Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/30/20230096

Neurophysiological and psychophysical references for trends in supervised VQA multimodal deep learning: An interdisciplinary meta-analysis

Jidong Qi

Center of brain and cognition science, University of Rochester, Rochester, United states

jqi9@u.rochester.edu

Abstract. Leading trends in multimodal deep learning for visual-question answering include Multimodal joint-embedding model, multimodal attention-based model, and multimodal external knowledge-based model. Several mechanisms and strategies are used in these models, including representation fusion methods, co-attention mechanisms, and knowledge base retrieval mechanisms. While a variety of works have comprehensively reviewed these strategies, a key gap in research is that there is no interdisciplinary analysis that connects these mechanisms with discoveries on human. As discussions of Neuro-AI continues to thrive, it is important to consider synergies among human level investigations and ANNs, specifically for using AI to reproduce higher order cognitive functions such as multisensory integration. Thus, Present meta-analysis aimed at the reviewing and connecting neurophysiological and psychophysical references to trends in VQA multimodal deep learning, focusing on 1) Providing back-up explanations for why several strategies in VQA MMDL leads to performances that are closer to human level and 2) Using VQA MMDL as an example to demonstrate how interdisciplinary perspective may foster the development of human level AI. The result of the meta-analysis builds connections between several sub-fields: Joint embedding mechanisms and SC neurons, multimodal attention mechanism and the retro-cue effect, and external knowledge base and engram mechanisms.

Keywords: multimodal deep learning, VQA, joint-embeddings, superior colliculus neurons, parallel and alternating attention.

1. Introduction

Multimodal learning involves the processing and integration of information from multiple modalities [1-3]. The modeling of human multimodal learning is interdisciplinary and benefits from computational, neurophysiological, and psychophysical approaches [2, 3]. Multimodal deep learning (MMDL) proposes the possibility of reproducing human multimodal learning process using artificial neuron networks (ANNs) [1, 2]. Contrast to unimodal ANNs, MMDL models interpret information steams from different sources and in different formats. Recent discoveries have shown the potential of MMDL models on solving a variety of multimodal tasks, including multimodal image and video description, multimodal visual question answering, multimodal speech synthesis, and multimodal emotion recognition [2].

DOI: 10.54254/2755-2721/30/20230096

Visual question answering (VQA) task involves answering natural language questions about a given image [4]. This task poses a rich set of challenges on multimodal recognition and categorization, also enabling a quantified correctness measurement of the answering agent's performances [4]. Several MMDL models were developed to accomplish VQA [2]. Trends of VQA MMDL models involve explorations of three model architectures: multimodal joint-embedding models (MMJEMs), multimodal attention-based models (MMAMs), and multimodal external knowledge based models (MMEKMs) [2]. MMJEM engages the construction of representative embedding for unimodal features and cross-modal correlations [5]. MMAM engages the construction of mechanisms that effectively guide attentional weightings across modalities [6]. MMEKM engages the construction of external fact-based accompanying database that functioned as sources of retrieval when the agent encounters VQA tasks [7]. Evidences have shown that all three types of architectures to some extent successfully capture multimodal correlational features as human beings do [2-7].

While VQA MMDL models are pursuing human level VQA learning performances, neurophysiological and psychological discoveries also contribute to multimodal VQA cognition modeling. Investigations in these fields cover similar discussions as those in the field of VQA MMDL. In relation to MMJEMSs, the representation of multimodal features were frequently studied among cortical and circuit level [8, 9] . In relation to MMAMs, modality-general attention was studied as a neurophysiological mechanism, and multimodal attention switching was specifically investigated through behavioral psychophysics [10]. In relation to MMEKMs, retrieving pre-existed multimodal knowledge base were studied under the context of cognitive neuroscience [11, 12].

However, although vast amount of works was done in each field, current review or meta-analysis lack of providing a combinational review of how these field-specific investigations may generate synergies. Discussions were made to focus on recent developments under the framework of one specific fields. Current criticisms toward interdisciplinary analytics mainly lies in the limited intersubjectivity of concepts, as in this case the different definition of 'attention' among neuroscience and computer science [13]. Nevertheless, such criticism completely rejects the intersubjectivity and force false rejections toward interdisciplinary synergies. Through selecting on one path and bracketing the other, analysis focuses on non-universal concepts takes will be able to integrate multidisciplinary information through using one concept system to prompt articulations of phenomena in another discipline [13]. Such multidisciplinary concept synthesis, with suitable number of context-related adaptions, have the potential to maintain the inner coherence in interdisciplinary analysis.

Thus, present meta-analysis aims at providing an interdisciplinary perspective on the development of VQA modeling. The analysis is organized according to the model-base framework of MMDL. Neurophysiology and psychophysics will be allied field of research that provide references to MMDL models. Discussion will surround the key debates or discoveries in MMDL mechanisms of joint embeddings, attention, and external knowledge base. In general, the primary goals of current meta-analysis are: 1) Providing back-up explanations for why several strategies in VQA MMDL leads to performances that are closer to human level; 2) Using VQA MMDL as an example to demonstrate how interdisciplinary perspective may foster the development of human level AI; 3) Generating possible ideas for future investigations. Present meta-analysis plans to emphasize more on the first two goals.

2. Multimodal joint-embedding models (MMJEM)

2.1. Mechanisms of MMJEM

$$h_m = f(W * [h_x, h_y, h_z]) \tag{1}$$

The mechanism of MMJEM involves embedding information from different modalities (e.g. text and images) into common representational spaces [14-21]. Inputs will be preprocessed through unimodal feature extractors (e.g., CNNs and RNNs), and processed unimodal representational vectors will be integrated through fusion methods to form multimodal representations. The multimodal representations will be used by layers of classifiers to solve VQA tasks. Optimizations will be performed in both

DOI: 10.54254/2755-2721/30/20230096

unimodal processing and multimodal integrations [14-21]. As a crucial part of MMJEM, the fusion of preprocessed unimodal feature vectors was frequently investigated, and several theories have been proposed. Earliest fusion strategies involved simple element wise computations and vector concatenations [14]. Later, to solve limitations of these early methods, for instance the limitations in capturing multimodal information complementariness, the bilinear pooling approach was proposed.

$$[a, b, c] + [d, e, f] = [a + d, b + e, c + f] \tag{2}$$

$$[a, b, c]||[d, e, f] = [a, b, c, d, e, f] \tag{3}$$

Element-wise computation and vector concatenation involve computing vectors to form a combination [15, 16]. A major difference is the combination through concatenation outputs vectors in different shapes. Examples of each are shown in formula 2 and 3. Multiple MMJEMs have been developed using element-wise computation and vector concatenation to fuse unimodal vectors and succeeded in VQA. Specifically, several studies have explored the effectiveness of element-wise computations. Antol et al. conducted one of the earliest attempts on VQA datasets during 2015. In their model, image features and LSTM encoded question features were fused via element-wise multiplication, and the model reaches 54.06% on VQA test standards [15]. Gao et al. proposed a similar mQA later the same year, with CNN as image encoder, two separate LSTM encoders for questions and answers, and element-wise addition as the fusing method [16].

Similarly, several studies have also explored the vector concatenation fusing, especially for the purpose of improving the model with element-wise computations [14,17,18]. MT Desta et al. developed an object-based reasoning architecture during 2018 and achieved an accuracy of 94.7% and 94.9 % on CLEVER count and exist tasks, with R- CNN performing visual attribute feature subtraction and LSTM performing Sentence sequential embedding subtraction [14]. Lu et al. developed an architecture during 2017 using Hierarchical question encoding, co-attention mechanism, and vector concatenation, and the model achieved 62.1% and 66.1% on open-end and multiple choice VQA tasks based on COCO-QA dataset [17]. Kevin et al. developed an architecture during 2015 using regional attention mechanism and dot product vector concatenation and achieved 58.94% on multiple-choice VQA questions based on MS-COCO [18].

$$y^{sc} = \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ijk} a_i^s b_j^c \tag{4}$$

Bilinear pooling involves taking outer products with unimodal vectors [19-21]. Unimodal characteristics vectors together forms dimensions of a multimodal representation space, enabling the multimodal representation space to capture a comprehensive view of feature to feature interactions among unimodal representations [19-21]. Mathematic expressions are shown above. The earliest concept of bilinear pooling was proposed by Tenenbaum et al. in 2000 [19]. Later, multiple MMDL architectures later took advantages of such approach and succeeded in VQA tasks, substituting the 'style' ('s' in formula (3))and 'content' ('c' in formula (3)) parameters proposed by Tenenbaum et al. with modalities and defining 'w' as an observational vector for multimodal synergies. H.Ben-Younes et al.have proposed the famous MUTAN model using Tucker decomposition, which is strictly equivalent to full bilinear projection. Experiments showed that MUTAN reaches a 67.36% accuracy on MSCOCO ground truth questions, outperforming previous models [5]. R.Cadene et al. designed the RUBi learning strategy in 2020, combining questions and specific visual feature regions using bilinear block fusion and reaching an accuracy level of 47.11 % on VQA-CP v2 dataset [20]. Fukui et al. also used bilinear modeling in their 2016 VQA model, demonstrating an 62.5% accuracy on 6W questions based on Visual Genome dataset [21].

2.2. Neurophysiological mulitsensory fusion

While MMDL approaches are aiming toward an artificial multimodal embedding, neurophysiological approaches reveal how separate modalities are integrated in human brain. Neuron imaging techniques

DOI: 10.54254/2755-2721/30/20230096

such as fMRI and EEG provided researchers insights on the neuron-bases of multimodal task performing [22]. Specifically, the activations of brain regions during multimodal learning may uncover how specific functional components are integrated within the MML processes [22]. Researchers have developed multiple multimodal integration theories under such framework, answering questions related to functional role of fusion, mechanism of fusion, and where the fusion happens [22-25].

To characterize the nature of multimodal computations, a variety of investigations have been done to compare the brain activation patterns during unimodal and multimodal tasks. Hypothesis of activation patterns includes multimodal activation magnitude less than (subadditive) , equal to (additive), or greater than (superadditive) unimodal combined [22]. Several studies have been done to examine these three hypotheses, and most of them show support towards the superadditive approach, which the neural responses elicited by multimodal neurons significantly exceeded sum of the neuron's two modality-specific responses [23].

However, Thomas et al.'s 2005 study have made significant discoveries regarding the manner of multi-sensory integration, showing that instead of just superadditive manner, neurons have the potential to change their manner of integration in response to external stimulus.They investigated the superior colliculus neurons (SC neurons), which have been long hypothesized serving a function of cross-sensory perception map. The result suggested that instead of a simple additive manner, different SC neurons incorporate different operations when integrating information from multiple senses. Furthermore, Stein et al. identified four types of SC neurons with distinctive integration operations: dual-modality dynamic range neurons (DMDR) that change monotonically in respond to changes in both auditory and visual stimuli intensity, single-modality dynamic range neurons (SMDR) that change just in respond to one modality, and no dynamic range neurons (NDR) that respond to none. In conclusion, Stein et al. 's study reveals that the inherent characteristics of SC neurons plays important roles in multi-sensory integration, providing insights on the complexity of integration strategies [24].

In addition, studies also showed that SC neurons not only have their varieties in long-term characteristics but also short-term plasticity and adaptations. In their 2009 study, Stein and his team continued to discover dynamic patterns of SC neuron activations. Specifically, the result of the experiment showed that SC neuron activations were negatively correlated with the repetition of pairs of stimuluses from different modalities. Stein and his team interpreted this result as showing an adaption or even possibly learning process that tends to integrate repeated multimodal stimulus in a larger extent and close to the response strength triggered by unimodal stimulus. In terms of general dynamic of integrations brought by SC neurons, this shows that the integration of unimodal stimulus may not fall strictly into fixed categories like subadditive, additive, or superadditive. Instead, the SC neuron integration is more of an experienced based phenomena that is highly flexible in respond to frequency and type of experience [25].

2.3. MMJEM and superior colliculus neurons

Relating back to embedding fusion methods of MMJEM, this neurophysiological evidence provides clues for the tradeoffs between bilinear pooling and simple vector computations. The evidence of SC neuron's superadditive manner possibly indicates simple vector additions may not be enough for representing multimodal synergies. The SC neuron plasticity introduced by Stein et al. in 2009 shows that the plasticity of SC neuron's integration strategies are crucial for MML, possibly suggesting that it is not only important to optimize parameters for representations as whole but also important to continuously optimize the manner of fusing.

$$I^{M1M2} = f(w_{ijt}), i \in I, j \in J, t \in T \tag{5}$$

In the case of VQA, Stein et al.'s discussion on SC neurons can be interpreted as more frequently a visual feature and a linguistic feature appear together, more deeply they will be fused by the multimodal agent. This plasticity can only be fulfilled with bilinear pooling, since bilinear pooling is able to record and evaluate comprehensively the feature-to-feature interaction and integration among unimodal representations. As the formula showed above, the combination pooling of two features from unimodal

DOI: 10.54254/2755-2721/30/20230096

representations can be feed into an evaluation function that considers several parameters, such as frequency of appearance, and the output of evaluation can be used for future optimization of integration strategies that operate feature-wise. Thus, this bilinear pooling method endows the model SC like plasticity in when encountering VQA information.

$$y^{sc} = \sum_{o=1}^{O}\sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij} a_i^s b_j^c L(o_{ij}), o_{ij} = I_{ij}^{M1M2}, i \in I, j \in J \tag{6}$$

$$L(o_{ij}) = \begin{cases} a_{d_1} DMDR + b_{d_1} SMDR + c_{d_1} NDR & \text{if} o_{ij} = A \\ a_{d_2} DMDR + b_{d_2} SMDR + c_{d_2} NDR & \text{if} o_{ij} = B \end{cases} \tag{7}$$

Moreover, the categorization of SC individual neuron responses, as what Stein and his colleague has suggested, may possibly provide an insight on how the optimization of feature fusing mentioned previously can be accomplished. The controlling of how deep two unimodal should be integrated can be done by manipulation on distribution of different types of neurons. For instance, the configuration of DMDR and NDR may have various activation distributions for specific integration demands. The mathematical formulas above show the trilinear space of I, J, and the evaluated index represented by o. The f(o) here represents a linear regression that take o as parameter and assign it with a distinctive distribution pattern of different types of neurons (The formula in this case uses the neuron types defined by Thomas et al. 2005). This population coding approach is more similar to human SC neuron configuration. Interestingly, such population coding may possibly promote the quantification of depth of integration. If there exists only one type of neuron within the population, when controlling the activation level, manipulations elicit on the number of neurons activated will have limited flexibility, since the largest step of manipulation equals the activation of an individual neuron. Thus, this population coding approach may have potentials in continuously optimizing parameters in embedding fusing. Further experiments are required to prove such advantage.

In conclusion, studies of SC neurons can possibly inspire the development of MMJEM from two perspectives: 1) Characteristics of SC neurons provide solid explanations on why certain strategies in MMJEM elicit more human-like performances. 2) Differences between SC population coding and MMJEM mechanisms may inspire ways to reduce computation cost and generate new optimization possibilities.

3. Multimodal attention based model (MMAM)

3.1. Mechanisms of MMAM

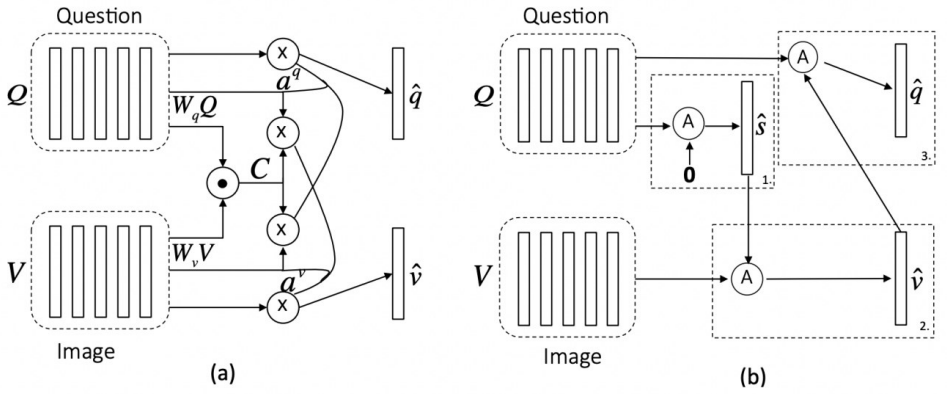

Figure 1. (a) Parallel co-attention mechanism. (b) Alternating co-attention mechanism. (a) and (b) retrieved from [14].

$$C = tanh(Q^T W_b V), V \in R^{d*N}, Q \in R^{d*T}, C \in R^{T*N}, W \in R^{d*d} \tag{7}$$

DOI: 10.54254/2755-2721/30/20230096

$$\begin{aligned} H^v &= tanh(W_v V + (W_q Q)C) \\ a^v &= softmax(w_{hv}^T H^v) \\ H^q &= tanh(W_q Q + (W_v V)C^T) \\ a^q &= softmax(w_{hq}^T H^q) \end{aligned} \tag{8}$$

$$W_v, W_q \in R^{k*d}, w_{hv}, w_{hq} \in R^k, a^v \in R^N, a^q \in R^T \tag{9}$$

The mechanisms of MMAM involves building co-attention mechanisms that optimizes and assigns distributions of attentional weightings to the input of multiple modalities. Several approaches have been proposed to accomplish the assignment of attention cross modalities, including the parallel co-attention and the alternating co-attention (see Figure 1) [14]. As defined by Lu et al. in 2016, parallel co-attention generates attention weightings for modalities simultaneously [17]. Mathematical formulas are shown above. In the formula, unimodal representation vectors V, Q and multimodal attention vector W were linked to affinity matrix C, forming a multimodal attention representations space based on spatial similarities. Using the affinity matrix, the representations of Q and V were projected to each other, and the attention vectors for each modality were assigned separately and optimized by soft-max.

Several studies have taken advantage of the parallel co-attention method. Dias et al. has incorporated Lu et al. 's parallel co-attention mechanism in 2022 and used BERT feature extractor for questions. The model achieves 57.84% accuracy on the validation set of the VQA2.0 dataset [26]. Another interesting work was done by Zhang et al. in 2021. Instead of constructing affinity matrix through spatial similarity as Lu et al. have done, Zhang et al. designed the parallel co-attention mechanism to focus not only on spatial dimension but feature wise paring. They proposed that for in questions like 'What is the color of the women's hair', it is important to attend on both the region 'hair' and the feature attribute 'color'. The model has a performance of 70.24% on VQA1.0 dataset and 61.30 on VQA 2,0 dataset [27].

$$\begin{aligned} H &= tanh(W_x X + (W_g G)I^T) \\ a_x &= softmax(w_{hx}^T H) \end{aligned} \tag{10}$$

On the other hand, as defined by Lu et al., the alternating co-attention generate attention weightings sequentially alternating among modalities. Formulas are shown above. In the formula, the X parameter will switch back and forth among V and Q (x = V, x = Q). When attention first transforms from Q to V, g equals an intermediate question feature, and when attention transforms back from V to Q, g equals the pre-attended image feature (X = Q, g = V). Thus, the attention alternates sequentially among visual and question features.

This alternative co-attention mechanism was supported by another study conducted by Yang et al. also during 2016. In their model, Yang et al. also proposed a mechanism that guides attention through accumulating query vectors between question and visual features. The modal achieves an outstanding accuracy rate in multiple datasets, including 29.3% for DAQUAR-ALL dataset, 46.2% for DAQUAR-REDUCED dataset, and 61.6% for COCO-QA dataset [28].

3.2. Psychophysics of attention guidance

Similar to MMDL, other fields such as psychophysics also provides perspectives towards how attention interacts with information and improve agents' performances, just as the discussions among parallel versus alternative mechanisms proposed by Lu et al. Psychophysics aim at investigating characteristic of human cognition, involving observing behavioral dependent variables in response to external independent variables. Several psychophysical investigations have been done targeting how human attention mechanism functions, specifically on how attention interacts with working memory and influence behavior accuracy [29-35]. One of the major attention phenomena that is studied over the past centuries is the retro-cue effect, which refers to the boost of working memory task accuracy when human subjects attend to relevant contents ahead [29]. Currently, trends of retro-cue effect studies cover five main aspects: time course, voluntary/strategic control, resistance to distraction, central attention

DOI: 10.54254/2755-2721/30/20230096

demands and splitting attention [29]. The following paragraphs review studies of splitting attention and voluntary control, which are the most relevant to the MMAM mechanism.

The study of splitting attention answers the question*: Whether retro-cueing multiple items yields retro-cue benefits compared to a no-cue baseline?* Several investigations have drawn different conclusion. Makovski et al. 's study during 2007 showed that only single-cue condition improves the performances [30]. However, later studies also indicated that retro-cuing multiple items can elicit improvements when the cue items are implicitly related to the characteristics of test items. For instance, evidence [31, 32] have proven that cues that indicates spatial characteristics (e.g., left-right retro-cue signaling the following stimulus comes from both sides) improves behavioral performances. Furthermore, Heuer et al. dived deeper into the discussion in 2016. Their study suggested that not only spatial cueing elicit benefits but also feature based cueing [33]. Moreover, Heuer et al revealed the spatial cueing has its benefits only when compared to spatially neural cues (i.e., the cues that close to each other in a random relationship), whereas feature based cues have benefits independent of spatial characteristics [33]. This possibly furthered the idea of a goal-directed attention, which the retro cues provide perceptual benefits through activating an aspect of 'query' that foreshadows the upcoming stimulus.

The study of voluntary/strategic control answers the question: *Whether attention is more stimulus-driven or goal driven?* Stimulus-driven hypothesis refers to the idea that attention automatically shifts to the cued location (I.e., spatial dimension is the fundamental guidance of attention), and goal-driven hypothesis refers to the idea that attention overtly shift according to the received information (i.e., The given information is the query, and the subject voluntarily guide attention to the possible answers). Research have supported the goal-directed hypothesis [33]. Specifically, the goal hypothesis is measured by cue-reliability, which is how likely the retro cue is valid for the following stimuli. It is hypothesized that if the cue-reliability is strongly correlated with the performance, the retro-cue attention is more likely to be driven by goals, since only in cue-reliable conditions will an effective 'query' be proposed. Gunseli et al. has confirmed such correlation in their 2015 study, showing that although retro-cue effect occurred in both reliable and unreliable condition, the effect was larger in the reliable condition [34]. Similar work was also done by Berryhill in 2012, with subjects showing more retro-cue effect in 100% reliable condition than in uninformative condition. Shimi et al. also had similar findings in 2014 [35].

3.3. MMAM attention mechanisms and retro-cue effects

Studies on attention splitting can possibly provide an insight on the discussion between parallel versus alternating hypothesis of cross-modal attention. As shown by a variety of studies, irrelevant single retro-cue promotes perceptual performance, while irrelevant multiple retro-cue do not. A potential interpretation of such results may be that cognitive load is essential for attention consideration. For single-cue conditions, subjects have enough cognitive sources to consider comprehensively the abstract (self-generated, not given) attention connections among representations, enabling them to develop a more accurate prediction of attention distribution. However, in the multiple-cue condition, subjects do not have enough cognitive resources. Thus, they will not be able to consider all distributions of attention mappings between representations, and instead only cues that have explicit guidance to the subsequent representation will be processed. This interpretation is potential parallel to Lu et al.'s testing result of alternating and parallel hypothesis, which alternating hypothesis is less expensive, and parallel hypothesis is more accurate. In the context of MMDL, the alternating co-attention narrows the focus region into explicit guidance path, thus having less computational costs, whereas the parallel co-attention searches through the whole picture between two elements (In the case of MMDL, the two elements are V and Q), thus having a higher accuracy. Similarly, in human, attentional guidance is essential for weaving complicated information (in the case of human, the two elements are the retro-cues and the subsequent stimulus), and a deeper consideration of the whole combination picture improves the ability for discovering hidden connections.

Furthermore, combining these discoveries provides us an inspiring perspective on attention guidance of MMAM. The studies on voluntary/strategic control of attention confirms the importance of a goal

DOI: 10.54254/2755-2721/30/20230096

directed mechanism. Projecting to the MMAMs mentioned previously, this favors the idea of feature-wise attention guidance proposed by Zhang et al. in 2021, since the feature-wise consideration has more flexibility in considering goals (e.g., 'hair color' and specific visual feature attribute further specifics the goal of query rather than 'hair' and the visual location of 'hair'). Also, similar to the conclusions drawn by investigations such as Berryhill et al. 2012, the feature-wise guidance also enables the consideration of spatial information, since spatial information can also be stored and considered as a type of feature attributes.

4. Multimodal external knowledge based model (MMEKM)

4.1. Mechanism of MMEKM

$$f^c = arg_{i \in j:rel(f_j)=r^c} maxS(g^F(f_i(a_i, r_i, b_i)), g^{NN}(x, Q)) \tag{11}$$

Multimodal external knowledge-based model (MMEKM) involves retrieving VQA facts from external database. The facts are described by visual entity (a), phrase entity (b), and logistic relationships between them (r). The formula above describes a process of selecting facts that best fits the visual input (x) and question (Q). The S(x) in the formula represents a scoring function that evaluates the capability between input representations and facts in knowledge base. It is worth mentioning that a,r,b here is preprocessed and stored in the external knowledge based and thus is non-parametrized. The answer is driven afterwards from the knowledge base (either entity a or entity b of selected optimized facts). Several studies have proven the effectiveness of MMEKM in accomplishing VQA tasks. Narasimhan et al.'s 2018 MMEKM model achieve's 62.20% accuracy on FVQA dataset [36]. Marino et al. have developed a more sophisticated model in 2012. In their model, they use MMBert mechanisms to interpret the inputs and transform the features into implicit knowledge, and then this implicit knowledge was used to operate visual and question symbols stored in external knowledge base [37]. They argued that the implicitness of input representation can be achieved through large transformer-based language models, as suggested by previous studies [38, 39]. The result showed that the model achieves 38.90% accuracy on OK-VQA dataset.

4.2. Neurophysiology of memory retrieval

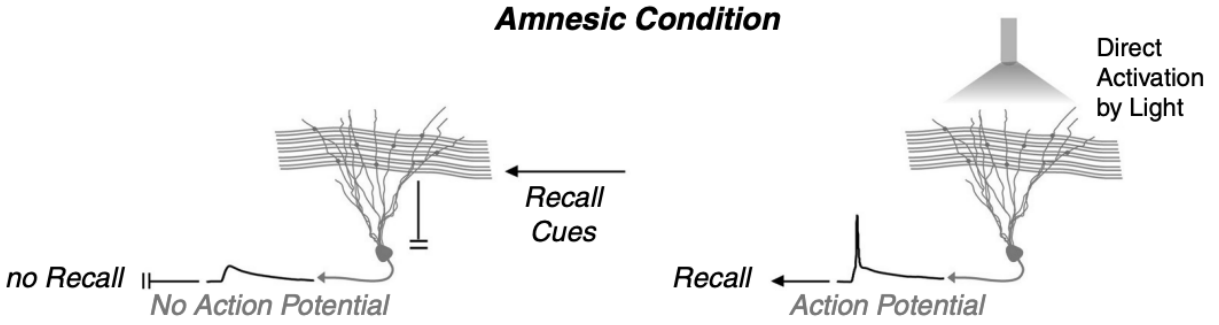

Figure 2. The inhibition and recover of engrams cells and the ability to retrieve memory. Retrieved from [40].

Several studies have indicated that engrams cells are crucial for memory retrieval. Specifically, impairment of engrams cells has shown to cause memory deficit, and the optogenetical evidence indicated that the deficit is correlated with memory retrieval [40]. This correlation is revealed by Nabavi et al.'s study in 2014. The study showed that optogenetically induced long-term-depression (LTD) of amygdala cells of rats impairs fear response, and optogenetically induced long-term-potentiation (LTP) re-evokes the fair response. Researchers interpret this result as the cell ensembles influenced, which are the engram cells, are thus responsible for memory retrieval [41]. Later, a variety of studies have

DOI: 10.54254/2755-2721/30/20230096

supported the result, with Tanaka et al. 2014 and Denny et al. 2014 suggested that reactivation of IEGs cfos and arc in hippocampal CA1 and DG is correlated with contextual fear memory recall [42, 43]. In addition, Figure 2 shows how direct stimulation of engram cell ensembles may restore the action potentials and lead to recovered recall [40].

Another recent significant discovery of engram cells is their relationships with forgetting. The discovery was reviewed by Ryans et al. in 2022. They suggested that as supported by a variety of evidence, forgetting in neural level may be characterized by remodeling mechanisms that switches engram cells from accessible states to inaccessible states [44]. Specifically, one mechanism that accomplishes this switching is the activation of RAC1, which a Rho GTPase that modulates lamellipodial extensions of growth cones in neurons [44]. Intuitively, RAC1 mediates engram cells through inhibiting excitatory synaptic inputs onto engram cells. Lv, L. et al.'s 2019 study has proven that the optogenetic activation of RAC1 in hippocampal CA1 after learning induces forgetting [45]. In addition to RAC1, studies also showed that another mechanism that may salience the engrams cells is the strengthening of inhibitory GABAergic inputs onto excitatory engram cells [46, 47].

4.3. MMEKM, engram inhibition, and forgetting

Combining the mechanism MMEKM and the engram inhibition, a key difference that MMEKM models lack a mechanism that modulates the knowledge base based on retrieval. Similar to the forgetting mechanism of engrams, fact representations in knowledge bases of MMEKM that are never used may be 'forgotten' to save the computational cost. As Narasimhan et al. mentioned when designing their 2018 MMEKM, the scanning process of the whole knowledge base is expensive and thus requires a multi-stage searching process. If a forgetting mechanism like engrams in incorporated, unnecessary combinations of entities that are rarely retrieved in the knowledge base may be silenced. Moreover, in another word, this is like imposing an attention mechanism when pairing facts and inputs, while more attentional weightings may be distributed to frequently retrieved contents.

5. Discussion

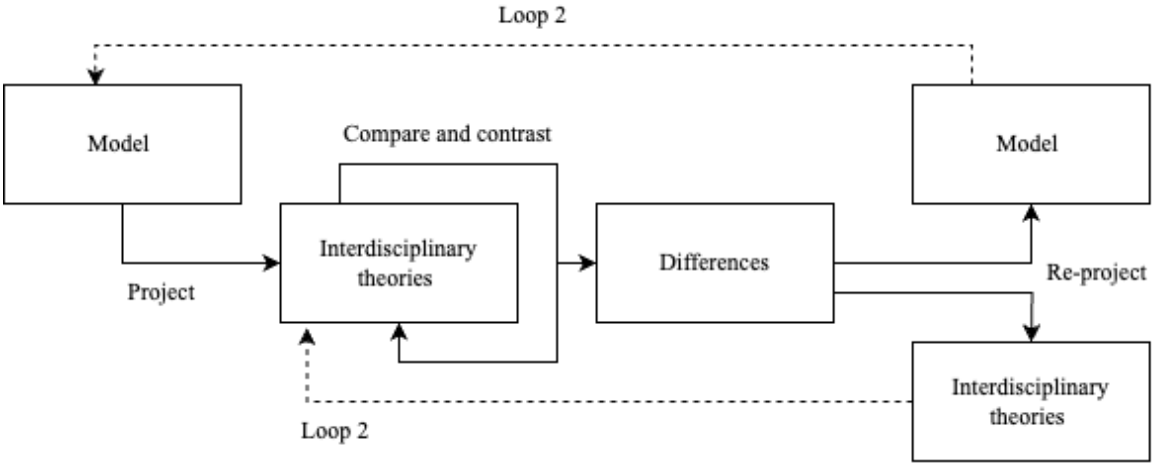

Figure 3. Interdisciplinary comparative framework for meta-analysis.

Furthermore, it is important to note that present meta-analysis does not aim at generating immediately practically feasible ideas or provide a comprehensive review on the picture of all three fields. Instead, the primary goal of current research is to evaluate the possibility of interdisciplinary analytics and call for more research in interdisciplinary perspectives, especially for complex and debatable topics like the modeling of multimodal cognition.

As shown in the Figure 3, present meta-analysis follows a specific logic when analyzing interdisciplinary materials. The structure of analysis follows the framework of deep learning studies,

DOI: 10.54254/2755-2721/30/20230096

which is model by model analysis. The mechanisms of models were then projected and compared to theories that were measured in human level with neurophysiology and psychophysics. To avoid common issues of terminologies in interdisciplinary analysis, a strategy that can be used specifically in this case is mathematical expressions, since mathematics have universal representative meanings in terms of logistics. Furthermore, an import aim of such projection is the identification of the key discussion across areas. For instance, the key discussion of MMAM is attention guidance. Through such identification will interdisciplinary analysis more coherent in terms of filling current research gaps. The differences after comparison can be re-propagated into subject-specific theories, both for the input model and the reference theories, and may even be used as reference for future revise. Such interdisciplinary comparison can be looped multiple times. Finally, this interdisciplinary model does not deny the value of field-specific developments as main contributors of field vitality and should more function as a supplementary strategy that inspires more research ideas for each field.

6. Conclusion

Present meta-analysis reviews the neurophysiological and psychophysical references for current trends in VQA multimodal deep learning models. Specially, present meta-analysis draws correlations and possible synergies between MMJEM and studies of Superior colliculus, MMAM and retro-cue effect, providing an insight on the interdisciplinary relationship of deep learning, neurophysiology, and psychophysics. The result of the analysis indicates the following: 1) Bilinear pooling MMJEM's performance on VQA can possibility be explained by human SC neurons' similar complexity. 2) Bilinear pooling enables possibilities to have SC neuron like plasticity when encountering VQA tasks.3) Attention psychophysics affirms the trade-off debate between parallel and alternating attention guiding in MMAM, providing insights on how such trade-off might be organized in terms of the complexity of connecting attention through elements. 4) A goal driven attention guidance is closer to how human operates. 5) The forgetting mechanism of engram cells in retrieval possibility inspires a 'retrieval weighting' distribution for MMEKM external knowledge base retrieval. More importantly, a key contribution of present met analysis is the connection drawn between several sub-fields: Joint embedding mechanisms and SC neurons, MMDL attention mechanism and the retro-cue effect, and external knowledge base and engram mechanisms.

References

[1] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 689-696).

[2] Ramachandram, D. and Taylor, G.W. (2017). Deep Multimodal Learning: A survey on recent advances and Trends. IEEE Signal Processing Magazine, 34(6), pp. 96–108. https://doi.org/10.1109/msp.2017.2738401.

[3] Murray, M.M. et al. (2016). Multisensory processes: A balancing act across the lifespan. Trends in Neurosciences, 39(8), pp. 567–579. https://doi.org/j.tins.2016.05.003.

[4] Antol, S. et al. (2015). VQA: Visual question answering. 2015 IEEE International Conference on Computer Vision (ICCV) [Preprint]. https://doi.org/10.1109/iccv.2015.279.

[5] Ben-younes, H. et al. (2017). Mutan: Multimodal Tucker Fusion for visual question answering. 2017 IEEE International Conference on Computer Vision (ICCV) [Preprint]. https://10.1109/iccv.2017.285.

[6] Wang, P. et al. (2017). The VQA-Machine: Learning How to use existing vision algorithms to answer new questions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Preprint]. https://doi.org/10.1109/cvpr.2017.416.

[7] Narasimhan, M. and Schwing, A.G. (2018). Straight to the facts: Learning knowledge base retrieval for factual visual question answering. Computer Vision – ECCV 2018, pp. 460–477. https://doi.org/10.1007/978-3-030-01237-3_28.

DOI: 10.54254/2755-2721/30/20230096

[8] Meredith, M.A. (2002). On the neuronal basis for Multisensory Convergence: A brief overview. Cognitive Brain Research, 14(1), pp. 31–40. https://doi.org/10.1016/s0926-6410(02)00059-9.
[9] Stein, B.E., Stanford, T.R. and Rowland, B.A. (2009). The neural basis of multisensory integration in the midbrain: Its organization and maturation.Hearing Research, 258(1–2), pp. 4–15. https://doi.org/10.1016/j.heares.2009.03.012.
[10] Talsma, D. et al. (2010). The multifaceted interplay between attention and multisensory integration. Trends in Cognitive Sciences, 14(9), pp. 400–410. https://doi.org/10.1016/j.tics.2010.06.008.
[11] Gabrieli, J. D. (1998). Cognitive neuroscience of human memory. Annual Review of Psychology, 49(1), 87–115. https://doi.org/10.1146/annurev.psych.49.1.87
[12] Josselyn, S. A., & Tonegawa, S. (2020). Memory engrams: Recalling the past and imagining the future. Science, 367(6473). https://doi.org/10.1126/science.aaw4325
[13] Outhwaite, W., & Turner, S. P. (2007). The SAGE handbook of social science methodology. Sage.
[14] Desta, M. T., Chen, L., & Kornuta, T. (2018). Object-based reasoning in VQA. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). https://doi.org/10.1109/wacv.2018.00201
[15] Agrawal, A., Lu, J., Antol, S. et al. (2016). VQA: Visual question answering. International Journal of Computer Vision, 123(1), 4–31. https://doi.org/10.1007/s11263-016-0966-6
[16] Haoyuan Gao, Junhua Mao. et al. (2015). Are you talking to a machine? Dataset and methods for multilingual image question answering. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'15). https://doi.org/10.48550/arXiv.1505.05612
[17] Jiasen Lu, Jianwei Yang. et al. (2016). Hierarchical question-image co-attention for visual question answering. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). https://doi.org/10.48550/arXiv.1606.00061
[18] Shih, K. J., Singh, S., & Hoiem, D. (2016). Where to look: Focus Regions for visual question answering. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.499
[19] Tenenbaum, J. B., & Freeman, W. T. (2000). Separating style and content with bilinear models. Neural Computation, 12(6), 1247–1283. https://doi.org/10.1162/089976600300015349
[20] R. Cadene, C. Dancette, H. Ben-younes. et al. (2019). RUBi: reducing unimodal biases for visual question answering. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc. https://doi.org/10.48550/arXiv.1906.10169
[21] Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/d16-1044
[22] Stein, B. E., Stanford, T. R., & Rowland, B. A. (2014). Development of multisensory integration from the perspective of the individual neuron. Nature Reviews Neuroscience, 15(8), 520–535. https://doi.org/10.1038/nrn3742
[23] Jiang, W., Wallace, M. T., Jiang, H., Vaughan, J. W., & Stein, B. E. (2001). Two cortical areas mediate multisensory integration in superior colliculus neurons. Journal of Neurophysiology, 85(2), 506–522. https://doi.org/10.1152/jn.2001.85.2.506
[24] Perrault, T. J., Vaughan, J. W., Stein, B. E., & Wallace, M. T. (2005). Superior colliculus neurons use distinct operational modes in the integration of multisensory stimuli. Journal of Neurophysiology, 93(5), 2575–2586. https://doi.org/10.1152/jn.00926.2004
[25] Yu, L., Stein, B. E., & Rowland, B. A. (2009). Adult plasticity in multisensory neurons: Short-term experience-dependent changes in the superior colliculus. The Journal of Neuroscience, 29(50), 15910–15922. https://doi.org/10.1523/jneurosci.4041-09.2009

DOI: 10.54254/2755-2721/30/20230096

[26] Dias, M., Aloj, H., Ninan, N., & Koshti, D. (2022). Bert based multiple parallel co-attention model for visual question answering. 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). https://doi.org/10.1109/iciccs53718.2022.9788253

[27] Zhang, S., Chen, M., Chen, J., Zou, F., Li, Y.-F., & Lu, P. (2021). Multimodal feature-wise co-attention method for visual question answering. Information Fusion, 73, 1–10. https://doi.org/10.1016/j.inffus.2021.02.022

[28] Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.10

[29] Souza, A. S., & Oberauer, K. (2016a). In search of the focus of attention in working memory: 13 years of the retro-cue effect. Attention, Perception, & Psychophysics, 78(7), 1839–1860. https://doi.org/10.3758/s13414-016-1108-5

[30] Makovski, T., & Jiang, Y. V. (2007). Distributing versus focusing attention in visual short-term memory. Psychonomic Bulletin & Review, 14(6), 1072–1078. https://doi.org/10.3758/bf03193093 H

[31] Holt, J. L., & Delvenne, J.-F. (2015). A bilateral advantage for maintaining objects in visual short term memory. Acta Psychologica, 154, 54–61. https://doi.org/10.1016/j.actpsy.2014.11.007

[32] Matsukura, M., Luck, S. J., & Vecera, S. P. (2007). Attention effects during visual short-term memory maintenance: Protection or prioritization? Perception & Psychophysics, 69(8), 1422–1434. https://doi.org/10.3758/bf03192957

[33] Heuer, A., & Schubö, A. (2016). The focus of attention in visual working memory: Protection of focused representations and its individual variation. PLOS ONE, 11(4). https://doi.org/10.1371/journal.pone.0154228

[34] Gunseli, E., van Moorselaar, D., Meeter, M., & Olivers, C. N. (2015). The reliability of retro-cues determines the fate of noncued visual working memory representations. Psychonomic Bulletin & Review, 22(5), 1334–1341. https://doi.org/10.3758/s13423-014-0796-x

[35] Berryhill, M. E., Richmond, L. L., Shay, C. S., & Olson, I. R. (2012). Shifting attention among working memory representations: Testing cue type, awareness, and strategic control. Quarterly Journal of Experimental Psychology, 65(3), 426–438. https://doi.org/10.1080/17470218.2011.604786

[36] Narasimhan, M., & Schwing, A. G. (2018). Straight to the facts: Learning knowledge base retrieval for factual visual question answering. Computer Vision – ECCV 2018, 460–477. https://doi.org/10.1007/978-3-030-01237-3_28

[37] Marino, K., Chen, X., Parikh, D., Gupta, A., & Rohrbach, M. (2021). Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr46437.2021.01389

[38] Z.Jiang, F. Xu, J. Araki, and G. Neubig. (2020). How can we know what language models know? TACL, 8:423–438.

[39] F. Petroni, T. Rocktaschel, S.Riedel. et al. (2019) Language models as knowledge bases? In EMNLP, pages 2463–2473.

[40] Tonegawa, S., Pignatelli, M., Roy, D. S., & Ryan, T. J. (2015). Memory engram storage and retrieval. Current Opinion in Neurobiology, 35, 101–109. https://doi.org/10.1016/j.conb.2015.07.009

[41] Nabavi, S., Fox, R., Proulx, C. D., Lin, J. Y., Tsien, R. Y., & Malinow, R. (2014). Engineering A memory with ltd and LTP. Nature, 511(7509), 348–352. https://doi.org/10.1038/nature13294

[42] Tanaka, K. Z., Pevzner, A., Hamidi, A. B., Nakazawa, Y., Graham, J., & Wiltgen, B. J. (2014). Cortical representations are reinstated by the hippocampus during memory retrieval. Neuron, 84(2), 347–354. https://doi.org/10.1016/j.neuron.2014.09.037

[43] Denny, C. A., Kheirbek, M. A., Alba, E. L., Tanaka, K. F., Brachman, R. A., Laughman, K. B., Tomm, N. K., Turi, G. F., Losonczy, A., & Hen, R. (2014). Hippocampal memory traces are

DOI: 10.54254/2755-2721/30/20230096

differentially modulated by experience, time, and adult neurogenesis. Neuron, 83(1), 189–201. https://doi.org/10.1016/j.neuron.2014.05.018

[44] Ryan, T. J., & Frankland, P. W. (2022). Forgetting as a form of adaptive engram cell plasticity. Nature Reviews Neuroscience, 23(3), 173–186. https://doi.org/10.1038/s41583-021-00548-3

[45] Lv, L., Liu, Y., Xie, J., Wu, Y., Zhao, J., Li, Q., & Zhong, Y. (2019). Interplay between α2-chimaerin and RAC1 activity determines dynamic maintenance of long-term memory. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-13236-9

[46] Das, S., Sadanandappa, M. K., Dervan, A., Larkin. et al. (2011). Plasticity of local GABAergic interneurons drives olfactory habituation. Proceedings of the National Academy of Sciences, 108(36). https://doi.org/10.1073/pnas.1106411108

[47] Stefanelli, T., Bertollini, C., Lüscher, C., Muller, D., & Mendez, P. (2016). Hippocampal somatostatin interneurons control the size of neuronal memory ensembles. Neuron, 89(5), 1074–1085. https://doi.org/10.1016/j.neuron.2016.01.024