Multimodal Machines from a perspective of humans
Ph.D Thesis Proposal

Sunit Bhattacharya
Charles University
Faculty Of Mathematics and Physics
Institute of Formal and Applied Linguistics
bhattacharya@ufal.mff.cuni.cz

Abstract

Deep Neural Networks have rapidly become the most dominant approach to solve many complicated learning problems in recent times. Although initially inspired by biological neural networks, the current deep learning systems are much more motivated by practical engineering needs and performance requirements. And yet, some of these networks exhibit a lot of similarities with human brains. This thesis proposal focuses on highlighting the differences in the learning mechanisms of humans and deep learning systems and explores yet how recent work has established similarities between representations learnt by deep learning systems and cognitive data collected from the human brain. Furthermore, we look into the benefits of using brain-inspired techniques and experiments to help build better systems for natural language processing applications and the results of the experiments done so far. Lastly, we outline the proposal to direct our future work towards the completion of the thesis.

1 Introduction

The AI renaissance (Tan and Lim, 2018) in the last few decades has been a chronicle of fantastic developments. Growing out of the idea of artificial neural networks organized in layers (McClelland et al., 1987), Deep Learning (Schmidhuber, 2015) is the most successful and profitable (Chui et al., 2018) AI technology at the present.

The incredible growth in Deep Learning based architectures right from the AlexNet (Krizhevsky et al., 2012) era to the revolution in Natural Language Processing with the Transformer (Vaswani et al., 2017) architecture, the last decade in AI has been a witness to many interesting developments. An interesting synthesis of such developments has manifested itself in the form of the state-of-the-art generalist models like GATO (Reed et al., 2022). The impact is such that sophisticated dialog models like LaMDA (Thoppilan et al., 2022) have made it to the news headlines with claims of it being 'sentient'. Hence, a pertinent question naturally emerges: is scaling existing architectures (Kaplan et al., 2020) the only way to solve all the problems in Artificial Intelligence? Recent work (Hoffmann et al., 2022; Chowdhery et al., 2022; Rae et al., 2021; Yu et al., 2022; Brown et al., 2020; Zhang et al., 2022) surely suggests that scaling helps. And also, more and more works using these huge models are demonstrating systems that beat expert human performance on an array of tasks that have been traditionally considered challenging. Popel et al. (2020) for instance demonstrated a system that matches (and in some situations surpass) the quality of human translation. The efficiency of recent diffusion models (Ho et al., 2020) like Imagen (Saharia et al., 2022) and DALL.E 2 (Ramesh et al., 2022) demonstrates the capability of AI systems to understand the nuances of language and combine that with the capability to understand images and generate realistic synthetic images[1]. On the question of models being proficient across tasks, the authors of GATO (Reed et al., 2022) report that in 450 out of 604 tasks that the model was trained on, the system performed at over 50% expert threshold. Here expert threshold refers to the performance of expert humans on the task. In other words, in around 75% of the tasks, GATO performed half as well as expert humans. Hence, the latest models do not aim to just be good at specific tasks in one domain, they also aim to be proficient in a multitude of different diverse tasks across domains. However, it should be pointed out that most of the systems that are trained to do multiple tasks while using a common underlying architecture are trained on the tasks in parallel. Most neural systems suffer from catastrophic forgetting (Parisi et al., 2019) when they are trained on tasks sequentially. The strive

[1] An implementation of the DALL.E system is available for demonstration at https://huggingface.co/spaces/dalle-mini/dalle-mini

to match or exceed human performance across a broad class of cognitive tasks is ultimately one of the hallmarks of the yet to be realized artificial general intelligence systems (AGI).

Concentrating specifically on domains such as Computer Vision and Natural Language Processing, many systems have surpassed human expert performance on tasks like image classification (He et al., 2015) and on benchmarks like SuperGLUE (He et al., 2020). However, there are criticisms at this practice of comparing human performance with machines. There is a risk that directly comparing performance accuracy between humans and machines may just "overstate machine performance" (Shankar et al., 2020). But still, it is widely accepted that humans are way better than the current AI systems at generalization. For instance, as several studies (Geirhos et al. (2018); Huber et al. (2021) show, the amazing performance of the Convolutional Neural Networks (CNN) based computer vision systems is heavily affected by out-of-distribution data. Similarly, Transformers too struggle with Out-Of-Distribution robustness. Albeit they fare better than CNNs (Bai et al., 2021).

While current AI systems may lack sufficient generalization capabilities and face problems with continual learning, studies show that their mechanisms work in a very similar way to actual human neural systems. And this is given the fact that modern neural systems are designed keeping engineering needs in mind and not the aspect of biological plausibility. It has been seen that CNN models show greater similarity to human and primate visual responses (Kalfas et al., 2018). They are in fact being considered as a model for the visual system (Lindsay, 2021). Recent evidence also suggests that Transformer based computer vision systems (such as Vision Transformers), employing self-attention do not just outperform CNNs on certain vision tasks, the errors they make are consistent with the errors that humans make (Tuli et al., 2021). However even though Transformers outperform LSTM based systems on NLP tasks, in a detailed study by Abnar et al. (2019), it was seen that LSTM based language models achieved a higher similarity score than the Transformer based models with human fMRI data on the same task. So, there is a streak of recent work showing that some neural systems do indeed have a lot of similarity with the human brains. In other words, some networks are more biologically plausible than others (Diehl et al., 2016; Bengio et al., 2015). However, it is also clear that models that are indeed performing better than others may not have greater similarity with humans. In other words, better neural models are not always biologically plausible [2].

In the real world, humans and all other animals, are continuously exposed to a stream of multimodal signals (via the different sensory organs) (Pollack, 2001). Cognitive scientists believe that this complex input space, in order to be reasonably processed by the brain, is converted into a manageable form (Kiebel et al., 2008). It is hypothesized that this transformation is achieved by exploiting the statistical regularities of the stimulus space (Chandrasekaran et al., 2009). Now, modern Deep Learning systems are also very effective at exploring the statistical regularities in the data fed to them (Sejnowski, 2020). Hence, if it is indeed just the capability to exploit statistical regularity in the multimodal data that enables humans to act 'intelligently' (measured by the performance on language tasks), an interesting question emerges. Apart from the performance, would neural networks exhibit similar cognitive biases(Goyal and Bengio, 2020) as humans? Multimodality is a relatively new field for artificial intelligence in particular and other disciplines in general. The overall trend in the context of studying multimodality in humans has been to study the modes in isolation rather than studying the synergy between them (Jewitt et al., 2016).

In the most fundamental sense, multimodality refers to the existence of more than one 'modality' within a given context. However, the definition of multimodality changes across disciplines. In a semiotic sense (Gibbons et al., 2012), the different modalities are considered as different semiotic modes (Siefkes, 2015). Under this framework, multimodal processes refers to the combination of various sign-systems such that the production and reception of such systems require interrelation of all the constituent sign-systems (Bateman, 2012). From a more cognitive and neuroscientific perspective, the idea of 'mode' is much more related to sensory organs (Forceville, 2021; Miralles, 2022). From this perspective, multimodal mental imagery is a crucial element for perception (Nanay, 2018).

In terms of multimodal literature, the definition of multimodality is similar to the notion of cognitive and neuroscientific literature (Parcalabescu

[2]The notion of biological plausibility here adapts the description of biological plausibility in Marblestone et al. (2016)

et al., 2021). However, the definitions of multimodality proposed in the machine learning literature is often task-relative and does not consider any general behavior as a whole. This has a couple of limitations. Natural Language Understanding is often considered to be the "Holy Grail" of NLP (Kiseleva et al., 2022) while human-like performance is considered to be the "Holy Grail" for most AI-applications (Ovchinnikova, 2012). So multimodal systems that are proficient in natural language understanding and exhibiting human-like performance is an important milestone in AI research. In this proposal, we consider a general definition of multimodality for machine learning that allows for the inclusion of any number of modalities and is task-independent.

So, current state of the art systems indeed exhibit traits that are similar to human brains while not being 'human-like' by design. At the same time, humans are good at generalization and learning new tasks while not completely forgetting the old tasks. This is something that modern neural systems struggle with.

And so, given these facts, this proposal concentrates on three major questions:

1. If humans and current AI systems were given the same multimodal tasks, how would their performance be compared.

2. How do we use multimodal deep learning systems to make predictions about observable human brain behaviour when handling multimodal tasks.

3. Does biological plausibility help in designing systems that exhibit human learning abilities.

This proposal is structured as follows. In Section 2, we review the existing literature looking into the different similarities and differences in the mechanisms of human learning. We follow that with reviewing how current techniques allow comparison between the representations learnt by neural models on specific tasks and human cognitive data (fMRI, EEG and so on) on the same task. We then explore catastrophic forgetting and multimodal learning in deep neural networks. In Section 4, we present the results of the experiments done so far. We first discuss about a novel dataset that we created by collecting the eye gaze, EEG and audio data from participants performing some language tasks as part of a psycholinguistic experiment. We then discuss about our experiments with using pretrained language models to predict human cognitive data. Then we go on to describe our experiments with exploring how different pretrained models encode different linguistic information in their layers. Finally, we describe a psycholinguistic experiment where we used a pretrained GPT-2 model to compare multimodal reading behaviour with human participants. Lastly, we describe our plan for future work in Section 5.

2 Related Work

Detailed exploration into the processes of human learning have been conducted with much depth and breadth from a perspective of neuroscience, cognitive science and psycholinguistics. The recent advances in deep learning have also contributed to an understanding about the mechanisms of learning in deep neural nets. In this section we take a closer look at both of those mechanisms to identify the points of similarity and differences between them.

2.1 Mechanisms of human and machine learning

Shuell (1986), in an early and influential description of learning from cognitive psychology perspective outlines human learning to be an "active and constructive" process that is mediated by higher-order processes in the brain. The hierarchical nature of psychological processes (Posner and Petersen, 1989) responsible for learning is hypothesized to be guided by *selective encoding*, *selective combination* and *selective comparison*. In other words, the process of learning involves the selection of relevant information from the stimuli (selective encoding) (Colegatef et al., 1973; Schotter et al., 2010), combining the selected information (selective combination) (Bartolomeo et al., 2012; Fernandez et al., 2019) and finally using the new encoded information by combining it with prior knowledge (selective comparison) (Heekeren et al., 2004). This general idea of hierarchical representation of knowledge is one of the core concepts driving representation learning (Bengio et al., 2013) with deep neural networks. Another concept that is at the heart of representation learning and thus deep learning is the back-propagation algorithm (Rumelhart et al., 1986). Ever since it's introduction, neuroscientists pondered on the biological plausibility of back-propagation (Stork, 1989). And recent evidence (Song et al., 2020; Lillicrap et al., 2020; Millidge et al., 2021) points to the fact that backpropagation might be possible in brain-learning mechanisms. However Bartunov et al. (2018) claim that back-propagation in its current form is impossible to implement in a real brain. The fact remains that we do not know exactly know how learning occurs in human brains.

In humans, it is the attention mechanism that selectively extracts information from the environment and relays it for further processing by the brain (Lindsay, 2020). Interestingly, the introduction of the "neural attention" (Bahdanau et al., 2014) and eventually "transformer attention" (Vaswani et al., 2017) was followed by major improvements in NLP applications. And although at a high-level both mechanisms of attention seem similar, they are not always correlated (Lai et al., 2020). In fact, there is a lack of research exploring the connection between human and artificial attention.

Also, current networks are data-hungry and lack in generalization performance in comparison to humans (Linzen, 2020). Training these models require tuning millions if not billions of parameters through multiple iterations on huge corpora. And current methods require operations on all the parameters of a neural model while learning (as well well as inference). In contrast, studies in neuroscience using multi-task fMRI data (Ramezani et al., 2014) shows that only a few regions in the brain are activated at the same time. In humans, it has been observed that single neurons respond selectively to the representations of the same concept across different sensory modalities (Quiroga et al., 2009). The same study also shows that the neurons grow less modality specific in the depths of certain brain areas. In other words, in certain deeper areas of the brain, concepts are represented in a way that neurons corresponding to those concepts exhibit activity whenever such concepts are referenced by any modality. To give an example, according to the study, the concept of "ice-cream" is represented in those brain areas in such a way that neurons corresponding to the 'concept' of ice-cream show activity whenever there is a reference to ice-cream by any modality (i.e. someone talking about ice-cream or a picture of an ice-cream). Goh et al. (2021) showed that the same phenomenon was observed in CLIP (Radford et al., 2021) based neural models.

Perconti and Plebe (2020) remarks, "Although deep learning models are grounded in the connectionist paradigm, their recent advances were basically developed with engineering goals in mind". The current focus of deep learning research is more applied and tailored to solve mostly practical industrial problems. However, as modern systems grow better at performing "complex cognitive tasks" (Knauff and Wolf, 2010), it makes it more interesting to compare not only their performance with respect to humans, but also to study the similarities and differences exhibited by their internal behaviour.

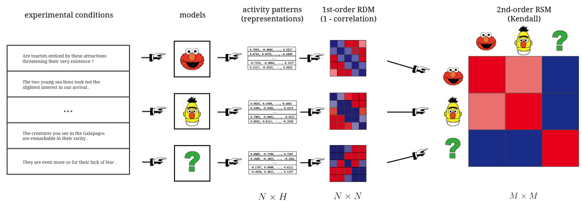

Figure 1: Using Representational Similarity Analysis (Kriegeskorte et al., 2008) to compare the representations of neural models and humans. Image taken from (Abdou et al., 2019)

2.2 Comparison of human and machine representations

In neuroscience and psycholinguistics, neural representation refers to the activity pattern in neurons associated with a particular experimental condition (Vilarroya, 2017). In the context of neural networks, representations of a model refer to the features that are extracted from the underlying data fed to the model. And given the fact that artificial neural networks were invented by keeping the biological neurons in mind (Sejnowski, 2020), there is now a growing attempt to compare the representations of both (Yang and Wang, 2020).

There are two main lines of research in this direction.

1. Comparing the representations of different models and their layers, trained on the same tasks, to get an understanding of the common features that the networks learn (Li et al., 2015; Kornblith et al., 2019; Nguyen et al., 2020; Kornblith et al., 2021).

2. Comparing the representations of trained artificial neural models with human brain data on the same task to determine their similarity with each other (Barrett et al., 2019).

Work on comparing representations between different convolution based neural models trained on different image datasets show that the layers closer to the input learn similar features (Wang et al., 2018). In contrast, Raghu et al. (2021) show that transformer based image models learn almost uniform representations across most their layers and that the representations diverge significantly in the last few layers (Grigg et al., 2021). From a more NLP standpoint, there is very limited work in this particular area. While geometrical features of the representations learnt by different models have been explored to some extent (Ethayarajh, 2019), the investigation of linguistic features captured by different models and their layers (Van Aken et al., 2019; Tenney et al., 2019; Mamou et al., 2020; Klafka and Ettinger, 2020; Maudslay and Cotterell, 2021), is an active field of research.

Work on comparing representations of neural models and human neurons is also a developing area of research. While major focus on this front has been to compare CNN based models with human cognitive data (Lindsay, 2020; Schrimpf et al., 2020; Xu and Vaziri-Pashkam, 2021), recent work has tried to extend the analysis for NLP models too. Banking on the success of large language models in various NLP tasks, there are now attempts to determine the ability of these models to 'capture' brain data (Schrimpf et al., 2021; Pasquiou et al., 2022). Apart from this, Abnar et al. (2019); Abdou et al. (2019, 2020, 2021) and (Eberle et al., 2022) have directly compared the representations from Transformer models with human cognitive data in the form of fMRI and gaze data. Figure 1 shows one way in which neural representations are compared with humans.

In a slightly different research direction, some recent works have explored the use of human cognitive data to augment deep learning models (Barrett and Hollenstein, 2020; Hollenstein et al., 2020) for diverse tasks like measuring text complexity, part-of-speech detection, named entity recognition and so on. Muttenthaler et al. (2020) demonstrate how to extract human language signals from EEG signals and inject that information into neural models. Futrell et al. (2019) demonstrated an interesting experimental paradigm of subjecting LSTM (Hochreiter and Schmidhuber, 1997) models through a controlled psycholinguistic experimental paradigm to shed light on the working of the models.

2.3 Catastrophic forgetting

Humans learn how to perform multiple tasks in succession over their lifespan. This capability of continual learning is difficult for current state of the art deep learning systems. However, it remains that learning to solve multiple tasks in a sequential manner is a key requirement for general AI (Legg and Hutter, 2007). In other words, it has been observed that when training a network on some task T1 is followed by training on some other task T2, the network optimizes its weights in a way that cater to solving T2 and thus forgetting the weights that it learnt to solve T1. This phenomenon has been called *catastrophic forgetting*. Catastrophic forgetting is a problem that was recognized way back when the first connectionist models appeared (McCloskey and Cohen, 1989; Ratcliff, 1990). As McCloskey and Cohen (1989) reason: the learning of new facts (interference) involves the building of new propositional structure[3] in the network. And since the new representations are separate from the other representations, the new adjustment of weights to encode the new input alters the network's response to other older inputs. French (1999) made a distinction between catastrophic interference and gradual interference. Gradual interference, i.e. forgetting the acquired knowledge gradually is something that occurs in humans too (McClelland et al., 1995). However what makes it truly catastrophic in artificial neural networks is that the new knowledge effectively wipes out the previous learning completely. In humans, the neocortical neurons are especially prone to catastrophic forgetting. But the neocortical learning system is complimented by the 'replay mechanism' of memories (experiences) from the hippocampus that, helps to perform tasks that have not been recently performed. Recent work in neuroscience later showed that animal brains may avoid catastrophic forgetting by storing the previously acquired knowledge in special neocortical circuits (Yang et al., 2009).

In the more relatively recent deep learning era, in one of the first works on catastrophic forgetting, Srivastava et al. (2013) argued that the choice of activation function has a significant effect on catastrophic forgetting. It was also found that when trained with dropout (Srivastava et al., 2014), net-

[3] Anderson and Bower (1974) defines a proposition as an associative configuration of elements which is abstract, structured according to certain rules of formation and has a 'truth value'(represents a particular concept/object)

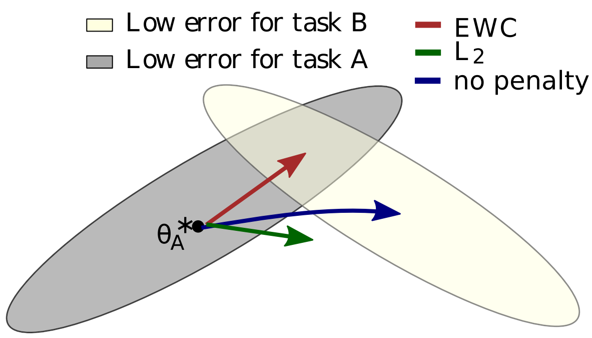

Figure 2: EWC regularizes the weight in such a way that the weights from task A is remembered while training on Task B. The image has been taken from (Kirkpatrick et al., 2017)

works learning similar tasks in a sequence suffer from lesser catastrophic forgetting than learning dissimilar tasks (Goodfellow et al., 2013). Luong et al. (2015) showed that training on two tasks simultaneously, the models optimize their weights to perform well on both tasks. A major development in the direction of reducing catastrophic forgetting has been the use of biologically inspired methods like elastic weight consolidation (EWC) (Kirkpatrick et al., 2017) and Aljundi et al. (2018), both of which rely on the Fisher Information Matrix (MacKay, 1992). EWC attempts to reduce catastrophic forgetting by penalizing the difference between the original and updated weights, with each weight scaled with respect to their importance to the original task. The general idea behind EWC rests on the hypothesis that there is a likely solution of some new task T2 in the vicinity of the weight space learnt while learning to solve an earlier task T1. EWC makes use of the fact that current models are overparameterized and multiple solutions to a particular task can be found in the huge parameter space. As a result of such regularization, the parameters are forced to stay in the vicinity of the weights learnt to solve the previous tasks. This effect is shown in Figure 2. Recent work in Automatic Speech Recognition (Eeckt et al., 2022) has shown that using EWC to protect adapter (Houlsby et al., 2019) weights from catastrophic forgetting leads to impressive gains in alleviating the problem of catastrophic forgetting in ASR systems in multi-task learning settings.

In terms of comparing catastrophic forgetting across model architecture families, Arora et al. (2019) found that LSTMs are more prone to catastrophic forgetting than CNNs and that increasing model capacity does not really help with reducing catastrophic forgetting. However the claim that increasing model capacity does not help with reducing catastrophic forgetting has been challenged by (Ramasesh et al., 2021) where they hsow that larger models suffer less from forgetting.

2.4 Multimodal Learning

Given the definition of multimodality in machines (as presented in Section 1), a major challenge in building efficient multimodal systems is to address the heterogeneity gap (Guo et al., 2019). In other words, since different model parts are trained on data of different modalities or different tasks, the features learnt by the individual parts reside on separate sub-spaces. And hence, the vector representations associated with similar semantics would be very different in different modalities.

Addressing this heterogeneity gap has led to the introduction of the concept of joint representation learning by projecting the representations from individual modalities into a common shared subspace. This idea of fusion (Gao et al., 2020) has been applied across multiple model architectures for a diverse set of problems. Recent work has extended the idea of multimodal fusion to the Transformer (Vaswani et al., 2017). Tsai et al. (2019) on the other hand propose a cross-modal architecture where the attention block of the Transformer (Vaswani et al., 2017) is modified to fuse data from two different modalities. This idea of cross-attention is further extended by Nagrani et al. (2021) by introducing a set of fusion 'bottlenecks' and achieving state of the art results on a number of benchmarks. (Zellers et al., 2022) on the other hand concatenate the representations obtained from modality specific encoders and process them via a vanilla transformer encoder.

Another technique that is widely used to alleviate the heterogeneity gap is by aligning the representations from different modalities to identify the relations between them (Baltrušaitis et al., 2018). Recently, models like CLIP (Radford et al., 2021), VATT (Akbari et al., 2021) and ALIGN (Jia et al., 2021) , all use a contrastive loss (Wang and Liu, 2021) to align the modality specific data and learn useful relations between them.

Recently however, there is a trend (Reed et al., 2022; Kaiser et al., 2017) to encode data from different modalities together and pass them through a common self attention mechanism. Although this methodology is still pretty new and relatively unexplored.

3 Methodology

In this section, we describe in details the tasks that we focus on with respect to the three questions posed in the Introduction (Section 1). Our focus lies primarily on three kinds of tasks:

- **Comparison of machine representations**: Different model architectures with different training objectives, often converge at comparable performance metrics. Our objective, using techniques like probing(Belinkov, 2022), is to investigate the learnt representations of such models and determine how different linguistic features are encoded in their layers.
- **Assessing the capability neural models to predict human multimodal behavior**: Merkx and Frank (2020) demonstrated how a "cognitively implausible model" such as the Transformer performs better at predicting cogitive data. Our goal is to extend this line of research by evaluating the performance of various models on prediction of cognitive data across a diverse range of tasks. Our principal concern is to identify if, for multimodal tasks, biological plausibility (banking on existing work in neuroscience) indeed translates to better performance for machines.
- **Compare humans and neural model performance on the same task**: The recent advances in deep learning has led to claims of neural models performing at par with humans on the same tasks. But as Borowski et al. (2019) and Funke et al. (2021) show, there are a couple of problems in the way that current research pits human performance against machine performance. Our goal is to address this problem by carefully designing a framework to test human performance against machines on multimodal tasks.

Our investigation into comparing human multimodal behavior with neural networks constitutes of studying language modelling and translation mechanisms under multimodal settings. To this end, we create a carefully designed psycholinguistic experiment to collect the behavioural data of humans. The experiment design is done to ensure that it can be replicated by a trained neural network. In the next few lines, we give a brief description of the tasks examined by the experiment.

3.0.1 Language Modelling

We frame the task of human reading as a language modelling task. Given a sentence s with N tokens such that $s = \{s_0, S_1, ..., s_N - 1\}$ in a corpus of sentences S, a language model M with parameters θ attempts to learn a distribution p_θ, such that p_θ is close to the real distribution p_{data}. In other words, the parameters of the model M, when optimized against a suitable loss function L (cross-entropy) learns to approximate how the words are distributed in the different sentences of S.

$$L(p_\theta, p_{data}) = \sum_{s \in S} p_\theta(s) p_{data}(s) \quad (1)$$

And hence when looked at the level of individual words in a sentence, p_θ can be written as:

$$p_\theta = \prod_{i=0}^{N-1} p_\theta(y_i | y_{i-1} ... y_0) \quad (2)$$

Our hypothesis is, given the same text to humans and neural models, word predictability statistics of humans are correlated with the probability associated by the models with the tokens in the text. Thus, effectively a language model learns to predict the occurrence of a token s_i in a sentence s given the occurrence of the tokens $\{y_{i-1}, ...y_0\}$ previously in the sentence.

We frame human reading as a language modelling task, where we posit that the probability associated with the prediction of the next token in a sentence translates to word predictability (Smith and Levy, 2013) in reading.

3.0.2 Machine Translation

Just as we train language models to predict the next token given a sentence context, in machine translation the goal is to predict the next token in a target language given the sentence context and a source language sentence. In other words, given a sentence $s = \{s_0, S_1, ..., s_N - 1\}$ in source-language $(L1)$ and its translation $t = \{t_0, t_1, ..., t_M - 1\}$ in the target language, the model is tasked with learning the distribution:

$$p_\theta = \prod_{i=0}^{N-1} p_\theta(t_i | t_{i-1} ... t_0, \mathbf{s}) \quad (3)$$

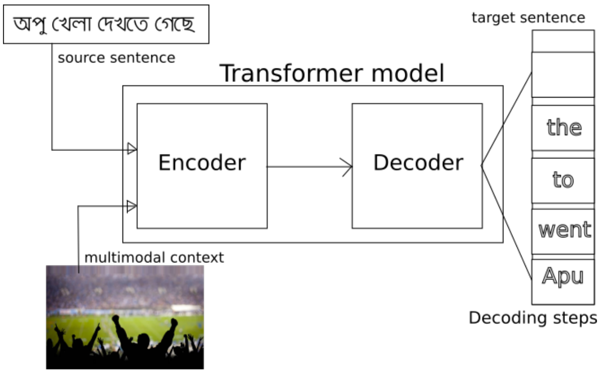

Figure 3: Multimodal machine translation model. The source sentence (Bengali in this example) and multimodal context are fed into an encoder that creates a joint representation of the both. This joint representation is then fed to the decoder that generates the translation of the source sentence (English in this example).

Similar to the approach outlined in the previous section, we test human and machine performance on a translation task (English to Czech).

3.0.3 Multimodal Language Modelling

To assess the role of additional multimodal information on the performance of language modelling, we modify the task of language modelling with an additional multimodal context such that the task translates to learning the distribution:

$$p_\theta = \prod_{i=0}^{N-1} p_\theta(y_i | y_{i-1} ... y_0, \mathbf{C}) \tag{4}$$

Here every token y_i is modelled as being conditioned on both the sentence context and the multimodal context (C) (shown in Figure 3.

3.0.4 Multimodal Machine Translation

Similar to multimodal language modelling, we extend the machine translation task with a multimodal context C to learning of the following distribution:

$$p_\theta = \prod_{i=0}^{N-1} p_\theta(t_i | t_{i-1} ... t_0, \mathbf{s}, \mathbf{C}) \tag{5}$$

3.1 Evaluation

To compare the performance of machines on the machine translation task and the multimodal machine translation task, we use commonly used metrics in machine translation literature like BLEU (Papineni et al., 2002) or METEOR (Lavie and Denkowski, 2009). However, since we ensure that both machine translation models are fed the same set of sentences, we then compare the outputs from

Figure 4: EMMT: Experiment setup with eyetracker (EEG and audio recorder not shown).

both using both automatic metrics and human annotators to evaluate the change in output quality. We use the same methodology for comparison of human outputs from translating with and without multimodal context. This gives us a framework for comparison. To compare the human and machine performance on the ‘reading’ tasks, we use the scores from the final softmax scores of the models to compare with metrics like surprisal (Monsalve et al., 2012).

4 Experiments

So far our experiments have concentrated on creating experimental frameworks to compare humans and neural models on the same tasks. We have also explored the nature of representation of different NLP models and their capabilities to predict cognitive data.

4.1 EMMT: A simultaneous eye-tracking, 4-electrode EEG and audio corpus for multi-modal reading and translation scenarios

In this section we describe our experiments with EMMT (Bhattacharya et al., 2022a), a dataset we created, containing monocular eye movement recordings, audio and 4-electrode electroencephalogram (EEG) data of 43 participants. The aim of the experiment was to collect cognitive data as responses of participants engaged in a number of language intensive tasks involving different text-image stimuli settings when translating from English to Czech. The experiment was designed in a way that it could be replicated by a neural system later (described in Section 5).

Each participant was exposed to 32 text-image stimuli pairs and asked to (1) read the English sentence, (2) translate it into Czech, (3) consult the image, (4) translate again, either updating or repeating the previous translation. Figure 5 shows how the four different stages were shown to the participants.

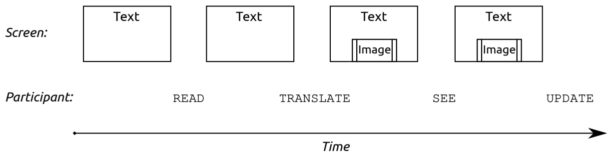

Figure 5: Visualization of the four experiment stages.

For the experiment, we used two sentence types (unambiguous and ambiguous) with three image stimuli types (related, unrelated and no image) in a within-subjects design, i.e., every participant is exposed to all conditions (but never on the same stimulus). This resulted in the following six configurations:

- **UR** (unambiguous sentence + related image)
- **UU** (unambiguous sentence + unrelated image)
- **UN** (unambiguous sentence + no image)
- **AR** (ambiguous sentence + related image)
- **AU** (ambiguous sentence + unrelated image)
- **AN** (ambiguous sentence + no image)

The related images (congruent stimuli) match the content of the text. The unrelated images (incongruent stimuli) are not relevant to the text. The “no image” condition refers to a control condition that is comprised of an image with white background and a text saying *No visual clue for this case*. Apart for these configurations, there was a pair of contrastive sentences (in each probe labelled as:

1. **AR** (ambiguous sentence + related image): *A person in a blue ski suit is racing two girls on skis.*

2. **UR** (unambiguous sentence + related image)): *A person in her blue ski suit is racing two girls on skis.*

The recordings were collected over a two week period and all the participants included in the study were Czech natives with strong English skills.

Data were recorded from each participant in a single session. Each experiment started with

Figure 6: `READ` (Stage 1) and `TRANSLATE` (Stage 2)

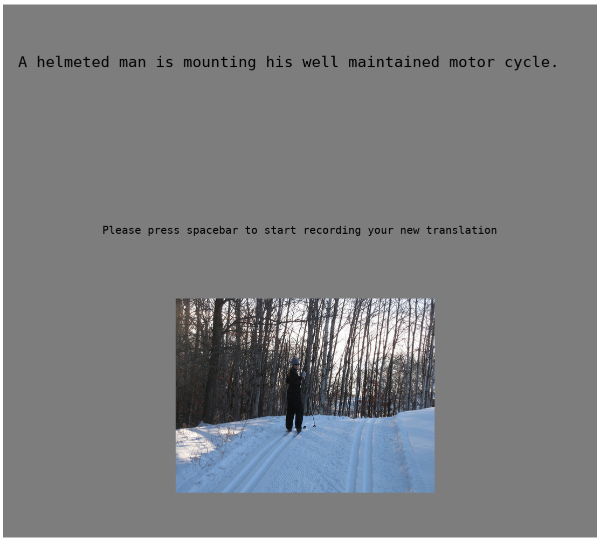

Figure 7: `SEE` (stage 3) and `UPDATE` (Stage 4).

the calibration and validation of the equipment involved (eye-tracker and EEG recorder). Each participant was then led through a practice round with four dummy stimuli, to get them acquainted with the procedures of the experiment. Being a self-paced experiment design, the participants were given an option to temporarily pause the experiment after completing the four stages of a stimulus to take a small break. If the participants opted to pause, the experiment would resume again with calibration and validation before starting from where it was stopped.

The average response times for each stage are shown in Table 1. In Stage 3 (`See`), when the image was first presented, the difference across conditions is very prominent, especially for unrelated and “no image” cases. The highest time is expended for the related images, followed by the unrelated images and finally the "no image" case. The same trend, with lower distinction, is repeated in Stage 4.

Condition	Stage 1	Stage 2	Stage 3	Stage 4
AR	11.41	8.89	8.46	7.47
AU	11.29	9.22	7.52	7.09
AN	11.44	8.98	5.95	6.98
UR	10.63	8.80	8.25	7.37
UU	11.20	8.81	6.92	6.73
UN	10.73	8.01	5.29	6.36

Table 1: Average duration of all stages (`READ`, `TRANSLATE`, `SEE`, `UPDATE`) and conditions.

The three image stimuli conditions can be thought of as a variant of the classic Stroop task (Stroop, 1935) involving the naming of coloured words (MacLeod, 1992). In this experiment, the stimuli in categories Condition 1 (AR, UR), Condition 2 (AU, UU) and Condition 3 (AN, UN) correspond to congruent, incongruent and neutral stimuli respectively. Table 2 shows t-test results for comparison of various pairs of stimuli conditions. The image and the textual stimuli, therefore, correspond to a variation of the classical visual-verbal stimuli condition of the original Stroop task.

Condition	t	p
AR-AU	1.441	0.150
AR-AN	4.085	<0.001
AU-AN	3.318	0.001
UR-UU	2.725	0.007
UR-UN	7.046	<0.001
UU-UN	4.618	<0.001

Table 2: T-Test results of case-wise comparisons in times in Stage 3.

We also found that participants spent a longer time translating the sentences with related images for both classes of sentences as there was significant cognitive effort required to integrate the visual information into the translation that they already had in their memory. For the incongruent stimuli, participants chose to disregard the visual information.

4.2 Eye-Tracking prediction using pretrained language models

This section describes our submission (Bhattacharya et al., 2022b) to the cognitive data prediction task at CMCL 2022 (Hollenstein et al., 2022). The task constituted of predicting eye-gaze attributes associated with reading sentences as a regression task. The data for the task was comprised of eye movements corresponding to reading sentences in six languages (Chinese, Dutch, English, German, Hindi, Russian). The training data for the task contained 1703 sentences while the development set and test set contained 104 and 324 sentences respectively. The data was presented in a way such that for each word in a sentence there were four associated eye-tracking features in the form of the mean and standard deviation scores of the Total Reading Time (TRT) and First Fixation Duration (FFD). The features in the data were scaled in the range between 0 and 100 to facilitate evaluation via the mean absolute average (MAE).

A total of 48 models of different configurations were trained with the data provided for the shared task. Thee models were primarily categorized as System-1 and System-2 models. For some word corresponding to a sentence in the dataset, System-1 models provided no additional context information. System-2 models on the other hand, contained the information of all the words in the sentence that preceded the current word, providing additional context. All systems under the System-1/2 labels were further trained as a BERT based system or a XLM based system. 1.

Corresponding to each such language models, the impact of different fine-tuning strategies on system performance was studied. Hence, for one setting, only the contextualized word representation was utilized by freezing the model weights and putting a learnable regression layer on top of the model output layer (classifier). Alternatively, the models were fine-tuned with the regression layer on top of them. Additionally, we also performed experiments with pooling strategies for the layer representations by either using the final hidden representation of the first sub-word encoding of the input or aggregating the representations of all subwords using mean-pooling or sum-pooling. The rationale behind using different pooling strategies was to have a sentence-level representation of the input tokens.

Our experiments demonstrated that the inclusion of context (previous words occurring in the sentence) helps the models to predict eye-tracking attributes better. We also found that XLM based models perform relatively better than the BERT based models. Our submissions achieved an average MAE of 5.72 and ranked 5^{th} in the shared task.

The average MAE showed further reduction to 5.25 in post task evaluation.

4.3 Other experiments

Our recent work involved probing pretrained language models (BERT(Devlin et al., 2019) and GPT-2(Radford et al., 2019)) to assess their ability to capture subtle linguistic traits like ambiguity, grammaticality and sentence complexity. We found that large pre-trained language models represent sentence ambiguity in a much less extractable way. We also documented that template-based datasets, such as BLiMP (Warstadt et al., 2020) used for sentence acceptability, are not good for probing because of surface-level artefacts. The experiment also showed that features relevant to the detection of ambiguity, complexity and grammaticality are more concentrated on the middle layers of the pre-trained models.

Another recent work of ours explored an extension to the well known Shannon's game (Shannon, 1951) by including an optional extra modality in the form of images and running it on human participants. We also replicated a version of this experiment on the GPT-2 family and compared the results with human counterparts. We observe that the GPT-2 model is able to make use of the extra modality to improve its prediction. We also observe that the GPT-2 model also exhibits some similar patterns to human annotators.

5 Future Plans

So far we have compiled the multimodal dataset recording human behavior on language modelling and machine translation tasks. We have also performed the initial experiments on comparing humans and pretrained language models like GPT-2 on a multimodal reading task. In addition, we have done some preliminary investigations into exploring how different pretrained models encode linguistic data across their layers and how suitable they are predict human cognitive data like eye-tracking statistics.

We plan to continue with the exploration of the similarities of human and machine behavior on multimodal tasks by first formulating a detailed account of multimodal processing in humans using the data collected in the form of the EMMT corpus. The investigation will also involve using recent state-of-the-art multimodal models on the stimulus from EMMT to gather data about machine performance on the data. We will then finally compare the accounts of the human behaviour and machine behavior across the models to understand how they fare against each other.

On the other front, we will continue with our experiments with investigating the layers of different pretrained models (including multimodal models) to gain an understanding of how different models encode linguistic information in their layers. We hope to combine the knowledge gained from this endeavour with the results from the results of comparison of human and machine behavior to identify the factors (architecture, optimizer, pre-training objective etc) that make some neural systems better at multimodal tasks. We also envisage to use this knowledge to determine if biological plausibility of neural models also translate to them being better at multimodal tasks. We would additionally attempt to determine if biologically inspired methods in neural architectures impact catastrophic forgetting in a multi-task learning scenario.

We would try to use the results from the experiments described above to pick the best performing models. Having already done some exploratory experiments in this direction, we would use the models to predict human cognitive data and thus extend the line of research in this area as described in (Laverghetta et al., 2022).

Finally, apart from using the state-of-the-art systems, we would also attempt to test some modifications to current models and new architectures for multimodal learning.

6 Conclusion

This proposal highlights three questions that emerge as deep neural networks get more powerful, sophisticated and perform more 'complex cognitive' tasks. We ask if human participants and state-of-the art neural systems trained on multimodal tasks were asked to perform the same multimodal language tasks, how would they fare against each other. We also ask how capable 'cognitively implausible' models are to predict observable human behavior. And these questions lead us to ask if biological plausibility in anyway impacts the performance of neural models. We highlight the relevant literature pertaining to these questions and the gap in there. Finally, we describe the experiments that we have performed so far in the attempts to find answers to the questions posed above.

References

Mostafa Abdou, Ana Valeria González, Mariya Toneva, Daniel Hershcovich, and Anders Søgaard. 2021. Does injecting linguistic structure into language models lead to better alignment with brain recordings? *arXiv preprint arXiv:2101.12608*.

Mostafa Abdou, Artur Kulmizev, Felix Hill, Daniel M Low, and Anders Søgaard. 2019. Higher-order comparisons of sentence encoder representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5838–5845.

Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. 2020. The sensitivity of language models and humans to winograd schema perturbations. *arXiv preprint arXiv:2005.01348*.

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154.

John R Anderson and Gordon H Bower. 1974. A propositional theory of recognition memory. *Memory & Cognition*, 2(3):406–412.

Gaurav Arora, Afshin Rahimi, and Timothy Baldwin. 2019. Does an lstm forget more than a cnn? an empirical study of catastrophic forgetting in nlp. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 77–86.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. 2021. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

David GT Barrett, Ari S Morcos, and Jakob H Macke. 2019. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current opinion in neurobiology*, 55:55–64.

Maria Barrett and Nora Hollenstein. 2020. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14(11):1–16.

Paolo Bartolomeo, Michel Thiebaut de Schotten, and Ana B Chica. 2012. Brain networks of visuospatial attention and their disruption in visual neglect. *Frontiers in human neuroscience*, 6:110.

Sergey Bartunov, Adam Santoro, Blake Richards, Luke Marris, Geoffrey E Hinton, and Timothy Lillicrap. 2018. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. *Advances in neural information processing systems*, 31.

John A Bateman. 2012. The decomposability of semiotic modes. In *Multimodal Studies*, pages 37–58. Routledge.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Yoshua Bengio, Dong-Hyun Lee, Jorg Bornschein, Thomas Mesnard, and Zhouhan Lin. 2015. Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*.

Sunit Bhattacharya, Věra Kloudová, Vilém Zouhar, and Ondřej Bojar. 2022a. Emmt: A simultaneous eye-tracking, 4-electrode eeg and audio corpus for multi-modal reading and translation scenarios. *arXiv preprint arXiv:2204.02905*.

Sunit Bhattacharya, Rishu Kumar, and Ondrej Bojar. 2022b. Team\'ufal at cmcl 2022 shared task: Figuring out the correct recipe for predicting eye-tracking features using pretrained language models. *arXiv preprint arXiv:2204.04998*.

Judy Borowski, Christina M Funke, Karolina Stosio, Wieland Brendel, T Wallis, and Matthias Bethge. 2019. The notorious difficulty of comparing human and machine perception. In *2019 Conference on Cognitive Computational Neuroscience*, pages 2019–1295.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chandramouli Chandrasekaran, Andrea Trubanova, Sébastien Stillittano, Alice Caplier, and Asif A Ghazanfar. 2009. The natural statistics of audiovisual speech. *PLoS computational biology*, 5(7):e1000436.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Michael Chui, James Manyika, Mehdi Miremadi, Nicolaus Henke, Rita Chung, Pieter Nel, and Sankalp Malhotra. 2018. Notes from the ai frontier: Insights from hundreds of use cases. *McKinsey Global Institute*, page 28.

Robert L Colegatef, James E Hoffman, and Charles W Eriksen. 1973. Selective encoding from multielement visual displays. *Perception & Psychophysics*, 14(2):217–224.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Peter U Diehl, Guido Zarrella, Andrew Cassidy, Bruno U Pedroni, and Emre Neftci. 2016. Conversion of artificial recurrent neural networks to spiking neural networks for low-power neuromorphic hardware. In *2016 IEEE International Conference on Rebooting Computing (ICRC)*, pages 1–8. IEEE.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309.

Steven Vander Eeckt et al. 2022. Using adapters to overcome catastrophic forgetting in end-to-end automatic speech recognition. *arXiv preprint arXiv:2203.16082*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Natalia B Fernandez, Wiebke J Trost, and Patrik Vuilleumier. 2019. Brain networks mediating the influence of background music on selective attention. *Social cognitive and affective neuroscience*, 14(12):1441–1452.

Charles Forceville. 2021. Multimodality. In *The Routledge Handbook of Cognitive Linguistics*, pages 676–687. Routledge.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Christina M Funke, Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas SA Wallis, and Matthias Bethge. 2021. Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, 21(3):16–16.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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)*, pages 32–42.

Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. 2018. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31.

Alison Gibbons, Mark Z Danielewski, Steve Tomasula, Stephen Farrell, Jonathan Safran Foer, and Graham Rawle. 2012. Multimodality. In *Multimodality, cognition, and experimental literature*, volume 3, pages 8–9. Routledge New York.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Anirudh Goyal and Yoshua Bengio. 2020. Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091*.

Tom George Grigg, Dan Busbridge, Jason Ramapuram, and Russ Webb. 2021. Do self-supervised and supervised methods learn similar visual representations? *arXiv preprint arXiv:2110.00528*.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.

In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hauke R Heekeren, Sean Marrett, Peter A Bandettini, and Leslie G Ungerleider. 2004. A general mechanism for perceptual decision-making in the human brain. *Nature*, 431(7010):859–862.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27, Marseille, France. European Language Resources Association.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2022. Cmcl 2022 shared task on multilingual and crosslingual prediction of human reading behavior. In *CMCL Shared Task on Multilingual and crosslingual prediction of human reading behavior*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Lukas S Huber, Robert Geirhos, and Felix A Wichmann. 2021. A four-year-old can outperform resnet-50: Out-of-distribution robustness may not require large-scale experience. In *SVRHM 2021 Workshop@ NeurIPS*.

Carey Jewitt, Jeff Bezemer, and Kay O'Halloran. 2016. Navigating a diverse field. In *Introducing multimodality*, page 2. Routledge.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *arXiv preprint arXiv:1706.05137*.

Ioannis Kalfas, Kasper Vinken, and Rufin Vogels. 2018. Representations of regular and irregular shapes by deep convolutional neural networks, monkey inferotemporal neurons and human judgments. *PLoS Computational Biology*, 14(10):e1006557.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Stefan J Kiebel, Jean Daunizeau, and Karl J Friston. 2008. A hierarchy of time-scales and the brain. *PLoS computational biology*, 4(11):e1000209.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, et al. 2022. Interactive grounded language understanding in a collaborative environment: Iglu 2021. *arXiv preprint arXiv:2205.02388*.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811.

Markus Knauff and Ann G Wolf. 2010. Complex cognition: the science of human reasoning, problem-solving, and decision-making.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. 2021. Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems*, 34:28648–28662.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Qiuxia Lai, Salman Khan, Yongwei Nie, Hanqiu Sun, Jianbing Shen, and Ling Shao. 2020. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 23:2086–2099.

Antonio Laverghetta, Animesh Nighojkar, Jamshidbek Mirzakhalov, and John Licato. 2022. Predicting human psychometric properties using computational language models. In *The Annual Meeting of the Psychometric Society*, pages 151–169. Springer.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2):105–115.

Shane Legg and Marcus Hutter. 2007. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4):391–444.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. 2015. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*.

Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. 2020. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346.

Grace W Lindsay. 2020. Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience*, 14:29.

Grace W Lindsay. 2021. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

David JC MacKay. 1992. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472.

Colin M MacLeod. 1992. The stroop task: The" gold standard" of attentional measures. *Journal of Experimental Psychology: General*, 121(1):12.

Jonathan Mamou, Hang Le, Miguel Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2020. Emergence of separable manifolds in deep language representations. *arXiv preprint arXiv:2006.01095*.

Adam H Marblestone, Greg Wayne, and Konrad P Kording. 2016. Toward an integration of deep learning and neuroscience. *Frontiers in computational neuroscience*, page 94.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do syntactic probes probe syntax? experiments with jabberwocky probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 124–131.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419.

James L McClelland, David E Rumelhart, PDP Research Group, et al. 1987. *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models*, volume 2. MIT press.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Danny Merkx and Stefan L Frank. 2020. Human sentence processing: Recurrence or attention? *arXiv preprint arXiv:2005.09471*.

Beren Millidge, Anil Seth, and Christopher L Buckley. 2021. Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2107.12979*.

David Miralles. 2022. Multi-modal self-adaptation during object recognition in an artificial cognitive system. *Scientific Reports*, 12.1:1–12.

Irene Fernandez Monsalve, Stefan L Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408.

Lukas Muttenthaler, Nora Hollenstein, and Maria Barrett. 2020. Human brain activity for machine attention. *arXiv preprint arXiv:2006.05113*.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213.

Bence Nanay. 2018. Multimodal mental imagery. *Cortex*, 105:125–134.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. 2020. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*.

Ekaterina Ovchinnikova. 2012. Natural language understanding and world knowledge. In *Integration of world knowledge for natural language understanding*, volume 3, page 16. Springer Science & Business Media.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Letitia Parcalabescu, Nils Trost, and Anette Frank. 2021. What is multimodality? In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 1–10, Groningen, Netherlands (Online). Association for Computational Linguistics.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

Alexandre Pasquiou, Yair Lakretz, John T Hale, Bertrand Thirion, and Christophe Pallier. 2022. Neural language models are not born equal to fit brain data, but training helps. In *International Conference on Machine Learning*, pages 17499–17516. PMLR.

Pietro Perconti and Alessio Plebe. 2020. Deep learning and cognitive science. *Cognition*, 203:104365.

Gerald S Pollack. 2001. Analysis of temporal patterns of communication signals. *Current opinion in neurobiology*, 11(6):734–738.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Michael I Posner and Steven E Petersen. 1989. The attention system of the human brain.

Rodrigo Quian Quiroga, Alexander Kraskov, Christof Koch, and Itzhak Fried. 2009. Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology*, 19(15):1308–1313.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128.

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. 2021. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Mahdi Ramezani, Kris Marble, H Trang, Ingrid S Johnsrude, and Purang Abolmaesumi. 2014. Joint sparse representation of brain activity patterns in multi-task fmri data. *IEEE Transactions on Medical Imaging*, 34(1):2–12.

Roger Ratcliff. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175*.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Elizabeth R Schotter, Raymond W Berry, Craig RM McKenzie, and Keith Rayner. 2010. Gaze bias: Selective encoding and liking effects. *Visual Cognition*, 18(8):1113–1132.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. 2020. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007.

Terrence J Sejnowski. 2020. The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(48):30033–30038.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. 2020. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR.

Claude E Shannon. 1951. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.

Thomas J Shuell. 1986. Cognitive conceptions of learning. *Review of educational research*, 56(4):411–436.

Martin Siefkes. 2015. How semiotic modes work together in multimodal texts: Defining and representing intermodal relations. *Living Linguistics*, 1.1:113–131.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Yuhang Song, Thomas Lukasiewicz, Zhenghua Xu, and Rafal Bogacz. 2020. Can the brain do backpropagation?—exact implementation of backpropagation in predictive coding networks. *Advances in neural information processing systems*, 33:22566–22579.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Rupesh K Srivastava, Jonathan Masci, Sohrob Kazerounian, Faustino Gomez, and Jürgen Schmidhuber. 2013. Compete to compute. *Advances in neural information processing systems*, 26.

David G Stork. 1989. Is backpropagation biologically plausible. In *International Joint Conference on Neural Networks*, volume 2, pages 241–246. IEEE Washington, DC.

J Ridley Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643.

Kar-Han Tan and Boon Pang Lim. 2018. The artificial intelligence renaissance: Deep learning and the road to human-level machine intelligence. *APSIPA Transactions on Signal and Information Processing*, 7.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. 2021. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*.

Betty Van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1823–1832.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Oscar Vilarroya. 2017. Neural representation. a survey-based analysis of the notion. *Frontiers in Psychology*, 8:1458.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John Hopcroft. 2018. Towards understanding learning representations: To what extent do different neural networks learn the same representation. *Advances in neural information processing systems*, 31.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Yaoda Xu and Maryam Vaziri-Pashkam. 2021. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications*, 12(1):1–16.

Guang Yang, Feng Pan, and Wen-Biao Gan. 2009. Stably maintained dendritic spines are associated with lifelong memories. *Nature*, 462(7275):920–924.

Guangyu Robert Yang and Xiao-Jing Wang. 2020. Artificial neural networks for neuroscientists: A primer. *Neuron*, 107(6):1048–1070.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.