coordinates, participants' gaze coordinates, and object bounding boxes through a subsequent processing approach based on advanced deep learning models [DYT+22b]. As depicted in Figure 3.1, the "OpenPose" model is utilized to capture critical points necessary for estimating the posture of both the left and right participants [CSWS17]. Meanwhile, the "Gaze360" model is used to capture the 3D gazes of both the left and right participants [KRS+19]. Furthermore, the "Detecto" model is employed to identify the bounding boxes encompassing the objects present on the tables [Ala19]. Additionally, the Object-Context Relation (OCR) matrix is created using information acquired from the training samples, which illustrates the interconnection between the contextual objects and the context-related object. This matrix is particularly important in the training since it serves as a type of previous knowledge about the links between the objects with each other.



(a) The Detecto model's object bounding box prediction overlay on a frame.



(b) The OpenPose model's participants' pose coordinates prediction overlay on a frame.

Figure 3.1: Visualization of Participants' pose and objects bounding box coordinates.

The BOSS dataset consists of individual frames, each measuring (1088, 1920, 3), featuring 5 contextual objects on the left table and an additional 5 contextually relevant objects on the right table. Every experiment within this dataset includes two participants, referred to as 'left' and 'right' participants. Initially, the participants' beliefs remain as "None" for certain frames. Subsequently, their belief corresponds to one of the 10 objects situated in one of the tables, guided by the contextual cues provided by Duan et al. Alongside this, Duan et al. utilized various state-of-the-art deep learning models to obtain additional input modalities such as the participants' poses, gazes, and

bounding boxes of the objects in each frame. The pose of both left and right participants is extracted by using the OpenPose (BODY_25) model, which outputs 25 sets of 3D joint coordinates (x, y, confidence score), forming a matrix of dimensions (2, 25, 3) per frame. The gaze of both left and right participants is captured by the Gaze 360 model, generating 3D eye coordinates (x, y, z) resulting in a size of (2,3). The gaze coordinates are defined in relation to the camera's perspective: the positive x-axis indicates a gaze towards the left, the positive y-axis denotes an upward gaze, and the positive z-axis signifies a gaze away from the camera. The coordinates of the bounding boxes of the objects in front of the participants' tables are extracted using the Detecto model. The objective of this model is to find the four coordinates of ten bounding boxes of the objects within a frame, resulting in a dimension of (10, 4). However, due to certain objects not being recognized by the model, the dimensions of these bounding boxes might vary per frame.

It is worth mentioning that the OpenPose and Gaze360 models do not require fine-tuning to accurately estimate the participants' poses and gazes for each frame in this dataset. These models can be directly applied to deduce the participants' poses and gazes. In contrast, the Detecto model requires fine-tuning since some objects in the dataset are absent from its pre-existing training data. To address this, Duan et al. manually annotated bounding boxes, corresponding to the objects, for a subset of experiments. These annotated labels were then used to fine-tune the Detecto model, enabling it to detect the objects for the remaining experiments of the dataset.

Additionally, Duan et al. utilized the training datasets to generate a matrix known as an Object-Context Relation (OCR) matrix [DYT+22b]. This matrix depicts the relationships existing between the various objects within the dataset, resulting in a matrix of dimensions (27, 27), comprising 26 objects and the "None" object. In each experiment within the training datasets, there is an implicit context provided by Duan et al. to the participants that involves a contextual object (those on the left table) and another from the context-related objects (those on the right table). By observing this implicit context, the OCR matrix values can be calculated for each and every object in the dataset. For example, consider the scenario where the object "Magazine" occurs in 20 experiments, with "Marker" appearing in 10 of those and "Scissors" in the remaining. In this scenario, the OCR matrix value between the "Magazine" and "Marker" would be 0.5, just as the value between the "Magazine" and "Scissors" would also be 0.5. Additionally, the value between "Magazine" and other objects (excluding marker and scissors) would be 0.

In order to better understand the OCR matrix, two key aspects should be noted. Firstly, within the experiment in this dataset, contextual objects are linked to one or more context-related objects, but not to other contextual objects. The reason is that each of the 15 contextual objects has no interrelationships among them. Finally, context-related objects are only associated with themselves, as they are not directly tied to the provided context. Given that the OCR matrix contains information about object-context relations, it can be used as a foundation for a novel cost function, instead of providing it as an input matrix to the model. The utilization of this matrix as a cost function will be further explained in the later sections.

4 Method

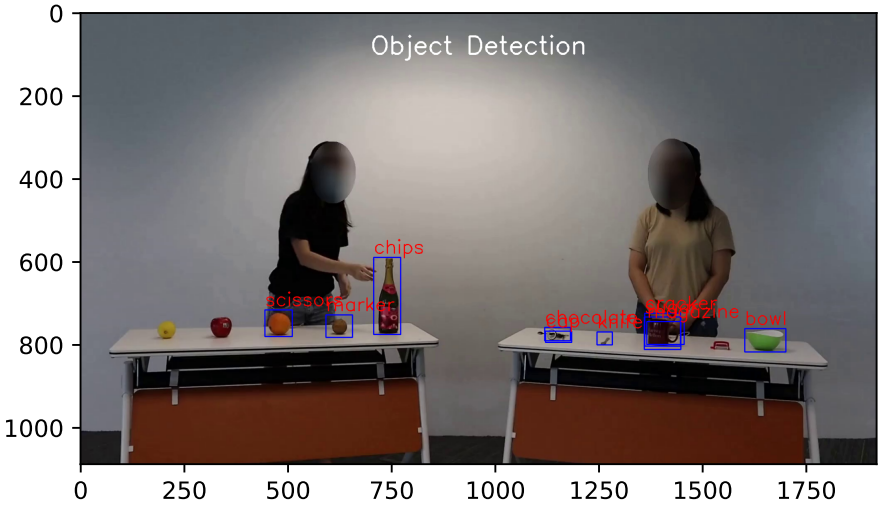
In this section, we will explore the approaches used in this study to increase the accuracy of predicting the beliefs of the participants using the BOSS dataset. As previously stated, the beliefs of the participants across all the experiments in the BOSS dataset correspond to one of the 27 objects, thereby making it a multiclass classification problem. We will assess the quality of the input modalities provided within the BOSS dataset. Furthermore, We will conduct a thorough examination of transformer-based models, namely the self-attention model and hierarchical cross-attention model, on top of a set of different feature extractors for input modalities. This will be discussed comprehensively in the below subsections.

4.1 Assessing the Quality of the Inputs Modalities

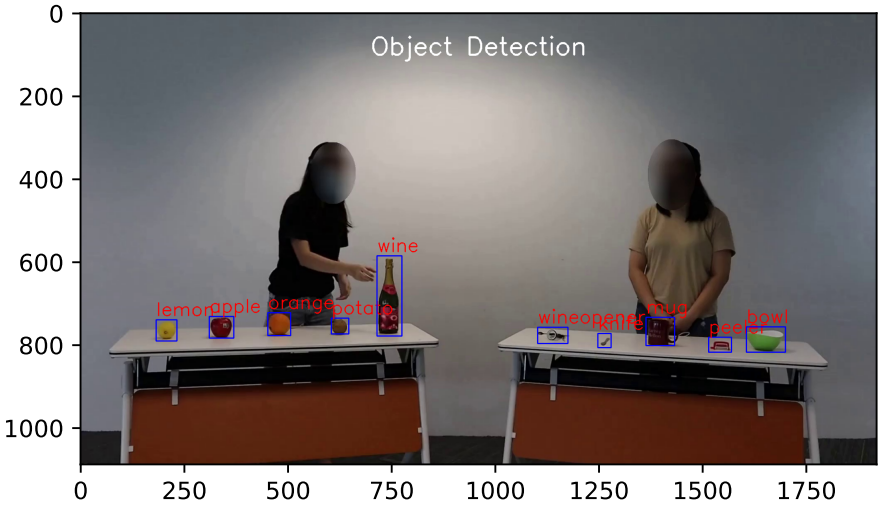
To assess the quality of the multimodal input modalities produced by state-of-the-art deep learning models, we processed to visualize these inputs by overlaying them onto the frame. This process of visualizing the inputs resulted in notable discoveries related to the quality of the inputs. Specifically, the participants' poses and gaze coordinates were observed to the satisfactory. However, there were instances of misidentification or non-detection pertaining to bounding boxes generated by the Detecto. Furthermore, a few mistakes were noted in the Object-Context Relation (OCR) Matrix derived from the training datasets. In the following, we will provide a comprehensive discussion about the challenges encountered with the Detecto model and OCR matrix along with the proposed solutions.

It is necessary to manually annotate the bounding boxes of the objects in one or a few videos because some objects used in the BOSS dataset are absent from the existing training data of the Detecto model. This annotation is necessary to enable the Detecto model to successfully detect the remaining objects in this dataset. Initial annotations made by Duan et al. may not fully consider the probabilistic distribution of the training data crucial for the performance of the Detecto model [DYT+22b]. As a result, this model struggles to accurately detect objects that lack association with annotated labeled data, resulting in misidentification and non-detection as shown in the Figure 4.1 (a). To address this issue, we annotated about 2,000 frames across various videos while taking the probabilistic distribution of the training data into account. Then, we used the state-of-the-art YOLOv5 [JSB+20] model for object detection, achieving 95.34% accuracy during testing. An example of revised bounding boxes of the objects extracted using the YOLOv5 is shown in Figure 4.1 (b).

The matrix representing the relationships between objects (Object-Context Relations or OCR matrix) exhibits certain mistakes as well. Initially, the training datasets comprise a total of 180 experiments, and as the OCR matrix is derived from these training datasets, it should logically sum up to 180. However, the sum amounts to 166, thus leaving 14 experiments unaccounted



(a) The Detecto model's object bounding box coordinates predictions: misidentifications and non-detections.

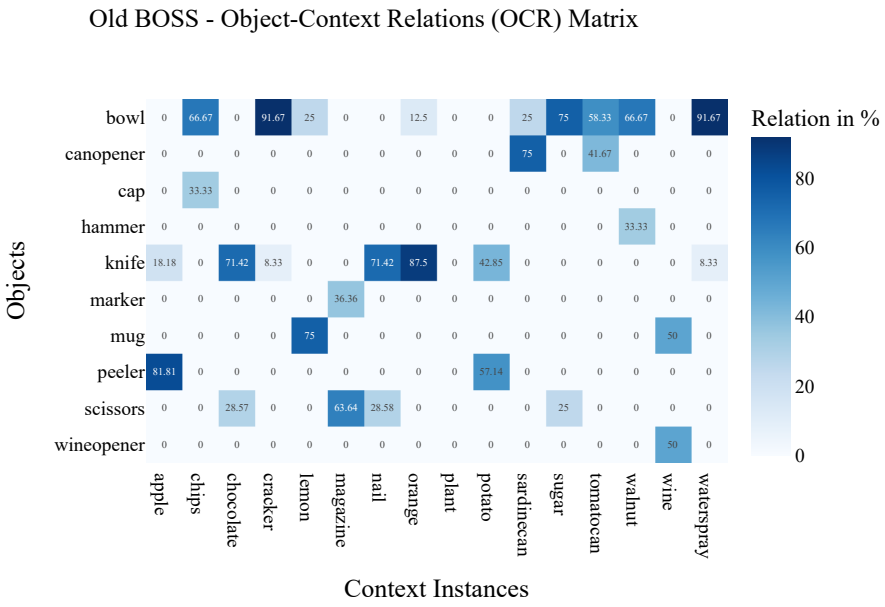


(b) The Revised YOLOv5 model object bounding box coordinates predictions.

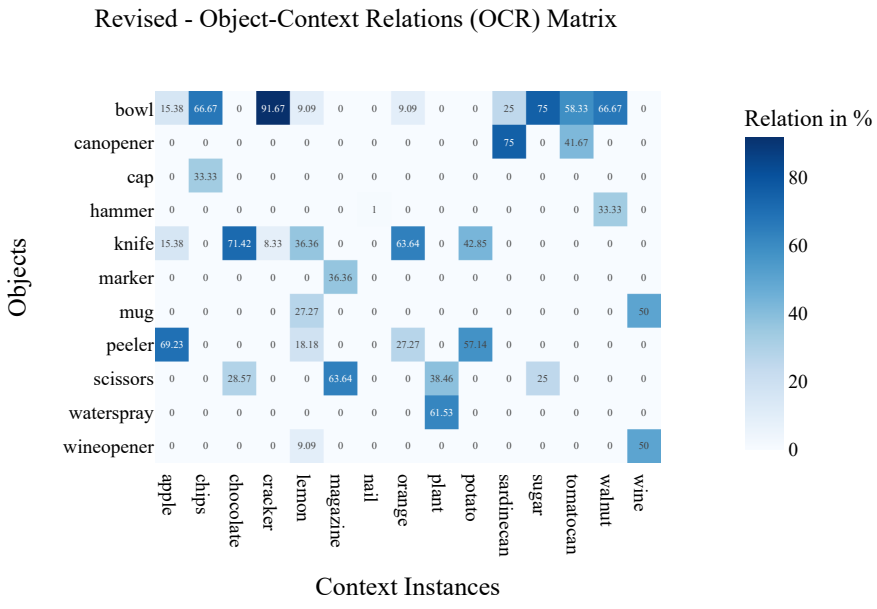
Figure 4.1: Bounding box coordinates prediction difference: Detecto model vs. YOLOv5 model.

for. Furthermore, a few mistakes are found in the values assigned to relationships between objects. For instance, in the case of the object "plant", the OCR matrix incorrectly shows no relation with any other objects. Yet, upon analyzing the training datasets, we recognize that the object "plant" is indeed related to "water spray" and "scissors" with a certain probability. Similarly, the object "nails" is linked to "bowl" and "knife" according to the matrix, but these connections do not match the experiments in the training datasets. To rectify these inaccuracies, a thorough examination of all experiments within the training samples was conducted. This involved manually observing the relationship between the objects in every experiment, which resulted in the

generation of a revised object-context relations (OCR) matrix with dimensions (27,27). The Figures 4.2 (a) and 4.2 (b) illustrate the difference between the initial OCR matrix and the revised OCR matrix.



(a) Inaccuracies in OCR Matrix from Duan et al. [DYT+22b].



(b) The Revised OCR matrix.

Figure 4.2: The OCR matrix difference: Duan et al vs. Revised.

4.2 Input Preprocessing and Feature Extraction

The BOSS dataset comprises a total of five multimodal inputs: frames, participants' pose coordinates, participants' gaze coordinates, object bounding boxes, and object-context relation (OCR) matrix. Prior to feeding into transformer-based architectures, a series of preprocessing steps are applied to ensure uniform scaling and distributions. This involves operations such as normalization and resizing. Additionally, features are extracted from these inputs using feature extraction techniques.

To begin with, each frame extracted from videos is resized to dimensions of (128, 128, 3) from its original size of (1088. 1920, 3). Following this, normalization is performed using specific mean and standard deviation to prepare the frames for use in a convolution-based network for feature extraction. This process uses a simple combination of convolution layers and max pooling layers to extract relevant features from the frame input instead of using large pre-trained convolution neural network models.

Regarding participants' poses, since participants stand behind their tables, certain joints and edges connected to the lower part of their bodies are blocked from view. As a result, the OpenPose model outputs zero value on these joints. These redundant joints are removed during preprocessing, resulting in dimensions of (2, 17, 3). Furthermore, the pose coordinates of the participants are normalized by dividing them by the original frame width and height. Additionally, we experimented with transforming the pose coordinates into image channels, as illustrated in [BZD+18], where each channel corresponds to a participant's joint and exhibits a Gaussian bump around the joint's position. This transformation results in an image with dimensions (16, 16, 17), where the 16*16 dimensions refer to the image's height and width. Feature extraction is carried out using feed-forward neural networks [DYT+22b], graph attention-based networks [BAY21; YRL+21] or convolution-based networks [BZD+18; ON15]. It's essential to note that separate feature extraction techniques are employed for each participant to handle them independently rather than as a single entity.

For participants' gaze coordinates, since the Gaze360 model outputs are already normalized, their original dimension of (2, 3) is kept. However, we used feed-forward neural networks as feature extractors for the flattened gaze coordinates. Similar to participants' poses, participants' gaze coordinates for each participant are processed individually.

Regarding the bounding box coordinates of each object, the YOLOv5 model provides normalized bounding box coordinates, requiring no further normalization. In contrast, the Detecto model, as provided within the BOSS dataset, outputs unnormalized bounding box coordinates. These unnormalized bounding box coordinates are normalized by dividing them by the original frame width and height. To accommodate the different number of objects detected by the YOLOv5 or Detecto model, which can be more or fewer than ten objects per frame due to misidentification or non-detection, zero-padding is applied to have output dimension (10, 4) for all frames throughout the experiments. Feed-forward neural networks are used as feature extractors for these bounding box coordinates after flattening. Additionally, the YOLOv5 model provides cropped images of detected objects, which are then subjected to a shallow combination of convolutional and max pooling layers for feature extraction.

When the OCR matrix is used as input, the OCR values, which already lie between 0 and 1, are flattened to be fed into feed-forward neural network-based feature extractors. When the OCR matrix serves as the ground truth for cost functions, the original dimension of (27, 27) is maintained, with

the simple addition that the OCR matrix value for any object in relation to itself is always 1. Further details regarding the use of the OCR matrix in the cost function will be clarified in the subsequent subsection.

The BOSS dataset consists of 300 videos or experiments, each with varying numbers of input modalities, which are split into 180 for training, 60 for validation, and 60 for testing purposes. To work with this in PyTorch [PGC+17], we utilized the dataset module, which stores the preprocessed input modality samples and their corresponding labels as PyTorch tensors. To efficiently manage variable-length sequences within the BOSS dataset, we implemented a custom dataset module called "dataset_shuffle". This module organizes the input modalities sequentially, based on frames across all the videos. We then grouped these sequences into batches of 32 frames, with this specific configuration established through the process of hyperparameter tuning. Moreover, to ensure randomness and variety in each epoch, we randomly shuffled all the videos and concatenated and arranged them again in batches of 32. It's worth noting that in each batch, the 32 frames always belong to the same video. We achieved this by incorporating a mechanism that includes a few frames from the previous batch without breaking the sequence if the current batch doesn't contain a full 32 frames from the same video. We employed the data loader module from Pytorch to facilitate easy access to these prepared samples, which are subsequently fed into our models. While this approach effectively handles the issue of varying video frame lengths, it is not the only viable option, as there are alternative methods. Ultimately, we opted for this approach because it significantly reduced GPU memory usage, a crucial factor in our study, while increasing RAM usage due to loading all videos at once.

4.3 Architectures

In the preceding subsection, the approaches to input preprocessing and the extraction of features from the multimodal inputs of the BOSS dataset are discussed. After extraction, the features are fed into a transformer-based architecture, followed by a classification layer to predict the beliefs of each participant for each frame. In particular, we utilized a self-attention transformer [KDW+21] and a hierarchical cross-attention transformer [CFP21]. We will look in detail at these two architectures below.

4.3.1 Self Attention Transformer Architecture

Within this architecture, a self-attention mechanism is applied across all the multimodal inputs. As illustrated in Figure 4.3, the multimodal inputs are pre-processed and fed into feature extractors, generating features from these inputs. These features are concatenated and passed into a projection layer. This projection layer embeds the inputs and encodes their positional information. These embeddings with positional encodings are then fed into the transformer encoder. The transformer encoder involves multi-head self-attention mechanisms, capturing distinct feature sets and relationships through attention heads, and position-wise feed-forward networks, learning nonlinear transformations per position via fully connected layers and activation functions. On top of the transformer, two MLP heads return separate left and right participants' belief outputs. The

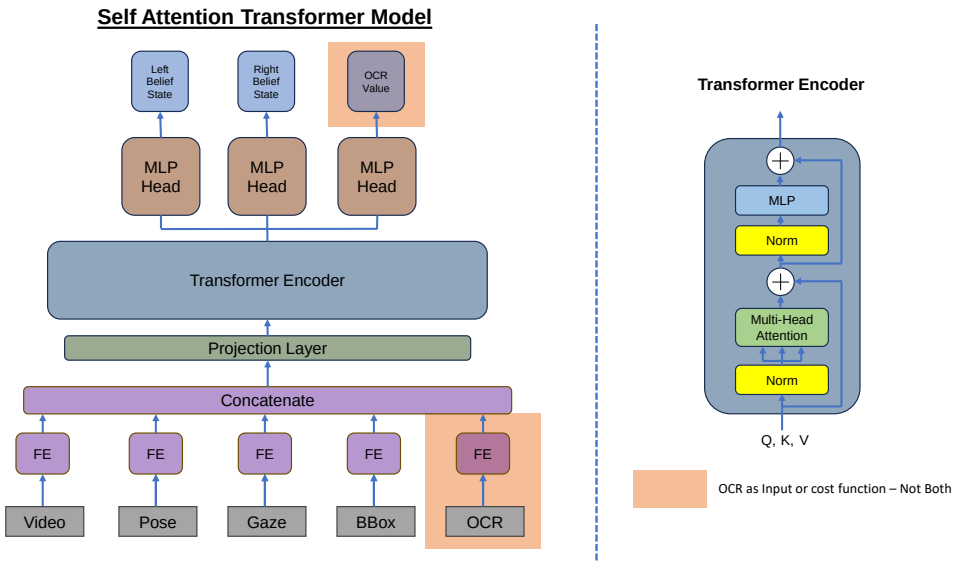


Figure 4.3: Proposed self-attention transformer architecture design for predicting BOSS dataset participants' beliefs.

architecture's output is a vector of dimensions (27, 1) for each participant's beliefs, corresponding to the 27 previously mentioned belief classes. When using the OCR matrix as a cost function instead of as an input, the model outputs an additional value of dimension (1,1).

4.3.2 Hierarchical Cross-Attention Transformer Architecture

This architecture operates through cross-attention mechanisms linking human communication input modalities (e.g., frames, participants' poses, participants' gazes) with object-context input modalities (e.g., frames, objects' bounding boxes, object context relations, or OCR matrix). As depicted in Figure 4.4, multimodal inputs are initially split into two categories: human communication and object-context input modalities. Following preprocessing and feature extraction, the features belonging to the same category are concatenated and then passed into a projection layer for embedding and positional encoding. Self-attention transformer encoders process these embeddings independently for each input modality category. The outputs of these self-attention transformer encoders are then fed into the cross-attention encoder. This specific encoder receives queries and keys from the self-attention transformer encoder associated with the human communication input modalities, while it obtains values from the object-context input modalities. This setup captures intricate relationships such as dependencies, associations, and alignments between human communication inputs and object-context inputs. Ultimately, the output from this cross-attention encoder is forwarded to MLP heads to independently predict the participants' beliefs. Similar to the self-attention transformer architecture, the output consists of a vector of dimension (27,1) per participant's beliefs. When the cost function incorporates the OCR matrix, the additional output has dimension (1, 1).

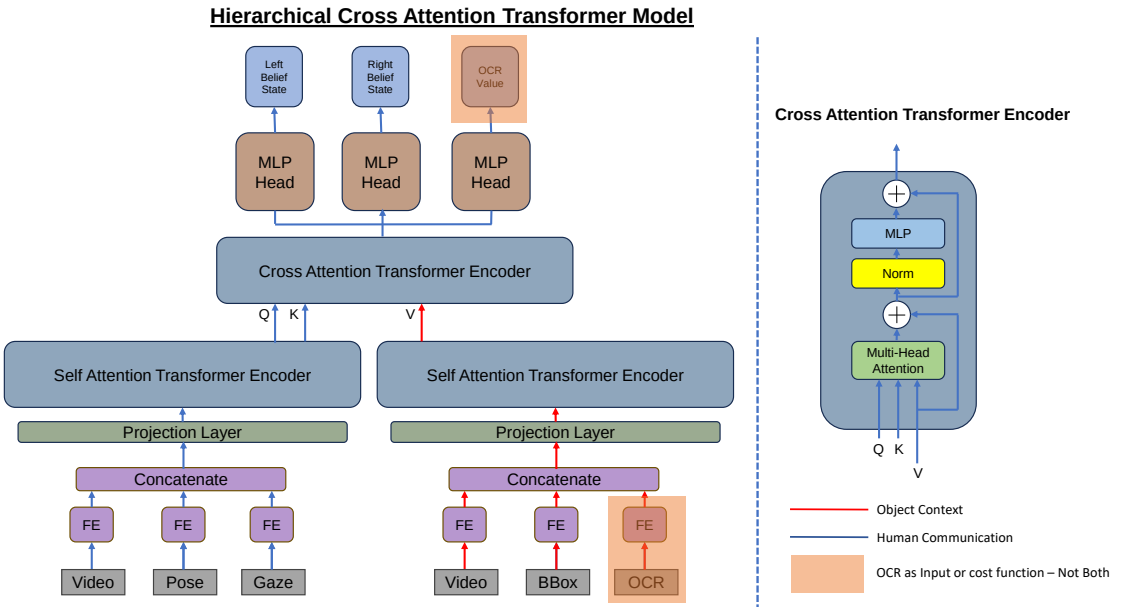


Figure 4.4: Proposed hierarchical cross-attention transformer architecture design for predicting BOSS dataset participants' beliefs.

4.4 Object-Context Cost Function

The BOSS dataset includes two labels that denote the beliefs of the left and right participants in each frame. For estimating the cost between the predicted and actual participants' beliefs, the cross-entropy function is used. In this work, we propose a novel cost function component that takes into account information provided by the object-context relation (OCR) matrix as an additional label. This addition aims to capture the relationship between the beliefs of the two participants, as the OCR matrix captures the interconnection between the contextual object (left participants table) and the context relation object (right participants table), corresponding to their respective beliefs. Our novel cost function is defined as

$$L(t_l, t_r, t_{ocr}, p_l, p_r, p_{ocr}) = CE(t_l, p_l) + CE(t_r, p_r) + \lambda L_{OCR}(t_{ocr}, p_{ocr}) \tag{4.1}$$

where t_l and t_r denote ground truth for left and right participant beliefs, p_l and p_r the predicted left and right participant beliefs, and λ is a hyperparameter that assigns a degree of significance to both cross-entropy and OCR cost. $CE(t, p)$ represents the cross-entropy cost function, and $L_{OCR}(t_{ocr}, p_{ocr})$ is an additional term that measures how well the model learns object-context relations:

$$L_{OCR}(t_{ocr}, p_{ocr}) = |t_{ocr} - p_{ocr}|_1 \tag{4.2}$$

with $t_{ocr} = OCR(t_l, t_r)$ representing the element corresponding to row t_l and column t_r of the OCR matrix.

In this research project, we evaluate the efficacy of this novel cost function by examining its effect on the performance of the above-mentioned architectures.

5 Experiments

5.1 Baseline

In the evaluation of the BOSS dataset, Duan et al. employed a range of baseline deep learning models with the primary objective of predicting participants' beliefs in each frame of the videos [DYT+22b]. This prediction task was accomplished using various multimodal inputs, including frames, participants' pose coordinates, participants' gaze coordinates, object bounding boxes, and an object-context relation (OCR) matrix. The baseline deep learning models used for this evaluation included a **Random** model that randomly assigns beliefs to individuals at each time step, a **Convolutional Neural Networks (CNN)** model that encoded frames as latent features using a pre-trained ResNet model and then passed them through two separate feed-forward networks for multitask classification, a **CNN + Conv1D** model that included a 1D convolutional (Conv1D) layer after the pre-trained ResNet model to capture short-term temporal dependencies, a **CNN + GRU** model that included a gated recurrent unit (GRU) layer after the pre-trained ResNet model to capture long-range temporal dependencies, a **CNN + LSTM** model that introduced a Long Short-Term Memory (LSTM) layer after the pre-trained ResNet model to capture long-range temporal dependencies.

Table 5.1: Table showcasing accuracy of BOSS belief prediction of various models using diverse input combinations as presented by Dual et al. [DYT+22b].

Methods	Feature Extractor	RGB+OCR+ObjDet	RGB+Pose+Gaze	All
Random	CNN, FNN	3.7%	3.7%	3.7%
CNN	CNN, FNN	23.26%	11.84%	24.23%
CNN+Conv1D	CNN, FNN	19.29%	13.27%	23.18%
CNN+GRU	CNN, FNN	21.55%	14.05%	15.24%
CNN+LSTM	CNN, FNN	19.8%	14.6%	13.99%

In each of these model configurations, the remaining input modalities (pose, gaze, object bounding boxes, and OCR matrix) were flattened and processed via a feed-forward neural network (FNN). In addition to the models trained on all input modalities, Duan et al. developed various model configurations, each accepting distinct combinations of inputs. These combinations involved frames, OCR matrix, and object bounding boxes, as well as combinations featuring frames, pose, and gaze, as detailed in Table 5.1. Among these diverse model configurations, the CNN model achieved the highest accuracy, reaching 24.23% accuracy when using all input modalities. It is noteworthy that these model configurations were trained using the Adam optimizer for just 5 epochs, with a fixed learning rate of 1e-3 and a batch size of 4 videos. The evaluation of the model's effectiveness was

based on metrics including cross-entropy cost function and classification accuracy. Furthermore, all experiments were conducted using NVIDIA A100-SXM4 GPUs (40GB) on Linux servers, with each individual run taking approximately 1.5 hours to complete.

5.2 Training Details

For all experiments, we use the 300 experiments or videos from the BOSS dataset. These videos are randomly divided into training, validation, and test sets, with 60% for the training set and 20% for the validation and test set each. We train the self-attention transformer models using the Adam optimizer with a weight decay of 0.0001 for 250 epochs, with a fixed learning rate of 0.0007, 4 attention heads, 4 transformer encoder layers, and batches consisting of 32 frames. Similarly, we train the cross-attention transformer models under identical configurations as the self-attention transformer, except with 2 attention heads and 1 transformer encoder. We implement our method using PyTorch. We evaluate the model's performance using metrics such as cross-entropy cost function and classification accuracy. Experiments are carried out on NVIDIA GeForce GTX 1070 GPU with 8GB of memory, running on Linux servers, and each run requires approximately 8 hours to complete.

5.3 Results

As a part of our research project, we conducted a series of experiments using transformer-based architectures, namely the self-attention transformer and the hierarchical cross-attention transformer. These architectures were applied on top of a collection of feature extraction layers, each dedicated to a distinct input modality, and were trained end-to-end using the BOSS dataset to predict the beliefs of participants, which encompass 27 distinct classes. It is important to note that in all of these experiments, we made use of all input modalities, including frames, participants' pose coordinates, participants' gaze coordinates, object bounding boxes, and an object-context relation (OCR) matrix, as this approach had proven effective by Duan et al. [DYT+22b].

As illustrated in Table 5.2, the results indicate that the self-attention transformer model, with a convolutional neural network-based (CNN) feature extractor for input frames and feed-forward neural network (FNN) for all other input modalities, achieved a peak accuracy of 62.05% on the test dataset. In contrast, the hierarchical cross-attention transformer with the same configuration only achieved a maximum accuracy of 45.36%. Similarly, when using a CNN to process the frames data, graph attention network (GNN) for participants' poses, and FNN for all other input modalities, the self-attention transformer model achieved the second-best accuracy of 55.21%. In contrast, the hierarchical cross-attention transformer with this configuration achieved a maximum accuracy of 45.85%. When employing a CNN-based feature extractor for frame data, participants' poses, and FNN for all other input modalities, the self-attention transformer model achieved an accuracy of 54.55%. The hierarchical cross-attention transformer with this configuration reached a maximum accuracy of 39.50%.

In the context of the self-attention transformer model using a CNN-based feature extractor for frames and cropped object images extracted from YOLOv5, the model displayed its lowest accuracy, registering only 10.10%. Meanwhile, the hierarchical cross-attention transformer with the same

Table 5.2: Comparison table of BOSS belief prediction accuracy between self-attention and hierarchical cross-attention transformer models using various feature extractors for input modalities.

S.No	Models	Feature Extractors				Test Accuracy
		Frame	Pose	Gaze OCR	ObjDct	
1.	Self-Attention_T Hierarchical CA_T	CNN	FNN	FNN	FNN	62.05% 45.36%
2.	Self-Attention_T Hierarchical CA_T	CNN	GNN	FNN	FNN	55.21% 45.85%
3.	Self-Attention_T Hierarchical CA_T	CNN	CNN	FNN	FNN	54.55% 39.50%
4.	Self-Attention_T Hierarchical CA_T	CNN	FNN	FNN	CNN	10.10% 15.31%

configuration managed to attain a slightly higher accuracy of 15.31% on the test dataset. This outcome suggests that the features extracted from cropped images of objects do not contribute significantly to the models' performance, unlike the object bounding box coordinates, which are considered a crucial input in enhancing model performance on this dataset.

In conclusion, the results consistently showed that the self-attention transformer model outperforms the hierarchical cross-attention transformer in all the experiments previously discussed. This difference in performance can be related to the cross-attention encoder's inability to find meaningful relationships between human communication and object context input modalities. Moreover, it is important to highlight that the self-attention and hierarchical cross-attention transformer architectures have approximately 582K and 300K trainable parameters respectively. These parameters are associated with the highest accuracy-achieving models from both transformer architectures. However, attempts to train the cross-attention transformer with a similar number of parameters failed to match the performance of its best configuration.

Furthermore, a dedicated study focused solely on the hierarchical cross-attention transformer had been conducted, exploring two distinct configurations. Initially, we configured the cross-attention encoder in the hierarchical cross-attention transformer to receive query and key from the human communication input modalities (HCI) and values from the object context input modalities (OCI). This configuration resulted in an accuracy of 45.36%, as shown in Table 5.3. Subsequently, we reversed the configuration of the cross-attention encoder to receive the query and key inputs from the OCI and values from the HCI, resulting in an accuracy of 45.12% on the test dataset. The relatively small difference in performance between these two configurations suggests that both are quite similar in achieving optimal results.

Lastly, it is worth noting that in all the experiments presented above, we have used the OCR matrix as an input for the models, rather than being incorporated into the cost function. In the next section, we will examine the performance of the model when the OCR matrix is used as an input versus when it is integrated into the cost function.

Table 5.3: Comparison table of results for hierarchical transformer architectures using different input configurations: Human communication input (HCI) as Query and Key, and Object context input (OCI) as value, and vice versa.

Models	Frame, Pose Gaze OCR ObjDct	Query, Key	Value	Test Accuracy
Hierarchical CA_T	CNN, FNN	HCI	OCI	45.36%
Hierarchical CA_T	CNN, FNN	OCI	HCI	45.12%

5.4 Ablation Study

5.4.1 Object bounding boxes: the Detecto model and the YOLOv5 model

In this subsection, we investigate the influence of revised bounding boxes, detected through YOLOv5, on the overall accuracy. Our approach involved implementing a self-attention transformer architecture coupled with a shallow combination of convolution and max pooling layers to process the frame data. Furthermore, we used feed-forward neural networks to process participants' poses, participants' gazes, the OCR matrix, and the object bounding boxes obtained using YOLOv5. This configuration achieved a maximum accuracy of 62.05% on the test dataset as shown in Table 5.4. Conversely, utilizing an identical configuration of the self-attention model, with the only difference being the use of bounding box coordinates from the Detecto model, as provided by Duan et al. [DYT+22b], resulted in a lower performance, with an accuracy of only 21.13% on the test dataset. Consequently, it becomes evident that the use of revised bounding boxes detected using YOLOv5 had a substantial impact on accuracy.

Table 5.4: Comparison table illustrating BOSS belief prediction accuracy of a self-attention model using objects bounding boxes from YOLOv5 [JSB+20] and Detecto models [Ala19].

Methods	Feature Extractor	ObjDct Model	Frame+Pose+Gaze+OCR+ObjDct
Self-Attention_T	CNN, FNN	YOLOv5	62.05%
Self-Attention_T	CNN, FNN	Detecto	21.13%

5.4.2 OCR Matrix: as an Input and as a Cost Function

In this subsection, we look into the performance of our models under two different configurations. First, we explore how they perform when we use a convolutional neural network-based features extractor to process frame data and feed-forward neural network-based feature extractors for all other input modalities, using the OCR matrix as input. Next, we explore our models' performance using the same configuration but without the OCR matrix as input, but rather adding the OCR matrix into the cost function.

In our experiments, when we employed the OCR matrix as an input with the self-attention transformer, it achieved a maximum accuracy of 62.05%, as detailed in previous sections. However, when we introduced the OCR matrix into the cost function while varying the hyperparameter λ with values of

1, 5, and 10, the model had an accuracy of 42.27%, 45.75%, and 50.21% respectively as depicted in the Table 5.5. It is worth noting that the OCR matrix values lie between 0 to 1, and until we assign a considerably higher value to the hyperparameter λ, this cost function has a limited impact on the overall cost function. Consequently, when λ is set to 1, the OCR cost function has minimal influence on the overall cost function, acting as minor noise within the cross-entropy cost function. In such a situation, the model primarily focuses on reducing the cross-entropy cost function between the true and predicted beliefs, rather than the OCR cost function. In contrast, when the hyperparameter λ is set to 10, the model is forced to reduce the OCR cost function, which represents the relationships between the objects associated with the beliefs. Based on the results, it can be concluded that the self-attention transformer performs well when the OCR matrix is used as an input instead of incorporating it into the cost function. However, when the OCR matrix is integrated into the cost function with a hyperparameter λ value of 10, the model attains its highest accuracy, indicating that when forced to emphasize the OCR cost function, it can achieve satisfactory accuracy, although still lower than when the OCR matrix is employed as an input.

Table 5.5: Table comparing BOSS belief prediction outcomes between the self-attention transformer and the hierarchical cross-attention transformer models using OCR matrix for input and incorporating OCR matrix into the cost function with different hyperparameter λ.

Models	Frame, Pose Gaze Obj_Dct	OCR as	Hyperparamtr λ	Test Accuracy
Self-Attention_T	CNN, FNN	Input	Not Used	62.05%
		Cost Func	1	42.27%
		Cost Func	5	45.75%
		Cost Func	10	50.21%
Hierarchical CA_T	CNN, FNN	Input	Not Used	45.36%
		Cost Func	1	44.09%
		Cost Func	5	35.75%
		Cost Func	10	40.63%

The hierarchical cross-attention transformer attains a peak accuracy of 45.36% when employing the OCR matrix as input. When using the OCR cost function with λ values set to 1, 5, and 10, this model demonstrates accuracy scores of 44.09%, 35.75%, and 40.63%, respectively as depicted in the Table 5.5. Thus, the hierarchical cross-attention transformer exhibits a similar performance pattern to that of the self-attention transformer, with one exception. The results highlight that when the hyperparameter λ is set to 1, the OCR cost function has little impact on the overall cost function which leads to the model attaining the maximum accuracy. This contrasts with the self-attention transformer, which responds differently under similar conditions. Nonetheless, it's important to note that this accuracy remains lower than when the OCR matrix is employed as an input.

6 Discussion

In this research project, we carried out an evaluation of transformer-based models trained on the BOSS dataset, providing valuable insights regarding the capabilities of different architectures and the impact of different input processing for predicting human beliefs in object-context scenarios.

We deployed two transformer-based models: the self-attention transformer and the hierarchical cross-attention transformer. These models utilized a range of feature extraction methodologies, encompassing convolution-based (CNN) feature extraction methods to process frame data, cropped object images sourced from the YOLOv5 model [JSB+20], and participants' pose representation derived using the Balakrishnan et al. [BZD+18] approach. Additionally, we employed feed-forward neural networks (FNN) to process participants' gaze coordinates, participants' pose coordinates, object bounding box coordinates obtained from the YOLOv5 model, and the object-context relation (OCR) matrix. Furthermore, we applied a graph attention network (GNN) to extract meaningful features from participants' pose coordinates. Notably, we experimented by integrating the OCR matrix into the novel cost function rather than including it as an input. Our research project involved the training of distinct models, with each model utilizing either the self-attention transformer or the hierarchical cross-attention transformer in conjunction with various combinations of the feature extraction techniques mentioned earlier. The performance of our models is effectively summarized in Figure 6.1.

Notably, the best-performing model in our study is the self-attention transformer, which utilizes CNN-based feature extraction techniques to process frame data and FNN for other input modalities with the OCR matrix as an input, outperforming its counterpart, the hierarchical cross-attention transformer model. This performance gap of roughly 17% suggests that the hierarchical cross-attention model struggles to establish meaningful relationships between human communication and object context input categories, a limitation that permits further investigation. Examining the number of trainable parameters, we observe that the self-attention transformer model possesses nearly twice as many parameters as the hierarchical cross-attention transformer model. Attempts to train the cross-attention transformer with a comparable number of parameters, however, failed to match the performance of its best configuration, highlighting the complexity of optimizing the hierarchical cross-attention architecture. The worst-performing models in our study are those that combined a self-attention transformer model with CNN-based feature extraction techniques to process frame data and cropped object images from YOLOv5 model, and FNN to process participants' pose, participants' gaze, and the OCR matrix. Similarly, the hierarchical cross-attention model using similar feature extraction techniques also demonstrated subpar performance. This suboptimal can be attributed to the utilization of CNN as feature extractors on cropped object images. This approach may have resulted in the loss of object location information within the frame, leading to poorer results.

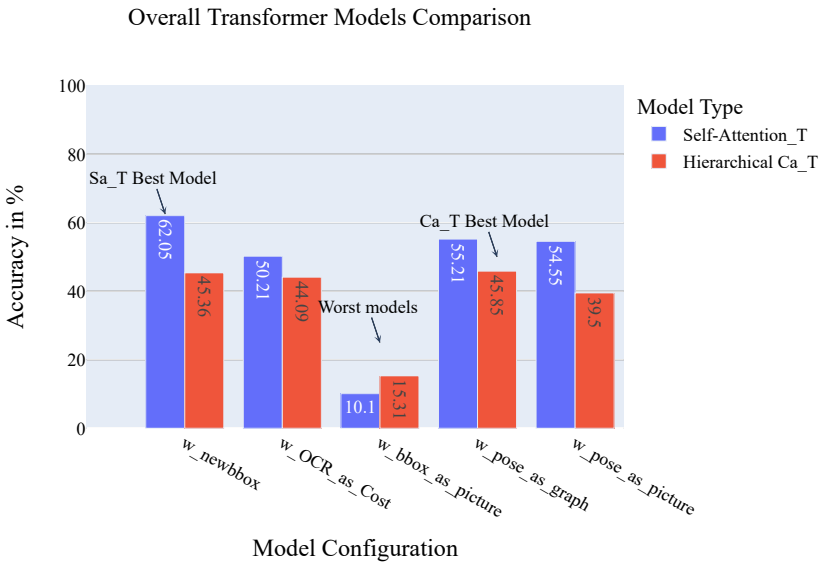


Figure 6.1: Figure depicting a comparison of transformer-based models employed in this research project for predicting participant beliefs using the BOSS dataset.

As illustrated in Figure 6.2, our project findings indicate a significant improvement in model performance compared to the models proposed by Duan et al [DYT+22b]. This improvement suggests that our models are better equipped to capture the nuances of human non-verbal cues within this dataset. This notable performance improvement can be primarily attributed to the use of revised object bounding box coordinates derived from the YOLOv5 model in contrast to the object bounding box coordinates provided within the boss dataset. When we trained our best-performing model with the revised object bounding box coordinates, it achieved the highest level of accuracy, showcasing an improvement of around 40% compared to the original object bounding box coordinated provided within the BOSS dataset.

In our analysis of the other input modalities, we found that using participants' pose coordinates as a 16*16 image in conjunction with the self-attention transformer model produced promising results as this approach leverages spatial information in the pose input [BZD+18], although it did not outperform the best-performing self-attention model. There is potential for improved performance by increasing the dimensionality of this image representation which could be a future work that we did not explore due to certain limitations. When we used a GNN to process the participants' pose coordinates alongside the hierarchical cross-attention transformer model, we achieved our best result with this model. This approach yielded a slightly higher accuracy compared to participants' pose coordinates processed using FNN alongside the hierarchical cross-attention transformer model. This suggests that further optimization may lead to substantial improvements, an area for future research. In terms of participants' gaze coordinates, we primarily used an FNN for processing across all of our models. The impact of participants' gaze coordinates on our model performance has not been extensively explored in our research project which remains an avenue for future work. Additionally, we introduced an innovative cost function based on an OCR matrix, aiming to capture the contextual relations between objects. We combined this with the cross-entropy cost function in an effort to improve belief prediction. Unfortunately, our research project findings suggest that

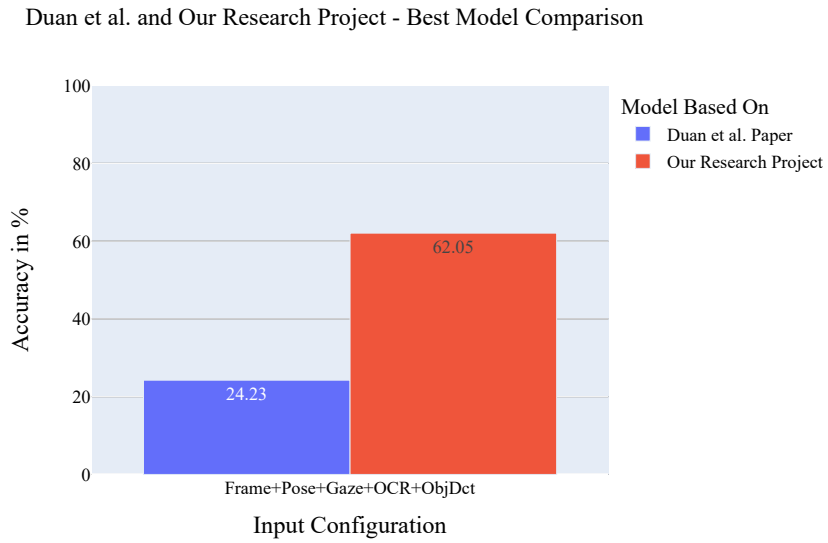


Figure 6.2: Figure depicting a comparison of the best model employed in Duan et al. [DYT+22b] and this research project for predicting participant beliefs using the BOSS dataset.

despite its novelty, this approach did not achieve the desired enhancements in predictive accuracy. This may be due to the fact that the OCR matrix contribution as an input was no longer available, as it was incorporated into the cost function. Additionally, it is possible that the model did not effectively learn how to leverage the OCR matrix-based cost function to improve participants' belief prediction. A potential avenue for future research could involve investigating the utilization of the OCR matrix both as an input and within the cost function, and exclusively employing the OCR matrix cost function without the inclusion of the cross-entropy cost function, to gain a more comprehensive understanding of the potential value of this novel cost function.

Additionally, future research avenues may include exploring combinations of input modalities in conjunction with the proposed transformer-based models to gain a deeper understanding of the impact of each input modality, as proposed by Duan et al. Additionally, investigating participant relationship types, as undertaken by Duan et al., and refining training methods could hold potential solutions for enhancing predictive accuracy. Nevertheless, it is essential to acknowledge certain limitations in further improving the belief prediction accuracy within the BOSS dataset. First, the method employed by Duan et al. for annotating belief states, while novel, may not be 100% accurate, potentially impacting models' training. Second, the task of modeling human belief through non-verbal communication remains inherently challenging, even for human observers let alone AI models, as it involves understanding and interpreting complex, context-dependent, and often subtle non-verbal cues and mental states that can vary widely among individuals.

To conclude, our research project represents a substantial advancement in the field of modeling human beliefs within the BOSS dataset. It has provided valuable insights into the optimization of model architectures, cost functions, and the selection of input modalities. While we have surpassed previous models' performance, the inherent complexity of the task suggests that continued research and innovation are essential in this field.

7 Conclusion

This research project explores the challenging task of predicting human beliefs within the BOSS dataset, as introduced by Duan et al [DYT+22b]. This dataset aims to model human belief states in an object context scenario, where participant pairs have to rely on non-verbal communication to accomplish collaborative tasks. Our investigation spans various aspects, including input modalities, architectural designs, and cost functions to enhance belief prediction accuracy using the BOSS dataset. We started our evaluation by assessing the diverse input modalities within the BOSS dataset, including participants' pose coordinates, gaze coordinates, object bounding boxes, and the Object-Context Relation (OCR) matrix. While the participants' pose and gaze coordinates are satisfactory, our evaluation showed that the original object bounding boxes from the BOSS dataset proved inadequate. Consequently, we trained a YOLOv5 object detection model and extracted new object bounding boxes. Furthermore, we found discrepancies in the OCR matrix and revised them through manual verification of the training data. Following that, we conducted experiments involving transformer-based models and various feature extraction techniques. Among the self-attention transformer model and hierarchical cross-attention transformer model illustrated in this project, the self-attention transformer model achieved the highest accuracy in predicting the beliefs of the participants within the BOSS dataset. This was accomplished by combining it with a convolution-based network to process frame data and using feed-forward neural networks to process all other input modalities. Moreover, we introduced a novel object-context-based cost function to train and evaluate the transformer-based models, enabling them to capture the complex relationships among participants' beliefs in addition to predicting those beliefs. Our findings indicate that this approach, while innovative, did not yield the desired improvements in predictive accuracy. To conclude, this research project not only sheds light on the complexities of modeling human beliefs in nonverbal communication scenarios, but also provides useful insights into the nuances of input selection, architectural design, and cost function formulation in pursuit of improved belief prediction accuracy within the BOSS dataset.

Bibliography

[Ala19] Alankbi. *Alankbi/Detecto: Build Fully-Functioning Computer Vision Models with Pytorch.* https://github.com/alankbi/detecto. 2019 (cit. on pp. 22, 36).

[BAY21] S. Brody, U. Alon, E. Yahav. "How attentive are graph attention networks?" In: *arXiv preprint arXiv:2105.14491* (2021) (cit. on p. 28).

[BGT08] C. L. Baker, N. D. Goodman, J. B. Tenenbaum. "Theory-based social goal inference". In: *Proceedings of the thirtieth annual conference of the cognitive science society*. Cognitive Science Society Austin, TX. 2008, pp. 1447–1452 (cit. on p. 19).

[Bin05] L. Binnie. "TOM goes to school: Theory of mind understanding and its link to schooling". In: *Educational and Child Psychology* 22.4 (2005), p. 81 (cit. on p. 18).

[BLF85] S. Baron-Cohen, A. Leslie, U. Frith. "Does the Autistic Child Have a Theory of Mind?" In: *Cognition* 21 (Nov. 1985), pp. 37–46. DOI: 10.1016/0010-0277(85)90022-8 (cit. on p. 15).

[BMA11] M. H. Bornstein, C. Mash, M. E. Arterberry. "Perception of object–context relations: Eye-movement analyses in infants and adults." In: *Developmental psychology* 47.2 (2011), p. 364 (cit. on p. 18).

[BS11] C. Baker, R. Saxe. "Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution". In: *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* (Jan. 2011) (cit. on p. 19).

[BST09] C. Baker, R. Saxe, J. Tenenbaum. "Action understanding as inverse planning". In: *Cognition* 113 (Dec. 2009), pp. 329–349. DOI: 10.1016/j.cognition.2009.07.005 (cit. on p. 19).

[BZD+18] G. Balakrishnan, A. Zhao, A. Dalca, F. Durand, J. Guttag. "Synthesizing Images of Humans in Unseen Poses". In: June 2018, pp. 8340–8348. DOI: 10.1109/CVPR.2018.00870 (cit. on pp. 28, 39, 40).

[CCT02] M. Carpenter, J. Call, M. Tomasello. "A new false belief test for 36-month-olds". In: *British Journal of Developmental Psychology, v.20, 393-420 (2002)* 20 (Sept. 2002). DOI: 10.1348/026151002320620316 (cit. on p. 15).

[CFP21] C.-F. R. Chen, Q. Fan, R. Panda. "Crossvit: Cross-attention multi-scale vision transformer for image classification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 357–366 (cit. on pp. 17, 29).

[CMS+20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229 (cit. on pp. 17, 18).

[CSW+15] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, A. M. Dollar. "The YCB object and Model set: Towards common benchmarks for manipulation research". In: *2015 International Conference on Advanced Robotics (ICAR)*. 2015, pp. 510–517. DOI: 10.1109/ICAR.2015.7251504 (cit. on p. 21).

[CSWS17] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh. "Realtime multi-person 2d pose estimation using part affinity fields". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7291–7299 (cit. on p. 22).

[CWG+21] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, W. Gao. "Pre-trained image processing transformer". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 12299–12310 (cit. on p. 17).

[DQGY10] P. Doshi, X. Qu, A. Goodie, D. Young. "Modeling recursive reasoning by humans using empirically informed interactive POMDPs". In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. 2010, pp. 1223–1230 (cit. on p. 19).

[DYT+22a] J. Duan, S. Yu, H. L. Tan, H. Zhu, C. Tan. "A survey of embodied ai: From simulators to research tasks". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 6.2 (2022), pp. 230–244 (cit. on p. 15).

[DYT+22b] J. Duan, S. Yu, N. Tan, L. Yi, C. Tan. "BOSS: A Benchmark for Human Belief Prediction in Object-context Scenarios". In: *arXiv preprint arXiv:2206.10665* (2022) (cit. on pp. 3, 15, 19, 21–23, 25, 27, 28, 33, 34, 36, 40, 41, 43).

[FBP+15] E. Fink, S. Begeer, C. C. Peterson, V. Slaughter, M. de Rosnay. "Friends, friendlessness, and the social consequences of gaining a theory of mind." In: *The British Journal of Developmental Psychology* 33.1 (2015), pp. 27–30 (cit. on p. 18).

[FQZ+21] L. Fan, S. Qiu, Z. Zheng, T. Gao, S.-C. Zhu, Y. Zhu. "Learning triadic belief dynamics in nonverbal communication from videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 7312–7321 (cit. on p. 19).

[FWCL21] A. Fuchs, M. Walton, T. Chadwick, D. Lange. "Theory of mind for deep reinforcement learning in hanabi". In: *arXiv preprint arXiv:2101.09328* (2021) (cit. on p. 15).

[Hin74] R. A. Hinde. *Biological bases of human social behaviour.* McGraw-Hill, 1974 (cit. on p. 15).

[Ho87] S.-B. Ho. *Representing and using functional definitions for visual recognition*. The University of Wisconsin-Madison, 1987 (cit. on p. 18).

[HRAD16] D. Hadfield-Menell, S. J. Russell, P. Abbeel, A. Dragan. "Cooperative inverse reinforcement learning". In: *Advances in neural information processing systems* 29 (2016) (cit. on p. 19).

[JSB+20] G. Jocher, A. Stoken, J. Borovec, NanoCode012, ChristopherSTAN, L. Changyu, Laughing, tkianai, A. Hogan, lorenzomammana, yxNONG, AlexWang1900, L. Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, F. Ingham, Frederik, Guilhen, Hatovix, J. Poznanski, J. Fang, L. Y., changyu98, M. Wang, N. Gupta, O. Akhtar, PetrDvoracek, P. Rai. *ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements*. Version v3.1. Oct. 2020. DOI: 10.5281/zenodo.4154370. URL: https://doi.org/10.5281/zenodo.4154370 (cit. on pp. 25, 36, 39).

[KDW+21] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, X. Zhai. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: 2021 (cit. on pp. 17, 29).

[KRS+19] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, A. Torralba. "Gaze360: Physically unconstrained gaze estimation in the wild". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6912–6921 (cit. on p. 22).

[KUT+13] J. Kiley Hamlin, T. Ullman, J. Tenenbaum, N. Goodman, C. Baker. "The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model". In: *Developmental science* 16.2 (2013), pp. 209–226 (cit. on p. 19).

[Lee17] J. Lee. "A survey of robot learning from demonstrations for human-robot collaboration". In: *arXiv preprint arXiv:1710.08789* (2017) (cit. on p. 15).

[MA11] C. Mash, M. Arterberry. "Perception of Object-Context Relations: Eye-Movement Analyses in Infants and Adults". In: *Developmental psychology* 47 (Mar. 2011), pp. 364–75. DOI: 10.1037/a0021059 (cit. on p. 15).

[NVNT20] D. Nguyen, S. Venkatesh, P. Nguyen, T. Tran. "Theory of mind with guilt aversion facilitates cooperative reinforcement learning". In: *Asian Conference on Machine Learning*. PMLR. 2020, pp. 33–48 (cit. on p. 15).

[OJG10] F. Osiurak, C. Jarry, D. Gall. "Grasping the Affordances, Understanding the Reasoning: Toward a Dialectical Theory of Human Tool Use". In: *Psychological review* 117 (Apr. 2010), pp. 517–40. DOI: 10.1037/a0019004 (cit. on p. 15).

[OJL10] F. Osiurak, C. Jarry, D. Le Gall. "Grasping the affordances, understanding the reasoning: toward a dialectical theory of human tool use." In: *Psychological review* 117.2 (2010), p. 517 (cit. on p. 18).

[ON15] K. O'Shea, R. Nash. "An introduction to convolutional neural networks". In: *arXiv preprint arXiv:1511.08458* (2015) (cit. on p. 28).

[PGC+17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer. "Automatic differentiation in PyTorch". In: (2017) (cit. on p. 29).

[PW78] D. Premack, G. Woodruff. "Does the chimpanzee have a theory of mind?" In: *Behavioral and Brain Sciences* 1.4 (1978), pp. 515–526. DOI: 10.1017/S0140525X00076512 (cit. on p. 18).

[RBHP21] M. D. M. Reddy, M. S. M. Basha, M. M. C. Hari, M. N. Penchalaiah. "Dall-e: Creating images from text". In: *UGC Care Group I Journal* 8.14 (2021), pp. 71–75 (cit. on p. 18).

[RPS+18] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. A. Eslami, M. Botvinick. "Machine theory of mind". In: *International conference on machine learning*. PMLR. 2018, pp. 4218–4227 (cit. on p. 15).

[Sax06] R. Saxe. "Uniquely human social cognition". In: *Current opinion in neurobiology* 16.2 (2006), pp. 235–239 (cit. on p. 15).

[Sla15] V. Slaughter. "Theory of mind in infants and young children: A review". In: *Australian Psychologist* 50.3 (2015), pp. 169–172 (cit. on p. 18).

[SPP18] R. K. Sinha, R. Pandey, R. Pattnaik. “Deep learning for computer vision tasks: a review”. In: *arXiv preprint arXiv:1804.03928* (2018) (cit. on p. 15).

[UBM+09] T. Ullman, C. Baker, O. Macindoe, O. Evans, N. Goodman, J. Tenenbaum. “Help or Hinder: Bayesian Models of Social Goal Inference.” In: vol. 22. Jan. 2009, pp. 1874–1882 (cit. on p. 19).

[VSP+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. “Attention is All You Need”. In: 2017. URL: https://arxiv.org/pdf/1706.03762.pdf (cit. on pp. 17, 18).

[WCW01] H. M. Wellman, D. Cross, J. Watson. “Meta-analysis of theory-of-mind development: The truth about false belief”. In: *Child development* 72.3 (2001), pp. 655–684 (cit. on p. 15).

[WKYL11] M. Wunder, M. Kaisers, J. R. Yaros, M. L. Littman. “Using iterated reasoning to predict opponent strategies.” In: *AAMAS*. 2011, pp. 593–600 (cit. on p. 19).

[XZC22] P. Xu, X. Zhu, D. A. Clifton. “Multimodal learning with transformers: A survey”. In: *arXiv preprint arXiv:2206.06488* (2022) (cit. on p. 19).

[YRL+21] Y. Yang, Z. Ren, H. Li, C. Zhou, X. Wang, G. Hua. “Learning dynamics via graph neural networks for human pose estimation and tracking”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 8074–8084 (cit. on p. 28).

[YRLW19] L. Ye, M. Rochan, Z. Liu, Y. Wang. “Cross-modal self-attention network for referring image segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 10502–10511 (cit. on p. 17).

[ZDY+21] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, J. Gao. “Multi-scale vision longformer: A new vision transformer for high-resolution image encoding”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 2998–3008 (cit. on p. 18).

[ZSL+20] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai. “Deformable detr: Deformable transformers for end-to-end object detection”. In: *arXiv preprint arXiv:2010.04159* (2020) (cit. on pp. 17, 18).

Declaration

I hereby declare that the work presented in this thesis is entirely my own. I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted hard copies.

---

place, date, signature