Are Visual-Linguistic Models Commonsense Knowledge Bases?

Hsiu-Yu Yang
University of Stuttgart
hsiu-yu.yang@ims.uni-stuttgart.de

Carina Silberer
University of Stuttgart
carina.silberer@ims.uni-stuttgart.de

Abstract

Despite the recent success of pretrained *language* models as on-the-fly knowledge sources for various downstream tasks, they have been shown to inadequately represent trivial common facts that *vision* typically captures. This limits their application to natural language understanding tasks that require commonsense knowledge. We seek to determine the capability of pretrained *visual–linguistic* models as knowledge sources on demand. To this end, we systematically compare *language-only* and *visual–linguistic* models in a zero-shot commonsense question answering inference task. We find that visual–linguistic models are highly promising regarding their benefit for text-only tasks on certain types of commonsense knowledge associated with the visual world. Surprisingly, this knowledge can be activated even when no visual input is given during inference, suggesting effective multimodal fusion during pretraining. However, we also reveal that there is still a huge scope for improvements towards better cross-modal reasoning abilities and pretraining strategies for event understanding.[1]

1 Introduction

Commonsense knowledge is essential in human life for task solving and communication. Being aggregated knowledge acquired from past experiences and communications, it is usually left implicit in human communication, and used subconsciously for reasoning and drawing inferences. This puts a challenge on natural language understanding systems, and a large body of work has put forward approaches to provide commonsense knowledge to models for downstream tasks (Yang et al., 2019; Chen et al., 2018; Mihaylov and Frank, 2018). While knowledge bases have been a popular way to provide relevant knowledge for a task at hand, more recently, pretrained language models (PTLMs) have become a popular mechanism to extract knowledge in free-form text on demand. Tasks range from, e.g., persona-grounded dialog (Majumder et al., 2020), narrative story generation (Ammanabrolu et al., 2020), to metaphor generation (Stowe et al., 2021). Shwartz and Choi (2020) and Bisk et al. (2020), however, suggest that text corpora alone may be insufficient for knowledge acquisition due to reporting bias found in them (Gordon and Durme, 2013). This has led to analyzing the knowledge that PTLMs possess through dedicated probing studies (Petroni et al., 2019; Singh et al., 2021; Zhou et al., 2020b)

Existing works on probing PTLMs have used a loose categorization of commonsense, which limits a comprehensive understanding of the types of commonsense they possess and lack, respectively. At the same time, the literature proposes vision as a promising knowledge source (Izadinia et al., 2015; Bagherinezhad et al., 2016; Sadeghi et al., 2015). It seems therefore straightforward to leverage visual–linguistic (VL) models for knowledge extraction on demand—these representation models are extensions of PTLMs to the visual–linguistic domain by pretraining LMs and image recognition models jointly on multimodal data (Tan and Bansal, 2019; Chen et al., 2020; Lu et al., 2019). Yet, the question on their capability to capture commonsense knowledge that can be activated through language only (Yun et al., 2021) is yet to be explored systematically.

In this work, we address this research gap by conducting a controlled comparison of text-only and VL models. Specifically, we extend a synthetic commonsense question answering (QA) dataset based on Ma et al. (2021)'s work, which structures knowledge relations into abstract types (called *dimensions* henceforth), and transform it to a QA inference task. We use the task to compare the models by applying them in a zero-shot manner,

[1] Our datasets, CWWV$_{Img}$ and CWWV$_{Clip}$, are provided at https://github.com/Mallory24/CS_Probing

Proceedings of the 29th International Conference on Computational Linguistics, pages 5542–5559
October 12–17, 2022.

and in their natural setting—masked language modeling.

The overarching question of our study is:

Do VL models learn to encode commonsense knowledge through multimodal pretraining, that can be activated during inference from textual input only?

In particular, we seek to empirically answer:

(**Q1**) Which dimensions of commonsense do VL models possess compared against text-only PTLMs?

(**Q2**) During pretraining, does explicit visual information (i.e., images) benefit commonsense knowledge encoding?

(**Q3**) During inference, is explicit visual observation (i.e., images) necessary for recalling commonsense knowledge?

(**Q4**) Do commonsense acquisition and retrieval depend on the architecture of VL models?

We address the questions by performing a range of experiments using various pretrained models and ablated variants. We find that existing VL models do complement PTLMs on certain commonsense dimensions, which are related to the visual world (*part-whole*, *spatial*, i.a.), and that they can be activated through language input only, making them promising for their use in natural language tasks that do not require explicit visual context. We also identify a range of limitations opening up several avenues for future work, including enhanced pretraining and modality integration strategies, and improved multimodal prompting (Shin et al., 2020; Zhong et al., 2021; Liu et al., 2021).

2 Related Work

Commonsense Knowledge Mining from Vision Although recent interest in commonsense knowledge mining remains text-based (Jastrzębski et al., 2018; Zhou et al., 2020b; Liang and McGuinness, 2021; Bosselut et al., 2019), several studies have explored the visual world: Chen et al. (2013) extract commonsense relationships from the web to improve visual understanding, while Zellers et al. (2018) exploit commonsense priors from visual resources (Krishna et al., 2017) for scene graph generation. Several works learnt specific types of commonsense, including object affordances (Goyal et al., 2017) and temporal causal knowledge (Zhang et al., 2020a). Only few works used VL models for purely text-based tasks (Cui et al., 2020; Tang et al., 2021) in a pipeline approach to extract commonsense from them.

The works above focus on explicit visual commonsense extraction. We, in contrast, seek to study the extent and types of commonsense knowledge that pretrained VL models implicitly capture and that complement pretrained text-based models.

Machine Commonsense Evaluation Commonsense knowledge evaluation is usually conducted with dedicated benchmarks specific for selected knowledge types. Existing formulations range from multiple choice question answering (Zellers et al., 2019; Zhou et al., 2019; Bisk et al., 2020; Richardson and Sabharwal, 2020) and machine reading comprehension (Huang et al., 2019) to knowledge base completion tasks (Petroni et al., 2019; Davison et al., 2019), which makes a systematic and comprehensive commonsense knowledge evaluation even more challenging (Santos et al., 2020). Recent works assess model consistency in commonsense reasoning by introducing linguistic perturbations, complementary counterparts, and logically-equivalent rephrased sentences (Zhou et al., 2020b; Singh et al., 2021; Zhou et al., 2020a). Akin to Ilievski et al. (2021), our goal is to present a comprehensive comparison of the commonsense knowledge resided in pretrained models. While previous research dominantly employs pretrained language-only models (PTLMs), we are not aware of any work like ours—a structured analysis of the types of commonsense knowledge implicitly encoded in pretrained VL models.

3 The QA Dataset CWWV$_{Img}$

To compare VL models against purely textual models with respect to the commonsense knowledge they capture, we extend Ma et al. (2021)'s procedure for creating a synthetic dataset of prompt–answer candidate instances (CWWV) to that of a *multimodal* commonsense dataset (CWWV$_{Img}$).

It provides a set of QA instances for various knowledge relations, structured into 10 dimensions of commonsense knowledge (e.g., *spatial*). Questions are in the form of filled prompts (Le Scao and Rush, 2021; Liu et al., 2021): an instance in CWWV$_{Img}$ has three natural language statements, each associated with a set of images. Each statement is a pair of a *prompt* (e.g., *Shade is not*) and one of three candidate answers (e.g., *sunny*). Table 1 shows an example for each dimension (associated images are omitted for space reasons).

CS dimension	Starting prompt	Answer candidates	# Instances
part-whole	Furry animals have	A_1: effect of chilling innovation. **A_2: millions of hair.** A_3: hole in.	1,165
taxonomic	Recruit is a way to	A_1: rate. **A_2: enlist.** A_3: slope.	1,323
distinctness	Shade is not	A_1: flat. A_2: postal worker. **A_3: sunny.**	828
similarity	Throw up is a synonym of	A_1: rutinic acid. A_2: random. **A_3: vomit.**	644
quality	A wet floor is	**A_1: slippery.** A_2: light brown. A_3: abbreviated to unido.	1,840
utility	A fork is used for	A_1: speed of transit. A_2: confuse voters. **A_3: picking up food.**	2,090
creation	Music is created by	A_1: olive oil mill. A_2: mapping process. **A_3: instruments.**	100
temporal	Going for a haircut requires	**A_1: finding barber.** A_2: hard examinations. A_3: write persuasively.	1,889
spatial	You are likely to find a document folder in	**A_1: file drawer.** A_2: madagascar jungle. A_3: minerals.	1,599
desire	You would thank someone because you want to	A_1: accomplish mutual goal. **A_2: feel good.** A_3: cool off.	1,781

Table 1: CWWV$_{Img}$: examples and their number per dimension ($13,259$ in total). **Correct answers** are in bold. Topic words with retrieved images are underlined (images not shown for space reasons).

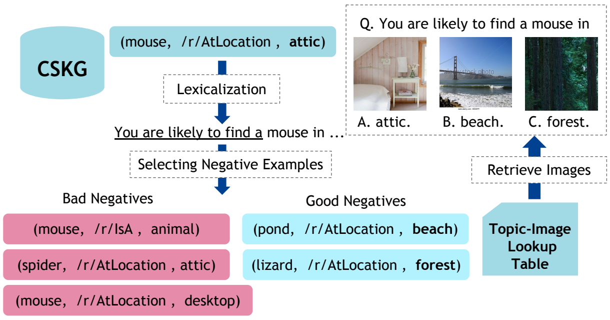

CSKG
(mouse, /r/AtLocation , attic)
Lexicalization
You are likely to find a mouse in ...
Selecting Negative Examples
Bad Negatives
(mouse, /r/IsA , animal)
(spider, /r/AtLocation , attic)
(mouse, /r/AtLocation , desktop)
Good Negatives
(pond, /r/AtLocation , beach)
(lizard, /r/AtLocation , forest)
Q. You are likely to find a mouse in
A. attic.
B. beach.
C. forest.
Retrieve Images
Topic-Image Lookup Table

Figure 1: CWWV$_{Img}$ construction pipeline.

In §3.1, we first describe Ma et al.'s (2021) pipeline, shown in Figure 1, to create the purely textual CWWV.[2] Then, to build our QA dataset CWWV$_{Img}$, we retrieve images for CWWV's instances through a topic-lookup table (Zhang et al., 2020b), as we explain in §3.2.

3.1 Generation of CWWV

The Commonsense QA data CWWV is automatically generated from a consolidated commonsense knowledge graph (Ilievski et al., 2020, CSKG). CSKG is an aggregation of 7 knowledge bases that represents knowledge statements as structured triples (h, r, t) of their start node h, relation label r, and end node t, as in $(\text{mouse}, /\mathtt{r}/\text{AtLocation}, \text{attic})$. All knowledge relations are categorized into one of 13 abstract dimensions, e.g., *spatial*, *similarity*, *temporal*. To create CWWV from CSKG, we only consider the knowledge bases ConceptNet (Speer et al., 2017), WordNet (Miller, 1995), Visual Genome (Krishna et al., 2017) and Wikidata (Vrandecic and Krötzsch, 2014), and 10 dimensions (30 knowledge relations, see App. A, Tab. 5 for the mapping).

Ground-Truth QA Generation Given a knowledge triple (h, r, t) in CSKG, a ground-truth prompt–answer pair is generated by treating the start node h and the end node t as prompt and correct answer, respectively, and applying pre-defined sentence templates to lexicalize them into a sentence. For example, the triple $(\text{mouse}, /\mathtt{r}/\text{AtLocation}, \text{attic})$ in Figure 1 is transformed into a sentence "*You are likely to find a* mouse *in an* attic". To prevent models from applying shortcuts, triples with overlapping content words between the start and end nodes are discarded; e.g., $(\text{bread slicer}, /\mathtt{r}/\text{UsedFor}, \text{slicing bread})$ won't be included. Uncommon concepts or named entities are also filtered out.[3]

Selecting Negative Candidates For each generated QA instance, Ma et al. (2021) select two negative answer candidates according to two principles: (i) the negative candidate is related to the prompt, and thus remains informative for decision, and (ii) it can be clearly discriminated from the correct one, and thus maintains fairness for the model.

To satisfy informativeness, negative candidates are randomly chosen from a pool of relation triples (h', r', t') with $r' = r$, i.e., the relation is the same as the original one. In this sense, a bad negative would be, e.g., $(\text{mouse}, /\mathtt{r}/\text{IsA}, \text{animal})$ (Fig. 1). To ensure fairness, the end node must not be the ground-truth one, $t' \neq t$, and h' must not share any overlapping tokens with h; e.g., $(\text{spider}, /\mathtt{r}/\text{AtLocation}, \text{attic})$ and $(\text{mouse}, /\mathtt{r}/\text{AtLocation}, \text{desktop})$ (Fig. 1) are discarded for violating these two heuristics.

We create CWWV by randomly sampling $2,500$ QA instances for each dimension (*creation* only has 141 samples), totalling $22,641$ instances.

[2] We use Ma et al.'s 2021 script provided at https://github.com/Mayer123/HyKAS-CSKG.

[3] Uncommon concepts are determined by low word frequency in a corpus https://pypi.org/project/wordfreq/ (accessed 9 September 2020) and named entities are identified through the capital letter.

3.2 Generation of CWWV$_{Img}$

Retrieving Images from Conceptual Captions The VL models we use in our experiments are pretrained on the training set of Conceptual Captions (Sharma et al., 2018, CC), a widely adopted dataset of weakly-associated image–text pairs collected from the web which may be regarded as a model's visual experience. We hence use the training set of CC as image retrieval pool to augment CWWV with images. As an efficient way to provide a visual environment, as realistic as possible, for a purely linguistic task, we perform an efficient retrieval method inspired by Zhang et al. (2020b): We first transform CC's image–caption pairs into a topic–image lookup table T. Given a prompt–answer candidate pair (statement) of CWWV, we then select several topic words based on TF–IDF weights, $QA_{topic} = \{t_1, t_2, ..., t_q\}$, and use them to query T for associated images; for example, "A wet floor is slippery" has a set of topic words $\{$wet, floor, slippery$\}$ (Further details can be found in Appendix. B.1). Recall that each QA instance has three statements. As shown in Figure 1, by querying T for the three candidate answers "attic", "beach", "forest", we retrieve their corresponding images. If no image can be retrieved for a statement of an instance, the instance is discarded altogether. The resulting image-grounded commonsense QA dataset, CWWV$_{Img}$, has $13,259$ QA instances in total with the image grounding rate of 58.6%.[1]

Quality of Retrieved Images We assessed the effectiveness of our simple retrieval approach through a human annotation study on Amazon Mechanical Turk (AMT), in which we asked workers to judge the association of image–word pairs. We sampled $1,000$ pairs from CWWV$_{Img}$ uniformly across the 10 commonsense dimensions, and asked AMT workers to judge each pair as either "associated" or "not associated". Details on the annotation methodology and data analyses are given in Appendix B.2. According to majority vote (2 out of 3 judges per pair), 64.2% of the pairs are associated, among which *part-whole* and *spatial* have more than average associated pairs (74.3% and 70.2%, respectively). The inter-annotator agreement under Fleiss' Kappa coefficient (Fleiss, 1971) is between $0.21 - 0.44$ across dimensions, which is only a fair to moderate agreement, indicating the high subjectivity of this task. We also found, unsurprisingly, that concrete words and nouns tend to get higher scores with their paired images.

4 Experiments: QA Task and Inference

Our goal is to assess the benefit of pretrained VL models for purely linguistic tasks underlying commonsense (CS) knowledge. Specifically, we seek to answer the questions **(Q1)** – **(Q4)** that we put forward in §1. To this end, we use our derived image-grounded dataset, CWWV$_{Img}$, and evaluate pretrained VL models against language-only PTLMs in a prompt-based QA task setting (§4.1). We stress that in order to solve CWWV$_{Img}$, only natural language understanding and commonsense knowledge is required, but no explicit visual input (images).

We perform our experiments in a *zero-shot* setting, i.e., without fine-tuning the models on the task, since our goal is to study the ability of *task-agnostic* pretrained models to capture commonsense knowledge (Tamborrino et al., 2020; Ma et al., 2021).

4.1 Task: Prompt-based Zero-Shot QA

Given an instance $T \in \mathcal{T}$ of CWWV$_{Img}$, comprising three natural language statements (and associated images I_i), $T_i = (Q||A_i), i = 1, \ldots, 3$, where Q is the prompt, and A_i a candidate answer. Let $t_j \in T_i, j = 1, \ldots, |T_i|$ denote the sequence of tokens in T_i. Then, the task is to determine which of the three statements is a true assertion (given visual context I_i or not). To mitigate the bias that some template prompts can favor one model over the other in terms of knowledge retrieval (Jiang et al., 2020), we use a two-stage inference procedure, namely a generative and a discriminative setting. During the generative stage, we test representative PTLMs and VL models, respectively, under their natural setting—masked language modeling (MLM) (as detailed in §4.3). Later in the discriminative stage, the ranking of the candidate answers A_i is determined by how well the model can reconstruct the masked tokens of the respective statement T_i by comparing the MLM loss.

4.2 Tested Models

Pretrained Language Models (PTLMs) We use BERT (Devlin et al., 2019) for our comparison, since this model serves as the linguistic backbone of the VL models that we study. We also compare against RoBERTa (Liu et al., 2019), which was pretrained on ten times more data than BERT (160GB vs. 16GB of text, resp.). We use the *BASE* models of the HuggingFace library (Wolf et al., 2019).

VL Models We select the single-stream model, UNITER (Chen et al., 2020), and the dual-stream variant, VILBERT (Lu et al., 2019), as the respective representative models of the two common VL architectures. UNITER is built to have unconstrained inter-modal and intra-modal attentions across all attention blocks, whereas VILBERT has certain attention blocks specifically constrained to perform inter-modal attention only.

We use the pretrained models of VOLTA (Bugliarello et al., 2021), which provides both architectures in an unified framework and under a controlled setup to allow a fair comparison.[4] The models were initialized with the pretrained BERT$_{BASE}$, and further trained on the training set of Conceptual Captions (CC) under three objectives: masked language modeling (MLM), masked object classification and image–text matching.

The visual input is preprocessed into a sequence form of visual tokens $v \in V$ consisting of an `[IMG]` feature and 36 object region features.[5] Each visual token is accompanied with its corresponding spatial encodings in a 5-d vector.[6]

4.3 Inference Variants

Common to all models is that they are queried with each of the three statements $T_i = (Q||A_i)$ of an instance $T \in \text{CWWV}_{Img}$, and the T_i that receives the lowest mean MLM loss $\text{S}(T_i)$ will be returned as answer. We explain the respective precise formulations of $\text{S}(T_i)$ that the textual and VL models apply in the following.

Inference in Language Models To compute $S(T_i)$ in the case of the language-only LMs, we sequentially mask out each token q_j in T_i's prompt Q of length L_Q, and compute its log-likelihood, conditioning on the remaining tokens $T_{i\setminus j}$[7]:

$$S(T_i) = -\frac{1}{L_Q}\sum_{j=1}^{L_Q} \log P(q_j|T_{i\setminus j}), q_j \in Q$$

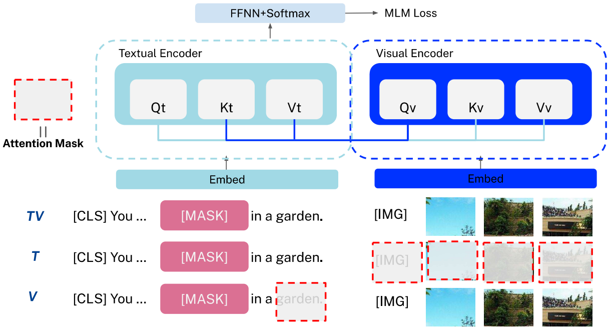


Figure 2: Different inference modes in the VL models (dual-stream shown here): **Retrieved** (TV), **Text-Only** (T), **Vision-Only** (V) **Modes**.

This way, the statement score is not affected by the internal bias of the answer's tokens, e.g., word frequency (Tamborrino et al., 2020).

Inference in VL Models Recall that CWWV$_{Img}$ provides multiple images to serve as visual context for each statement (§4.1). Apart from examining the behavior of the VL models when they get exactly the same input as the textual LMs (**Text-Only**), we further analyse them on modes differing in their input (Fig. 2): (i) their natural setting with multimodal input (**Retrieved**), (ii) **Vision-Only**, and (iii) visual noise with random images (**Dummy**).

Text-Only Mode To examine **Q3** (see §1), if explicit, situated visual input is required for using VL models for purely language-based tasks underlying CS knowledge, we apply inter-modal attention masks on the models' visual input. This way only the representations of the linguistic encoder affect the model decision. Hence, we can test if the VL models encode aggregated visual exposures which they can activate through language only input. Models under this mode are suffixed $*_T$, e.g., **UNITER**$_T$.

Retrieved Mode Given a statement in sequence form, $T_i = (t_1, t_2, \ldots, t_{|T_i|})$, with its associated images $I_i = \{e_1, \ldots, e_q\}$ based on the q topic tokens t_l identified in T_i (see §3.2). Instead of testing the models under a *situated* context like the standard VQA task, our goal is to examine their effectiveness in activating world knowledge from explicit visual input during inference. We thus deviate from the conventional setup and do not provide a single, but multiple images that represent general visual concepts to the models. We apply a threshold τ on the object detection score to choose

[4] We choose UNITER and VILBERT since they perform the best among each respective variant on a wide range of VL benchmarks (Bugliarello et al., 2021).

[5] Region features are extracted by a Faster R-CNN with a ResNet-101 backbone (Anderson et al., 2018) trained on Visual Genome with 36 regions of interest; `[IMG]` is the mean-pooling of the 36 features (Bugliarello et al., 2021).

[6] $(x_1, y_1, x_2, y_2, w * h)$: normalized left/top/right/bottom coordinates and the area.

[7] For convenience, we use the notation $X_{i\setminus j}$ to refer to a sequence of elements $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{|X|})$.

row	part-whole	taxonomic	distinctness	similarity	quality	utility	creation	temporal	spatial	desire	All
1 RoBERTa	68.5	61.8	**80.2**	**67.4**	69.7	**74.2**	72.0	**60.9**	54.8	**65.9**	**67.5**
2 BERT	62.8	**71.2**	80.1	54.8	68.1	72.4	**74.0**	53.7	52.4	60.4	65.0
3 BERT$_{CC}$	68.4	62.0	66.6	51.1	66.0	65.4	62.0	53.6	**63.7**	58.3	61.9
4 UNITER$_T$	**70.9**	59.8	71.3	51.2	**69.9**	71.5	71.0	52.7	61.5	62.5	64.0
5 VILBERT$_T$	63.9	60.3	64.9	46.7	66.1	71.2	58.0	52.2	61.0	62.8	60.7

Table 2: Accuracy on CWWV$_{Img}$ under Text-Only inference mode (The full table can be seen in App. C, Tab.7).

a *set* of salient regions from each individual image[8], resulting in a set of extracted visual tokens $V_{e_i} \cup$ {[IMG]} which we feed into the model. Now, during MLM inference, we need to ensure fairness to the LMs by avoiding information leakage from the visual modality. Hence, when masking out token q_j in T_i's prompt Q, we also mask out the subset of visual tokens V_{q_j} associated to $q_j \in Q$:

$$S(T_i) = -\frac{1}{L_Q}\sum_{j=1}^{L_Q}\log P(q_j|(T_{i\setminus q_j}\|V_{e_i\setminus V_{q_j}})$$

To reconstruct q_j, it is conditioned on the concatenation of the unmasked textual tokens $T_{i\setminus q_j}$ and the remaining visual tokens $V_{e_i\setminus V_{q_j}}$ that are not associated to q_j. Models evaluated under this mode are denoted as **UNITER**$_{TV}$ and **VILBERT**$_{TV}$.

Vision-Only Mode We ablate the textual modality via inter-modal attention masking, the VL models can hence rely only on visual observation to reconstruct masked tokens in the prompt Q of a candidate statement. For example, to reconstruct the masked "vegetables" token in prompt "You are likely to find [MASK] in" (Fig. 2), answer candidate A_3 ("garden") is not given as observable input, but only its respective associated image I_3. Note that also the visual tokens associated to "vegetables" are masked:

$$S(T_i) = -\frac{1}{L_Q}\sum_{j=1}^{L_Q}\log P(q_j|(Q_{i\setminus q_j}\|V_{e_i\setminus V_{q_j}})$$

This posits a challenging task on testing the models' cross-modal integration ability of non-aligned concepts without the help of the corresponding textual part as a bridge (e.g., relate the appearance of "garden" to the concept of "vegetables"). The variants are suffixed $*_V$, e.g., **VILBERT**$_V$.

Dummy Mode We sample 3 images from the CC image pool that are not in CWWV$_{Img}$ to serve as random visual input for each QA instance. During inference, we allow the VL models to fully observe this visual input $\tilde{V}_{e_i}$ since it is not or barely related to any textual token $q_j \in Q$:

$$S(T_i) = -\frac{1}{L_Q}\sum_{j=1}^{L_Q}\log P(q_j|X_{i\setminus q_j}\|\tilde{V}_{e_i})$$

, where $X = T$ in the Retrieved Mode, and $X = Q$ for Vision-Only. This is an adversarial test of the cross-modal reasoning ability of the VL models and we expect a lower performance. In addition, it enables us to assess in how far the VL models can deal with noise. These variants are suffixed $*_{(T)\tilde{V}}$, e.g., **UNITER**$_{T\tilde{V}}$.

5 Results

To address the overarching question we put forward in §1, namely if VL models can serve as commonsense knowledge base, and to what extent they complement pure linguistic model, we first examine the encoded knowledge of the VL models on individual commonsense dimensions, i.e., **(Q1)** in §1. We report the models' effectiveness by measuring their mean accuracy in selecting the correct answer out of the three statement candidates of each QA instance in CWWV$_{Img}$. We declare outperformance if $p < 0.05$ according to the paired student's t-test (Fisher, 1949) for statistical significant differences between any two accuracy scores.[9]

Table 2 shows the effectiveness of all models per commonsense dimension and overall when they are given the exact same input, i.e., natural language statements. Comparing the VL models, UNITER$_T$ and VILBERT$_T$, against their linguistic backbone BERT, we see that both are more effective on the **part-whole**, **spatial**, and **desire** dimensions.[10]

On **spatial**, both UNITER$_T$ and VILBERT$_T$ even outperform RoBERTa, which was pretrained on an order of magnitude more data than BERT

[8] Starting at $\tau \geq 0.7$, we decrease τ by 0.1 steps until at least one region is found.

[9] We used Anderson-Darling's (Anderson and Darling, 1954) method to test for normal distribution.

[10] We also evaluate UNITER initialized with BERT weights, without further pretraining on CC, called UNITER_BERT$_T$. It yields similar results as BERT (see App. C, Tab. 7+8) and indicates pretraining on visual data may lead to a catastrophic forgetting on some CS dimensions that require linguistics.

row		part-whole		spatial		taxonomic		distinctness	
		CWWV$_{Img}$	CWWV$_{Clip}$	CWWV$_{Img}$	CWWV$_{Clip}$	CWWV$_{Img}$	CWWV$_{Clip}$	CWWV$_{Img}$	CWWV$_{Clip}$
1	UNITER$_T$	**70.9**	**76.6**	**61.5**	**59.4**	59.8	58.0	**71.3**	**72.5**
2	VILBERT$_T$	63.9	70.7	61.0	58.8	**60.3**	**59.4**	64.9	62.6
3	UNITER$_{TV}$	63.0	68.1	57.4	54.1	54.0	55.9	65.9	68.1
4	VILBERT$_{TV}$	55.0	58.0	52.9	**59.4**	49.9	58.0	55.9	62.6

Table 3: Model accuracy on CWWV$_{Img}$ and CWWV$_{Clip}$. **Bold** represents the highest score per commonsense dimension.

and the VL models (§4.2). The benefit of visual–linguistic pretraining for spatial (and concrete part–whole) relations is in accordance with what we would expect, and in line with existing work on the spatial dimension (Yatskar et al., 2016; Cui et al., 2020). UNITER$_T$ is also on par with RoBERTA on **part–whole** and **quality**. On the other hand, the VL models failed to retain knowledge associated with other dimensions during VL pretraining, performing significantly worse ($p < .05$) than BERT in particular on **taxonomic** and **distinctness**. Regarding **taxonomic** (and **similarity**), we observe that the VL models tend to struggle with visually non-depictable concepts (e.g., *speculate*, *remember*). And contemplating **distinctness** (e.g., *flood* vs. *drought*) may be challenging for the VL models due to the unnatural, simultaneous co-occurence of opposite concepts in a single image.

Regarding **temporal**, where events are expressed as verbal phrases (e.g., *checking vital signs*, *wait on tables*), the ability of the VL models to leverage the potential benefit of visual information is limited by their pretraining on isolated images (and regions) instead of, e.g., videos, and with objectives (§4.2) which essentially stipulate the models to learn modality alignments, and in particular region-level recognition (Chen et al., 2020), which limits their ability to capture inter-object interactions, relevant for verb-centric and event understanding (see also Hendricks and Nematzadeh 2021).

6 Analysis

To answer our questions **(Q2)**-**(Q4)** we put forward in §1, we first examine **(Q2)**, the benefit of visual information during *pretraining*. We then investigate **(Q3)**, the role of explicit visual input during *inference*. Lastly, we look into **(Q4)**, whether the process of knowledge acquisition and retrieval act consistently across VL models.

6.1 Role of Visual Input during Pretraining

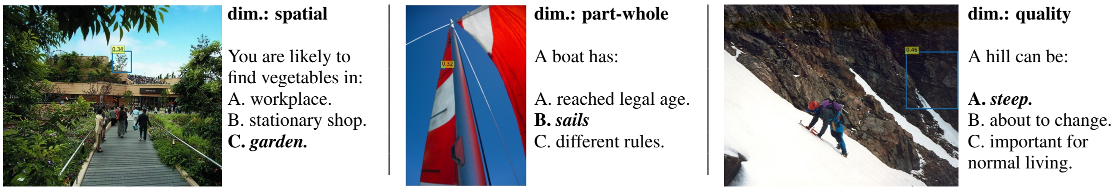
The VL models show the ability to learn certain types of commonsense knowledge that complements that in text-based models by leveraging information during visual–linguistic pretraining (cf. Cui et al. 2020), which they can activate even when no visual information is given during inference. While we found that this does not attribute to the mere size of the training data, it is not clear if the models benefit from the explicit visual information (i.e., visual features), or from the weakly associated verbalizations (i.e., captions) of its visual data. The latter would just be an effect of the domain shift to the visual world, providing information that is typically not found in text corpora.

To examine the contribution of visual features, we further pretrained BERT on the textual part of the VL models' training data (CC captions)[11], referred to as **BERT$_{CC}$**. The model's effectiveness drops on all dimensions except **part–whole** and **spatial** (rows 2+3, Tab. 2), so the verbalizations are indeed beneficial for these dimensions, but detrimental for the others. Notably, UNITER$_T$, being pretrained additionally on images, overall obtains a higher accuracy than BERT$_{CC}$, and outperforms it on **part–whole**. For **spatial**, though, the captions seem to serve as sufficient surrogate of visual spatial relationships, while explicit visual information is of less benefit for UNITER$_T$. We see these effects only for single-stream UNITER$_T$, while double-stream VILBERT$_T$ falls short against BERT$_{CC}$. We will return to the aspect of the architecture differences in the following section. With regard to **(Q2)**, our results indicate that *pretraining* on explicit visual input is indeed crucial for encoding certain commonsense dimensions.

6.2 Role of Visual Input during Inference

§5 showed promising results regarding the benefit of VL models for *text-based* tasks underlying certain types of CS knowledge. The fact that the VL models were designed to receive multimodal input raises the question if their inference ability benefits from being fed both, textual and visual input. We analyse the VL models when we feed in the

[11] Code from Frank et al. (2021), with MLM and 5 epochs.



dim.: spatial

You are likely to find vegetables in:
A. workplace.
B. stationary shop.
C. ***garden.***

dim.: part-whole

A boat has:

A. reached legal age.
B. ***sails***
C. different rules.

dim.: quality

A hill can be:

A. ***steep.***
B. about to change.
C. important for normal living.

Table 4: Positive examples of VILBERT$_V$'s inference with visual answer tokens only. Images correspond to the visual input for the correct textual answer tokens (i.e., "garden", "sails", "steep"). The bounding boxes mark the highly attended (> 0.3) visual tokens of VILBERT$_V$ on the last inter-modal layer.

images along the textual prompts of CWWV$_{Img}$. We also compare against a subset of CWWV$_{Img}$, CWWV$_{Clip}$, to study how the strength of image–text association may affect the models. To obtain CWWV$_{Clip}$, we estimate the association quality of the image–word pairs of every QA instance in CWWV$_{Img}$ by measuring their CLIPScore (Hessel et al., 2021), and keep those instances with an average score of > 0.6.[12] The proportion of concrete words (concreteness scores > 4, Brysbaert et al., 2014) in CWWV$_{Clip}$ vs. CWWV$_{Img}$ are 35% vs. 27%, respectively.

Table 3 provides results for selected CS dimensions (we refer to Tab. 7+8 in App. C for all results). We see that while the retrieval-based multimodal models, UNITER$_{TV}$ and VILBERT$_{TV}$, do integrate information from the visual input, they both perform noticeably worse than their text-only counterparts, UNITER$_T$ and VILBERT$_T$, that do not get visual input during inference. Partially, this seems to be an effect of noise in the visual stream in the form of weakly-associated or abstract words. For some dimensions, including **part–whole**, **taxonomic** and **distinctness**, both UNITER$_{TV}$ and VILBERT$_{TV}$ yield higher accuracy scores on CWWV$_{Clip}$ which has more strongly associated and concrete image–text pairs than on CWWV$_{Img}$ (see Tab. 3). These performance gains cannot solely be attributed to a differing intrinsic difficulty level of CWWV$_{Clip}$, since the purely text-based models, in contrast, consistently perform worse on CWWV$_{Clip}$ than on CWWV$_{Img}$ (with a mean accuracy drop on CWWV$_{Clip}$ of -4.7pp and -1.5pp for RoBERTa and BERT, respectively; results shown in App. C, Tab. 7+8). In sum, our findings support our hypothesis that abstract concepts and weak cross-modal associations affect the inference ability of VL models, an issue we observed for **taxonomic** and **distinctness**.

Does Visual Input Alone Activate CS Knowledge? We address the question of whether visual context alone can provide substantial and informative cues with the *Vision-Only* mode, which disentangles the contribution of the visual from the textual modality. Both, UNITER$_V$ and VILBERT$_V$ perform much worse than when being also fed textual input (results not shown, see App. C). Yet, we find that visual context does play a beneficial role in some cases as opposed to under $\tilde{V}$-only mode, where the accuracy drops to random (UNITER) or close above random (VILBERT). Table 4 illustrates several cases for which visual cues alone can provide reasonable and sufficient information.

Regarding **(Q3)**, the visual stream of the VL models does not seem to play a dominant role during inference; nevertheless, in the extreme case where of missing textual information, the VL models rely on visual input for decision making.

6.3 Role of VL Model Architectures

If visual noise is indeed the reason for low inference abilities, then the models should fail when they receive only noisy, non-sensible visual input. We observe this effect in the single-stream UNITER$_{T\tilde{V}}$, where we see slight effectiveness drops when we feed in a set of unrelated (*dummy*) images $\tilde{V}_e$ along the textual prompt Q (§4.3) (-1.3pp/-2.4pp for All on CWWV$_{Img}$/CWWV$_{Clip}$, resp.; results shown in App. C, Tab. 7+8). Noise is only part of the reason, though, since overall, UNITER$_{TV}$ yields lower scores on CWWV$_{Clip}$ than on CWWV$_{Img}$ (-.6pp, see Tab. 7+8, App. C for all scores). Unexpectedly, the difference between T and TV is even larger on CWWV$_{Clip}$ than on CWWV$_{Img}$, this amplification may be explained by a higher distribution of concrete concepts in CWWV$_{Clip}$: It may be in particular the visual information of concrete concepts that the model can most effectively learn to ground better in the linguistic encoder, and then activate textually even if no visual inputs are given

[12] We determined this threshold with the mean CLIPScore under the image–word pair group that has absolute association according to human evaluation (§3.2).

during inference (Park and Myaeng, 2017). So, as can be assumed in a linguistic task, the textual stream seems to be the driving force for successful inference. An analysis of the *Modality Importance* (MI) score (Cao et al., 2020) of UNITER$_{TV}$ further supports this. The MI measures the average attention traces of the masked tokens in prompt Q during inference to determine the relevance of textual input vs. visual input (see App. D, Fig. 5 for visualizations and calculation details). The visualization of the average MI scores clearly shows a higher attention density on the textual than on the visual modality.

VILBERT, in turn, which fell short against BERT and BERT$_{CC}$, behaves differently: On CWWV$_{Clip}$, VILBERT$_{TV}$ more effectively includes visual input for decision making, with an overall accuracy that is closer to UNITER$_{TV}$ than it is on CWWV$_{Img}$ (+2.5pp, see Tab. 7+8). And on **taxonomic** and **spatial**, VILBERT$_{TV}$'s accuracy is even higher than UNITER$_{TV}$'s (Tab. 3; the proportion of concrete concepts in CWWV$_{Clip}$ is 32% for taxonomic, 47% for spatial, and 34% for part-whole vs. 25%, 44%, and 32%, resp., in CWWV$_{Img}$). Finally, *Dummy-Mode* VILBERT$_{T\tilde{V}}$ remarkably outperforms VILBERT$_{TV}$ across all commonsense dimensions on CWWV$_{Img}$ (+4.9pp on All) and notably on **taxonomic**, **spatial** and **part–whole**, but is on par on CWWV$_{Clip}$ ($p < .05$). We examine the aspect of noise further with samples where VILBERT$_{T\tilde{V}}$ predicts correctly but VILBERT$_{TV}$ fails. We constantly find VILBERT$_{T\tilde{V}}$ to only pay substantial attentions (> 0.3) to the same single visual token across all samples as well as in all CS dimensions (see App. E).

In summary, regarding **(Q4)**, in the case of single stream UNITER, it indicates that (explicit) visual input is beneficial for pretraining, but not for inference. In contrast, double-stream VILBERT seems to be more dependent on receiving signals from both, text and vision, during inference, but also on a strong semantic image–text association, which it can more effectively use if provided. In case of noise, it relies on textual input for decision-making.

7 Conclusion

Regarding our research questions put forward in §1, our findings strongly suggest that the VL models learn to encode certain visual knowledge in their textual streams during multimodal pretraining, in particular for concrete concepts (see also Kiela et al., 2018), which they can activate from purely textual input during inference, i.e., visual information is not required, or not even beneficial. The fact that the textual stream is the driving force for inference is promising, given that we examined the benefit of VL models for a purely linguistic task.

Regarding the dependence of the architecture for commonsense acquisition and activation, we conclude that the examined single-stream seems to be better suited for text-only QA tasks, while double-stream seems to require some form of signal in the visual stream during inference (but cannot leverage it properly to the extent that it would be better than text-only input).

In summary, we find VL models to be promising regarding their potential use for natural language tasks requiring commonsense knowledge. We also identified a range of limitations for future work: The ability to handle visual noise, to understand events and verbs, and to integrate inconsistent modalities towards metaphorical, rather than situated understanding. Future work lies also on multimodal prompt-engineering for improved knowledge retrieval on commonsense intensive tasks.

Acknowledgments

We are grateful to Filip Ilievski and his co-authors for answering our questions regarding the consolidated commonsense knowledge base and the anonymous reviews for their constructive and valuable feedback.

References

Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O. Riedl. 2020. Automated storytelling via causal, commonsense plot ordering. *CoRR*, abs/2009.00829.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society.

T W Anderson and D A Darling. 1954. A test of goodness of fit. *J. Am. Stat. Assoc.*, 49(268):765–769.

Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2016. Are elephants bigger than butterflies? reasoning about sizes of objects. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3449–3456. AAAI Press.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *CoRR*, abs/2005.07310.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. NEIL: extracting visual knowledge from web data. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1409–1416. IEEE Computer Society.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Wanqing Cui, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2020. Beyond language: Learning commonsense from images for reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4379–4389, Online. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ronald A Fisher. 1949. The design of experiments.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction, AKBC@CIKM 13, San Francisco, California, USA, October 27-28, 2013*, pages 25–30. ACM.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5843–5851. IEEE Computer Society.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L. McGuinness, and Pedro A. Szekely. 2021. Dimensions of commonsense knowledge. *CoRR*, abs/2101.04640.

Filip Ilievski, Pedro A. Szekely, Jingwei Cheng, Fu Zhang, and Ehsan Qasemi. 2020. Consolidating commonsense knowledge. *CoRR*, abs/2006.06114.

Hamid Izadinia, Fereshteh Sadeghi, Santosh Kumar Divvala, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2015. Segment-phrase table for semantic segmentation, visual entailment and paraphrasing. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 10–18. IEEE Computer Society.

Stanislaw Jastrzębski, Dzmitry Bahdanau, Seyedarian Hosseini, Michael Noukhovitch, Yoshua Bengio, and Jackie Cheung. 2018. Commonsense mining as knowledge base completion? a study on the impact of novelty. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*, pages 8–16, New Orleans, Louisiana. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. Learning visually grounded sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 408–418, New Orleans, Louisiana. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Zhicheng Liang and Deborah L. McGuinness. 2021. Commonsense knowledge mining from term definitions. *CoRR*, abs/2102.00651.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *CoRR*, abs/2107.13586.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. *AAAI*, 35(15):13507–13515.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online. Association for Computational Linguistics.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Joohee Park and Sung-hyon Myaeng. 2017. A computational study on word meanings and their distributed representations via polymodal embedding. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 214–223, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Kyle Richardson and Ashish Sabharwal. 2020. What does my QA model know? devising controlled probes using expert knowledge. *Transactions of the Association for Computational Linguistics*, 8:572–588.

Fereshteh Sadeghi, Santosh Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1456–1464. IEEE Computer Society.

Henrique Santos, Minor Gordon, Zhicheng Liang, Gretchen Forbush, and Deborah L. McGuinness. 2020. Exploring and analyzing machine commonsense benchmarks. *CoRR*, abs/2012.11634.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 883–898, Online. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. 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)*, pages 6724–6736, Online. Association for Computational Linguistics.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887, Online. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *CoRR*, abs/2107.02681.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357, Florence, Italy. Association for Computational Linguistics.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198, San Diego, California. Association for Computational Linguistics.

Tian Yun, Chen Sun, and Ellie Pavlick. 2021. Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4357–4366, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5831–5840. IEEE Computer Society.

Hongming Zhang, Yintong Huo, Xinran Zhao, Yangqiu Song, and Dan Roth. 2020a. Learning contextual causality from time-consecutive images. *CoRR*, abs/2012.07138.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020b. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Pei Zhou, Rahul Khanna, Bill Yuchen Lin, Daniel Ho, Xiang Ren, and Jay Pujara. 2020a. Can BERT reason? logically equivalent probes for evaluating the inference capabilities of language models. *CoRR*, abs/2005.00782.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020b. Evaluating commonsense in pretrained language models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9733–9740. AAAI Press.

A Details to the Creation of CWWV

A.1 Data Sources

Following Ma et al. (2021), we use the knowledge sources whose relations can be mapped to ConceptNet relation labels, viz. ConceptNet (Speer et al., 2017), WordNet (Miller, 1995), Visual Genome (Krishna et al., 2017) and Wikidata (Vrandecic and Krötzsch, 2014). ConceptNet represents commonsense knowledge in a graph structure of concept nodes connected by relational edges. WordNet focuses on lexical taxonomic knowledge. Visual Genome is a resource of images densely annotated with region descriptions that describe the depicted objects, their attributes and relationships, and which can be represented as scene graphs. Wikidata is a relational knowledge base of entities.

A.2 Relations

Since, in contrast to Ma et al. (2021), our goal is not to pretrain models on selected knowledge relations, but to reach a high coverage of relations for model evaluation, we consider more relations (30 in total), covered by 10 dimensions (Ma et al. (2021) used 14 and 7, resp., refer to Table 5 for the mappings between the commonsense dimensions and knowledge relations evaluated here). We do not consider the relations *lexical*, *comparative*, and *relational-other*. According to (Ma et al., 2021)'s categorization, ConceptNet does not support any relation type that can be mapped to *comparative*. *Relational-other* clusters noisier relations, which may counteract a clean evaluation. *Lexical* requires understanding of formal linguistic knowledge, which is not our target here.

B Retrieving Images of CWWV$_{Img}$

B.1 Topic-Image Lookup Table

Topic words are identified through term frequency-inverse document frequency (TF-IDF) weight. For a preprocessed caption containing non-stop words only $C_j = \{w_1, w_2, ..., w_l\}$. The TF-IDF weight $w_{i,j}$ of each word w_i in a caption C_j is computed:

$$w_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|C|}{1 + |C_j : w_i \in C_j|}$$

where the term frequency of w_i in C_j is calculated by dividing its appearances $n_{i,j}$ by the total number of words in C_j; inverse document frequency is computed by taking the inverse proportion of the number of captions in which w_i occurs $|C_j : w_i \in C_j|$ within a batch of captions $|C|$.

Dimension	Relation Type	Template
part-whole	/r/PartOf	h is a part of t
	/r/HasA	h has a t
	/r/MadeOf	h can be made of t
taxonomic	/r/IsA	h is a t
	/r/InstanceOf	h has an instance of t
	/r/MannerOf	h is a way to t
distinctness	/r/Antonym	h is the opposite of t
	/r/DistinctFrom	h is not t
similarity	/r/Synonym	h is a synonym of t
	/r/SimilarTo	h is similar to t
	/r/DefinedAs	h is the t
quality	/r/HasProperty	h is/are t
	/r/NotHasProperty	h is/are not t
	/r/SymbolOf	h is a symbol of t
utility	/r/ReceivesAction	h can be t
	/r/UsedFor	h is/are for t
	/r/CapableOf	h can t
	/r/NotCapableOf	h do not t
creation	/r/CreatedBy	h is created by t
temporal	/r/HasFirstSubevent	The first thing you do when you h is t
	/r/HasLastSubevent	The last thing you do when you h is t
	/r/HasSubevent	Something that might happen when you h is t
	/r/HasPrerequisite	h requires t
	/r/Causes	The effect of h is t
spatial	/r/AtLocation	You are likely to find h in t
desire	/r/CausesDesire	h would make you want to t
	/r/MotivatedByGoal	You would h because you want t
	/r/Desires	h wants to t
	/r/NotDesires	h doesn't want t
	/r/ObstructedBy	h is obstructed by t

Table 5: Knowledge dimensions with their clustered knowledge relation types and the corresponding lexicalized templates evaluated in this work.

Each caption is now a sequence of topic words sorted according to their TF-IDF weights, we take the top-k topic words to represent the new caption $C'_j = \{t_1, t_2, ..., t_k\}$. We save the lemma form of t_i and its paired image I_j accompanied by its computed TF-IDF weight into the topic-image lookup table T, where each topic is mapped to several images because of its multiple occurrences in different image-caption pairs. Under the assumption that there exists alignment between each image-caption pair, the TF-IDF weights can be further treated as an approximation of how relevant the paired image depicts the topic. The higher the TF-IDF weight, the better the paired image captures the theme of a topic.

B.2 Image-word Pair Association Analysis

We assess the effectiveness of our simple retrieval approach through a human annotation study on Amazon Mechanical Turk (AMT) in which we ask workers to judge the association of image–word pairs. To this end, we sample $1,000$ pairs from CWWV$_{Img}$ uniformly across the 10 commonsense dimensions. Each HIT comprises a random sequence of 10 image–words pairs, one for each dimension. For each HIT, we ask 3 AMT workers to judge each pair as either “associated” or “not associated”. We define an image as “associated” to its paired word when it can “successfully capture the word’s meaning by either containing the object, picturing the event, depicting the action, or characterizing the appearance, emotion, or manner that the word can describe". For polysemous words (e.g., *clean* can refer to an action or appearance), the workers are encouraged to judge whether the image can capture at least one sense of the word.

In total, 37 workers[13] participated; we paid 0.20 per HIT with an hourly wage of \$12. According to majority vote (2 out of 3), 64.2% of the pairs are associated, among which *part-whole* and *spatial* have more than average associated pairs (74.3% and 70.2% respectively). The inter-annotator agreement under Fleiss’ Kappa coefficient (Fleiss, 1971) is between $0.21 - 0.44$ across dimensions, which is only a fair to moderate agreement, indicating the high subjectivity of this task.

Since some words are inherently more visualizable, we further analyze the pair association score from different facets, such as POS tags and word concreteness (Brysbaert et al., 2014). Table 6 shows the analysis of majority association score and Fleiss Kappa score, broke down into three categories: commonsense dimensions, POS tags and word concreteness. POS tags are recognized using spacy-nlp package while word concreteness is categorized based on the concreteness ratings (5 point scale from abstract to concrete) provided by Brysbaert et al. (2014). Brysbaert et al. (2014) defined concrete words as those that can be directly experienced through senses (e.g. sweetness can be experienced through tasting) whereas abstract words can only be inferred from the linguistic context; they surveyed at least 25 annotations for each word in the list (37,058 words and 2,896 two-words expressions). To obtain a categorical analysis, we define words that receive mean concreteness rating above 4 as *concrete*.

Figure 3 and Figure 4 display concrete words distribution and POS tags distribution over each commonsense dimension respectively, where we see that both *spatial* and *part-whole* dimensions contain more concrete words.

C Results of CWWV$_{Img}$ and CWWV$_{Clip}$

The full results of CWWV$_{Img}$ and CWWV$_{Clip}$ can be found in Tables 7 and 8, respectively.

[13] Workers must reside in the UK, USA, CA, NZ, or AU.

	Majority Association (Absolute Association)	Fleiss Kappa
Commonsense Dimensions		
part-whole	70.3% (42.9%)	0.32
taxonomic	64.9% (31.6%)	0.21
distinctness	63.2% (33.7%)	0.38
similarity	61.1% (35.8%)	0.27
quality	51.5% (24.7%)	0.23
utility	63.2% (32.6%)	0.27
creation	59.4% (40.6%)	0.44
temporal	64.9% (39.4%)	0.33
spatial	74.2% (44.3%)	0.25
desire	69.1% (41.5%)	0.21
POS tags		
NOUN	68.6% (42.9%)	0.33
VERB	58.4% (27.9%)	0.24
ADJ	62.1% (30.1%)	0.24
Word Concreteness		
Conc.	81.2% (57.0%)	0.30
Non-Conc.	58.5% (29.9%)	0.28 0.28
All	64.2% (36.7%)	0.30

Table 6: Majority association measures how often the image-word pair is annotated as “ associated” by the majority of the annotators (2 out of 3) whereas absolute association refers to the whole agreement. Fleiss Kappa score is the inter-annotator agreement.

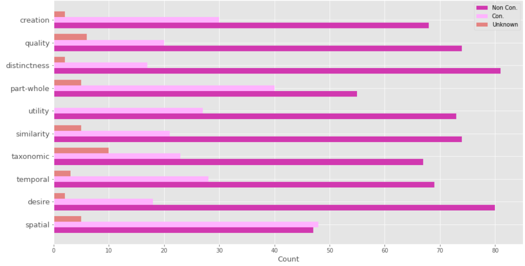


Figure 3: concrete words distribution across commonsense dimensions

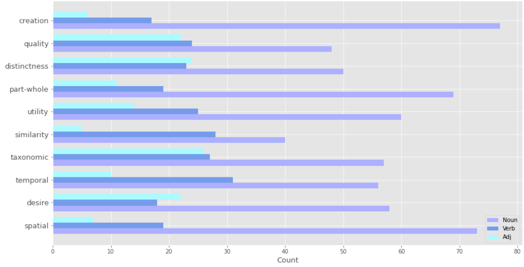


Figure 4: POS tags distribution across commonsense dimensions

D Modality Importance Scores

Following Cao et al. (2020), we analyze which modality of single-stream model (*textual* v.s. *visual*) is more dominant during inference by examining the modality importance score (MI score). In particular, we are interested in the average attention traces on the [MASK] tokens that refer to the head h of the original knowledge triple (h, r, t) before it is transformed into a QA statement (i.e., [MASK] tokens that represent the prompt template, e.g., “You are likely to find X in’, are not considered). Similar to Cao et al. (2020), for the textual modality we disregard the attention values spent on the two special tokens [CLS] and [SEP]; analogously, for the visual modality, the attention value paid to the [IMG] is also ignored. Therefore, the MI scores of the modalities do not sum up to 1.

For a sequence of bimodal tokens, $S = (\texttt{[CLS]}, t_1, ..., t_m, \texttt{[SEP]}, v_1, ..., v_n)$, where $t_1, ..., t_m$ refer to the textual tokens, and $v_1, ..., v_n$ denote the visual ones, the average MI score $\overline{I_{M,j}}$ for each attention head j is calculated as follows:

$$\overline{I_{M,j}} = \frac{1}{L_h} \sum^{L_h} \sum_{i \in S} \mathbb{1}(i \in M) \cdot \alpha_{i,j}$$

$\alpha_{i,j}$ refers to the attention score of the [MASK] token spends on the token i at head j The MI score of each respective commonsense dimension can be seen in Figure 6; Figure 5 gives the mean MI scores across all dimensions.

The visualization of the average MI scores shows a clearly higher attention density on the textual than on the visual modality (Fig. 6). We also observe

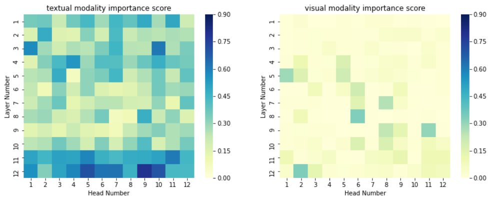


Figure 5: Visualization of MI scores of UNITER$_{TV}$ across 144 attention heads for the cases where UNITER$_{TV}$ is correct and BERT incorrect. Left: textual MI; right: visual MI.

that the MI scores vary across commonsense dimensions. We see a *higher* attention density on the *lower textual* layers, and a *low* density on the *visual* parts on **spatial, temporal, desire**; on the **other dimensions**, we see a higher density on the *upper textual* layers and overall a *higher* density on the *intermediate visual* layers.

E Results with Dummy Images

The visualization of attention traces of VILBERT$_{T\tilde{V}}$ is displayed in Figure. 7.

row	Images	part-whole	taxonomic	distinctness	similarity	quality	utility	creation	temporal	spatial	desire	All
		1, 165	1, 323	828	644	1, 840	2, 090	100	1, 189	1, 599	1, 781	13, 259
1 RoBERTa	–	68.5	61.8	80.2	**67.4**	69.7	**74.2**	72.0	**60.9**	54.8	**65.9**	**67.5**
2 BERT	–	62.8	71.2	80.1	54.8	68.1	72.4	74.0	53.7	52.4	60.4	65.0
3 BERT$_{CC}$	–	68.4	62.0	66.6	51.1	66.0	65.4	62.0	53.6	**63.7**	58.3	61.9
4 UNITER_BERT$_{T}$	–	70.1	**74.5**	**81.4**	62.4	**72.0**	73.8	**79.0**	54.5	53.9	61.5	66.5
5 UNITER$_{T}$	–	**70.9**	59.8	71.3	51.2	69.9	71.5	71.0	52.7	61.5	62.5	64.0
6 ViLBERT$_{T}$	–	63.9	60.3	64.9	46.7	66.1	71.2	58.0	52.2	61.0	62.8	60.7
7 UNITER$_{TV}$	retrieved	63.0	54.0	65.9	46.4	62.4	65.4	62.0	49.2	57.4	58.5	58.4
8 ViLBERT$_{TV}$	retrieved	55.0	49.9	55.9	42.2	57.4	60.5	52.0	47.2	52.9	56.6	53.0
9 UNITER$_{T\tilde{V}}$	dummy	61.5	51.6	63.4	42.2	63.6	66.4	55.0	49.4	58.2	59.7	57.1
10 ViLBERT$_{T\tilde{V}}$	dummy	60.4	58.9	64.9	43.9	63.4	65.5	55.0	48.4	56.8	62.0	57.9
11 UNITER$_{V}$	retrieved	36.4	36.6	40.1	38.5	34.2	36.6	32.0	34.8	36.2	34.3	36.0
12 ViLBERT$_{V}$	retrieved	37.8	35.1	37.7	39.8	36.8	35.7	41.0	33.0	37.6	34.0	36.8
13 UNITER$_{\tilde{V}}$	dummy	30.8	26.3	45.7	28.6	29.2	28.7	19.0	28.7	29.6	30.7	29.7
14 ViLBERT$_{\tilde{V}}$	dummy	34.8	35.8	50.5	40.4	30.4	31.1	30.0	29.4	33.5	30.1	34.6

Table 7: Model accuracy on CWWV$_{Img}$. **Bold** represents the highest score per commonsense dimension across all models and settings; underlined scores denote the best model under the same setting per commonsense dimension.

row	Images	part-whole	taxonomic	distinctness	similarity	quality	utility	creation	temporal	spatial	desire	All
		170	85	86	188	143	120	8	154	144	91	1, 189
1 RoBERTa	–	70.2	60.1	**75.8**	**61.6**	64.3	68.3	75.0	**56.5**	52.4	57.6	62.8
2 BERT	–	61.2	70.6	73.6	45.3	69.5	65.8	**87.5**	52.9	48.2	57.6	63.2
3 BERT$_{CC}$	–	71.8	61.5	65.9	44.2	66.9	64.2	75.0	42.4	**61.8**	52.1	60.6
4 UNITER_BERT$_{T}$	–	69.6	**72.7**	**75.8**	55.8	**70.8**	68.3	**87.5**	**56.5**	47.7	59.0	**64.3**
5 UNITER$_{T}$	–	**76.6**	58.0	72.5	52.3	66.9	**70.0**	75.0	44.7	59.4	54.9	63.0
6 ViLBERT$_{T}$	–	70.7	59.4	62.6	53.5	63.6	**70.0**	62.5	43.5	58.8	**59.7**	60.5
7 UNITER$_{TV}$	retrieved	68.1	55.9	68.1	43.0	61.0	64.2	62.5	42.4	54.1	58.3	57.8
8 ViLBERT$_{TV}$	retrieved	58.0	58.0	62.6	38.4	57.8	56.7	62.5	41.2	59.4	54.2	54.9
9 UNITER$_{T\tilde{V}}$	dummy	66.0	51.7	62.6	45.3	62.3	69.2	37.5	42.4	59.4	57.6	55.4
10 ViLBERT$_{T\tilde{V}}$	dummy	62.8	54.5	63.7	41.9	58.4	63.3	50.0	44.7	56.5	59.0	55.5
11 UNITER$_{V}$	retrieved	33.5	38.5	45.1	40.7	37.0	31.7	37.5	35.3	35.3	34.0	36.9
12 ViLBERT$_{V}$	retrieved	39.4	34.3	41.8	37.2	35.1	40.0	25.0	31.8	40.6	31.2	35.6
13 UNITER$_{\tilde{V}}$	dummy	29.3	19.6	48.4	36.0	23.4	26.7	25.0	34.1	28.2	27.8	29.8
14 ViLBERT$_{\tilde{V}}$	dummy	38.8	30.8	49.5	38.4	29.9	30.0	25.0	28.2	34.1	29.2	33.4

Table 8: Model accuracy on CWWV$_{Clip}$. **Bold** represents the highest score per commonsense dimension across all models and settings; underlined scores denote the best model under the same setting per commonsense dimension.

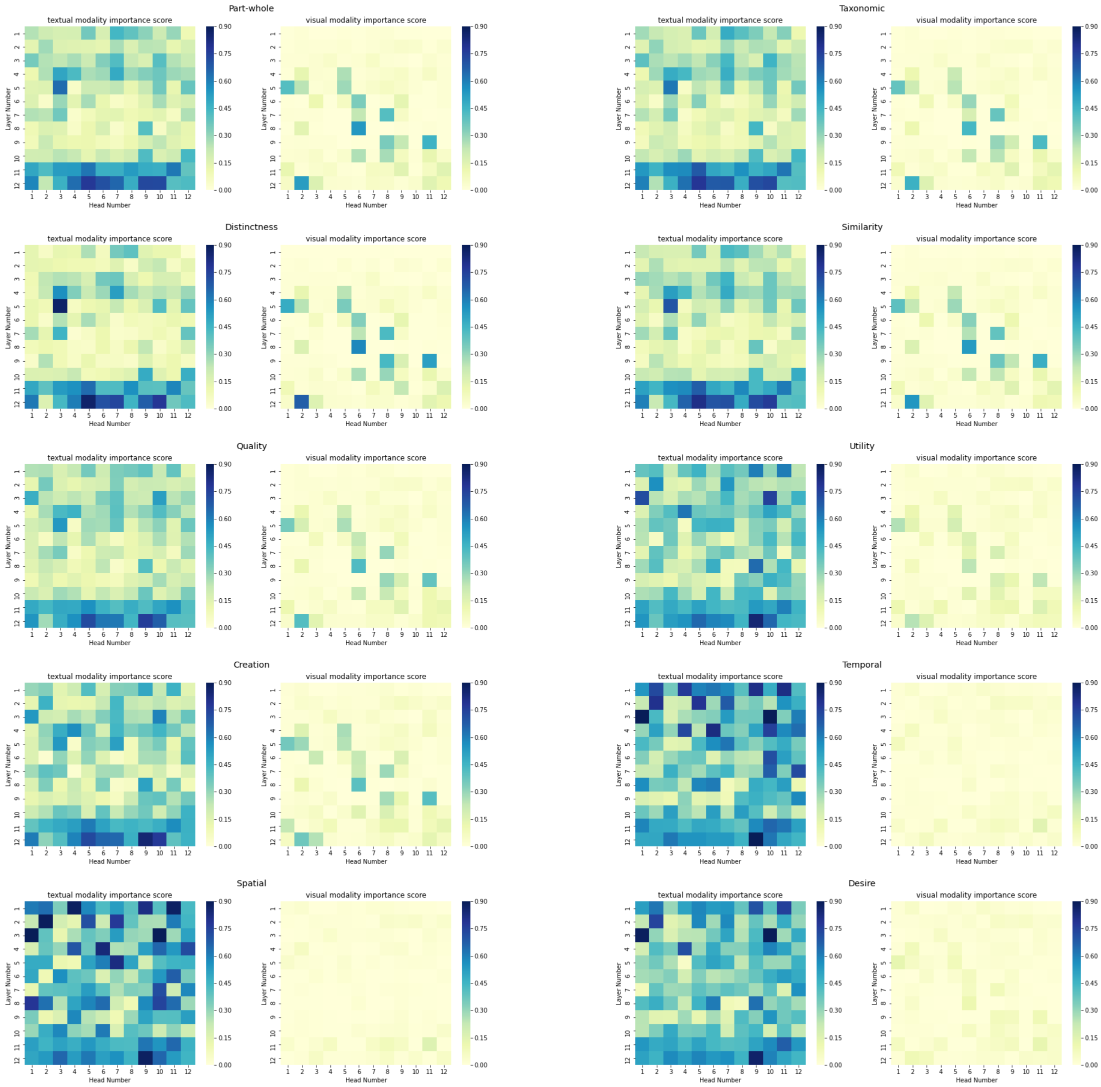
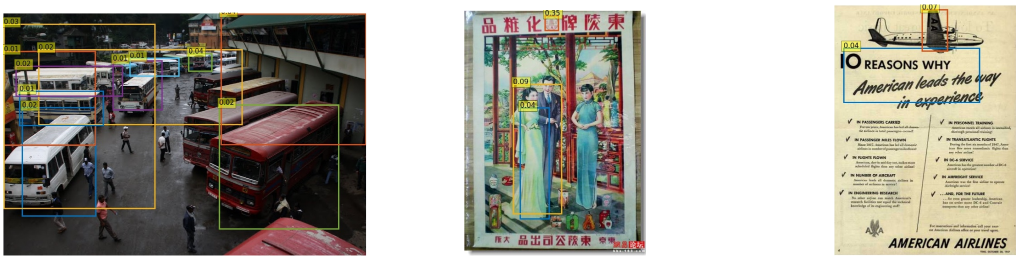


Figure 6: Visualization of the average modality importance scores of UNITER$_{TV}$ across 144 attention heads and across commonsense dimensions under the cases where UNITER$_{TV}$ predicts correctly whereas BERT predicts incorrectly. The order (from top to bottom and from left to right) of commonsense dimension is: part-whole, taxonomic, distinctness, similarity, quality, utility, creations, temporal, spatial, desire.



(a) Dummy Image 1

(b) Dummy Image 2

(c) Dummy Image 3

Figure 7: Attention traces of VILBERT$_{T\tilde{V}}$ last inter-modal layer averaged across cases where VILBERT$_{T\tilde{V}}$ predicts correctly. The bounding boxes represent salient visual tokens of each dummy image and the values in yellow boxes refer to the averaged attention scores. VILBERT$_{T\tilde{V}}$ overly pays attention (0.35) to one single visual token in the second dummy image (7b).