Knowledge enhancement of multimodal tasks

Maria Lymperaiou

National Technical University of Athens

Introduction

Multimodal learning: combining different modalities or types of information for improving performance on several tasks.

- Example: images are usually associated with tags and text explanations
- Conflicts and noise between modalities
- Modalities have different quantitative influence over the prediction output
- Multimodal applications provide more accuracy and robustness than single modality applications as they combine information from multiple sources

[1] https://towardsdatascience.com/multimodal-deep-learning-ce7d1d994f4

[2] Multimodal Co-learning: Challenges, Applications with Datasets, Recent Advances and Future Directions

Introduction

State of the art multimodal frameworks rely on multimodal transformers.

- For example, vision and language are combined in VL Transformer

Transformer-based multimodal learning comprises of two stages:

- **Unsupervised pre-training** on large amounts of unlabelled data
- **Supervised fine-tuning** on labelled datasets dedicated for different downstream tasks

General Challenges

- **Representation**: need to embed data in a unified manner, no matter their 'natural' representation (symbolic, signals etc).
- **Translation**: how to map data from one modality to another: ambiguous and subjective. Example: Many ways to describe an image but maybe no ideal way to translate text.
- **Alignment**: direct relations between (sub)elements from two -or more- different modalities.
- **Fusion**: join information from two -or more- modalities to perform a prediction.
- **Co-learning**: transfer knowledge between modalities, their representation, and their predictive models.

[1] Multimodal Machine Learning: A Survey and Taxonomy

Modality representation

- Text:
  1. Recurrent NNs (RNNs, LSTMs, GRUs)
  2. Distributed word representations (Word2Vec, GloVe, ELMo)
  3. Language transformers (BERT, RoBERTa)
- Image:
  1. CNNs
  2. Image Transformers

Multimodal Transformer: Inputs

- Input token [CLS]
- Word tokenizer (ex Wordpiece): w=w1, w2, ..., wn
- Segment token [SEP]
- Visual token [IMG]
- Visual features
- End token [END]
- Token embedding a sequence [CLS][w1, w2, ..., wn][SEP][IMG][END]
- Segment embeddings indicate the source of each input element by assigning a unique label to each of them
- Position embeddings

Pre-training datasets

Datasets: large-scale paired datasets with images and text for pre-training:

- Conceptual Captions (CC): text & image
- COCO captions: text & image
- Visual Genome: text & image
- SBU Captions: text & image
- Other datasets: Visual Genome QA, GQA, VQAv2, BookCorpus, English Wikipedia

Modality interaction

Modality interaction: aggregating information from modalities via an encoder.

- Fusion encoder
  - Single-stream
  - Dual-stream
- Dual encoder

[1] A Survey of Vision-Language Pre-Trained Models

Steps for end-to-end VL architectures: Encoder

A. Fusion Encoder: takes text embeddings and image features as input and use several fusion approaches.

- The last layer will be treated as the fused representation of different modalities
- Two types of fusion schemes:
  - **Single stream**: assumes simple potential correlation and alignment between modalities
  - **Dual stream**: cross-attention mechanism to model V-L interaction, where the query vectors are from one modality while the key and value vectors are from the another.
    A cross-attention layer usually contains two unidirectional cross-attention sublayers: one from language to vision and another from vision to language.

[1] A Survey of Vision-Language Pre-Trained Models

Steps for end-to-end VL architectures: Encoder

B. Dual Encoder

- Uses two single-modal encoders to encode two modalities separately. No cross-attention!
- Then, it adopts simple methods such as shallow attention layer or dot product to project the image embedding and text embedding to the same semantic space for computing V-L similarity scores.
- Modeling strategy is much more efficient comparing to Fusion encoder, as no heavy cross interaction transformer network is needed.
- Feature vectors of images and text can be pre-computed and stored - even more effectiveness -in retrieval tasks-!

[1] A Survey of Vision-Language Pre-Trained Models

Steps for end-to-end VL architectures: Encoder

- **Fusion** encoder performs better on VL **understanding** tasks
- **Dual** encoder performs better on **retrieval** tasks
- Some VL transformers combine both benefits in a unified model

[1] A Survey of Vision-Language Pre-Trained Models

Pre-training approaches: Fusion encoder

- **Single-stream Architectures**: BERT-like models where they incorporate an Image Embedder, a Text Embedder, and a multi-layer Transformer.
  Models: VisualBERT, VL-BERT, Pixel-BERT, InterBERT, VLP, OSCAR, B2T2, Unified VLP, ViLT, VL-T5, XGPT Unicoder-VL, UNITER, X-UNITER (generative), SIMVLM (zero-shot), SOHO, VideoBERT (zero-shot),

Pre-training approaches: Fusion encoder

- **Two-stream Architectures**: two independent encoders for learning visual and text representations.
  Models: ViLBERT, ALBEF, Visual Parsing, WenLan LXMERT, X-LXMERT (generative), UNIMO (single- and multi-modal)

Pre-training approached: Dual encoder

- CLIP (zero-shot),
- DALL-E (zero-shot, generative)
- ALIGN

Some models can do both fusion encoder and dual encoder:

- VLMO
- FLAVA

Example: Single-stream fusion encoder architectures

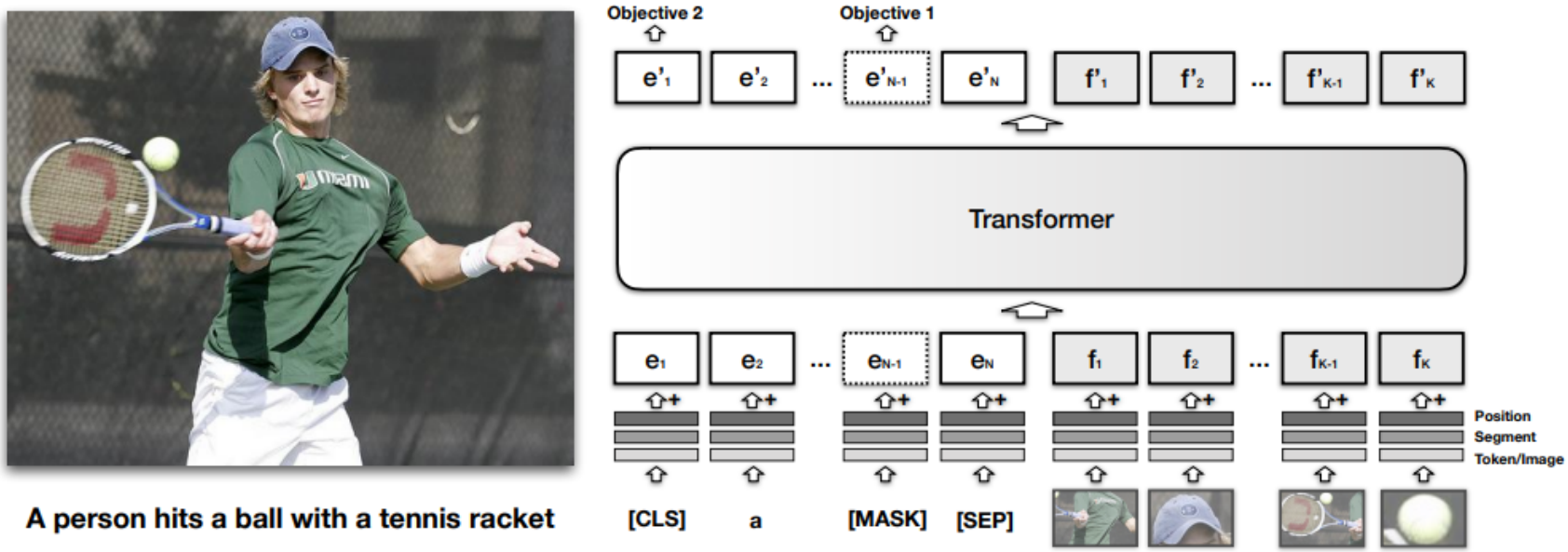

Figure 1: VisualBERT architecture.

Example: Two-stream fusion encoder architectures

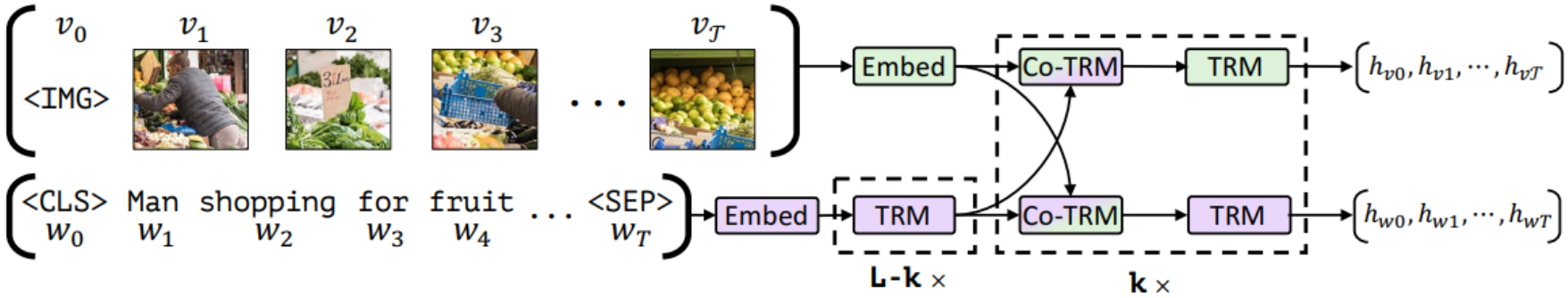

Figure 2: ViLBERT architecture: two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers.

Example: Dual encoder architectures

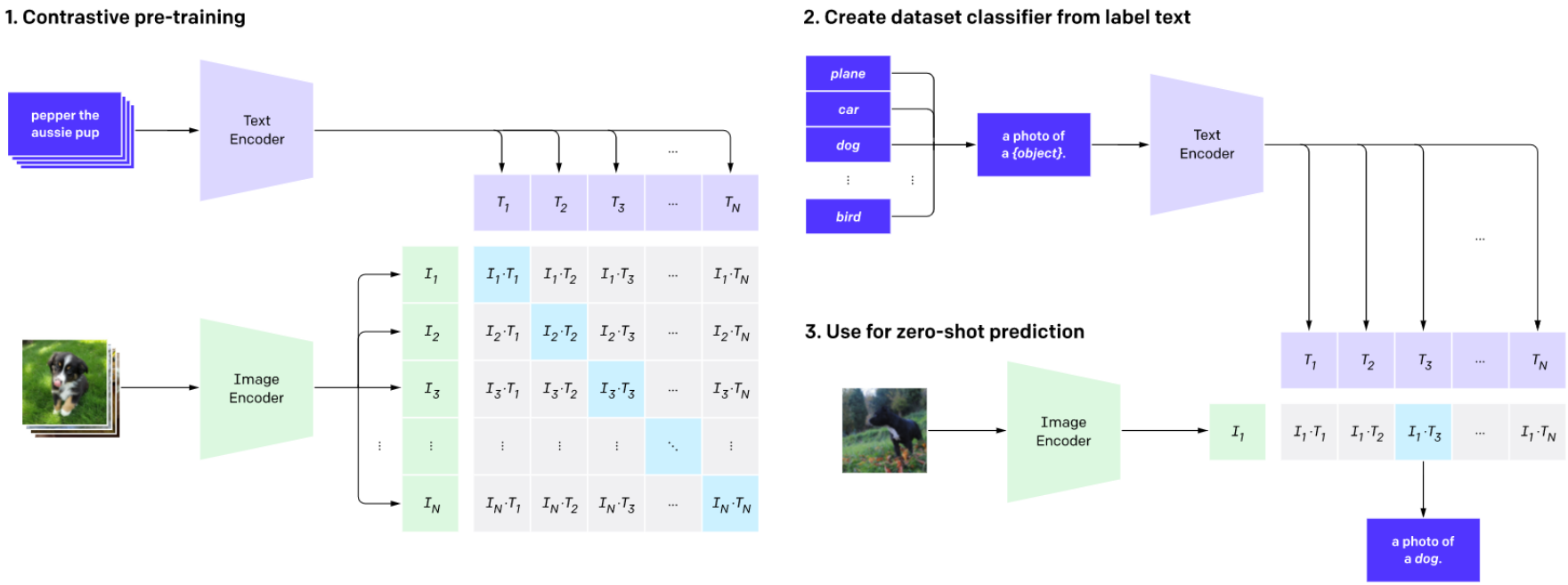

Figure 3: CLIP architecture: pre-trains an image encoder and a text encoder to predict which images were paired with which texts in the dataset.

Steps for end-to-end VL architectures: Pre-training tasks

- Input images and texts have been encoded as vectors and fully interacted
- Next step: design **pre-training tasks**
- Important step: defines what a VL model can learn from data

[1] A Survey of Vision-Language Pre-Trained Models

Pre-training tasks

Language tasks

- **Masked Language Modeling (MLM)**: recover the corrupted tokens in a corpus based on the unmasked ones.
- **Prefix Language Modeling (PrefixLM)**: PrefixLM differs from the standard LM such that it enables bi-directional attention on the prefix sequence.
- **Next sentence prediction (NSP)**:

Pre-training tasks

Vision tasks

- **Masked Object Prediction (MOP)**: randomly mask objects (i.e., masking RoI features with zeros) with a probability and asking the model to predict proprieties of these masked objects.
  - RoI-Feature Regression
  - Detected Label Classification
- **Random Pixel Sampling**: randomly sample feature pixels, similar to Dropout mechanism

Pre-training tasks

Cross-modality tasks

- **Unidirectional (seq2seq) Language Modeling (seq2seqLM)**: masked tokens can attend only to previous ones
- **Masked Multimodal Modelling (MMM)**: reconstruct either image or text from their corrupted versions considering both modalities
- **Cross-Modality Matching (Image-Text Matching-ITM)**: For each sentence, with a probability, replace it with a mismatched sentence, and classify image-text pairs.
- **Image Question Answering (IQA)**: We ask the model to predict the answer to image-related questions when the image and the question are matched (i.e., not randomly replaced)

Pre-training tasks

Cross-modality tasks

- **Cross-Modal Masked Region Prediction (MRP)**: infer object relationships from other unmasked regions and learn V-L alignments by inferring from the text.
  - **Masked Region Classification (MRC)**: learns to predict the semantic class of each masked region from text
  - **Masked Region Feature Regression (MRFR)**:learns to regress the hidden masked region feature to its corresponding original region feature based on text
- **Cross-Modal Contrastive Learning (CMCL)**: learn universal vision and language representation under the same semantic space by pushing the embeddings of matched image-text pairs together while pushing the non-matched ones apart.

Some remarks

- Choosing the right pre-training setup is crucial.
- Pre-training using data in a domain close to the downstream task is a better choice (even if the dataset is synthetic).
- In any case, pre-training with more data is better, but size is **not** the most influential factor: domain relevance is more important than size.
- In case of significant discrepancy between pre-training/downstream datasets, pre-training does **not** help: better train end-to-end.
- More diverse datasets yield increased transferability to downstream tasks.

[1] Are we pretraining it right? Digging deeper into visio-linguistic pretraining

Some remarks

- Inferior annotations is a big issue. Solution: generated in-domain data.
- Best pretrained model $\longrightarrow$ best downstream model **if** finetuned **but** it may not be the most transferable and generalizable pre-trained model otherwise.
- Right pre-training can boost performance without architectural changes

[1] Are we pretraining it right? Digging deeper into visio-linguistic pretraining

Steps for end-to-end VL architectures: Adapting pre-trained models

- Pre-training tasks are able to help VL models to learn general visual and linguistic features
- This knowledge can now be applied in certain tasks, called **downstream tasks**

[1] A Survey of Vision-Language Pre-Trained Models

Vision-Language downstream tasks

Extension of NLP Tasks

- **Image-Text Retrieval**: given a caption and a pool of images, retrieve the target image that is best-described by the caption. Also, inverse task of **Text-Image Retrieval**
- **Visual Question Answering**: answering questions about visual information.

[1] Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods

Vision-Language downstream tasks

Extension of NLP Tasks

- **Visual Dialog**: creating a meaningful dialog in a natural and conversational language about a visual content.
- **Visual Referring Expression**: identifying the object in an image referred to by a natural language expression.[2]
- **Visual Entailment**: predicting whether the image semantically entails the text.
- **Multimodal Machine Translation**: translation from source language(s) to target language(s) by leveraging the visual information along with the text in source language(s).

[1] Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods

[2] Visual Referring Expression Recognition: What Do Systems Actually Learn?

Vision-Language downstream tasks

Extension of CV Tasks

- **Visual Generation**: generation of visual content (image or sequence of images) by conditioning on the text.
- **Visual Reasoning**: provide a relationship between detected objects by generating an entire visual scene graph. The scene graph is leveraged to reason and answer questions about visual information. It can also be used to reason about whether a natural language statement is true regarding a visual input.
- **Visual Commonsense Reasoning**: answer questions given an image and provide a rationale for the choice of answer using commonsense knowledge.[2]

[1] Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods

[2] From Recognition to Cognition: Visual Commonsense Reasoning

Vision-Language downstream tasks

Extension of both NLP and CV Tasks

- **Vision-and-Language Navigation**: interpreting a natural language navigation instruction on the basis of what you see.[2]

Vision-Language downstream tasks

Generative tasks

- **Visual Description Generation (Captioning)**: Given non-linguistic information (e.g., image or video), the goal is to generate a human-readable text snippet that describes the input.
- **Image Generation from Text**: Given a textual description, generate an image (usually with GANs or generative transformers - check Text-to-Image Generation)
- **Visual Storytelling**: instead of dealing with a single visual input, a sequence of visual inputs is used to generate a narrative summary based on the text aligned with them.
- **Story Visualization**: given a sequence of textual or semantic inputs, generate corresponding image frames.

[1] Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods

Zero-shot extensions

- Zero-shot discriminative tasks
  - Example: zero-shot cross-modal retrieval with CLIP
- Zero-shot generation (generate unseen classes or combinations). Most state-of-the-art models somehow exploit CLIP:
  - DALL-E stepping upon CLIP's zero-shot retrieval capabilities, drives generation towards unseen instances
    Or CLIP 'steers' a GAN latent space based on -probably unseen- input text:
  - Big Sleep: BigGAN + CLIP
  - Fuse Dream: CLIP+GAN latent space optimizations
  - VQGAN+CLIP

[1] Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods

[2] Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments

Task-specific datasets

- MS-COCO (image retrieval)
- Flickr 30k (image retrieval)
- Visual Genome (captioning, VQA)
- Visual7W (VQA)
- VQAv2 (VQA)
- GQA (VQA, visual reasoning)
- NLVR (visual reasoning)
- CLEVR (visual reasoning)
- CLEVR-Ref+ (referring expressions)
- RefCOCO, RefCOCO+, RefCOCO-g (referring expressions)
- SNLI-VE (visual entailment)
- GuessWhat (multimodal dialog)
- VCR (visual commonsense reasoning)

Steps for end-to-end VL architectures: fine-tuning

- Supervised fine-tuning stage
- Based on the downstream task we prefer, we fine tune the pre-trained model on an appropriate dataset
- i.e. we train for a few epochs in order to re-adjust the weights towards the downstream task
- This is the power of Transfer Learning!

Steps for end-to-end VL architectures: Evaluation

Metrics are defined as per downstream task:

- **Recall@k**: Image-Text retrieval
- **Accuracy**: Visual Referring Expression, Visual Question Answering, Visual Reasoning with text, Visual Entailment, Visual Commonsense Reasoning
- **Precision**: Visual Referring Expression
- **Inception score (IS), Frechet Inception Distance (FID)**: Text-Image Generation
- **BLEU, METEOR, CIDEr, SPICE**: Image Captioning
- **BLEU, METEOR** Multimodal Machine Translation

Introduction

G=(V, E)
Edges may be:

- Weighted vs unweighted
- Directed vs undirected
- With vs without features

Graph types

- **Multi-relational graphs**: Different types of relationships
- **Heterogeneous graphs**: Different node types and different edge types
- **Multipartite graphs**: heterogeneous graph that exclusively contain edges that connect nodes of different types.

Knowledge graphs

What is a knowledge graph (KG)?

- structured representation of facts F
- consist of entities E, relationships R and semantic descriptions
- directed and heterogeneous structure
- human knowledge in the form of triplets: (h, r, t) or (s, p, o)

Knowledge graphs

- existing edges express known facts
- Two scenarios for missing edges:
  1. Open World Assumption (OWA) assumes that unobserved facts are either missing or false
  2. Closed World Assumption (CWA) assumes that all unobserved facts are false

Graph representation

- Node embeddings (node2vec, DeepWalk, Large-scale Information Network Embedding (LINE), Graph2vec etc)
- Graph Neural Networks (GNNs)
- Graph Convolutional Networks (GCN)
- Graph Attention Networks (GATs)
- Graph Transformers

Why using external knowledge?

- Pre-trained transformers can 'understand' what they have learned before - but not outside of that
  - For example, an NLP transformer trained in general vocabulary cannot understand medical documents. Something similar can happen with VL too
- Repeating pre-training on extra data is prohibitively expensive
- Even then, some rare concepts may not be captured effectively
- Commonsense and factual knowledge is well represented in knowledge graphs, but probably is not 'seen' in pre-training datasets
  - And in the same time such types of knowledge can be useful, even necessary for some tasks!
- Why not just fine tune instead of adding a KG?
  - Not sure

Why using external knowledge?

- Pre-training unstructured data may contain of noise, inconsistencies, errors, biases, which are sometimes hard to be captured and resolved
  - Ensuring clean and unbiased data could be a field of study itself...
- KGs are well structured, you know what you are getting. This is very important, especially in critical domains (like medical)
- Lack of transparency during pre-training: can we be sure what a transformer has learned at all?
  - Even if a model has 'seen' some concepts, we are not sure how it will exploit such information
  - KGs can provide reasoning paths, and therefore explainable insights over how an answer has been derived

Knowledge usage in existing literature

Why previous works in VL models have incorporated external knowledge?

- Performance boosting
- Explainability
- Extendability of tasks: addressing problems needing commonsense, factual knowledge, named entities, events etc

Knowledge usage in existing literature

VL tasks with external knowledge so far:

- Visual Question Answering (VQA)
- Visual Commonsense Reasoning (VCR)
- Visual Reasoning (VR)
- Visual Storytelling (VS)/Story Visualization (SV)
- Image Retrieval from Text (IRT)
- Multiple tasks at once using a single VL transformer

What is this knowledge we are talking about?

- **Hierarchy**: Cat is an animal
- **Lexical**: 'Run' is a verb
- **Commonsense knowledge**: Sugar is sweet, if I go out in the rain I'll get wet
- **Factual knowledge**: Athens is the capital of Greece
- **Temporal/event knowledge**: COVID-19 appeared in 2019, Christmas day is on the 25 Dec each year.
- **Named entities**: Company named Google
- **Part-whole**: A wheel is part of a car
- **Comparative**: Cities are larger than villages
- **Utility**: Spoon is used for eating
- **Visual**: an image with food on a plate

External knowledge sources

We can identify three types of external knowledge sources:

- Implicit knowledge (non-symbolic)
- Explicit knowledge (Knowledge graphs and ontologies)
- Web crawled knowledge

Also, there is internal or *self-knowledge* which relies on extracting knowledge from existing data

External knowledge sources - Implicit knowledge

- Non-symbolic form (weights between neurons)
- Can be obtained via unsupervised pre-training of transformers
- Unstructured knowledge bases
- Advantages:
  - No manual annotations and no need for supervision
  - Many pre-trained models are open source and ready to be used
  - Scalability: Incorporate more knowledge automatically by re-training or fine-tuning
- Disadvantages:
  - Computationally expensive knowledge acquisition: pre-training requires powerful hardware and several days of training
  - Not always sufficient to handle tasks requiring knowledge outside their pre-training 'vocabulary' or concept drifts
  - Data biases will be reflected in the final answer
  - Black-box: we don't know **what** we learn, we don't know **how** we learn

External knowledge sources - Explicit knowledge

- Structured knowledge bases
- Advantages:
  - High-quality annotations, even in fields requiring expertise
  - Transparency: we know **what** we learn, we know **how** we learn
  - Biases can be tackled
  - Concept drifts can be handled by updating concepts
  - We can tune and measure what each kind of knowledge offers (hierarchies, commonsense, encyclopedic knowledge)
- Disadvantages:
  - Manual construction
  - Experts are necessary for certain knowledge bases (ex: medical)

External knowledge sources - Explicit knowledge

Widely used structured KGs:

- Wordnet (hierarchical)
- ConceptNet (commonsense)
- DBPedia (Hierarchical, Encyclopedic/Factual)
- Wikidata (Encyclopedic/Factual)
- WebChild (Commonsense)
- HasPartKB (Commonsense, part-whole)
- Visual Genome (Visual)

External knowledge sources - Web knowledge

- Unstructured data crawled from the web
- Advantages:
  - No manual labelling in contrast to explicit
  - No expensive pre-training in contrast to implicit
  - Easily obtained: this can open the way for a standard external knowledge source, which helps standardizing the benchmarking of knowledge-enhanced VL models
- Disadvantages:
  - Quality of scrapped knowledge - internet contains anything
  - Biases will be reflected in result
- Medium transparency: we have access to the external information component contributing to an answer (**what** it learns), even though this info is not as explicit as a structured KG node (**how** it learns)

Knowledge sources

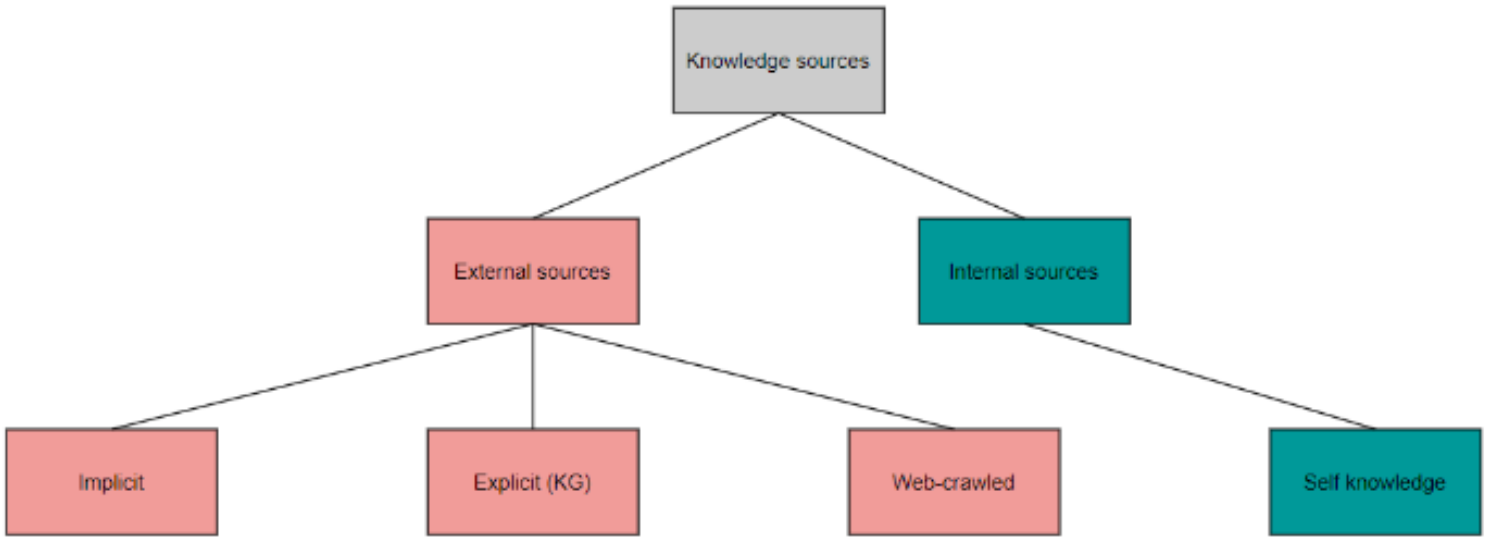

Figure 1: Overview of knowledge sources

Figure 4: Overview of knowledge sources

Knowledge-enhanced tasks

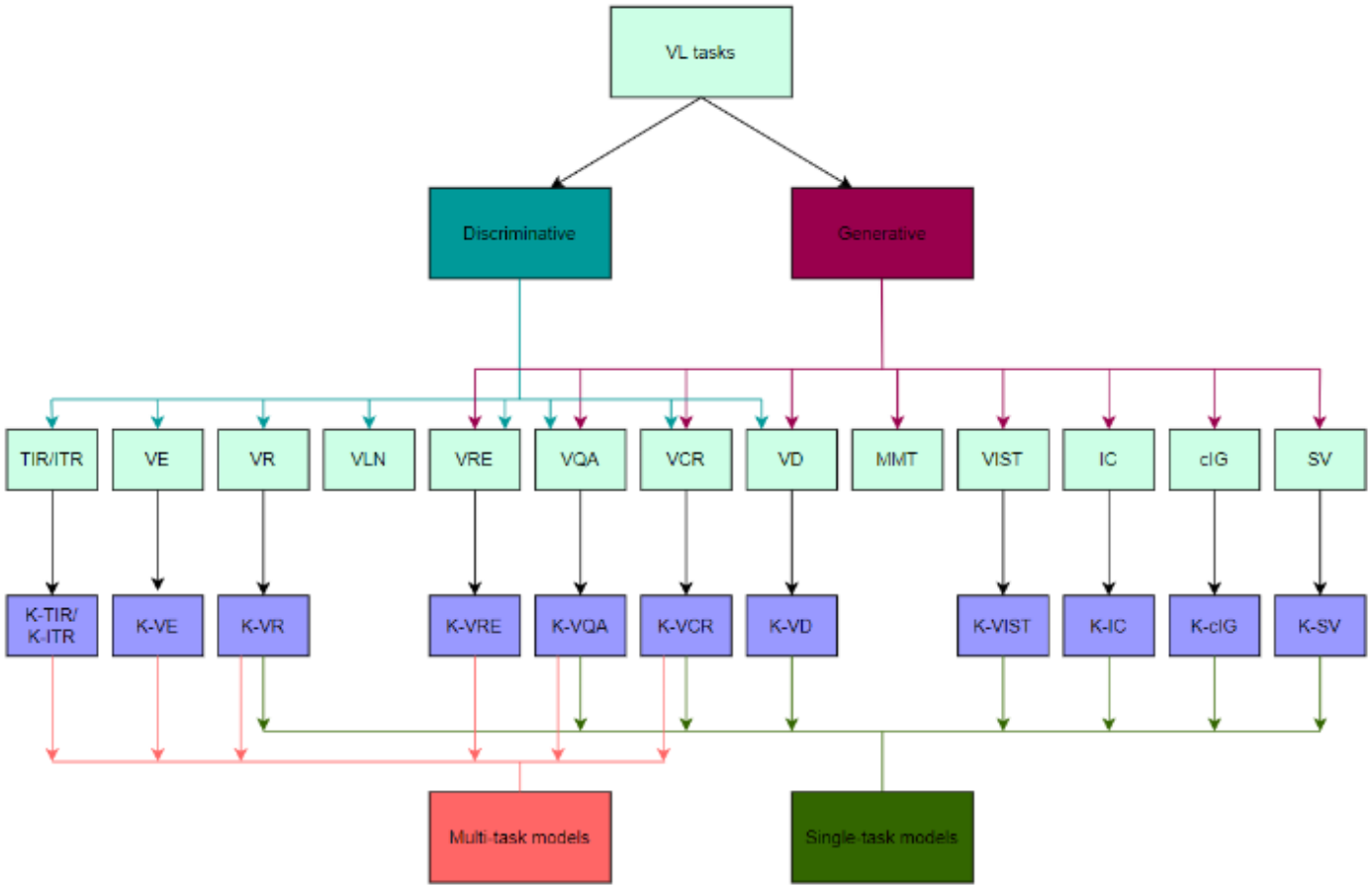

Figure 5: A taxonomy of VL tasks with knowledge

VQA outline

1. Input:
   - Image (a scene containing multiple objects and relationships)
   - Question in natural language
2. Task: Find an answering on the question by reasoning on the image
3. Answer:
   - Candidate answers in natural language (classification problem)
   - Generation of answer in natural language (NLG)
   - KG node (knowledge enhanced VQA)

Visual Question Answering

Figure 6: Example of the VQA task: Image and two questions. (From the VQA dataset)

[1] VQA: Visual Question Answering

VQA: questions

Note the two types of reasoning required.

- The first question needs visual info only: *count number of pizza slices on image*.
- The second one needs something more: *see what ingredients you have, check if all are vegeterian, and if yes answer yes, else no*. Here external knowledge can help.

Knowledge VQA: datasets

- KB-VQA
- FVQA (factual VQA)
- KVQA (knowledge-aware VQA)
- OK-VQA (outside-knowledge VQA)
- Text-KVQA (scene text+knowledge)
- Visual7W+KB
- S3VQA
- ZS-F-VQA (Zero-shot Fact VQA)

Knolwedge VQA: modality encoding

1. Visual modality: Faster-RCNN, ResNet and similar
2. Text modality:
   - Keyword extraction
   - Static distributed representations (Word2Vec)
   - Language Transformers
3. Visual-language encoding:
   - No joint reasoning
   - Joint reasoning with VL transformers
4. Graph encoding
   - No encoding (SPARQL queries)
   - Encoding of extracted text (GloVE, BERT)
   - Graph embeddings using GNNs

Knowledge VQA: Taxonomy

1. Keyword-based explicit KG querying
2. Sequential language models for question encoding
   - Embedding-based fact retrieval from KG
   - Multi-modal graphs
3. Transformer-based question encoding
   - Transformers for language encoding
   - Joint multimodal encoding with VL transformers

Knowledge VQA: keyword based

1. [1]: KB built upon rules extracted from image classes, attributes and actions; diverse graph nodes and edges
2. [2]: SPARQL queries from image attributes on DBPedia help retrieving relevant knowledge which is embedded using Doc2Vec. Knowledge, image captions and image attributes are fed in LSTM.
3. Wang et al [3]: Explainability via reasoning paths. Scene classes and attributes, as well as question key-phrases are linked with DBPedia entities.

Knowledge VQA: towards embedding representations

- No more SPARQL queries from keyword extraction
- Vector representation of involved modalities
- Initially, KG facts were ranked based on embedding similarity (dot product etc)
- Later on, improvements were made thanks to GNNs

Knowledge VQA: RNN encoding - fact retrieval

1. [4]: Learn a mapping between question and KG queries using LSTM encoding. The reasoning process is revealed by obtaining the fact connecting Q and A.
2. [5]: Map image and question features in the same vector space via an MLP. Also, extract facts and embed them via GloVe. Dot similarity ranks facts according to input question-image pairs.
3. [6]: Consider multiple relevant facts in the form of a subgraph. GCN reasons on the subgraph to provide the final answer.

Knowledge VQA: RNN encoding - fact retrieval

1. [7]: Named entities extracted from question and image fetch facts from Wikidata. Retrieved facts and image coords are embedded, and so does the question. MLP defines the answer based on both representations.
2. [8]: Scene text, image and question retrieve facts from external KG. Those components construct a multi-relational graph, upon which a GGNN is applied to define the answer.

Knowledge VQA: RNN encoding - multimodal graphs

1. [9]: Multiple views of the same image: visual, semantic and factual. Intra-modal graph convolutions focus on the most relevant parts of each modality; Cross-modal convolutions aggregates information from different modalities. GCN is applied to derive the answer.
2. [10]: Multi-Modal Heterogeneous Graph construction and reasoning. A heterogeneous GCN extracts question relevant information.
3. [11]: Construction of a subgraph from question and image related concepts, which act as anchor points to ConceptNet and Wikidata. Similarity between anchor entity embeddings and question embeddings results in the answer.

Knowledge VQA: Multiple Feature Spaces

1. [12]: Multiple feature spaces used for alignment between image/question and KG. An alignment process based on mapping (like BERT token mapping) of KG triples guides -unseen- answer prediction.
   - Semantic space (focuses on question - creates relations)
   - Object space (image and question - creates entities)
   - Knowledge space (answers)

   Zero-shot capabilities are achieved, by capturing semantics outside training data by using a KG. Interpretability regarding the support entity and relation is also addressed.

Knowledge VQA: Transformers

- Success of transformers over RNN/LSTM/GRU should follow VL tasks
- Either
  1. extract text from image (captioning) and use a language transformer
  2. or keep modalities as they are and use a VL transformer
- In any case, text embeddings can be obtained from a transformer

Knowledge VQA: Transformers - Single modality

1. [13]: Scene graph is constructed from visual and question embeddings. Concept graph is constructed from joint embeddings and external knowledge. SBERT is used to obtain representations for all modalities. GAT selects most relevant nodes from scene graph and concept graph.
2. [14]: GPT-3 as knowledge base, few-shot task understanding: provide few image captions, questions and answers and then ask for a novel answers based on new question-caption pairs. Answers are generated, not selected.
3. [15]: Unimodal pre-trained transformers yield better generalization capabilities over multimodal approaches of comparable size when external knowledge is necessary. BERT acts as implicit KB and thus receives captioned image and question. Unimodal and multimodal transformers are proven to have complementary capabilities.

Knowledge VQA: Transformers - Multi modal

1. [16]: ConceptBERT contains one dual-stream (ViLBERT) vision-language module and one concept-language module (bidirectional Transformer). Their outputs jointly form a concept-vision-language representation.
2. [17]: weakly-supervised framework on web knowledge first retrieves knowledge in natural language. Then language and vision are encoded using LXMERT.
3. [18]: KRISP only considers external knowledge during inference stage but not during training. Implicit and explicit (DBPedia, ConceptNet, VisualGenome and hasPart) knowledge sources are exploited.

Visual QA: ConceptBERT

How much does knowledge help?

Dataset	L	VL	CL	CVL
VQA 2.0	26.68	67.9	38.24	**69.95**
OK-VQA	14.93	31.35	22.12	**33.66**

Figure 7: Evaluation results on VQA 2.0 and OK-VQA validation sets for different modalities

This indicates that knowledge among with V+L boosts VQA results comparing to other combinations.

[1] ConceptBert: Concept-Aware Representation for Visual Question Answering

Knowledge VQA: Transformers - Multi modal

1. [19]: Scene text, image and GKB facts are inserted in a multimodal transformer, enabling interaction through attention mechanisms between the different modalities.
2. [20]: MAVEx utilizes multi-granular queries to retrieve external knowledge from various sources. A finetuned ViLBERT creates a pool of candidate answers. Retrieved knowledge instances are matched with the queries to acquire the highest ranked supporting fact.
3. [21]: Instead of using pre-selected candidate answers, passage retrieval can fetch the answer. Questions and images are jointly encoded in dense vectors using LXMERT.

Knowledge VQA: Transformers - Multi modal

1. [22]: LXMERT language encoder input is modified to incorporate factual knowledge from Wikipedia by aligning Wikipedia2Vec embeddings with BERT wordpiece vectors. Only fine-tuning is required.
2. [23]: S3, presented with the S3VQA dataset, targets to answer visual question based on all the participating modalities simultaneously. Entity spans from the question are selected to be matched with objects of scene graphs -usually matched with external knowledge- corresponding to images.

Visual QA: ConceptBERT

Evaluation Datasets:

- VQA 2.0
- Outside Knowledge-VQA (OK-VQA)

Model	Overall	Yes/No	Number	Other
Up-Down	59.6	80.3	42.8	55.8
XNM Net	64.7	-	-	-
ReGAT	67.18	-	-	-
ViLBERT	67.9	82.56	54.27	67.15
SIMPLE	67.9	82.70	54.37	67.21
CONCAT	68.1	82.96	54.57	68.00
ConceptBert	**69.95**	**83.99**	**55.29**	**70.59**

Figure 8: ConceptBERT results on VQA 2.0.

[1] ConceptBert: Concept-Aware Representation for Visual Question Answering

Knowledge VQA: Evaluation

1. Answer accuracy: how many times ground truth answer was predicted. Many ablations regarding individual components accuracy were reported. Similarly, *fact* accuracy is reported.
2. WUPS: steps to traverse on WordNet tree until the least common subsumer node.
3. Precision (@ 1/5/10): number of relevant answers among the ones considered.
4. Recall (@ 1/5/10): number of total relevant answers captured. Similarly, *fact* recall is reported.
5. Human evaluation on returned answers and on facts contributing to the returned answer.

The task of Visual Reasoning (VR)

The goal in visual reasoning is to learn a model that comprehends the visual content by reasoning about it.

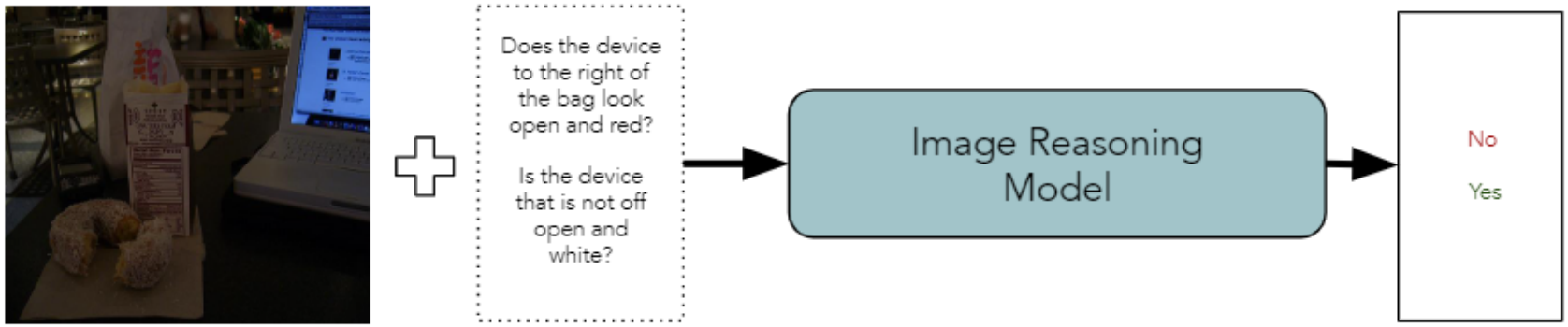

Figure 9: A VR model needs to comprehend an image so that an answer is inferred.

The task of Visual Reasoning (VR)

- Extension of VQA task
- It has been observed that VQA models struggle when comparing the attributes of objects, or when novel at-tribute combinations needs to be recognized (such as in compositional reasoning).
- Discovers a relationship between detected objects by generating an entire visual scene graph
- Neuro-symbolic approaches come as natural
- However, not many works have leveraged existing KGs
- The scene graph can also be used to reason about whether a natural language statement is true or not regarding a visual input (visual entailment)

The task of Visual Reasoning (VR)

Datasets containing spatial knowledge:

- CLEVR
- Visual Genome

There are no dedicated datasets for 'knowledge-enhanced' version of visual reasoning, in contrast with VQa.

The task of Visual Reasoning (VR)

Impression, Sunrise

Q: *What is on the boat?*
A: ***Person***.

Equestrian Portrait of Charles I

Q: *What animal is this?*
A: ***Horse***.

Reasoning about Abstraction

Q: *Who did coin the nickname "Impressionist" in his/her satirical article?* **A:** ***Louis Leroy***.

Q: *Who is wearing greenwich-made armor?*
A: ***King Charles***.

Reasoning about Cultural Context

Visual Reasoning with knowledge in Artistic Domains

- Difficulties arising from abstract styles and cultural contexts
- A deep relational model captures and memorizes the relations among different samples
- Hierarchical knowledge is incorporated into the meta-learning based model.
- The automatic analysis of artistic images can provide auxiliary information for art appreciations
- This study opens the door for mechanical imitation of human aesthetic
- AQUA dataset is crafted for this purpose

[1] Knowledge is Power: Hierarchical-Knowledge Embedded Meta-Learning for Visual Reasoning in Artistic Domains

Visual Reasoning with knowledge in Artistic Domains

- Challenges:
  1. Artistic images are represented heterogeneously due to their abstractions and artistic styles
  2. Artistic images have humanistic and social dimensions due to their cultural and historical contexts.
- KG based on YAGO3
- Hierarchical-Knowledge Embedding captures implicit graph relationships
- A multi-modal fusion model merges I and Q representations
- Knowledge-Based Representation Learning: merges fused I-Q embedding with the knowledge embedding
- A relation model predicts the textual answer for the corresponding questions

[1] Knowledge is Power: Hierarchical-Knowledge Embedded Meta-Learning for Visual Reasoning in Artistic Domains

The task of Visual Entailment (VE)

- Visually-grounded version of the Textual Entailment task
- An image is augmented with textual premise and hypothesis: predict whether the image semantically *entails* the text, given image-sentence pairs, where the *premise* is defined by an image instead of a natural language sentence, while the *hypothesis* remains linguistic.
- A visual entailment model has to predict whether a hypothesis *entails*, *contradicts* or remains *neutral* with respect to the premise.

Example of visual entailment

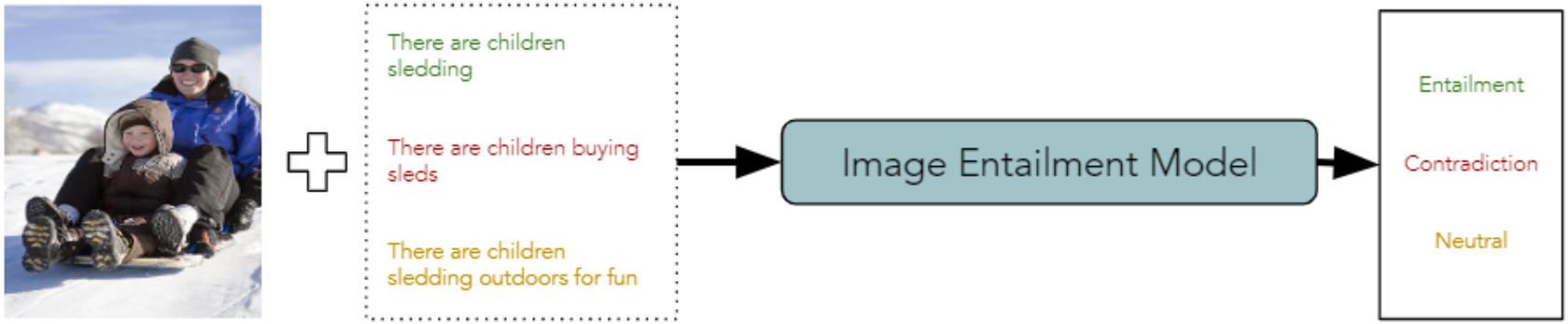

Figure 11: Example of the VE task: Given an image as a *premise* and a natural language text as a *hypothesis*, an Image Entailment Model predicts whether the hypothesis is an entailment, contradiction,or neutral by understanding the evidence(s) present in the image.

The task of Visual Entailment (VE)

- There are no works so far exploiting external knowledge for the VE task
- Comparing to VQA, it is a much underexplored task
- Only two datasets have been developed:
  - V-SNLI
  - SNLI-VE (extension from Flickr30K dataset)
- Just a couple of models are addressing the VE task
- Lots of room for development and incorporation of external knowledge
- Maybe we can develop a 'visual commonsense entailment' task? - as an analog to the VCR task, analyzed in the following section-

Visual Commonsense Reasoning

- Humans can infer people' s actions, goals, and mental states from a scene.
- For machines this process requires higher-order cognition and commonsense reasoning about the world.
- Extension of VQA task: given a question, an image and a set of candidate answers, predict the correct answer together with a *rationale* justifying this answer choice.
- Enhanced explainability of VQA via VCR: *why* was an answer returned?
- External knowledge can contribute naturally to this task (check following image)

Knowledge for commonsense reasoning

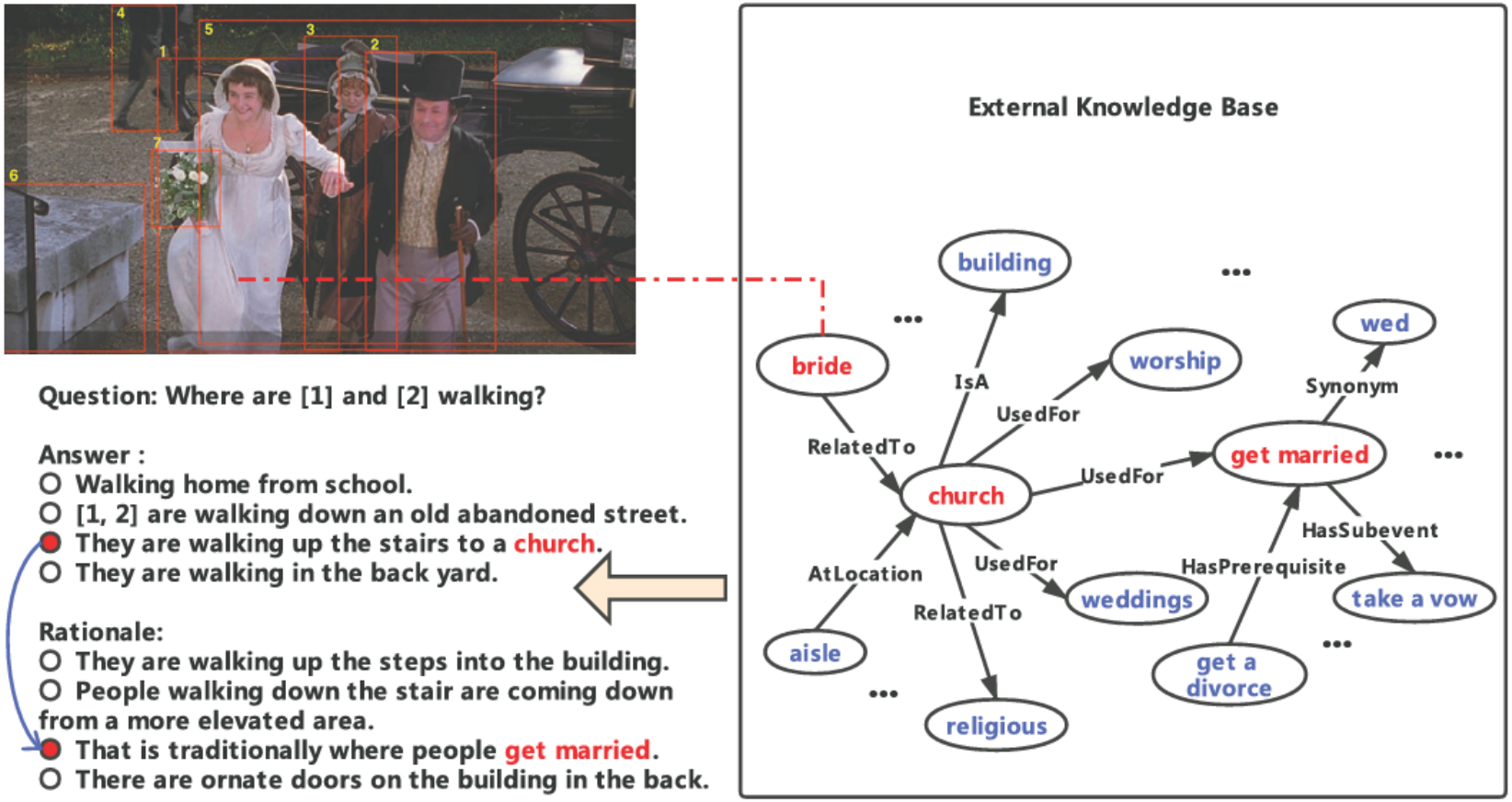

Figure 12: Example of the VCR task: Image, question, candidate answers and knowledge given, selected answer and rationale are returned.

Datasets for VCR

- Visual Commonsense Reasoning (VCR): 10k images,290k multiple choice questions and correspondingly 290k correct answers and rationales
- Visual COMmonsense rEasoning in Time (Visual COMET): contains reasoning about events in present, before and after, as well as intents

Also useful knowledge for commonsense in **SWAG**: grounded commonsense inference, unifying natural language inference and commonsense reasoning. For example: Given a partial description like "she opened the hood of the car," humans can reason about the situation and anticipate what might come next ("then, she examined the engine").

VCR results

VCR can be decomposed in three parts:

- Answering a question: Q $\longrightarrow$ A
- Providing rationale: QA $\longrightarrow$ R
- Predict both answer and rationale: Q $\longrightarrow$ AR

Motivation for knowledge in VCR

- Models cannot produce interpretable reasoning paths.
- Object intra-relationships are limited to homogeneous graphs, ignoring the cross-domain semantic alignment among visual concepts and linguistic words.

Key remarks for VCR

- Transformer models are used widely, at least for language representation
- Also abundant usage of Graph Neural Nets (GNNs)

Knowledge enhanced VCR approaches

1. [24]: Heterogeneous Graph Learning (HGL) framework for seamlessly integrating the intra-graph and inter-graph reasoning in order to bridge vision and language domain.
2. [25]: Pretrained vision–language–knowledge embedding module, which co-embeds multimodal data including images, natural language texts, and knowledge graphs into a single feature vector.
3. [26]: Hierarchical semantic enhanced directional graph network. First, aggregate the hierarchical vision-language relationships, followed by a module which dynamically selects entities in each reasoning step, according to the importance of these entities.

Knowledge enhanced VCR approaches

1. [27]: Multi-level counterfactual contrastive learning network jointly models the hierarchical visual contents and the inter-modality relationships. Instance-level, image-level, and semantic-level contrastive learning are used to extract discriminative features, wtih generated informative factual and counterfactual samples for contrastive learning.
2. [28]: KVL-BERT incorporates commonsense knowledge into the cross-modal BERT. External commonsense knowledge extracted from ConceptNet is integrated into the multi-layer Transformer. Relative position embedding and mask-self-attention to weaken the effect between the injected commonsense knowledge and other unrelated components in the input sequence.

Knowledge enhanced VCR approaches

1. [29]: Zero-shot commonsense question answering. Focuses on commonsense knowledge integration where contextually relevant knowledge is often not present in existing knowledge bases.
2. [30]: Cell-level, layer-level and attention-level joint information transfer via multi-level knowledge transfer network for effectively capturing knowledge from different perspectives

Knowledge for commonsense reasoning: Heterogeneous Graph Learning

Open question from state-of-the-art models: Main ideas of Heterogeneous Graph Learning (HGL):

- Intra-graph and inter-graph reasoning in order to bridge vision and language domain.
- Consists of a primal vision-to-answer heterogeneous graph (VAHG) module and a dual question-to-answer heterogeneous graph (QAHG) module to interactively refine reasoning paths for semantic agreement.
- Integrates a contextual voting module to exploit long-range visual context for better global reasoning

[1] Heterogeneous Graph Learning for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: Heterogeneous Graph Learning

HGL consists of two parts:

- Part 1: Build a heterogeneous graph given visual and linguistic node representations. Given a question and an image, the relevant node is located on the graph via utilize heterogeneous graph reasoning.
- Utilize the evolved heterogeneous graph to guide the answer selection.
- Part 2: Semantics with no correlation with the second modality (such as 'rainy day') are benefited thanks to a contextual voted module (CVM) for visual scene understanding with a global perspective at the low-level features.

[1] Heterogeneous Graph Learning for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: Heterogeneous Graph Learning

How HGL works:

- Taking the image, question and candidate answers with four-way multiple choices as input.
- A ResNet50 tailed with the CVM is used to obtain the visual representation V with global reasoning.
- BERT extracts question & candidate answer representations.
- The three representations (V-Q-A) form the input of heterogeneous graph module.
- A primal VAHG module with a dual QAHG module is used to construct heterogeneous graph relationship to align semantics among vision, question and answer.
- Output: two evolved representations.
- The two representations are fed into a parser and classification for the final result (right answer choice).

[1] Heterogeneous Graph Learning for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: Heterogeneous Graph Learning

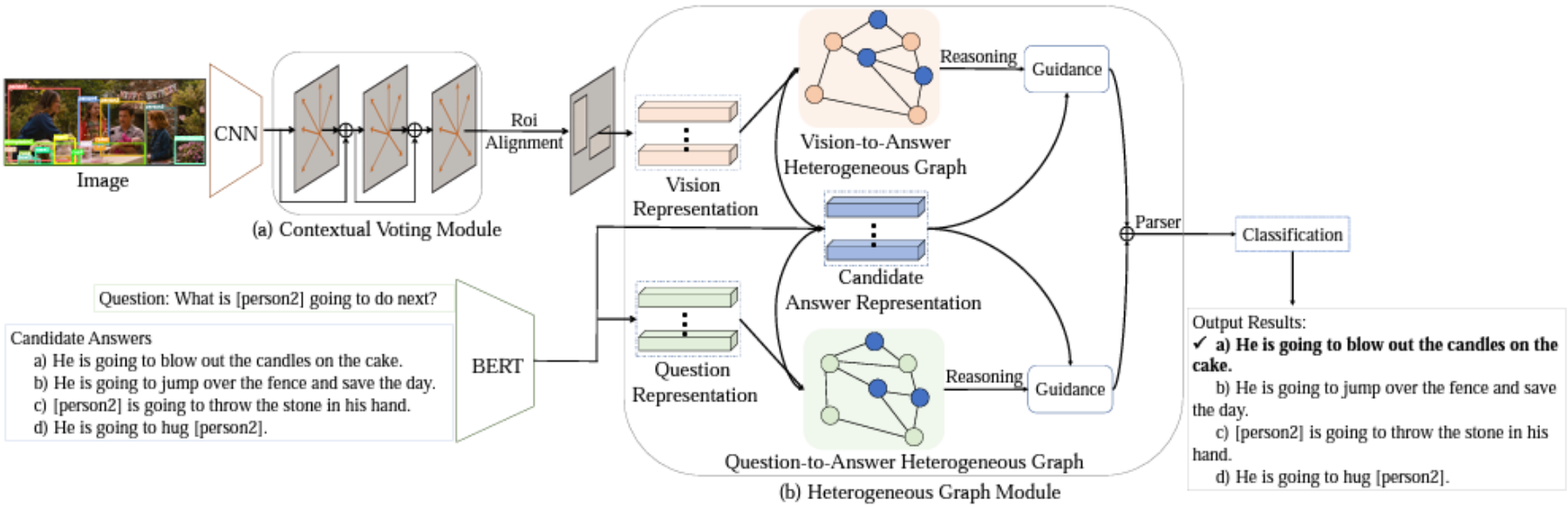

Figure 13: HGL framework

[1] Heterogeneous Graph Learning for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: Heterogeneous Graph Learning

Results (on VCR benchmark):

- Accuracy improvement to state-of-the-art: $Q \longrightarrow A$ +5%, $QA \longrightarrow R$ +3.5%, $Q \longrightarrow AR$ +5.8%.
- Still around 20% behind human perception.

[1] Heterogeneous Graph Learning for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: Multi-level

Key ideas:

1. Multi-level knowledge transfer network: injects external knowledge captured from different 'perspectives'.
2. Knowledge based reasoning approach: relates the transferred knowledge to visual content and compose the reasoning cues to derive the final answer. It contains two components:
   - Knowledge-enriched visual attention: aligns knowledge and vision, exploiting their correlations.
   - Reasoning composition: composes the aligned cues to provide the final answer from the perspective of fine-grained and global context.

This approach is not based on training a multi-modal Transformer.

[1] Multi-Level Knowledge Injecting for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: Multi-level

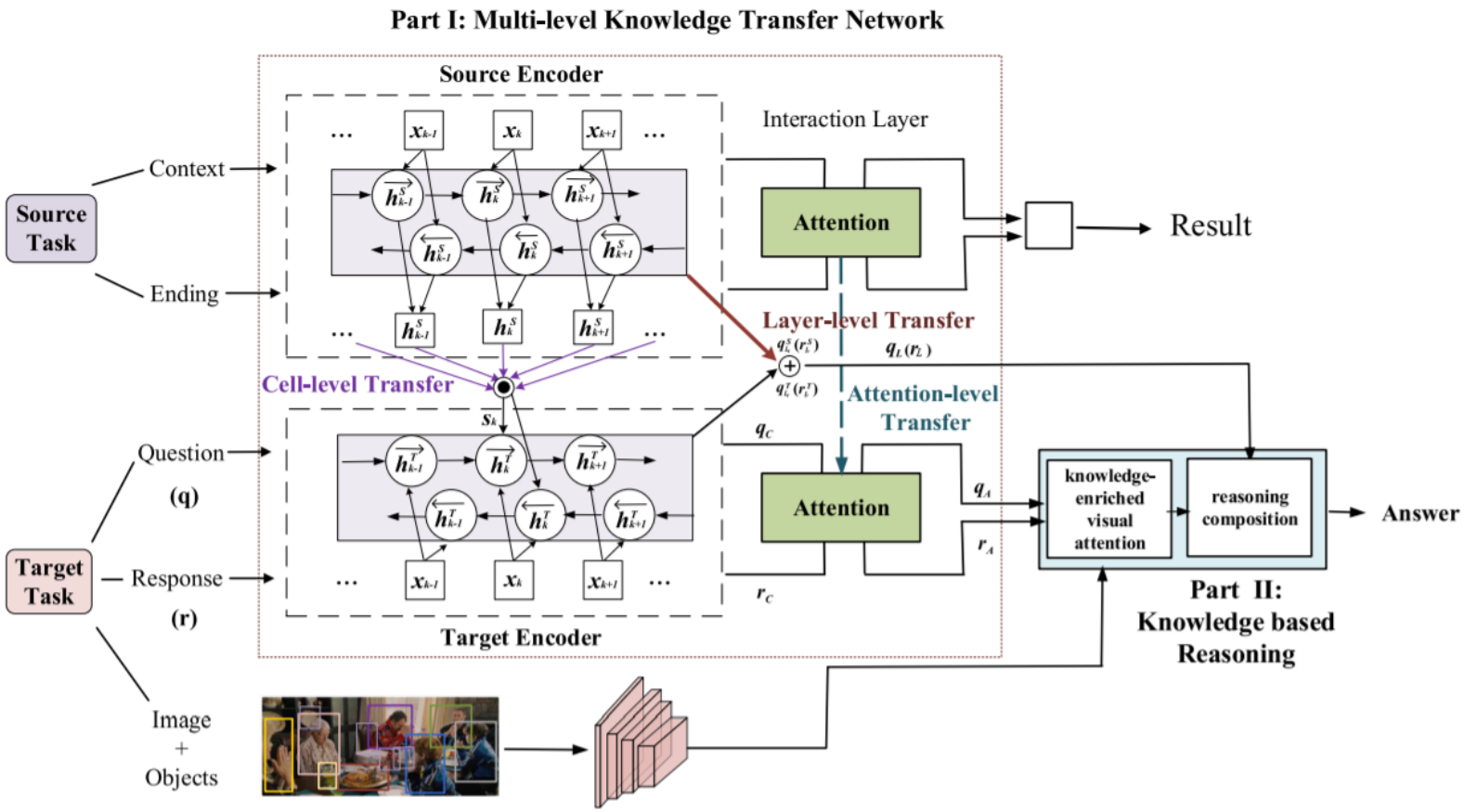

Figure 14: Knowledge Transfer Network with knowledge based reasoning

[1] Multi-Level Knowledge Injecting for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: Multi-level

Commonsense knowledge:

- Source Task: Grounded Commonsense Inference. For given context, choose ending which best describes what will happen next.
- Source Knowledge: SWAG. Contains various commonsense facts about the everyday world, such as 'drive motorcycle with a helmet'. Inputs include context c and endings V per context.

Knowledge transfer network:

- Commonsense Knowledge Learner encoder: apply BERT on SWAG data (c and V), encode using a BiLSTM.
- Interaction Layer: determine the fine-grained correlation between c and V.
- Knowledge Transfer Process: Cell-Level, Layer-Level and Attention-Level Knowledge Transfer.

[1] Multi-Level Knowledge Injecting for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: Multi-level

Results:

- Necessity of external knowledge is confirmed.
- Source domain selection is important, as this approach is based on transfer learning.
- VCR metrics are still behind human performance, this task has way to go.

[1] Multi-Level Knowledge Injecting for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: KVL-BERT

- Incorporate commonsense knowledge by injecting relevant entities extracted from ConceptNet.
- Injected knowledge should be visible only to related tokens: RMGSR (Relative-position-embedding and Mask-self-attention Guided Semantic Representations) algorithm.
- Relative position embeddings on the enriched sentence preserves readability and structure of the original sentence.
- Mask-self-attention mechanism preserves the semantic and visual representations of the original input.
- VL-BERT acts as the backbone pre-trained model: aligns input text tokens/image regions, and learns a joint VL representation.

[1] KVL-BERT:Knowledge Enhanced Visual-and-Linguistic BERT for visual commonsense reasoning

Knowledge for commonsense reasoning: KVL-BERT

VL-BERT inputs:

- Token embedding of the knowledge-enriched sentence
- Its special position embedding
- Segment embedding
- Visual feature embedding
- Visible matrix (from RMGSR)

Two extra KVL-BERT variants:

1. Inject ConceptNet entities into sentences, use same positional embedding for text & relevant knowledge extracted token.
2. Knowledge embeddings using TransE and word embeddings are inserted to VL-BERT.

[1] KVL-BERT:Knowledge Enhanced Visual-and-Linguistic BERT for visual commonsense reasoning

Knowledge for commonsense reasoning: KVL-BERT

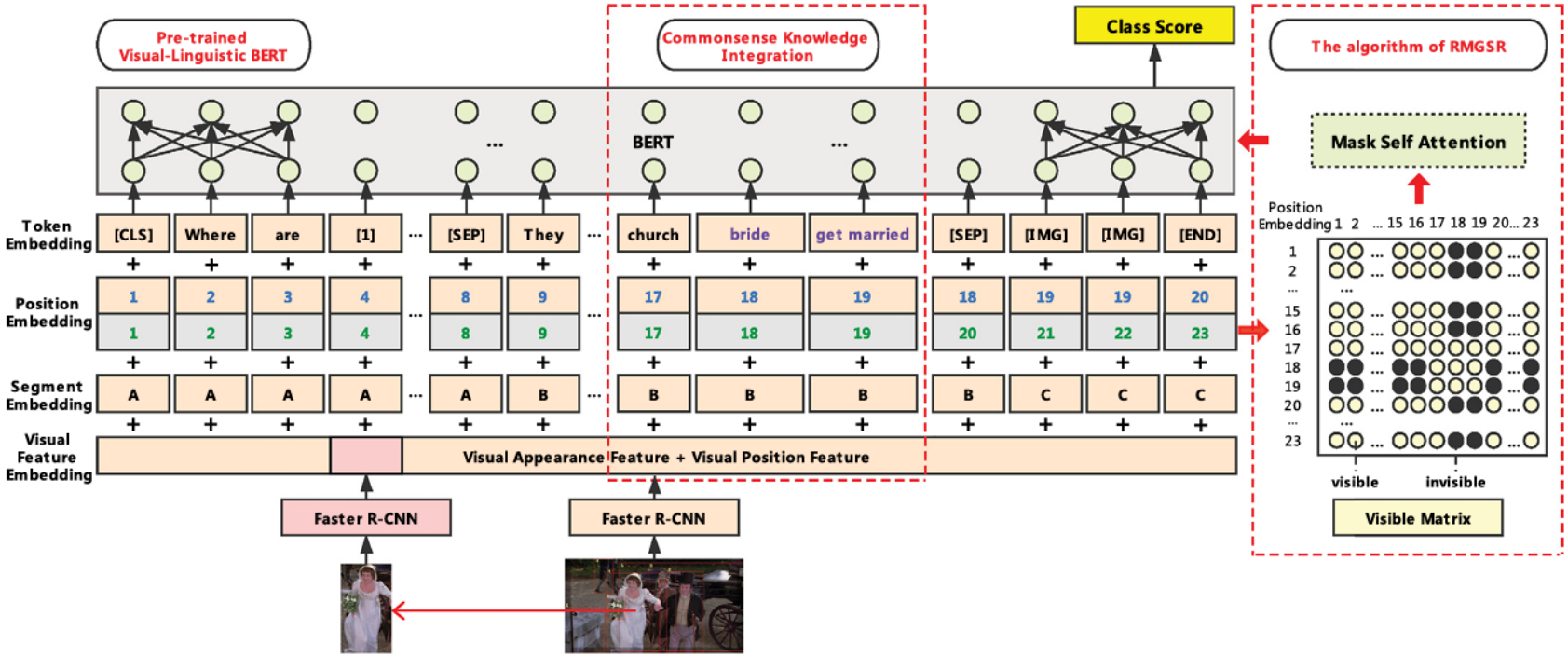

Figure 15: KVL-BERT architecture.

[1] KVL-BERT:Knowledge Enhanced Visual-and-Linguistic BERT for visual commonsense reasoning

Knowledge for commonsense reasoning: KVL-BERT

Pre-training datasets:

- Conceptual Captions (text, image)
- BooksCorpus (text)
- English Wikipedia (text)

Results on VCR benchmark val set / knowledge-related sub set comparing to VL-BERT:

- $Q \longrightarrow A$: +0.2% / +7.4%
- $Q \longrightarrow A$: +0.6% / +4.4%
- $Q \longrightarrow AR$: +0.4% / +10.4%

[1] KVL-BERT:Knowledge Enhanced Visual-and-Linguistic BERT for visual commonsense reasoning

Knowledge for commonsense reasoning: VLK co-embeddings

- Input: Image, a question in NLP and candidate answers
- Challenges: knowledge acquisition and multimodal alignment
- Proposal: Vision–Language–Knowledge Co-embedding (ViLaKC) model that extracts knowledge graphs relevant to the question from ConceptNet, and uses them together with the input image to answer the question.
- Images, natural language texts, and knowledge graphs are embedded into a single feature vector

[1] Vision–Language–Knowledge Co-Embedding for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: VLK co-embeddings

- How is knowledge embedded?
  1. Graph convolutional neural network (GCN) layer embeds the knowledge graph
  2. Multi-head self-attention layer co-embeds graph embedding with the image and natural language question
- Three modules in total:
  1. Knowledge extraction module (KEM)
  2. Vision–language (question and candidate responses)–knowledge embedding module (VLKEM)
     - **1st stage**: each modality embedded **independently**. Image with ResNet101, language with BERT, knowledge with GCN.
     - **2nd stage**: independently embedded multimodal vectors are integrated into one with a co-embedder pretrained with the VCR v1.0 dataset.
  3. Answer determination module (ADM)

[1] Vision–Language–Knowledge Co-Embedding for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: VLK co-embeddings

How knowledge is retrieved:

- Object labels $\mathcal{L}(v)$ detected in the input image and concept words extracted from the question q and candidate responses r are used as keywords for pre-fetching and storing relevant knowledge k.
- Top-50 relevant retrieved triplets are converted in sentences and their cosine similarity with the question q is calculated.

[1] Vision–Language–Knowledge Co-Embedding for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: VLK co-embeddings

Pre-training in two-stages:

1. Task agnostic stage on more than 1.3 million data:
   V+L with pre-training tasks:
   - Masked Language Modelling (MLM) with image (MLMI)
   - Masked Object Classification (MOC) with text (MOCT)
   - Image-Text Matching (ITM)
2. Task-specific stage on 200k data:
   additional knowledge from VCR v1.0 dataset and ConceptNet with pre-training tasks:
   - Masked Language Modelling (MLM) with image and knowledge (MLMIK)
   - Masked Object Classification (MOC) with text and knowledge (MOCTK)
   - Image-Text-Knowledge Matching (ITKM)

[1] Vision–Language–Knowledge Co-Embedding for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: VLK co-embeddings

Implementation and results:

- Evaluation on the VCR v1.0 dataset.
- GCN is the most effective method for graph embeddings.
- Three co-embedding experiments:
  1. Early fusion (EF): Concat knowledge with vision (f_{vk}) and with text (f_{lk}) and then co-embed those 2 vectors.
  2. Late fusion (LF): Co-embed vision (f_v) and language (f_l) in a vector f_{vl} and then concat with knowledge.
  3. Concurrent fusion (CF): Co-embed vision (f_v), text (f_l) and knowledge (f_k) together.
- CF performs better.
- Performance improvement over SOTA V+L: 0.8% for the Q $\longrightarrow$ A, 0.9% for the QA $\longrightarrow$ R, 1.3% for the Q $\longrightarrow$ AR.

[1] Vision–Language–Knowledge Co-Embedding for Visual Commonsense Reasoning

Knowledge for commonsense reasoning: KM-BART

Knowledge Enhanced Multimodal BART (KM-BART):

- Sequence-to-sequence Transformer model.
- Generative BART architecture adapts to a multimodal model with visual and textual inputs.
- Getting from understanding tasks to multimodal generation tasks.
- Visual Commonsense Generation (VCG) task: generate commonsense inferences about what might happen before/after, and the present intents of characters.

[1] KM-BART:Knowledge Enhanced Multimodal BART for Visual Commonsense Generation

Knowledge for commonsense reasoning: KM-BART

KM-BART pre-training:

- Introduces novel pretraining tasks: Knowledge-based Commonsense Generation (KCG). Leverages commonsense knowledge from a large language model pretrained on external commonsense knowledge graphs.
- Other pre-training tasks:
  - Masked Language Modeling (MLM)
  - Masked Region Modeling (MRM)
  - Attribution Prediction (AP)
  - Relation Prediction (RP)
- Pre-training Datasets: Conceptual Captions, SBU, Microsoft COCO and Visual Genome.

[1] KM-BART:Knowledge Enhanced Multimodal BART for Visual Commonsense Generation

Knowledge for commonsense reasoning: KM-BART

How KM-BART works:

- Adapt NLP BART to cross-modal inputs (V+L).
- Visual Feature Extractor: pretrained Masked R-CNN for visual embeddings.
- Cross-Modal Encoder: based on a multi-layer bidirectional Transformer, special tokens introduced to handle to multimodal tasks.
- Special tokens:
  - $< before >, < after >, or < intent >$: Knowledge-Based Commonsense Generation.
  - $< region\ caption >$: Attribution Prediction and Relation Prediction.
  - $< caption >$: Masked Language Modeling and Masked Region Modeling.

[1] KM-BART:Knowledge Enhanced Multimodal BART for Visual Commonsense Generation

Knowledge for commonsense reasoning: KM-BART

How KM-BART works:

- Decoder: unidirectional (autoregressive generative decoder).
- Knowledge-Based Commonsense Generation: knowledge from COMET, a large language model pretrained on external commonsense knowledge graphs. Given a natural language phrase and a relation as inputs, COMET generates natural language phrases as commonsense descriptions.
- Self-Training Based Data Filtering: examples in the COMET-based filtered dataset closely resemble the examples in the VCG dataset.

[1] KM-BART:Knowledge Enhanced Multimodal BART for Visual Commonsense Generation

Knowledge for commonsense reasoning: KM-BART

Event and image	Task	Model	Generated Sentence
2 is holding an envelope 	intent	without event§	**give 1 some bad news** **reassure 1** **contemplate what 1 is saying to her**
		with event†	**see what the letter said** **give mail to 1** **open the envelope**
		ground truth	receive the envelope from 1 see what's inside the envelope
	before	without event§	**walk up to 1** **have seen 1 in the distance** **be interested in what 1 has to say**
		with event†	**pick the envelope up** **call 1 to meet him** **walk to 1**
		ground truth	receive mail be given an envelope bring the envelope with her
	after	without event§	**finish telling 1 she has a difficult time** **ask 1 what the papers are for** **let go of 1**
		with event†	**open the envelope** **hand the envelope to 1** **embrace 1**
		ground truth	read the contents of the envelope to 1 hand the envelope to 1 read the love letter

Figure 16: VCG dataset example with inference sentences from KM-BART

[1] KM-BART:Knowledge Enhanced Multimodal BART for Visual

Knowledge for commonsense reasoning: KM-BART

	Modalities	Event	BLEU-2	METEOR	CIDER	Unique	Novel
Park et al. (2020)[a*]	Image+Event+ Place+Person	N	**10.21**	**10.66**	**11.86**	*33.90*	*49.84*
Park et al. (2020)[b*]	Image	N	6.79	7.13	5.63	26.38	46.80
Ours[§]	Image	N	*9.04*	*8.33*	*9.12*	**50.75**	**52.92**
Park et al. (2020)[c*]	Image+Event+ Place+Person	Y	*13.50*	**11.55**	*18.27*	*44.49*	*49.03*
Park et al. (2020)[d*]	Image+Event	Y	12.52	10.73	16.49	42.83	47.40
Ours[†]	Image+Event	Y	**14.21**	*11.19*	**21.23**	**57.64**	**58.22**

Figure 17: Results on VCG validation set against SOTA

[1] KM-BART:Knowledge Enhanced Multimodal BART for Visual Commonsense Generation

Knowledge for commonsense reasoning: KM-BART

Models	Event	Before	After	Intent	Total
Park et al. (2020)[c*]	N	38.7	31.3	30.7	33.3
Ours[§]	N	**61.3**	**68.7**	**69.3**	**66.7**
Park et al. (2020)[c*]	Y	48.0	48.0	38.7	44.9
Ours[†]	Y	**52.0**	**52.0**	**61.3**	**55.1**

Figure 18: Human evaluation results from KM-BART

[1] KM-BART:Knowledge Enhanced Multimodal BART for Visual Commonsense Generation

Image captioning - Intro

Similar to the taxonomy of **VQA**, image captioning follows either the RNN-based language *generation* idea, or else the transformer-based language *generation*.

Image captioning

1. [31]: Feed external knowledge (ConceptNet) to RNN structure and produce caption via LSTMs. Knowledge entities are extracted with the help of detected objects in image.
2. [32]: Concepts absent from the image can be inferred from the visual information and external knowledge via commonsense reasoning. A semantic graph is derived from the learned features with the guidance of the external knowledge graph. A GCN reasons over the graph, producing a Relation-aware graph, from which the caption is extracted. The captioning result refines the knowledge graph in turn.

Image captioning

1. [33]: In order to reveal incomprehensible intentions that cannot be expressed straightforwardly by machines, external knowledge extracted from knowledge graph is injected into the encoder-decoder framework, producing novel and meaningful captions. Word attention indicates the most important words that should be included in a caption.
2. [34]: Iterative learning algorithm that alternates between 1) commonsense reasoning for embedding visual regions into the semantic space to build a semantic graph and 2) relation reasoning for encoding semantic graphs to generate sentences. This paper focuses on reasoning of object relationships, mostly relying on Visual Genome graphs.

Image captioning meets commonsense reasoning

1. [35]: Visual Commonsense generation (VCG): between image captioning and VCR. KM-BART reasons about commonsense knowledge from multimodal inputs of images and texts. A new pretraining task of Knowledge-based Commonsense Generation (KCG) boosts model performance by leveraging commonsense knowledge from a large language model pretrained on external commonsense knowledge graphs.

Visual Commonsense generation

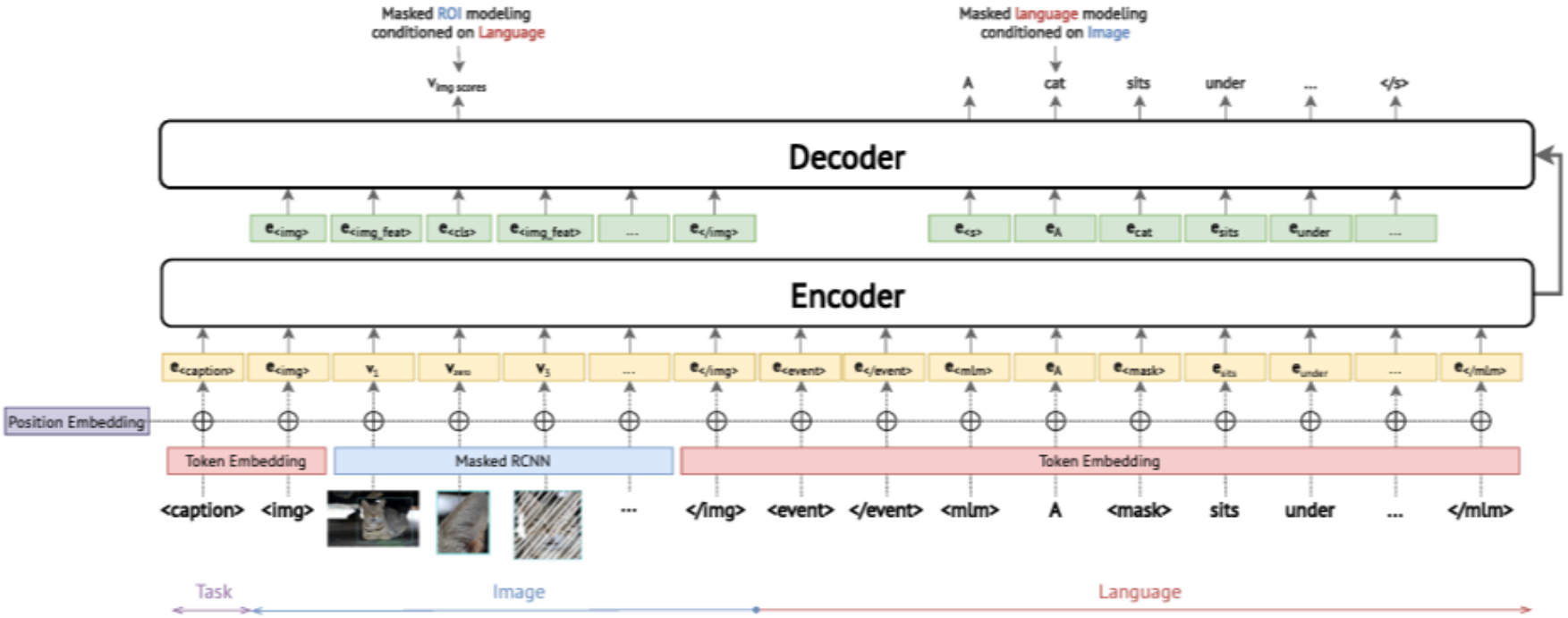

Figure 19: KM-BART architecture

[1] KM-BART:Knowledge Enhanced Multimodal BART for Visual Commonsense Generation

Visual Storytelling

- Create textual description from image
- Instead of describing a single image, describe a sequence of them
- Narrative summary based on textual descriptions
- Different from visual captions, stories contain not only factual descriptions, but also imaginary concepts that do not appear in the images.

Visual Storytelling

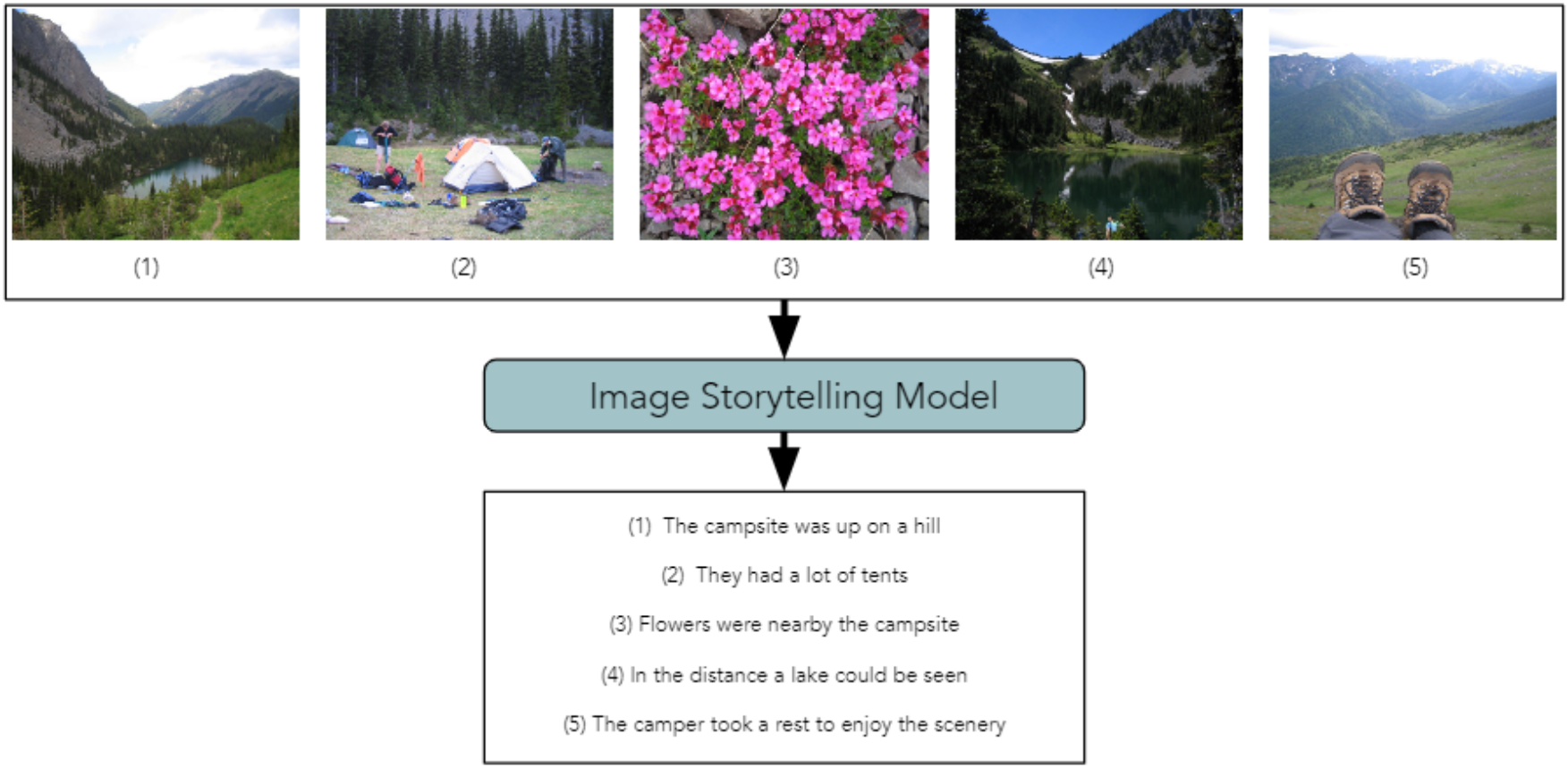

Figure 20: Example of the VS task: Sequence of images with generated coherent textual descriptions.

Knowledge in visual storytelling

1. [36]: Most relevant knowledge KGs from a knowledge base are extracted for commonsense integration. Extracted image features per frame are encoded via a GRU and matched with knowledge graphs to produce a sequence of captions.
2. [37]: KG-Story tackles the observed monotonous and limited in vocabulary generated stories. KG-Story distills a set of representative words from the input prompts, enriches the word set by using external knowledge graphs, and finally generates stories based on the enriched word set. A Transformer-GRU is used as the term predictor.

Knowledge in visual storytelling

1. [33]: First, a multimodal imagining module is leveraged to learn the imaginative storyline explicitly, improving the coherence and reasonability of the generated story. Second, a relational reasoning module is employed to fully exploit the external knowledge (commonsense knowledge base) and task-specific knowledge (scene graph and event graph) with a relational reasoning method based on the storyline.
2. [38]: Goal: increasing the diversity of the generated stories while preserving the informative content from the images. A concept selection module that suggests a set of concept candidates is supposed to boost diversity. Then, a large scale pre-trained model is utilized to convert concepts and images into full stories. To enrich the candidate concepts, a commonsense knowledge graph is created for each image sequence from which the concept candidates are proposed.

Story Visualization

- A Language-to-Image Generation task - Inverse visual storytelling
- Given a textual 'story' description, corresponding images should be generated
- Images should be coherent, based on the previous ones in the sequence

Story Visualization with knowledge

- [39]: Knowledge learned from Visual Genome can improve the spatial structure of images from a different target domain without needing fine-tuning.
  Captions do not explicitly mention some semantics: the caption 'two people are standing outside in a sunny day' does not explicitly mention 'sky', but it should be represented visually, guided by commonsense knowledge (here from ConceptNet).

Knowledge Enhanced Vision-Language Representations

- Challenge: Lack of detailed semantic alignment prevents V+L models from learning fine-grained information of real scenes.
- ERNIE-VL: Structured knowledge from scene graphs.
- Builds detailed semantic connections (objects, attributes, relationships) across vision and language, aligning the two modalities.
- Scene graph prediction tasks: predict nodes of different types in the scene graph parsed from the sentence.
- Instead of masking out random words, mask corresponding scene graph nodes, so that the model understands semantically significant information.

[1] ERNIE-ViL: Knowledge Enhanced Vision-Language Representations through Scene Graphs

Knowledge Enhanced Vision-Language Representations

- Input: Sentence and an image. Those are embedded:
  - Sentence embeddings with BERT.
  - Region visual features from object detector summed with location features from bounding box info.
  - V+L two-stream encoder for joint representation.
- Pre-training datasets:
  - Out-of-domain: Conceptual Captions & SBU Captions.
  - In-domain: MS-COCO & Visual Genome.
- Task specific datasets:
  - Visual Question Answering (VQA 2.0)
  - Visual Commonsense Reasoning (VCR)
  - Region-to-Phrase Grounding (RefCOCO+)
  - Image-text Retrieval / Text-image Retrieval (Flickr30K)

[1] ERNIE-ViL: Knowledge Enhanced Vision-Language Representations through Scene Graphs

Knowledge Enhanced Vision-Language Representations

- Knowledge obtained via scene graphs.
  - Scene graph parsed from text.
  - Scene graph prediction tasks: object/attribute/relationship prediction
  - Then, the model can learn correct semantics when masked word or masked object prediction is ambiguous.
- Pre-training tasks:
  - Masked Object Prediction (Graph+Language)
  - Masked Relationship Prediction (Graph+Language)
  - Masked Attribute Prediction (Graph+Language)
  - Masked Language Modelling (Language)
  - Masked Region Prediction (Vision)
  - Image-text Matching (Vision+Language)

[1] ERNIE-ViL: Knowledge Enhanced Vision-Language Representations through Scene Graphs

Knowledge Enhanced Vision-Language Representations

Results for ERNIE-VL with comparison to BERT:

initialized text stream parameters	pre-training tasks	VCR Q→AR (dev)	VQA dev	RefCOCO+ val	IR R@1 (dev)	TR R@1 (dev)
BERT	w/o **SGP**	59.06	72.38	72.81	70.74	85.00
BERT	w/ **SGP**	59.92	73.04	73.50	72.96	87.40
ERNIE-2.0	w/ **SGP**	61.24	73.18	74.02	73.58	87.80

Figure 21: Results of downstream vision-language tasks for ERNIE-ViL pre-trainging with/without Scene Graph Prediction (SGP) tasks

[1] ERNIE-ViL: Knowledge Enhanced Vision-Language Representations through Scene Graphs

Resources

- Recent Advances in Vision and Language Pre-trained Models (VL-PTMs) github repository: https://github.com/yuewang-cuhk/awesome-vision-language-pretraining-papers
- Awesome Vision-and-Language github repository: https://github.com vision-and-language

[1] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei.
Building a large-scale multimodal knowledge base system for answering visual queries, 2015.

[2] Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton 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), pages 4622–4630, 2016.

[3] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel.
Explicit knowledge-based reasoning for visual question answering.
In *IJCAI*, 2017.

[4] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel.
Fvqa: Fact-based visual question answering.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 40:2413–2427, 2018.

[5] Medhini Narasimhan and Alexander G. Schwing.
Straight to the facts: Learning knowledge base retrieval for factual visual question answering.
ArXiv, abs/1809.01124, 2018.

[6] Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing.
Out of the box: Reasoning with graph convolution nets for factual visual question answering, 2018.

[7] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar.
Kvqa: Knowledge-aware visual question answering.
In *AAAI*, 2019.

[8] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty.
From strings to things: Knowledge-enabled vqa model that can read and reason.
2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 4601–4611, 2019.

[9] J. Yu, Zihao Zhu, Yujing Wang, Weifeng Zhang, Yue Hu, and Jianlong Tan.
Cross-modal knowledge reasoning for knowledge-based visual question answering.
ArXiv, abs/2009.00145, 2020.

[10] Zihao Zhu, J. Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu.
Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering.
In *IJCAI*, 2020.

[11] Guohao Li, Xin Wang, and Wenwu Zhu.
Boosting visual question answering with context-aware knowledge aggregation.
Proceedings of the 28th ACM International Conference on Multimedia, 2020.

[12] Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z. Pan, Zonggang Yuan, and Huajun Chen.
Zero-shot visual question answering using knowledge graph.
In *SEMWEB*, 2021.

[13] Maryam Ziaeefard and Freddy Lécué.
Towards knowledge-augmented visual question answering.
In *COLING*, 2020.

[14] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang.
An empirical study of gpt-3 for few-shot knowledge-based vqa.
ArXiv, abs/2109.05014, 2021.

[15] Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa Etxabe, and Eneko Agirre.
Image captioning for effective use of language models in knowledge-based visual question answering.
ArXiv, abs/2109.08029, 2021.

[16] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lécué.
Conceptbert: Concept-aware representation for visual question answering.
In *FINDINGS*, 2020.

[17] Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral.
Weakly-supervised visual-retriever-reader for knowledge-based question answering.
In *EMNLP*, 2021.

[18] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Kumar Gupta, and Marcus Rohrbach.
Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa.
2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14106–14116, 2021.

[19] Arka Ujjal Dey, Ernest Valveny, and Gaurav Harit.
External knowledge enabled text visual question answering.
2021.

[20] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi.
Multi-modal answer validation for knowledge-based vqa.
ArXiv, abs/2103.12248, 2021.

[21] Chen Qu, Hamed Zamani, Liu Yang, William Bruce Croft, and Erik 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.

[22] Diego Garcia-Olano, Yasumasa Onoe, and Joydeep Ghosh.
Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection.
2021.

[23] Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti.
Select, substitute, search: A new benchmark for knowledge-augmented visual question answering.
Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021.

[24] Weijiang Yu, Jingwen Zhou, Weihao Yu, Xiaodan Liang, and Nong Xiao.
Heterogeneous graph learning for visual commonsense reasoning.
In *NeurIPS*, 2019.

[25] JaeYun Lee and Incheol Kim.
Visual commonsense reasoning with vision-language co-embedding and knowledge graph embedding.
2020.

[26] Mingyan Wu, Shuhan Qi, Jun Rao, Jiajia Zhang, Qing Liao, Xuan Wang, and Xinxin Liao.
Hierarchical semantic enhanced directional graph network for visual commonsense reasoning.
Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing, 2021.

[27] Xi Zhang, Feifei Zhang, and Changsheng Xu.
Multi-level counterfactual contrast for visual commonsense reasoning.
Proceedings of the 29th ACM International Conference on Multimedia, 2021.

[28] Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao.
Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning.
Knowl. Based Syst., 230:107408, 2021.

[29] Antoine Bosselut, Ronan Le Bras, and Yejin Choi.
Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering.
In *AAAI*, 2021.

[30] Zhang Wen and Yuxin Peng.
Multi-level knowledge injecting for visual commonsense reasoning.
IEEE Transactions on Circuits and Systems for Video Technology, 31:1042–1054, 2021.

[31] Yimin Zhou, Yiwei Sun, and Vasant G Honavar.
Improving image captioning by leveraging knowledge graphs.
2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 283–293, 2019.

[32] Jingyi Hou, Xinxiao Wu, Yayun Qi, Wentian Zhao, Jiebo Luo, and Yunde Jia.
Relational reasoning using prior knowledge for visual captioning.
ArXiv, abs/1906.01290, 2019.

[33] Feicheng Huang, Zhixin Li, Shengjia Chen, Canlong Zhang, and Huifang Ma.
Image captioning with internal and external knowledge.
Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020.

[34] Jingyi Hou, Xinxiao Wu, Xiaoxun Zhang, Yayun Qi, Yunde Jia, and Jiebo Luo.
Joint commonsense and relation reasoning for image and video captioning.

In *AAAI*, 2020.

[35] Yiran Xing, Z. Shi, Zhao Meng, Yunpu Ma, and Roger Wattenhofer.
Km-bart: Knowledge enhanced multimodal bart for visual commonsense generation.
In *ACL/IJCNLP*, 2021.

[36] Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu 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*, pages 5356–5362. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[37] Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao 'Kenneth' Huang, and Lun-Wei Ku.
Knowledge-enriched visual storytelling, 2019.

[38] Arushi Goel, Basura Fernando, Thanh-Son Nguyen, and Hakan Bilen.
Injecting prior knowledge into image caption generation.
In *ECCV Workshops*, 2020.

[39] Adyasha Maharana and Mohit Bansal.
Integrating visuospatial, linguistic and commonsense structure into story visualization, 2021.

[40] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha.
Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions, 2021.

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

[42] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow.
Trends in integration of vision and language research: A survey of tasks, datasets, and methods.
Journal of Artificial Intelligence Research, 71:1183–1317, Aug 2021.

[43] Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick.
Visual referring expression recognition: What do systems actually learn?, 2018.

[44] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel.
Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, 2018.

[45] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi.
From recognition to cognition: Visual commonsense reasoning, 2019.

[46] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang.
Visualbert: A simple and performant baseline for vision and language, 2019.

[47] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai.
Vl-bert: Pre-training of generic visual-linguistic representations, 2020.

[48] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu.
Uniter: Universal image-text representation learning, 2020.

[49] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou.
Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training, 2019.

[50] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao.
Unified vision-language pre-training for image captioning and vqa, 2019.

[51] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao.
Oscar: Object-semantics aligned pre-training for vision-language tasks, 2020.

[52] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee.
Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

[53] Hao Tan and Mohit Bansal.
Lxmert: Learning cross-modality encoder representations from transformers, 2019.

[54] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid.
Videobert: A joint model for video and language representation learning, 2019.

[55] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter.
Fusion of detected objects in text for visual question answering, 2019.

[56] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu.
Pixel-bert: Aligning image pixels with text by deep multi-modal transformers, 2020.

[57] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang.
Interbert: Vision-and-language interaction for multi-modal pretraining, 2021.

[58] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei.
Vlmo: Unified vision-language pre-training with mixture-of-modality-experts, 2021.

[59] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee.
12-in-1: Multi-task vision and language representation learning, 2020.

[60] Wonjae Kim, Bokyung Son, and Ildoo Kim.
Vilt: Vision-and-language transformer without convolution or region supervision, 2021.

[61] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi.
X-lxmert: Paint, caption and answer questions with multi-modal transformers, 2020.

[62] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao.
Vinvl: Revisiting visual representations in vision-language models, 2021.

[63] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang.
Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning, 2021.

[64] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu.
Seeing out of the box: End-to-end pre-training for vision-language representation learning, 2021.

[65] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao.
Simvlm: Simple visual language model pretraining with weak supervision, 2021.

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.
Learning transferable visual models from natural language supervision, 2021.

[67] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.
Zero-shot text-to-image generation, 2021.

[68] Amanpreet Singh, Vedanuj Goswami, and Devi Parikh.
Are we pretraining it right? digging deeper into visio-linguistic pretraining, 2020.

[69] JaeYun Lee and Incheol Kim.
Vision-language-knowledge co-embedding for visual commonsense reasoning.
Sensors, 21(9), 2021.

[70] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum.
Neural-symbolic vqa: Disentangling reasoning from vision and language understanding, 2019.

[71] Noa García and Yuta Nakashima.
Knowledge-based video question answering with unsupervised scene descriptions.
ArXiv, abs/2007.08751, 2020.

[72] JaeYun Lee and Incheol Kim.
Vision–language–knowledge co-embedding for visual commonsense reasoning.
Sensors (Basel, Switzerland), 21, 2021.

[73] Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang.
Km4: Visual reasoning via knowledge embedding memory model with mutual modulation.
Inf. Fusion, 67:14–28, 2021.

[74] Chunpu Xu, Min Yang, Chengming Li, Ying Shen, Xiang Ao, and Ruifeng Xu.
Imagine, reason and write: Visual storytelling with graph knowledge and relational reasoning.
In *AAAI*, 2021.

[75] Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama.
Commonsense knowledge aware concept selection for diverse and informative visual storytelling.
In *AAAI*, 2021.

[76] Integrating rule-based entity masking into image captioning.
2020.

[77] Feicheng Huang, Zhixin Li, Haiyang Wei, Canlong Zhang, and Huifang Ma.
Boost image captioning with knowledge reasoning.
ArXiv, abs/2011.00927, 2020.