

The Contribution of Knowledge in Visiolinguistic Learning: A Survey on Tasks and Challenges

Maria Lymperaiou¹, Giorgos Stamou¹

¹*AILS Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens*

Abstract

Recent advancements in visiolinguistic (VL) learning have allowed the development of multiple models and techniques that offer several impressive implementations, able to currently resolve a variety of tasks that require the collaboration of vision and language. Current datasets used for VL pre-training only contain a limited amount of visual and linguistic knowledge, thus significantly limiting the generalization capabilities of many VL models. External knowledge sources such as knowledge graphs (KGs) and Large Language Models (LLMs) are able to cover such generalization gaps by filling in missing knowledge, resulting in the emergence of hybrid architectures. In the current survey, we analyze tasks that have benefited from such hybrid approaches. Moreover, we categorize existing knowledge sources and types, proceeding to discussion regarding the KG vs LLM dilemma and its potential impact to future hybrid approaches.

Keywords

Visiolinguistic Learning, Transformers, Knowledge Graphs, Large Language Models, Hybrid Architectures

1. Introduction

Visiolinguistic (VL) learning has been one of the fastest evolving fields of artificial intelligence, especially after the emergence of the Transformer [1], which enabled a variety of powerful architectures. Popular VL tasks such as Visual Question Answering (VQA) [2], Visual Reasoning (VR) [3], Visual Commonsense Reasoning (VCR) [4], Visual Entailment (VE) [5], Image Captioning (IC) [6], Image-Text Retrieval (ITR) and inversely Text-Image Retrieval (TIR) [7], Visual-Language Navigation (VLN) [8], Visual Storytelling (VIST) and Visual Dialog (VD) [9] have been significantly benefited from recent transformer-based advancements which follow the *pre-train fine-tune* learning framework. *Pre-training* is responsible of fusing generic information regarding visual and linguistic patterns, as well as how those two modalities interact, based on information present in large-scale datasets. Self-supervised objective functions are employed to help the VL model learn interdependencies between vision and language during pre-training. For example, masking out words from image captions enforces learning how to fill them based on visual cues; reversely, image regions can be masked out, with language guiding


In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), *Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023)*, Hyatt Regency, San Francisco Airport, California, USA, March 27-29, 2023.

✉ marialymp@islab.ntua.gr (M. Lymperaiou); gstam@cs.ntua.gr (G. Stamou)

🆔 0000-0001-9442-4186 (M. Lymperaiou); 0000-0003-1210-9874 (G. Stamou)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

their reconstruction. Task-specific *fine-tuning* steps upon this basic understanding of vision and language, by refining the neural weights of the trained model to adapt to each specific task at a time, upon which the final evaluation is performed.

Despite the rich VL knowledge acquired during this process, current transformer-based VL models [10, 11, 12, 13, 14, 15, 16, 17, 18] lack generalization to several concepts and scenarios that require **commonsense knowledge**, or knowledge of **abstract entities**, **facts** and **real-world events**. Of course, this is somehow expected, since neither pre-training nor fine-tuning VL datasets contain or demand perceiving concepts beyond visual descriptions. Figure 1 presents some examples of this claim: questions (Q) about the image (I) require some knowledge beyond the visual domain, so that the correct answer (A) can be inferred.



Q: What days might I most commonly go to this building? A: Sundays.



Q: In which continent was the person in the image born? A: North America.



Q: Who among the people in the image is the eldest? A: Person in the left.



Q: What is the name of the object used to eat this food? A: Chopsticks.

Figure 1: External knowledge is required to answer these visual questions [19, 20, 21].

The first image of Figure 1 requires knowledge about **human culture and history** [19], to combine with **visual** information: the object in the image is a *church*, and *people usually go to the church on Sundays*. The second image [21] requires one more reasoning step, since it is not only required to detect that this is a postage stamp containing the photo of a *person* (**visual** information), but also who this person is. Knowledge about **named entities** recognizes this person as *Alexander Hamilton*. Further **factual** knowledge provides that *Alexander Hamilton was born in today's Saint Kitts and Nevis* and *Saint Kitts and Nevis is in North America*. The combination of these two facts derives the final answer *Alexander Hamilton was born in North America*. The third image [20] requires the **visual** extraction of the two *people* present in it. Then, **named entities** knowledge assigns the identities *Serena Williams* and *Venus Williams* to these two people. Their *age* is provided as a combination of **named entities** and **factual** knowledge, yielding the **comparative** knowledge fact that *Serena Williams is older than Venus Williams*. Finally, **spatial** knowledge derives that *Serena Williams is the person in the left*. The overall combination of **named entities**, **comparative** and **spatial** knowledge returns the final answer *The person in the left*. It becomes obvious that answering these question requires more knowledge from external sources, which is extracted and combined to infer an answer.

Thus, the incorporation of external knowledge in earlier or later stages of the *pre-training/fine-tuning* process is necessary to enhance the capabilities of VL models, so that they are able to respond to more real-world scenarios. Such knowledge is typically represented using entities, relationships and semantic descriptions [22] stored in structured Knowledge Graphs (KGs)

[23, 24, 25, 26]. Language Models (LMs) such as BERT [27] have been proven capable of storing relational knowledge learned from linguistic data during pre-training, introducing the LM-as-KB scenario [28]. This knowledge can then be retrieved by constructing queries as fill-the-blank statements, which the LM is tasked to complete. Further works validate the abilities of LMs for world-knowledge storage and retrieval, while showcasing their scaling capacity according to the number of parameters [29]. There are some prerequisites for LM to successfully serve as knowledge bases; accessing the data similarly to KG querying, updating outdated facts while trespassing the risk of catastrophic forgetting, unlocking their rather obscure reasoning capabilities and measuring the degree of their interpretability and explainability are still open challenges [30]. More recently, the impressive results of Large Language Models (LLMs) [31, 32, 33, 34, 35] in various linguistic tasks greatly inspire their possible usage as rich and simultaneously vast knowledge bases (KBs) to aid VL learning.

Prior surveys in VL learning [36, 37, 38, 39, 40, 41, 42] do not focus on the collaboration between knowledge and deep learning VL models. An exhaustive presentation of the knowledge-enhanced VL (KVL) topic was presented in [43] for the first time. In the current survey paper, we focus on state-of-the-art endeavors involving transformer models for the VL representation, leading to *hybrid* approaches when combined with external knowledge. Finally, we discuss around potential trends regarding the external knowledge assisting VL models and how it is expected to affect future applications in the field.

2. Knowledge and Reasoning

2.1. Types of external knowledge

External knowledge sources are divided in two main categories, *explicit* and *implicit* [43]. They are both capable of providing **factual**, **commonsense**, **temporal**, **lexical** or other knowledge senses [44] missing from pre-trained VL models. The type of the external knowledge source used significantly defines the way of retrieving and harnessing knowledge for VL models.

Explicit knowledge refers to the knowledge stored in KGs in a structured format. Such knowledge is symbolically represented in the form of triplets (h, r, t) , which contain entities (h, t) and their in-between relationships r . Extracting an answer from a KB is a fully transparent process, and the path followed can be deterministically recovered. This is crucial especially when evaluating multi-step and compositional reasoning, so that the factuality of the reasoning path followed is guaranteed. Nevertheless, crafting and maintaining KGs requires manual effort or supervision, therefore hindering the automatic extension of such KBs.

Popular open-source knowledge graphs that have contributed to VL learning are ConceptNet [24], DBPedia [26], Wikidata [25], YAGO [45] and others. The retrieval of KG facts is primarily based on SPARQL queries, a clear and deterministic query language designed for this purpose. Many open-source knowledge graphs also provide APIs for frequent queries. Due to the constraints SPARQL querying imposes, bridging the gap between natural language user queries and SPARQL has been a useful venture [46].

Knowledge graph representation learning provides low-dimensional distributed vectors, following the popular strategy of linguistic embeddings [22]. This approach allows the application

of Deep Learning techniques on KGs, while a better-suited communication between KGs and VL models is established, since all three modalities can be viewed as numerical vectors.

Implicit knowledge covers unstructured knowledge stored in neural weights, depicting facts and relationships learned during model pre-training. This process allows the integration of multiple and large-scale data sources without the need for human supervision. Existing knowledge can be extended and updated by re-training or fine-tuning the existing neural network (LM/LLM). However, this process is computationally prohibitive for most research institutions, while the factuality, fairness and trustworthiness of learned knowledge and reasoning are questionable, since they significantly depend on the quality of the training data used.

There is an ongoing list of LMs-as-KB, since any language model pre-trained under self-supervised learning objectives can potentially serve as a KB. Retrieving knowledge from LMs is not as straightforward as for KGs, due to the opaque LM structure. Therefore, there are two main ways to access LM knowledge, *indirect access (fine-tuning)* and *direct access (prompting)*.

Fine-tuning has been the most concrete way to exploit LM knowledge, even if it does not actually retrieve existing knowledge. Similarly to how fine-tuning works for VL models (Section 1), LM fine-tuning refers to adapting neural weights towards a downstream linguistic task by training for a few epochs on a small labelled linguistic dataset, appropriate for the desired task.

Prompting obeys to the pre-train, prompt, predict pipeline [47], which allows direct access to information stored in a pre-trained LM, should an appropriate prompt is designed. This is where the difficulty of this approach lies: composing an optimal prompt is an open research topic, and sub-optimal prompt templates may just denote the lower bound of knowledge contained within LMs [48, 49]. Prompts in textual format, called *discrete prompts* are quite intuitive to humans, therefore hints on how to craft them can be based on human conversational behavior. Therefore, mining templates from large corpora [49], paraphrasing of existing prompts [49, 50], fill-blank via language generation [51, 52] and others have been proposed as viable directions. On the other hand, *soft prompts* circumvent interpretability in the sake of efficiency, by directly accessing the LM’s embedding space. Prefix-tuning using continuous task-specific vectors in a frozen LM [53], soft prompting based on discrete prompting initialization [54] and others have demonstrated promising results. The non-trivial search for prompt-based knowledge retrieval is rewarded with few-shot or even zero-shot reasoning, able to revolutionize KVL models.

2.2. Reasoning in Knowledge Graphs and Large Language Models

Reasoning has been one of the milestones regarding the capacity of LLMs to sufficiently perform as KBs. It is regarded as the capability of drawing conclusions and making decisions based on given information, and can be divided in *formal* and *informal*. *Formal* reasoning refers to following a set of rules in a logical and deterministic manner. On the other hand, *informal* reasoning is mostly based on a generic experience and intuition of the world, thus being prone to errors, while however being more flexible [55].

KGs tend to approach the *formal* reasoning path, due to their structured nature and the determinism accompanying decision-making. *Informal* reasoning is mostly interconnected with unstructured KBs, which have acquired a more probabilistic sense of the world. LMs have demonstrated some adequate reasoning capabilities, as long as they are large enough, thus

belonging to LLMs [56]. Nevertheless, the landscape of the full potential of LLM reasoning has not yet been entirely explored. Prompting has been utilized towards unlocking reasoning capabilities of LLMs, encouraging them to reveal their Chain-of-Thought (CoT) instead of merely providing the final answer [57, 58]. CoT has been proven successful towards unveiling hidden reasoning capabilities, either in the few-shot setting [58], where the LLM is prompted with some exemplars of the desired reasoning, together with an instructive phrase, or in the zero-shot setting [57], where the model receives an instructive phrase without exemplars. Another line of work proposes the evaluation of LLMs on downstream tasks exploiting well-crafted datasets, each of which is dedicated on different reasoning senses. To this end, various tests have stressed LLM capabilities on arithmetic [59], symbolic [58], commonsense [60], and other types of reasoning. Overall, the findings occurring from the aforementioned endeavors suggest that indeed, LLMs present *emergent* reasoning capabilities simulating human thinking patterns, though being incapable of tackling complex reasoning challenges. Nevertheless, such evidence cannot conclude whether LLM present *real* reasoning capabilities or if they can just perfectly overfit on the vast information they receive [55]. So far, knowledge-enhanced VL literature trusts the LM-as-KB paradigm for several downstream tasks, demonstrating successful results.

3. VL tasks with knowledge

3.1. Visual Question Answering (VQA)

In Visual Question Answering (VQA), a model receives an image I and a textual question Q referring to the image, and predicts a textual answer A . The answer A can be either selected among pre-defined candidates, viewing VQA as a *classification* problem, or else be generated, thus placing VQA (free-text VQA) in the language *generation* family. Knowledge guidance can assist in addressing scenarios where standalone visual information is not adequate, as the ones presented in Figure 1. A variety of knowledge-demanding datasets for knowledge-driven VQA (K-VQA) have been developed [19, 20, 61, 62, 63, 64, 65], setting a good starting point for relevant model implementations. Early attempts in K-VQA were heavily relying on exact matching between visual or textual concepts and KG nodes via SPARQL queries [66, 61, 67], inducing errors in cases when such explicit concept mapping does not exist. Embedding representations provide a more flexible solution by retrieving similar KG facts to visual and textual concepts [62, 68, 20]. Nevertheless, the context-free nature of classic word embedding methods [69, 70] can only serve a limited amount of cases, impeding generalization to scenarios when contextualization is necessary. Transformers leveraged on the linguistic side allowed further improvements on K-VQA models [71, 72] paving the way for consequent end-to-end VL approaches.

ConceptBERT [73] was the first breakthrough towards a unified KVL transformer-based architecture, achieved by considering all three modalities in a joint representation, hence being able to incorporate **commonsense** knowledge in the reasoning process. **Factual** knowledge injection following the unified KVL strategy was also explored, only requiring fine-tuning to leverage the contribution of external KGs [21]. Multiple knowledge senses can be fused in unified KVL architectures, such as **factual** and **visual** knowledge, in order to cross-check the validity of predicted answers [74]. The dynamic incorporation of external knowledge regardless its type transforms knowledge injection to a passage retrieval problem, offering advanced

adaptability to relevant architectures [75]. Passage retrieval from Wikipedia is also followed in [76], where visual cues of different granularities (global captions, image labels and scene text) are combined to retrieved facts; consequently, all linguistic information is provided to a T5 transformer model [77], which generates the final answer A . Similarly, [78] resorts to passage retrieval from external sources, but suggests joint training of the retrieval module and the T5 answer generator, contrary to prior works. Generic information obtained via passage retrieval from external knowledge sources is deemed inadequate to answer targeted visual questions. To this end, knowledge acquisition is refined by focusing on common entities present in queries, retrieved passages and images [79]. The combination of several KGs [24, 26, 80, 81] under a unified larger KG can facilitate knowledge retrieval, which is performed based on the linguistic similarity between contextualized questions (questions enhanced with visual captions and scene text as context) and KG facts. The enriched input is provided to a T5 transformer which finally generates the final free-text answer A [82]. Multimodal passage retrieval is addressed via the proposed Multimodal Inverse Cloze Task as pre-training objective, which learns the alignment between visual and textual information from Wikipedia entries. This technique can aid K-VQA involving named entity recognition [83].

Diverging from the usage of KGs as the knowledge source at hand, first works exploiting the LM-as-KB paradigm were introduced for K-VQA. Specifically, GPT-3 [31] can be used to provide facts in a few-shot manner, receiving visual captions as prompts [84], similar to how traditional KGs receive SPARQL queries. Performance gains can be achieved by utilizing multiple captions as prompts to a variety of pre-trained LLMs, enabling zero-shot reasoning [85]. Using again linguistic captions as a modality mediator, [86] leverages frozen LLMs to address zero-shot VQA. Chain of Thought (CoT) prompting of LLMs is another interesting direction, which enhances explainability of the answer derivation pipeline by revealing intermediate reasoning steps [87]. Instead of resorting to the linguistic modality to obtain unimodal LLM prompts, other approaches opt to fine-tune a visual encoder jointly with the LLM, so that aligned LLM-VL representations are achieved [88].

There are a few implementations combining explicit and implicit knowledge sources to enjoy advantages of both worlds. KRISP [89] leverages several external KGs [24, 26, 81], visual knowledge from Visual Genome [90], as well as implicit knowledge from BERT [27]. REVIVE [91] deploys several visual features to retrieve knowledge from various sources, such as Wikidata and GPT-3. Visual feature guidance was proven critical towards improving the knowledge retrieval process. Fusing both implicit and explicit knowledge in the VL reasoning process is also followed in KAT, using a refined framework that fetches information from Wikidata and GPT-3 upon which joint reasoning is performed. A transformer decoder receives the output of the reasoning module to generate the final answer [92].

3.2. Visual Commonsense Reasoning (VCR)

Visual Commonsense Reasoning (VCR) is a task closely related to VQA. Given a challenging question Q regarding an image I , a VCR model is tasked to predict the answer A , accompanied by a rationale R [4]. **Commonsense** knowledge can be provided from large-scale KGs, such as ConceptNet [24] and ATOMIC [93], or dedicated datasets, such as SWAG [94], which contains descriptions about sequences of events.

Transformer-based endeavors for knowledge-assisted VCR (K-VCR) naturally utilize BERT [27] as the backbone architecture to construct end-to-end KVL models. In KVL-BERT [95], the input Q together with candidate answers A guide the retrieval of relevant commonsense facts [24], resulting in a knowledge-enriched linguistic input. Then, visual features among with this enriched input are inserted in a BERT-like VL model (VL-BERT [96]) so that the correct A is selected. Consequently, inferring R requires feeding VL-BERT with the predicted A , candidate rationales R and visual features. Aligning independent modality representations within a single multimodal embedding is proposed in [97]. The same work introduces extensions of VL pre-training objectives [43] to incorporate commonsense knowledge from [24] as an extra modality, therefore enforcing learning KVL interrelationships. Dynamic commonsense augmentation of image-text training data is a suggested direction, accompanied by learning to reconstruct hidden visual labels based on knowledge facts retrieved from commonsense KBs [98].

Implicit knowledge sources have been gaining popularity in recent K-VCR literature. GPT-2 [99] has assisted dynamic reasoning over images, inferring **temporal** hypotheses regarding what might have happened before and what might happen after the depicted situation [100]. Chain of Thought (CoT) reasoning is inherently tied to VCR, as reasoning paths are highly associated with selecting rationales R . The rise in popularity of CoT techniques for linguistic tasks is highly interconnected with the development of LLMs, which have been proven able to reveal intermediate reasoning steps [57]. There are not yet many works in the VL direction, even though the introduction of novel appropriate datasets with grounded answer rationales highlight the prospects of such an approach [101]. Specifically, [101] tackles VCR by captioning the image, and then feed the caption together with the existing linguistic input to the LLM. Another promising work in this direction introduces Multimodal-CoT without using language as the mediating modality, proposing a two-stage process to separately infer the answer A and the rationale R , while stating that a LM with less than 1B parameters is adequate for state-of-the-art performance [102]. It is expected that the rapid rise of popularity of LLMs in complex linguistic QA reasoning [103] may soon give rise to more LLM-augmented VCR approaches, addressing more aspects of reasoning.

3.3. Image Captioning (IC)

Image Captioning (IC) is a widespread VL task, asking from a model to generate a caption c for an image I . Several knowledge-enhanced IC (K-IC) techniques employ recurrent neural networks (RNNs) or relevant variants such as Long Short Term Memory (LSTM) networks, while leveraging knowledge sources for **commonsense**-enhanced captioning [104, 105, 106, 107].

The integration of external knowledge in IC using transformers, was first explored in [108], where **event** and **named-entity** knowledge is fused together with textual and visual data in a Transformer encoder [1] to generate entity/event-aware captions. **Commonsense** descriptions derived from ConceptNet [24] and ATOMIC [93] are able to assist visual commonsense generation (VCG), a challenging task that requires inferring **intents** and **temporal** sequence of events [109]. This is achieved by incorporating **commonsense** descriptions to BART, a powerful language generation model [110]. **Geographical** information guiding **factual** knowledge retrieval to assist IC was first explored in [111], where visual features together with the extracted facts are inserted in a Transformer encoder-decoder structure, ultimately generating the caption c .

Apart from external knowledge considerations, IC faces additional challenges as a *language generation* task: VL transformers are not well-suited for generative tasks, even though they excel in understanding tasks, where an answer has to be selected among a set of pre-defined options. XGPT tackles this challenge by adapting generative pre-training [99, 31] for VL tasks [112], which is achieved by introducing novel generative pre-training objectives. The collaboration of GPT-2 [99] with CLIP [15] is viewed as highly promising, since both models have been trained on an abundance of web-data, thus incorporating numerous knowledge senses in an *implicit* manner. ClipCap [113] leverages this collaboration without re-training CLIP or GPT-2; instead, a lightweight transformer-based mapping module is trained to match CLIP representations to GPT-2, which eventually generates the caption c . The CLIP-GPT-2 combination was also followed in VC-GPT [114]. Another lightweight improvement combining a pre-trained CLIP visual encoder and a frozen GPT-2 text decoder further boosts performance of low-resource approaches [115]. A cross-modal filter that selects the most relevant visual information, so that captioning errors are reduced is proposed in [116], still respecting the frozen CLIP-GPT-2 framework.

A factor that overshadows the knowledge-enhanced IC capabilities is the lack of dedicated datasets for testing. So far, IC models are evaluated on classic datasets containing images and captions, such as COCO [117] and Flickr [118], which however are not challenging in terms of external knowledge required. The construction of appropriate datasets that would follow the paradigm of knowledge-demanding VQA datasets, such as OK-VQA [19], K-VQA [20], FVQA [62], KB-VQA [61], or VCR datasets [4] will give prominence to the abilities of K-IC models.

3.4. Sequential Generation

There are two tasks touching sequential generation, depending on which modality is generated at a time. Visual Storytelling (VIST) refers to generating captions c_1, c_2, \dots, c_N for a visual story comprised of frames I_1, I_2, \dots, I_N , as an extension of IC for sequences. The reverse task of Story Visualization (SV) involves synthesizing visual frames I_1, I_2, \dots, I_N from textual captions c_1, c_2, \dots, c_N . For both tasks, consistency throughout the story and relevance between the two modalities are required.

Harnessing **commonsense** knowledge for VIST was first attempted in [119], followed by [120, 121], which however utilize RNN-bases structures for text generation. The usage of transformer-based models is explored in [122], where visual concepts are enriched through ConceptNet. All relevant enriched concepts are provided to BART, which ultimately outputs appropriate captions. Regarding SV, there is a different architectural line followed [123, 124, 125, 126] based on Generative Adversarial Networks (GANs) [127]. **Commonsense** and **spatial** knowledge considerations were primarily addressed in [125], demonstrating encouraging results in favor of the usage of external knowledge. Nevertheless, some recent approaches follow transformer-based approaches enhanced with *implicit* knowledge. Specifically, Story-DALL-E [128] is able to even synthesize unseen stories in a zero-shot fashion. It leverages DALL-E [129] as a multimodal unstructured knowledge base that provides high-quality visual synthesis from text.

Similar to IC, sequential generation tasks face the lack of appropriate datasets, upon which the external knowledge contribution would be more evident and meaningful.

3.5. Multi-taskers

Knowledge-free VL models have already achieved incorporating a variety of VL tasks under the same pre-training body, only requiring fine-tuning on smaller labelled text-image datasets. This way, there is no need to design and implement a separate architecture per independent task, but rather exploit visiolinguistic relationships present in large scale datasets used for pre-training, such as COCO [117], Visual Genome [90], Conceptual Captions [130] and SBU [131].

Reasoning tasks, including visual question answering, visual-text entailment and visual commonsense reasoning can be easily incorporated under the same model, due to the high similarity of the inferences these tasks have to make. Visual cues are enhanced with higher-level cognition provided by VisualCOMET [100] and among with textual inputs, they are fed in a GPT-2 model that generates free-text rationales for all three tasks, therefore providing explainability of answers [132]. The same reasoning tasks were also explored in [133], testing results on knowledge-demanding datasets [19, 62] as well. Knowledge embeddings representing facts from external knowledge sources [24, 25] are aligned with textual descriptions, which are inserted together with the knowledge features in a multimodal transformer. KB-VLP [134] is another multi-task model tackling visual question answering and visual reasoning, enhanced with **commonsense** and **logical** capabilities. Entity extraction from images and text is performed to map VL concepts to Wikidata [25] entries, based on knowledge graph embeddings, thus finally assisting the retrieval of the most relevant Wikidata facts.

Few-shot and zero-shot learning via LLMs in low-resource scenarios can be very practical for generative VL tasks. In [88], an encoder-decoder Transformer model is the most appropriate option for text generation, addressing free-form VQA and IC, while various prompts are leveraged in inference time to enforce generating a suitable textual prediction. PromptCap [135] is designed to accurately generate fine-grained and controllable captions based on GPT-3 prompting. The generated captions can be leveraged to provide context for VQA. The Socratic model framework [136] is a novel and promising direction, since it demonstrates that the complementary knowledge obtained during pre-training from VL and LM can be combined towards several multimodal tasks with the help of multimodal prompting.

There are some significant observations arising from the construction of multi-task VL models regarding the usage of external knowledge. For example, in [133], incorporation of Wikidata failed to enhance reasoning capabilities as expected. The authors attribute this deteriorated performance to ambiguities and noise existing in the large Wikidata KB. Furthermore, comparing to task-specific implementations for VQA, IC etc, it is obvious that multi-task implementations are significantly fewer. This indicates that the incorporation of external knowledge is not as straightforward for multiple tasks as it may be for task-specific models.

3.6. The future of knowledge in VL

Recent literature around KVL research reveals several gaps that need to be covered. In our opinion, the most prominent gap is the lack of appropriate knowledge-demanding datasets for most tasks, which limits the extend to which KVL models are evaluated. Apart from VQA [19, 20, 61, 62, 63, 64, 65] and VCR [4], other VL tasks are tested on classic benchmark datasets; therefore, the contribution of knowledge in downstream performance is not as prevalent as in

situations where various senses of knowledge are explicitly required.

At the same time, we have observed that several efforts are dedicated towards constructing *linguistic* benchmarks questioning reasoning capabilities in LMs, including mathematics [59], symbolic reasoning [58], implicit reasoning based on strategies [137], commonsense understanding [138], temporal, causal, linguistic understanding and others [139]. We argue that such attempts could assist the creation of appropriate VL datasets, which would incorporate visual and linguistic challenges, so that knowledge contribution would be more concrete.

With the surge of larger and larger models, explainability concerns are raised in the broader AI community [140], especially when black-box VL models are combined with black-box unstructured KBs. Older KVL architectures were often addressing explainability [67], as an immediate result of incorporating KGs for answer prediction, since the path leading to the final answer could be retrieved. Later works, even though widely exploiting KGs, mainly focused on improving downstream performance rather than enhancing the interpretability of reasoning towards these results. Currently, KVL approaches have totally deviated from the pursue of explainability, especially since opaque LLMs started acting as KBs for VL models.

Another emerging issue is prompt design for KVL architectures that exploit LLMs as their external knowledge source. Similar challenges to linguistic prompt search also arise for the VL setting (especially since many models use language as a mediator between vision and language), thus inducing some ambiguity regarding the quality of results and the reasoning process followed to return these results. Multimodal prompt design is still an unexplored but crucial field towards unlocking the full potential of LLM-VL models.

Ultimately, we view that the future of KVL research is highly interconnected with the current trends in LLMs, both in terms of designing appropriate knowledge-demanding benchmarks for KVL tasks, as well as answering the KG vs LLM ongoing dilemma, analyzed in Section 4.

4. Knowledge Graphs or Large Language Models?

Throughout our analysis we recognize some potential trends towards selecting the type of knowledge source to assist VL models towards hybrid approaches. Even though KGs clearly dominate previous VL implementations, we can assume some focus shifting towards LMs-as-KB, due to the rapid development of L(L)Ms and generally their ever increasing popularity in recent NLP literature. For example, the advent of ChatGPT¹ created an unprecedented hype, opening several discussions regarding the future of AI as a whole. The almost human-level capabilities of ChatGPT and relevant implementations have created an abundance of opportunities in NLP literature, while sparking a lot of unavoidable criticism. Since VL learning tends to follow the advancements in NLP, as it happened with KG-based knowledge boosting [141, 142, 143, 144], it is highly likely that more LLM-based VL implementations will soon emerge.

However, this trend comes at a cost: LLMs have already scaled to billions and even trillions [34] of parameters, a procedure that requires massive training. Concerns regarding the cost of pre-training language models has been raised even before this tremendous parameter scaling [145]; apart from computational budget, issues such as fair access and environmental impact of

¹ChatGPT

relevant implementations should question the reliance and preference of the research community towards them.

In the meanwhile, the completely opaque learning and decision-making process may be more harmful than beneficial; even though known biases, errors and inaccuracies² are reported in LLM publications, their actual usage raises doubts regarding the quality and the trustworthiness of the knowledge provided to the final task. Early on the introduction of LLMs, such as GPT-3 [31], papers exposing failure cases regarding mathematical reasoning, logic and ethical requests [146] gained lots of attention. Probing how LLMs can be fooled, sheds some light to their reasoning process and how they are being confused by misleading inputs [147, 148], which can be a promising starting point towards defeating logical brittleness. Relevant endeavors uncover LLM deductive reasoning capabilities [149], proving that exhaustive memorization compensates for LLM inability to learn to reason. In total, reasoning capabilities of LLMs pose several open questions [55], such as whether heuristics conceal reasoning incapability, or whether reasoning steps can be trustworthy, given that inconsistencies and false rationales are sometimes provided with certainty. More refined challenges reveal that LLMs’ world-knowledge suffers from robustness issues as well, since they confuse likely and unlikely situations, even though they are capable of recognizing impossible events [150].

Even if aforementioned concerns are somehow addressed in consequent versions of relevant LLMs, the search for the optimal fact retrieval process via prompts still remains an open challenge [47]. Moreover, papers promoting certain prompts to LLMs avoid describing the process behind discovering such optimal prompts, and merely provide some experimental comparison between similar prompt phrases [57]. Therefore, if the golden prompt [57] *let’s think step by step* has been defined via extended experimentation, there is no guarantee regarding its optimality and reliability; on the other hand, if there is a certain methodology behind, it seems that it has not been fully unlocked (or at least released to the public). Low-performance instructive prompts targeting CoT reasoning are semantically relevant to the golden prompt, thus raising doubts regarding the consistency of LLM-occurring answers and their sensitivity to -slight or more intense- input variations.

Overall, the aforementioned shortcomings are expected to be transferred to VL implementations if LLMs eventually replace KGs to some extent, thus raising a crucial question: is it worth it to quickly adapt to the trend or better wait?

5. Conclusion

In this survey, we analyzed the collaboration of external knowledge with VL learning. Existing models and datasets drive several challenges in this upcoming field, since the full potential of knowledge-enhanced approaches has not been yet unlocked. Knowledge graphs and large language models can both serve as knowledge bases for VL, posing different advantages and disadvantages. To this end, the current paper devotes an extended discussion over the KB vs LLM dilemma for VL, highlighting significant open issues tied to the current state of the LM-as-KB paradigm. All in all, we hope that our work can introduce researchers to the knowledge-enhanced VL exploration, while denoting challenges of the knowledge adoption process.

²Birds are not real

Acknowledgments

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the 3rd Call for HFRI PhD Fellowships (Fellowship Number 5537).

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 30, Curran Associates, Inc., 2017. URL: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
- [2] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, D. Parikh, Vqa: Visual question answering, 2016. *arXiv:1505.00468*.
- [3] F. He, Y. Wang, X. Miao, X. Sun, Interpretable visual reasoning: A survey, *Image and Vision Computing* 112 (2021) 104194. URL: <https://www.sciencedirect.com/science/article/pii/S0262885621000998>. doi:<https://doi.org/10.1016/j.imavis.2021.104194>.
- [4] R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, From recognition to cognition: Visual commonsense reasoning, 2019. *arXiv:1811.10830*.
- [5] N. Xie, F. Lai, D. Doran, A. Kadav, Visual entailment task for visually-grounded language learning, *arXiv preprint arXiv:1811.10582* (2018).
- [6] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, R. Cucchiara, From show to tell: A survey on deep learning-based image captioning, 2021. *arXiv:2107.06912*.
- [7] S. R. Dubey, A decade survey of content based image retrieval using deep learning, *IEEE Transactions on Circuits and Systems for Video Technology* (2021) 1–1. URL: <http://dx.doi.org/10.1109/TCSVT.2021.3080920>. doi:10.1109/tcsvt.2021.3080920.
- [8] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, A. van den Hengel, Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, 2018. *arXiv:1711.07280*.
- [9] A. El-Nouby, S. Sharma, H. Schulz, D. Hjelm, L. E. Asri, S. E. Kahou, Y. Bengio, G. W. Taylor, Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction, 2019. *arXiv:1811.09845*.
- [10] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. *arXiv:1908.02265*.
- [11] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, Oscar: Object-semantics aligned pre-training for vision-language tasks, 2020. *arXiv:2004.06165*.
- [12] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, D. Kiela, Flava: A foundational language and vision alignment model (2021).
- [13] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, 2021. *arXiv:2102.03334*.
- [14] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, Y. Cao, Simvlm: Simple visual language

- model pretraining with weak supervision, 2021. URL: <https://arxiv.org/abs/2108.10904>. doi:10.48550/ARXIV.2108.10904.
- [15] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021.
 - [16] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, J. Fu, Seeing out of the box: End-to-end pre-training for vision-language representation learning, 2021. URL: <https://arxiv.org/abs/2104.03135>. doi:10.48550/ARXIV.2104.03135.
 - [17] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, P. Zhang, Florence: A new foundation model for computer vision, 2021. URL: <https://arxiv.org/abs/2111.11432>. doi:10.48550/ARXIV.2111.11432.
 - [18] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision (2021). URL: <https://arxiv.org/abs/2102.05918>. doi:10.48550/ARXIV.2102.05918.
 - [19] K. Marino, M. Rastegari, A. Farhadi, R. Mottaghi, Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3190–3199.
 - [20] S. Shah, A. Mishra, N. Yadati, P. P. Talukdar, Kvqa: Knowledge-aware visual question answering, in: AAAI, 2019.
 - [21] D. Garcia-Olano, Y. Onoe, J. Ghosh, Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection, 2021.
 - [22] S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu, A survey on knowledge graphs: Representation, acquisition and applications, IEEE transactions on neural networks and learning systems PP (2021).
 - [23] C. Fellbaum, Wordnet: An electronic lexical database (1998).
 - [24] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: AAAI, 2017.
 - [25] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85.
 - [26] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. G. Ives, Dbpedia: A nucleus for a web of open data, in: ISWC/ASWC, 2007.
 - [27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, ArXiv abs/1810.04805 (2019).
 - [28] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, S. Riedel, Language models as knowledge bases?, 2019. URL: <https://arxiv.org/abs/1909.01066>. doi:10.48550/ARXIV.1909.01066.
 - [29] C. Wang, X. Liu, D. Song, Language models are open knowledge graphs, 2020. URL: <https://arxiv.org/abs/2010.11967>. doi:10.48550/ARXIV.2010.11967.
 - [30] B. AlKhamissi, M. Li, A. Celikyilmaz, M. Diab, M. Ghazvininejad, A review on language models as knowledge bases, 2022. URL: <https://arxiv.org/abs/2204.06031>. doi:10.48550/ARXIV.2204.06031.
 - [31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan,

- R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- [32] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. v. d. Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, L. Sifre, Training compute-optimal large language models, 2022. URL: <https://arxiv.org/abs/2203.15556>. doi:10.48550/ARXIV.2203.15556.
- [33] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, Palm: Scaling language modeling with pathways, 2022. URL: <https://arxiv.org/abs/2204.02311>. doi:10.48550/ARXIV.2204.02311.
- [34] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. Bosma, Z. Zhou, T. Wang, Y. E. Wang, K. Webster, M. Pellat, K. Robinson, K. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. V. Le, Y. Wu, Z. Chen, C. Cui, Glam: Efficient scaling of language models with mixture-of-experts, 2021. URL: <https://arxiv.org/abs/2112.06905>. doi:10.48550/ARXIV.2112.06905.
- [35] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. v. d. Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. d. L. Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, L. Sifre, Improving language models by retrieving from trillions of tokens, 2021. URL: <https://arxiv.org/abs/2112.04426>. doi:10.48550/ARXIV.2112.04426.
- [36] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, 2017. arXiv:1705.09406.
- [37] K. Kafle, R. Shrestha, C. Kanan, Challenges and prospects in vision and language research, 2019. arXiv:1904.09317.
- [38] W. Guo, J. Wang, S. Wang, Deep multimodal representation learning: A survey, *IEEE Access* 7 (2019) 63373–63394. doi:10.1109/ACCESS.2019.2916887.
- [39] A. Mogadala, M. Kalimuthu, D. Klakow, Trends in integration of vision and language research: A survey of tasks, datasets, and methods, *Journal of Artificial Intelligence Research* 71 (2021) 1183–1317. URL: <http://dx.doi.org/10.1613/jair.1.11688>. doi:10.1613/jair.1.11688.
- [40] C. Zhang, Z. Yang, X. He, L. Deng, Multimodal intelligence: Representation learning, information fusion, and applications, *IEEE Journal of Selected Topics in Signal Processing*

- 14 (2020) 478–493. URL: <http://dx.doi.org/10.1109/JSTSP.2020.2987728>. doi:10.1109/jstsp.2020.2987728.
- [41] S. Uppal, S. Bhagat, D. Hazarika, N. Majumdar, S. Poria, R. Zimmermann, A. Zadeh, Multimodal research in vision and language: A review of current and emerging trends, 2020. arXiv:2010.09522.
 - [42] J. Cao, Z. Gan, Y. Cheng, L. Yu, Y.-C. Chen, J. Liu, Behind the scene: Revealing the secrets of pre-trained vision-and-language models, in: ECCV, 2020.
 - [43] M. Lymperaiou, G. Stamou, A survey on knowledge-enhanced multimodal learning, 2022. URL: <https://arxiv.org/abs/2211.12328>. doi:10.48550/ARXIV.2211.12328.
 - [44] F. Ilievski, A. Oltramari, K. Ma, B. Zhang, D. L. McGuinness, P. Szekely, Dimensions of commonsense knowledge (2021). URL: <https://arxiv.org/abs/2101.04640>. doi:10.48550/ARXIV.2101.04640.
 - [45] T. Tanon, G. Weikum, F. Suchanek, Yago 4: A reason-able knowledge base (2020) 583–596. doi:10.1007/978-3-030-49461-2_34.
 - [46] W. Zheng, M. Zhang, Automated query graph generation for querying knowledge graphs, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 2698–2707. URL: <https://doi.org/10.1145/3459637.3482235>. doi:10.1145/3459637.3482235.
 - [47] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (2023). URL: <https://doi.org/10.1145/3560815>. doi:10.1145/3560815.
 - [48] L. Adolphs, S. Dhuliawala, T. Hofmann, How to query language models?, 2021. URL: <https://arxiv.org/abs/2108.01928>. doi:10.48550/ARXIV.2108.01928.
 - [49] Z. Jiang, F. F. Xu, J. Araki, G. Neubig, How can we know what language models know?, 2019. URL: <https://arxiv.org/abs/1911.12543>. doi:10.48550/ARXIV.1911.12543.
 - [50] A. Haviv, J. Berant, A. Globerson, Bertese: Learning to speak to bert, 2021. URL: <https://arxiv.org/abs/2103.05327>. doi:10.48550/ARXIV.2103.05327.
 - [51] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3816–3830. URL: <https://aclanthology.org/2021.acl-long.295>. doi:10.18653/v1/2021.acl-long.295.
 - [52] H. Guo, B. Tan, Z. Liu, E. P. Xing, Z. Hu, Efficient (soft) q-learning for text generation with limited good data, 2021. URL: <https://arxiv.org/abs/2106.07704>. doi:10.48550/ARXIV.2106.07704.
 - [53] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 4582–4597. URL: <https://aclanthology.org/2021.acl-long.353>. doi:10.18653/v1/2021.acl-long.353.
 - [54] Z. Zhong, D. Friedman, D. Chen, Factual probing is [mask]: Learning vs. learning to

- recall, 2021. URL: <https://arxiv.org/abs/2104.05240>. doi:10.48550/ARXIV.2104.05240.
- [55] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, *ArXiv abs/2212.10403* (2022).
 - [56] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, 2022. URL: <https://arxiv.org/abs/2206.07682>. doi:10.48550/ARXIV.2206.07682.
 - [57] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, 2022. URL: <https://arxiv.org/abs/2205.11916>. doi:10.48550/ARXIV.2205.11916.
 - [58] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Hsin Chi, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, *ArXiv abs/2201.11903* (2022).
 - [59] A. Amini, S. Gabriel, S. Lin, R. Koncel-Kedziorski, Y. Choi, H. Hajishirzi, MathQA: Towards interpretable math word problem solving with operation-based formalisms, 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2357–2367. URL: <https://aclanthology.org/N19-1245>. doi:10.18653/v1/N19-1245.
 - [60] P. Bhargava, V. Ng, Commonsense knowledge reasoning and generation with pre-trained language models: A survey, 2022. URL: <https://arxiv.org/abs/2201.12438>. doi:10.48550/ARXIV.2201.12438.
 - [61] Q. Wu, P. Wang, C. Shen, A. R. Dick, A. van den Hengel, Ask me anything: Free-form visual question answering based on knowledge from external sources, 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) 4622–4630.
 - [62] P. Wang, Q. Wu, C. Shen, A. R. Dick, A. van den Hengel, Fvqa: Fact-based visual question answering, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2018) 2413–2427.
 - [63] A. K. Singh, A. Mishra, S. Shekhar, A. Chakraborty, From strings to things: Knowledge-enabled vqa model that can read and reason, 2019 *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019) 4601–4611.
 - [64] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, J. Tan, Cross-modal knowledge reasoning for knowledge-based visual question answering, *ArXiv abs/2009.00145* (2020).
 - [65] Z. Chen, J. Chen, Y. Geng, J. Z. Pan, Z. Yuan, H. Chen, Zero-shot visual question answering using knowledge graph, in: *SEMWEB*, 2021.
 - [66] Y. Zhu, C. Zhang, C. Ré, L. Fei-Fei, Building a large-scale multimodal knowledge base system for answering visual queries, 2015. *arXiv:1507.05670*.
 - [67] P. Wang, Q. Wu, C. Shen, A. R. Dick, A. van den Hengel, Explicit knowledge-based reasoning for visual question answering, in: *IJCAI*, 2017.
 - [68] M. Narasimhan, A. G. Schwing, Straight to the facts: Learning knowledge base retrieval for factual visual question answering, *ArXiv abs/1809.01124* (2018).
 - [69] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, 2014, pp.

- 1532–1543. URL: <https://aclanthology.org/D14-1162>. doi:10.3115/v1/D14-1162.
- [70] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, 2013. [arXiv:1310.4546](#).
 - [71] M. Ziaeeafard, F. Lécué, Towards knowledge-augmented visual question answering, in: COLING, 2020.
 - [72] A. Salaberria, G. Azkune, O. L. de Lacalle, A. S. Etxabe, E. Agirre, Image captioning for effective use of language models in knowledge-based visual question answering, [ArXiv abs/2109.08029](#) (2021).
 - [73] F. Gardères, M. Ziaeeafard, B. Abeloos, F. Lécué, Conceptbert: Concept-aware representation for visual question answering, in: FINDINGS, 2020.
 - [74] J. Wu, J. Lu, A. Sabharwal, R. Mottaghi, Multi-modal answer validation for knowledge-based vqa, [ArXiv abs/2103.12248](#) (2021).
 - [75] C. Qu, H. Zamani, L. Yang, W. B. Croft, E. G. Learned-Miller, Passage retrieval for outside-knowledge visual question answering, Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021).
 - [76] F. Gao, Q. Ping, G. Thattai, A. N. Reganti, Y. Wu, P. Natarajan, Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 5057–5067.
 - [77] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of Machine Learning Research* 21 (2020) 1–67. URL: <http://jmlr.org/papers/v21/20-074.html>.
 - [78] W. Lin, B. Byrne, Retrieval augmented visual question answering with outside knowledge, in: Conference on Empirical Methods in Natural Language Processing, 2022.
 - [79] J. Wu, R. J. Mooney, Entity-focused dense passage retrieval for outside-knowledge visual question answering, in: Conference on Empirical Methods in Natural Language Processing, 2022.
 - [80] N. Tandon, G. de Melo, G. Weikum, Acquiring comparative commonsense knowledge from the web, *Proceedings of the National Conference on Artificial Intelligence* 1 (2014) 166–172.
 - [81] S. Bhakthavatsalam, K. Richardson, N. Tandon, P. Clark, Do dogs have whiskers? a new knowledge base of haspart relations, 2020. [arXiv:2006.07510](#).
 - [82] Z. Chen, Y. Huang, J. Chen, Y. Geng, Y. Fang, J. Z. Pan, N. Zhang, W. Zhang, Lako: Knowledge-driven visual question answering via late knowledge-to-text injection, Proceedings of the 11th International Joint Conference on Knowledge Graphs (2022).
 - [83] P. Lerner, O. Ferret, C. Guinaudeau, Multimodal inverse cloze task for knowledge-based visual question answering, [ArXiv abs/2301.04366](#) (2023).
 - [84] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, L. Wang, An empirical study of gpt-3 for few-shot knowledge-based vqa, [ArXiv abs/2109.05014](#) (2021).
 - [85] A. M. H. Tiong, J. Li, B. Li, S. Savarese, S. C. H. Hoi, Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training, in: Conference on Empirical Methods in Natural Language Processing, 2022.
 - [86] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, S. Hoi, From images to textual prompts: Zero-shot vqa with frozen large language models, [ArXiv abs/2212.10846](#) (2022).
 - [87] Z. Chen, Q. Zhou, Y. Shen, Y. Hong, H. Zhang, C. Gan, See, think, confirm: Interactive

- prompting between vision and language models for knowledge-based visual reasoning, ArXiv abs/2301.05226 (2023).
- [88] W. Jin, Y. Cheng, Y. Shen, W. Chen, X. Ren, A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models, 2021. URL: <https://arxiv.org/abs/2110.08484>. doi:10.48550/ARXIV.2110.08484.
 - [89] K. Marino, X. Chen, D. Parikh, A. K. Gupta, M. Rohrbach, Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 14106–14116.
 - [90] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, F.-F. Li, Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. arXiv:1602.07332.
 - [91] Y. Lin, Y. Xie, D. Chen, Y. Xu, C. Zhu, L. Yuan, Revive: Regional visual representation matters in knowledge-based visual question answering, 2022. URL: <https://arxiv.org/abs/2206.01201>. doi:10.48550/ARXIV.2206.01201.
 - [92] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, J. Gao, KAT: A knowledge augmented transformer for vision-and-language, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 956–968. URL: <https://aclanthology.org/2022.naacl-main.70>. doi:10.18653/v1/2022.naacl-main.70.
 - [93] J. D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, Y. Choi, Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs, in: AAAI, 2021.
 - [94] R. Zellers, Y. Bisk, R. Schwartz, Y. Choi, Swag: A large-scale adversarial dataset for grounded commonsense inference, 2018. URL: <https://arxiv.org/abs/1808.05326>. doi:10.48550/ARXIV.1808.05326.
 - [95] D. Song, S. Ma, Z. Sun, S. Yang, L. Liao, Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning, Knowl. Based Syst. 230 (2021) 107408.
 - [96] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, Vl-bert: Pre-training of generic visual-linguistic representations, 2020. arXiv:1908.08530.
 - [97] J. Lee, I. Kim, Vision-language-knowledge co-embedding for visual commonsense reasoning, 2020.
 - [98] S. Ye, Y. Xie, D. Chen, Y. Xu, L. Yuan, C. Zhu, J. Liao, Improving commonsense in vision-language models via knowledge graph riddles, 2022. URL: <https://arxiv.org/abs/2211.16504>. doi:10.48550/ARXIV.2211.16504.
 - [99] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.
 - [100] J. S. Park, C. Bhagavatula, R. Mottaghi, A. Farhadi, Y. Choi, Visualcomet: Reasoning about the dynamic context of a still image, in: In Proceedings of the European Conference on Computer Vision (ECCV), 2020.
 - [101] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL: <https://arxiv.org/abs/2209.09513>. doi:10.48550/ARXIV.2209.09513.
 - [102] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, A. Smola, Multimodal chain-of-thought reasoning in language models, 2023. URL: <https://arxiv.org/abs/2302.00923>.

doi:10.48550/ARXIV.2302.00923.

- [103] X. Daull, P. Bellot, E. Bruno, V. Martin, E. Murisasco, Complex qa and language models hybrid architectures, survey, 2023.
- [104] Y. Zhou, Y. Sun, V. G. Honavar, Improving image captioning by leveraging knowledge graphs, 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (2019) 283–293.
- [105] J. Hou, X. Wu, Y. Qi, W. Zhao, J. Luo, Y. Jia, Relational reasoning using prior knowledge for visual captioning, ArXiv abs/1906.01290 (2019).
- [106] F. Huang, Z. Li, S. Chen, C. Zhang, H. Ma, Image captioning with internal and external knowledge, Proceedings of the 29th ACM International Conference on Information & Knowledge Management (2020).
- [107] D. K. Aditya Mogadala, Xiaoyu Shen, Integrating rule-based entity masking into image captioning, 2020.
- [108] W. Zhao, Y. Hu, H. Wang, X. Wu, J. Luo, Boosting entity-aware image captioning with multi-modal knowledge graph, 2021. URL: <https://arxiv.org/abs/2107.11970>. doi:10.48550/ARXIV.2107.11970.
- [109] Y. Xing, Z. Shi, Z. Meng, Y. Ma, R. Wattenhofer, Km-bart: Knowledge enhanced multimodal bart for visual commonsense generation, in: ACL/IJCNLP, 2021.
- [110] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. URL: <https://arxiv.org/abs/1910.13461>. doi:10.48550/ARXIV.1910.13461.
- [111] S. Nikiforova, T. Deoskar, D. Paperno, Y. Winter, Generating image captions with external encyclopedic knowledge, ArXiv abs/2210.04806 (2022).
- [112] Q. Xia, H. Huang, N. Duan, D. Zhang, L. Ji, Z. Sui, E. Cui, T. Bharti, X. Liu, M. Zhou, Xgpt: Cross-modal generative pre-training for image captioning, 2020. URL: <https://arxiv.org/abs/2003.01473>. doi:10.48550/ARXIV.2003.01473.
- [113] R. Mokady, A. Hertz, A. H. Bermano, Clipcap: Clip prefix for image captioning, ArXiv abs/2111.09734 (2021).
- [114] Z. Luo, Y. Xi, R. Zhang, J. Ma, A frustratingly simple approach for end-to-end image captioning, 2022.
- [115] R. P. Ramos, B. Martins, D. Elliott, Y. Kementchedjieva, Smallcap: Lightweight image captioning prompted with retrieval augmentation, ArXiv abs/2209.15323 (2022).
- [116] Z. Luo, Z. Hu, Y. Xi, R. Zhang, J. Ma, I-tuning: Tuning frozen language models with image for lightweight image captioning, 2022. URL: <https://arxiv.org/abs/2202.06574>. doi:10.48550/ARXIV.2202.06574.
- [117] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755.
- [118] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics 2 (2014) 67–78. URL: <https://aclanthology.org/Q14-1006>. doi:10.1162/tac1_a_00166.

- [119] P. Yang, F. Luo, P. Chen, L. Li, Z. Yin, X. He, X. Sun, Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 5356–5362. URL: <https://doi.org/10.24963/ijcai.2019/744>. doi:10.24963/ijcai.2019/744.
- [120] C.-C. Hsu, Z.-Y. Chen, C.-Y. Hsu, C.-C. Li, T.-Y. Lin, T.-H. K. Huang, L.-W. Ku, Knowledge-enriched visual storytelling, 2019. [arXiv:1912.01496](https://arxiv.org/abs/1912.01496).
- [121] C. Xu, M. Yang, C. Li, Y. Shen, X. Ao, R. Xu, Imagine, reason and write: Visual storytelling with graph knowledge and relational reasoning, in: AAAI, 2021.
- [122] H. Chen, Y. Huang, H. Takamura, H. Nakayama, Commonsense knowledge aware concept selection for diverse and informative visual storytelling, in: AAAI, 2021.
- [123] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, J. Gao, Storygan: A sequential conditional gan for story visualization, 2019, pp. 6322–6331. doi:10.1109/CVPR.2019.00649.
- [124] A. Maharana, D. Hannan, M. Bansal, Improving generation and evaluation of visual stories via semantic consistency, [ArXiv abs/2105.10026](https://arxiv.org/abs/2105.10026) (2021).
- [125] A. Maharana, M. Bansal, Integrating visuospatial, linguistic, and commonsense structure into story visualization, [ArXiv abs/2110.10834](https://arxiv.org/abs/2110.10834) (2021).
- [126] N. Tsakas, M. Lymperaiou, G. Filandrianos, G. Stamou, An impartial transformer for story visualization, 2023. URL: <https://arxiv.org/abs/2301.03563>. doi:10.48550/ARXIV.2301.03563.
- [127] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, volume 27, 2014. URL: <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>.
- [128] A. Maharana, D. Hannan, M. Bansal, Storydall-e: Adapting pretrained text-to-image transformers for story continuation, 2022. URL: <https://arxiv.org/abs/2209.06192>. doi:10.48550/ARXIV.2209.06192.
- [129] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, 2021. [arXiv:2102.12092](https://arxiv.org/abs/2102.12092).
- [130] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: ACL, 2018.
- [131] V. Ordonez, G. Kulkarni, T. Berg, Im2text: Describing images using 1 million captioned photographs, in: J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 24, Curran Associates, Inc., 2011. URL: <https://proceedings.neurips.cc/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7e9bea9-Paper.pdf>.
- [132] A. Marasović, C. Bhagavatula, J. S. Park, R. L. Bras, N. A. Smith, Y. Choi, Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs, in: FINDINGS, 2020.
- [133] V. Shevchenko, D. Teney, A. R. Dick, A. van den Hengel, Reasoning over vision and language: Exploring the benefits of supplemental knowledge, [ArXiv abs/2101.06013](https://arxiv.org/abs/2101.06013) (2021).
- [134] K. Chen, Q. Huang, Y. Bisk, D. J. McDuff, J. Gao, Kb-vlp: Knowledge based vision and

language pretraining, 2021.

- [135] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, J. Luo, Promptcap: Prompt-guided task-aware image captioning, *ArXiv abs/2211.09699* (2022).
- [136] A. Zeng, A. S. Wong, S. Welker, K. Choromanski, F. Tombari, A. Purohit, M. S. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, P. R. Florence, Socratic models: Composing zero-shot multimodal reasoning with language, *ArXiv abs/2204.00598* (2022).
- [137] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, J. Berant, Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies, *Transactions of the Association for Computational Linguistics* 9 (2021) 346–361. URL: https://doi.org/10.1162/tacl_a_00370. doi:10.1162/tacl_a_00370.
- [138] A. Talmor, J. Herzig, N. Lourie, J. Berant, CommonsenseQA: A question answering challenge targeting commonsense knowledge, 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4149–4158. URL: <https://aclanthology.org/N19-1421>. doi:10.18653/v1/N19-1421.
- [139] A. Srivastava, A. Rastogi, A. Rao, A. Shueb, A. Abid, A. Fisch, A. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. Kocurek, A. Safaya, A. Tazarv, Z. Wu, Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022. doi:10.48550/arXiv.2206.04615.
- [140] S. R. Islam, W. Eberle, S. K. Ghafoor, M. Ahmed, Explainable artificial intelligence approaches: A survey, 2021. URL: <https://arxiv.org/abs/2101.09429>. doi:10.48550/ARXIV.2101.09429.
- [141] I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Matsumoto, Luke: Deep contextualized entity representations with entity-aware self-attention, in: *EMNLP*, 2020.
- [142] Y. Qin, Y. Lin, R. Takanobu, Z. Liu, P. Li, H. Ji, M. Huang, M. Sun, J. Zhou, Erica: Improving entity and relation understanding for pre-trained language models via contrastive learning, in: *ACL/IJCNLP*, 2021.
- [143] N. Poerner, U. Waltinger, H. Schütze, E-BERT: Efficient-yet-effective entity embeddings for BERT, in: *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Online, 2020, pp. 803–818. URL: <https://aclanthology.org/2020.findings-emnlp.71>. doi:10.18653/v1/2020.findings-emnlp.71.
- [144] L. Bauer, L. Deng, M. Bansal, Ernie-nli: Analyzing the impact of domain-specific external knowledge on enhanced representations for nli, in: *DEELIO*, 2021.
- [145] O. Sharir, B. Peleg, Y. Shoham, The cost of training nlp models: A concise overview, 2020. URL: <https://arxiv.org/abs/2004.08900>. doi:10.48550/ARXIV.2004.08900.
- [146] L. Floridi, M. Chiriatti, Gpt-3: Its nature, scope, limits, and consequences, *Minds and Machines* 30 (2020) 681–694.
- [147] N. Kassner, H. Schütze, Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly, 2019. URL: <https://arxiv.org/abs/1911.03343>. doi:10.48550/ARXIV.1911.03343.
- [148] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Hsin Chi, N. Scharli, D. Zhou, Large language models can be easily distracted by irrelevant context, *ArXiv abs/2302.00093*

(2023).

- [149] Z. Yuan, S. Hu, I. Vulic, A. Korhonen, Z. Meng, Can pretrained language models (yet) reason deductively?, *ArXiv abs/2210.06442* (2022).
- [150] C. Kauf, A. A. Ivanova, G. Rambelli, E. Chersoni, J. S. She, Z. Chowdhury, E. Fedorenko, A. Lenci, Event knowledge in large language models: the gap between the impossible and the unlikely, *ArXiv abs/2212.01488* (2022).