Visual Commonsense Generation & its incorporation into a Multimodal Topic Modeling algorithm

Felipe González-Pizarro felipegp@cs.ubc.ca
Sahithya Ravi sahiravi@cs.ubc.ca
Aditya Chinchure aditya10@cs.ubc.ca

Abstract

The task of commonsense knowledge generation is largely limited to the language domain, with models such as COMET (for explicit knowledge) and GPT-3 (for implicit knowledge). Moreover, VisualCOMET, a commonsense generation model that utilizes the visual context, is limited to three people-centric relations. Since commonsense generation on entire scenes, or parts of a scene, can be helpful in several downstream multimodal tasks, including visual question answering, story-telling, and topic modeling, we propose a general-purpose visual commonsense generation model, VisualCOMET+, by extending VisualCOMET with four diverse inference relations. Using the clue-rationale pairs from a visual abductive reasoning dataset, we successfully train our commonsense generation model by creating ground-truth structured commonsense triplets. Then, we show that we can get coherent and more diverse topics by incorporating generated commonsense inferences and visual features into a novel multimodal topic modeling algorithm, Multimodal CTM.

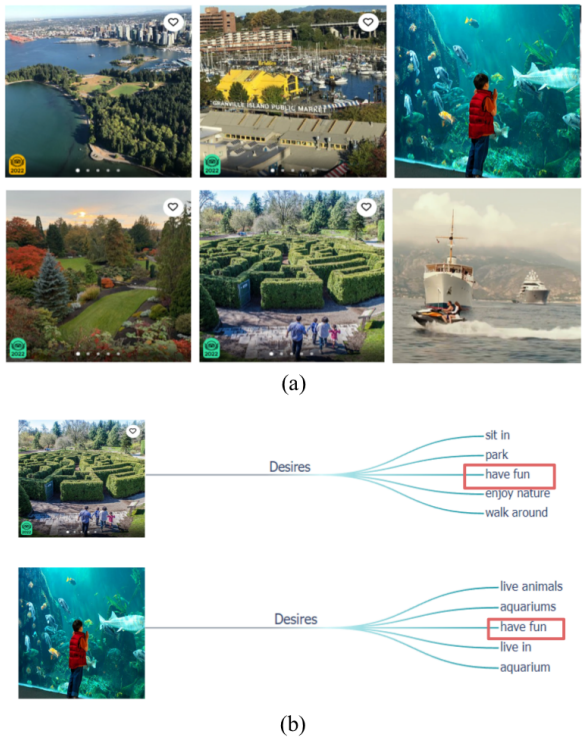


Figure 1: An example of Topic Modeling that requires commonsense reasoning.

1 Introduction

Visual commonsense generation is a recent area of research that is fundamental to many real-world tasks. Visual reasoning in humans is driven not only by the visual cues we observe but also by imagining the context accompanying those cues and our commonsense knowledge and reasoning abilities. For example, given the images in Figure 1(a) and the question: "What is common among all these images?", we can use our commonsense knowledge to deduce that these are all images of fun activities in a tourist destination. Therefore these images belong to the common topic of "tourist attractions". Incorporating such commonsense knowledge into machine learning models, as shown in Figure 1(b), is essential for developing models that can not only recognize visual cues but also reason about them.

Recent work in commonsense generation is primarily focused on the natural language domain. Large-scale knowledge bases such as ATOMIC (Sap et al., 2019) and ConceptNet (Speer et al., 2017a) contain commonsense knowledge as a graph of nodes representing objects or entities connected by relation edges. The popularity of the Transformer model has further led to the development of COMET (Hwang et al., 2021), a commonsense transformer model trained on such knowledge graphs. Furthermore, recent advancements with large-scale models such as GPT-3 (Brown et al., 2020) have brought commonsense generation capabilities but are largely inaccessible due to their size and cost. In the visual domain, VisualCOMET (Park et al., 2020) introduced a model

that attempts to predict what happens *before*, *after*, and *intent* of a person in a movie frame. This is quite limited, as the model does not reason about the overall scene, and therefore cannot be applied to many downstream tasks like VQA.

Contributions: Motivated by previous limitations of natural language processing techniques in the visual context, our work makes the following contributions:

- **VisualCOMET+:** A model that can generate commonsense inferences on provided image + textual cues. We extend the existing VisualCOMET model by introducing diverse relations that go beyond people-centric knowledge, such as *HasProperty*, *HasContext*, *indicates*, and *AtLocation*. We hypothesize that this model will be useful for multimodal downstream applications such as topic modeling, story-telling, and VQA, and dialog.
- **Multimodal CTM**: We introduce a multimodal topic modeling algorithm that takes as input texts, images, and inferences from VisualCOMET+. We show that incorporating image features and VisualCOMET+ inferences allows us to obtain a better representation of the input documents and identify coherent and more diverse topics. To the best of our knowledge, this is the first multimodal neural topic modeling algorithm.

2 Related Work

This section provides background information on commonsense generation, topic modeling algorithms, and the incorporation of commonsense knowledge into topic models.

2.1 Commonsense generation

Reasoning about events and entities has long been of interest to AI research. In the field of NLP, structured large-scale knowledge bases (KBs) like ConceptNet (Speer et al., 2017a) and ATOMIC (Sap et al., 2019) are widely used to provide additional commonsense knowledge to models. ConceptNet contains 3.4 million assertions focusing on concepts and their taxonomic and lexical relations (e.g., *RelatedTo*, *Synonym*, *IsA*), and physical commonsense knowledge (e.g., *MadeOf*). ATOMIC, on the other hand, contains 880,000 triplets focusing on event-centric social commonsense about *causes*, *effects*.

However, incorporating knowledge directly from KBs suffers from two limitations: lack of coverage and consideration for context. A commonsense Transformer, COMET (Hwang et al., 2021), attempts to alleviate these issues by fine-tuning pre-trained language models on KBs. COMET can generate contextualized commonsense inferences dynamically and generalize to unseen inputs. COMET has been successfully used for generating knowledge in language tasks (Majumder et al., 2020; Tian et al., 2021; Chakrabarty et al., 2022; Shwartz et al., 2020).

Several variants of COMET have subsequently been released. The most relevant to our work is VisualCOMET (Park et al., 2020), which generates temporal inferences for causes and effects of the events in an image. We believe that the event-specific nature of this model makes it less applicable to datasets that require knowledge about entities rather than events. Another recent work, KM-BART (Xing et al., 2021), proposes novel pre-training regimes for visual commonsense generation, but is also limited to event-specific knowledge. In this work, we propose an extension to VisualCOMET, VisualCOMET+, that supports additional relations to reason beyond people and events, and obtain more general-purpose, diverse inferences.

2.2 Topic modeling

The creation of vast amounts of data has led to the development of various techniques designed to summarize and understand textual data (Peter et al., 2015). A well-known method is topic modeling, a robust approach for extracting core themes or *topics* from large collections of documents. When a topic modeling is applied to a corpus of documents (e.g., a collection of news articles), the output will include a list of topics (e.g., “politics”, “economics”, “sports”). Usually, each topic is represented by a collection of terms that make sense together (e.g., {“tropical”, “storm”, “hurricane”, “cyclone”, “weather”, and “rain”}) (Zhao et al., 2021).

From a practical standpoint, topic modeling can be viewed as an extreme form of multi-document summarization, where it can be used to understand the underlying general themes presented in a large collection of documents (Blei et al., 2010; Boyd-Graber et al., 2017). However, studies have shown that the output of topic models do not always accurately represent the characteristics of the analyzed

document collections (El-Assady et al., 2019) or make sense to the end users (Hoque and Carenini, 2015). Part of this problem is because most topic modeling approaches focus on the co-occurrence of terms as the primary signal to detect the topical structure among them (Harrando and Troncy, 2021). As a result, these methods do not capture semantic and pragmatic relations between terms in the corpus (Harrando and Troncy, 2021; Song et al., 2020; Hong et al., 2020).

Prior work has suggested using external knowledge to overcome this drawback (Hong et al., 2020), and commonsense knowledge (i.e., relations between concepts) is one promising alternative (Harrando and Troncy, 2021). In this work, we explore how commonsense knowledge could improve the performance of a popular and well-known neural topic model.

2.3 Topic Modeling and Commonsense

The developments of deep neural networks has led to the development of several neural topic models (NTMs) to address probabilistic topic model limitations in terms of performance, efficiency, and usability (Zhao et al., 2021). One of the most popular neural topic modeling algorithms is *Contextualized Topic Models* (CTM) (Bianchi et al., 2021a,b), which uses external word representation (e.g., SBERT (Reimers and Gurevych, 2019)) to get more coherent topic than previous popular approaches (e.g., ProdLDA (Srivastava and Sutton, 2017), and LDA (Blei et al., 2003)).

Topic modeling algorithms such as *CTM* (Bianchi et al., 2021a) show that adding contextual information to neural topic models significantly improves the resulting topics' coherence. Recent work has explored expanding these contextualized representations by injecting external knowledge, such as commonsense knowledge (see (Bosselut et al., 2019)), to improve their performance. Injecting commonsense knowledge into topic modeling algorithms might help to obtain a more semantically meaningful representation of the input document (Shah et al., 2021) and, therefore, topics more aligned with commonsense relations.

Commonsense knowledge has already been used for different tasks such as question answering (Bauer et al., 2018), sentiment analysis (Ghosal et al., 2020; Ravi et al., 2021), and dialogue (Young et al., 2018). However, only a few attempts to incorporate commonsense knowledge into topic modeling algorithms exist (Rajagopal et al., 2013; Harrando and Troncy, 2021). One of these approaches is the *Commonsense Topic Model (CSTM)* (Harrando and Troncy, 2021). This recently proposed topic modeling technique augments clustering with knowledge extracted from ConceptNet (Speer et al., 2017b) to find more interpretable topics by humans. After evaluating this approach on several datasets, the authors claim their proposal generally finds more coherent topics than the traditional LDA.

Considering these promising results, in this project, we explore if we find more coherent and diverse topics by injecting contextual commonsense inferences based on the image, into a neural topic modeling algorithm. To the best of our knowledge, we are the first to inject commonsense knowledge in a multimodal setting for a neural topic modeling algorithm.

3 Method

Figure 2 shows the architectural overview of our proposed approach. Given an image, we first generate commonsense inferences relevant to the image using VisualCOMET+. We feed these inferences, along with text and image features of documents, into our proposed Multimodal CTM. We expect to find coherent and more diverse topics. We describe VisualCOMET+ in Section 3.1, how we generate visual commonsense inferences in Section 3.1.1, and our multimodal topic modeling algorithm (Multimodal CTM) in Section 3.2.

3.1 VisualCOMET+

The first part of our pipeline is generating visual commonsense inferences. Our model architecture is based on VisualCOMET (Park et al., 2020). Given an image containing an event (e.g., a person drowning), VisualCOMET can generate inferences about what happened before (e.g., the ship sank) and after (e.g., he called for help). VisualCOMET has been trained on 60,000 images and three event relations (before, after, and intent). Our goal is to extend VisualCOMET and support new relations, including *HasProperty* (properties of an object such as what it is used for, where it is found, etc.); *AtLocation* (where an object/event is usually found); *HasContext* (what contexts are similar to the given input); and *Indicates* (what does this imply).

Our architecture is an adaptation of Visual-

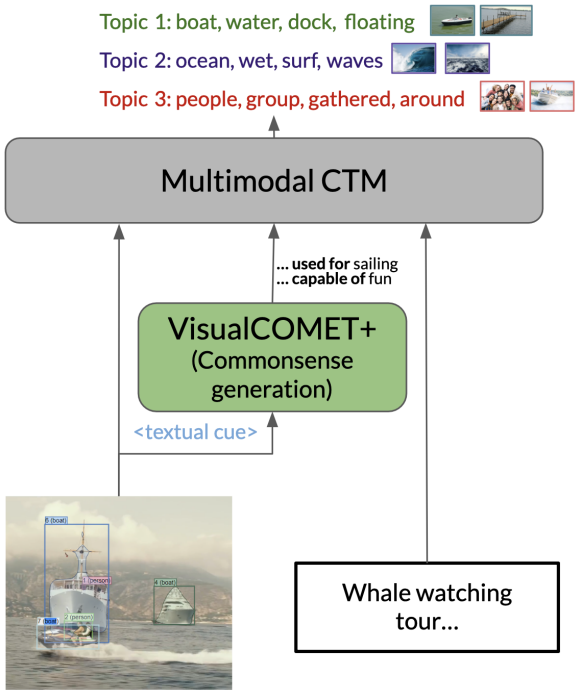


Figure 2: Overall architecture

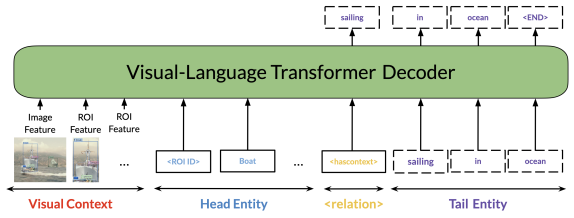


Figure 3: VisualCOMET+: Vision-Language Transformer for our approach. We feed the image and ROI tokens, a head entity with the ROI ID and text cue, and a relation. The transformer decoder then generates an inference.

COMET's transformer architecture based on BART (Lewis et al., 2020) and is shown in Figure 3. Our input sequence is composed of the visual context (regions of interest (ROIs)), the ROI ID token, the text cue language tokens, and the relation we are interested in. We expect our model to generate relevant inferences to the object and relation provided in the input sequence. We modify the text-to-image grounding mechanism in VisualCOMET+ with an additional ROI ID token, a number that is appended to the start of each text cue, to signify to the model which ROI visual feature is being referred to.

During training, we utilize the usual seq2seq negative log-likelihood loss, as specified in VisualCOMET (Park et al., 2020). Note that we do not train with the EP Loss proposed in VisualCOMET because it is out of the scope of the course project. We train our model for 5 epochs, starting from the VisualCOMET checkpoint.

3.1.1 Commonsense Knowledge Acquisition

To train VisualCOMET+, we extract commonsense triplets from Sherlock (Hessel et al., 2022), a visual abductive reasoning dataset. This dataset is based on images from VCR and VisualGenome, but we only sample from the part of the dataset that uses VCR images, as the original VisualCOMET dataset is also based on VCR images.

In this dataset, an image is annotated with multiple descriptive clues describing the most important regions of the image. Each clue is denoted by a bounding box and a textual description of that region of interest. In addition, each clue is annotated with a rationale which *explains* the clue. Figure 4 shows an example of a clue-rationale pair, and its associated bounding box, on the left. This dataset is an ideal candidate to extract commonsense triples for our model because the clue-rationale pairs are true to the image context.

We construct triplets from clue-rationale pairs by mapping to ConceptNet (Speer et al., 2017b), a commonsense knowledge base. We match the clue description and rationale with ConceptNet nodes using text processing and sequence matching methods for a given clue-rationale pair. Then, we find the shortest path from clue keywords → rationale keywords using Yen's shortest path finding algorithm[1]. See Figure 4 for an example.

In order to map ConceptNet relations to ours, we aggregate certain ConceptNet relations that best fit into each of our proposed relations, *HasProperty*, *HasContext* and *AtLocation*. A list of all the mappings is provided in the appendix Section 9.1. For example, the ConceptNet relations <usedfor>, <partof> and <hasproperty> are mapped to *HasProperty*, and a text form of the ConceptNet relation is prepended to the inference. Therefore, the ConceptNet triplet [*saddle*, <usedfor>, *riding*] becomes [*saddle*, *HasProperty*, *used for riding*] in our training data. In addition, we directly connect the clue to rationale with a new relation *indicates*. This is a new generic relation that provides an explanation for the given clue.

In this way, we construct 50K triplets from Sherlock and use 80% for training and 20% for testing.

[1]We use the `shortest_simple_paths` method from NetworkX (https://networkx.org)

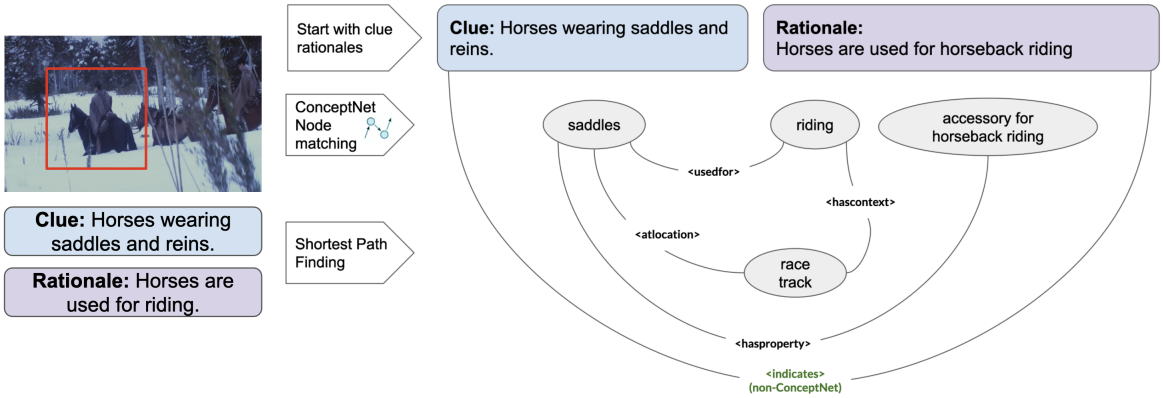


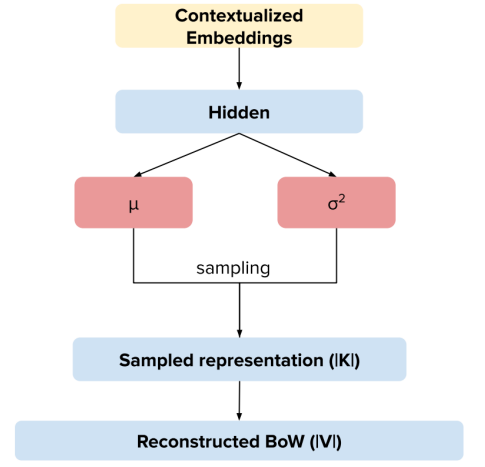
Figure 4: Building commonsense triplets using the clue and rationale from Sherlock (Hessel et al., 2022).

3.2 Multimodal CTM

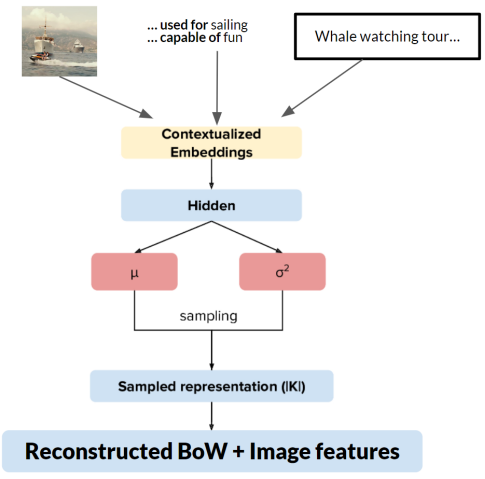
Having obtained commonsense inferences, we incorporate them into Multimodal CTM to find coherent and more diverse topics than current neural topic modeling approaches. We propose a new topic modeling algorithm because, to the best of our knowledge, there is no neural topic modeling algorithm that takes as input visual and textual features. Incorporating features from multiple modalities, as well as commonsense inferences, provides the model with an improved context to model topics.

We develop Multimodal CTM by extending the neural variational topic model CTM (Bianchi et al., 2021b) (see Figure 5 (a)). This variational autoencoder model takes as input a pre-trained representation of text documents (e.g., by using SBERT (Reimers and Gurevych, 2019)) to get a rich syntactic and semantic representation between tokens (Zhao et al., 2021). The model adjusts its latent space by reconstructing the bag-of-words from the documents. The topics (i.e., a set of keywords) are extracted from its latent space.

We extend this well-known topic modeling algorithm by allowing it to take as an input the image and the text of the document. For each document, we also add as an input the relevant inferences (in textual format) from VisualCOMET+ (see Figure 5(b)). The commonsense inferences from VisualCOMET+ are concatenated with the initial textual content of the posts. We embed the visual and textual features using OpenAI's CLIP model (Radford et al., 2021). This allows us to obtain a common representation between the two modalities.



(a)



(b)

Figure 5: (a) High-level scheme of the architecture for CTM described in (Bianchi et al., 2021b) (b) Architecture of our proposed Multimodal CTM.

		
DII	A group of people that are sitting next to each other.	Adult male wearing sunglasses lying down on black pavement.
SIS	Having a good time bonding and talking.	[M] got exhausted by the heat.

Figure 6: VIST dataset. Example descriptions of images in isolation (DII) and stories of images in sequence (SIS).

3.2.1 Dataset for Topic Modeling - VIST

We demonstrate the utility of our proposed methods through multiple evaluation metrics on the Visual Story-Telling dataset (VIST)[2]. This dataset contains 81,743 unique photos in 20,211 sequences, aligned to descriptive caption (DII) and story language (SIS) (see Figure 6). We use the DIIs as text cues to VisualCOMET+ for generating commonsense inferences, and use the SIS as the text document in Multimodal CTM. We choose this dataset because (1) the images in this dataset are general and diverse, and (2) the SIS text is not simply describing the image, but rather is an extension to the image, making this analogous to social media content, that topic modeling is useful for.

3.2.2 Generating Commonsense Inferences for VIST

We sample 17,000 instances of unique images, DII captions, and story triplets from VIST. We feed into VisualCOMET+ the image feature of the full image and the image's caption as the textual cue, in order to generate inferences for the four new relations that we added: *HasProperty*, *AtLocation*, *HasContext*, and *indicates*. We obtain one inference from each and join them into a single sentence. An example of this is provided in Figure 7.

4 Evaluation

We evaluate the quality of commonsense inferences generated from VisualCOMET+ by using metrics that indicates N-gram overlap. This is the same procedure used for evaluating VisualCOMET (Park et al., 2020). We mainly focus on the BLUE-2 score (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). For BLEU-2, we measure the 2-gram overlap between the generated inferences and ground truth triplets created from clue-rationale pairs, as described in Section 3.1.1.

By injecting commonsense into the topic modeling algorithm, we expect to find coherent and more diverse topics. Thus, we will evaluate topic models on five metrics: three for topic coherence (NPMI (Lau et al., 2014); C_v (Röder et al., 2015a), and WECO (Ding et al., 2018)) and two to quantify the diversity of the resulting topics (TD (Dieng et al., 2020); and I-RBO (Bianchi et al., 2021a)). See Section 9.3 for a detailed description of these metrics.

5 Results

We first discuss results on VisualCOMET+ using the triplets from the Sherlock dataset. Then, we discuss the topic modeling results after injecting commonsense inferences into our Multimodal CTM.

5.1 VisualCOMET+ results

On the test set of triplets from the Sherlock dataset, we obtain a BLEU-2 score of 0.306, which exceeds the BLEU-2 score of 0.135 reported for VisualCOMET (Park et al., 2020). Likewise, we obtain a score of 0.175 on the METEOR metric, whereas VisualCOMET achieves 0.115. However, it is unreasonable to compare scores directly, because we use the new relations to conduct our experiments, and, triplets from Sherlock are completely different from triplets from the VisualCOMET dataset. However, the scores indicate that the model has learned to generate inferences on the new relations, without any significant drop in performance.

We show qualitative examples in the appendix, in Section 9.2. We see that in Figure 10 the model is able to generate commonsense inferences for all of the seven supported relations. We also see that the model is able to reason beyond the image and the text cue, with phrases such as "messy person", "used for dishes", "kitchen", and "left the food". In Figure 11, we see that providing different ROIs and text cues for the same image can lead to diverse generations.

[2]https://visionandlanguage.net/VIST/

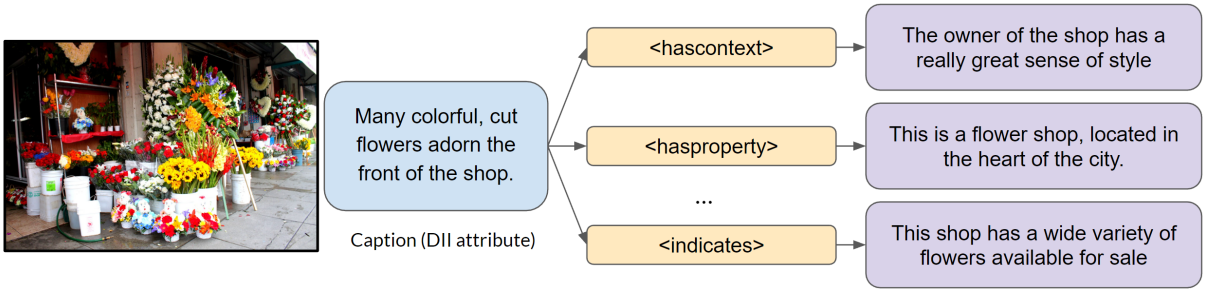


Figure 7: Example of commonsense inferences for an (image, DII) pair of VIST. We feed VisualCOMET+ with the image features of the full image and its DII as textual cue. We obtain inferences using four relations: *HasProperty*, *AtLocation*, *HasContext*, and *indicates* to use in Multimodal CTM.

Model	Relations	Test Size	BLEU-2	METEOR
VisualCOMET	*before, after, intent*	145k	0.135	0.115
VisualCOMET+	*HasProperty*, *AtLocation*, *HasContext*, *indicates*	10k	0.306	0.175

Table 1: Evaluation of VisualCOMET+ on the test set generated using Sherlock. VisualCOMET (Park et al., 2020) scores are provided as reference. While we cannot compare BLEU-2 and METEOR scores directly, because our model uses different relations and a smaller test set, we can say that the model is learning to generate inferences for the new relations.

Documents embeddings	Coherence			Diversity	
	NPMI	C_v	WECO	TD	IRBO
Text	-0.04	0.38	0.21	0.62	0.98
Text-Image	-0.04	0.38	**0.22**	0.67	**0.99**
Text-Inferences	**-0.03**	**0.39**	**0.22**	0.62	0.98
Text-Image-Inferences (Multimodal CTM)	**-0.03**	**0.39**	**0.22**	**0.68**	**0.99**

Table 2: Comparison of topics' coherence and topics' diversity between document representations. Each result averaged over 11 runs. We compute all the metrics for 25 topics. Best results are bold.

5.2 Topic modeling results

Table 2 shows the effects of using different input representations in our topic modeling algorithm. We compute topic coherence and diversity metrics for 25, 50, and 75 topics (see Section 9.4 for more details). We average results for each metric over 11 runs of each model.

Our results suggest that by incorporating image and VisualCOMET+ inferences, we can obtain more diverse topics with similar or slightly higher coherence. We hypothesize that the scores obtained for topics' coherence, predominantly based on word occurrences in the corpus, are justified by the fact that the top words identified by Multimodal CTM do not explicitly co-occur more in the corpus but are rather semantically related through the external knowledge. A qualitative analysis of these results (e.g., by using word intrusion tasks (Chang et al., 2009)) can provide more insights into the differences between the topics found using various documents' representations. We leave this analysis for future work.

As an example, Figure 8 shows the most relevant documents associated with one topic identified by Multimodal CTM. The most relevant keywords associated with this topic are related to *weddings* (e.g., "bride", "husband", and "love"). We show the top 7 (image, story) pairs from VIST most related to this topic. For each document, we display the image, the story, and the probability of this document belonging to this topic. All these documents seem to be highly related. These results highlight the usefulness of our algorithm on a multimodal dataset.

We also used an interactive topic modeling visualization tool to get a better interpretation of our results. LDAvis (Sievert and Shirley, 2014), projects topics into a two-dimensional space. Circles represent topics, and the similarity between topics determines their positions. The circle size indicates a topic's prevalence in the corpus. Figure 9 shows a visualization of a topic model. The most

Figure 8: Example of the most relevant documents to a topic in Multimodal CTM.

relevant keywords of the currently selected topic and its similarity with others allow us to interpret this theme as *weddings*.

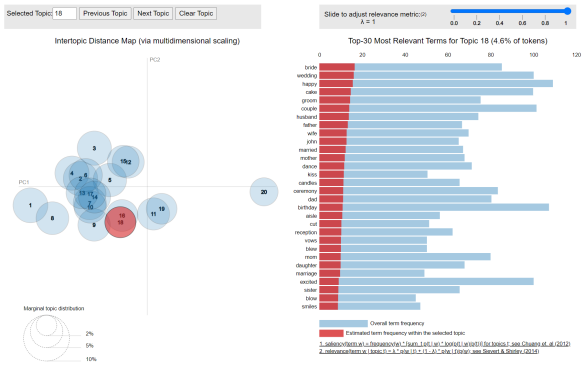


Figure 9: Visualization of topics from Multimodal CTM. On the left, a global view of topics is provided. On the right, the most relevant keywords from the selected topic appear.

6 Discussion & Limitations

VisualCOMET+ does a reasonable job of generating meaningful inferences given image and textual cues, as shown in qualitative examples. With VIST, we attempted to generate inferences on a completely different dataset, since both Sherlock and VisualCOMET datasets are based on VCR images, whereas VIST is not. Even then, the model was able to generate good quality inferences which enhance the topic model. That said, VisualCOMET+ suffers from a few drawbacks. First, inference diversity is limited, where the inferences generated for different relations are very similar. Second, inferences may contain information that is incorrect with respect to the image. Improving image features and using more training may alleviate these issues.

Multimodal CTM is the first neural topic model that takes into account commonsense inferences and visual features to identify the main themes of a corpus. Our results show that images and VisualCOMET+ inferences can result in more coherence and diverse topics. In future work, we would like to analyze if the resulting topics match the corpus and if the granularity of those is adequate for real-world applications. Experiments in other multimodal datasets can also provide insights in terms of the performance of our algorithm.

While the performance of MultiModal CTM is adequate, there are venues for improving its performance. For example, in the current version of the algorithm, the decoder of the variational autoencoder only reconstructs the bag-of-words of the document's textual content. We hypothesize that by adding an additional task to the decoder, such as reconstructing the image features of the documents, its performance can be boosted.

7 Takeaways

1. We discovered that extracting commonsense knowledge from existing VL datasets is a promising and *less expensive* alternative to acquiring human annotations.

2. Extending VisualCOMET helped us in understanding and implementing a *basic grounding mechanism* to tie an image region to a corresponding textual cue.

3. We gained a deeper understanding of *neural topic models* and how to adapt them to a multimodal setting.

4. We experimented with different metrics to evaluate the quality of topics, and recognized some limitations of automatic metrics (e.g., measuring topic diversity only on the top ten keywords of topics might not be very insightful).

5. We learned to improve our algorithms in terms

of *space and time complexity*. Currently, we can evaluate topic models five times faster than in the previous two weeks.

6. We learned how to divide a larger idea into *non-overlapping components*, which helped us deliver results on time.

8 Conclusions

We introduced VisualCOMET+, Vision Language transformer that can generate commonsense inferences on not only people-centric relations (e.g., what a person did "before") but more diverse relations that encompass the properties of objects (e.g., "HasProperty") and provide more context and rationale (e.g., "Indicates"). We have also presented Multimodal CTM, a new multimodal topic modeling algorithm that incorporates text, images, and commonsense inferences from VisualCOMET+, to find coherent and diverse topics.

References

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230, Brussels, Belgium. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David Blei, Lawrence Carin, and David Dunson. 2010. Probabilistic topic models. *IEEE signal processing magazine*, 27(6):55–65.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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.

Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science : Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics (TACL)*.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 830–836, Brussels, Belgium. Association for Computational Linguistics.

Mennatallah El-Assady, Rebecca Kehlbeck, Christopher Collins, Daniel Keim, and Oliver Deussen. 2019. Semantic concept spaces: Guided topic model refinement using word-embedding projections. *IEEE transactions on visualization and computer graphics*, 26(1):1001–1011.

Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3198–3210, Online. Association for Computational Linguistics.

Ismail Harrando and Raphaël Troncy. 2021. Discovering interpretable topics by leveraging common sense knowledge. In *Proceedings of the 11th on Knowledge Capture Conference*, pages 265–268.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning. In *ECCV*.

Yang Hong, Xinhuai Tang, Tiancheng Tang, Yunlong Hu, and Jintai Tian. 2020. Enhancing topic models by incorporating explicit and implicit external knowledge. In *Asian Conference on Machine Learning*, pages 353–368. PMLR.

Enamul Hoque and Giuseppe Carenini. 2015. Convisit: Interactive topic modeling for exploring asynchronous online conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 169–180.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. *Advances in Neural Information Processing Systems*, 34:2018–2033.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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 *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

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

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *ECCV*.

Jessica Peter, Steve Szigeti, Ana Jofre, and Sara Diamond. 2015. Topicks: Visualizing complex topic models for user comprehension. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 207–208. IEEE.

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

Dheeraj Rajagopal, Daniel Olsher, Erik Cambria, and Kenneth Kwok. 2013. Commonsense-based topic modeling. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, New York, NY, USA. Association for Computing Machinery.

Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. 2021. Vlc-bert: Visual question answering with contextualized commonsense knowledge.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015a. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015b. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019.

Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Adnan Muhammad Shah, Xiangbin Yan, Samia Tariq, and Mudassar Ali. 2021. What patients like or dislike in physicians: Analyzing drivers of patient satisfaction and dissatisfaction using a digital topic modeling approach. *Information Processing & Management*, 58(3):102516.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629. Association for Computational Linguistics.

Carson Sievert and Kenneth Shirley. 2014. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.

Dandan Song, Jingwen Gao, Jinhui Pang, Lejian Liao, and Lifei Qin. 2020. Knowledge base enhanced topic modeling. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pages 380–387. IEEE.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017a. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017b. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021. Word embedding-based topic similarity measures. In *International Conference on Applications of Natural Language to Information Systems*, pages 33–45. Springer.

Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. 2021. HypoGen: Hyperbole generation with commonsense and counterfactual knowledge. In *Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4).

Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. KM-BART: Knowledge enhanced multimodal BART for visual commonsense generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 525–535, Online. Association for Computational Linguistics.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4713–4720. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

9 Appendix

9.1 ConceptNet relations to VisualCOMET+ relations

We map ConceptNet relations to our new relations, *HasProperty*, *HasContext* and *AtLocation* in the following manner:

1. <usedfor>, <hasproperty>, <capableof>, <partof>, <madeof>, <hasa> are mapped to *HasProperty*.
2. <hascontext>, <similarto>, <etymologicallyrelatedto>, <mannerof> are mapped to *HasContext*.
3. <atlocation> is mapped to *AtLocation*.
4. All other relations from ConceptNet are ignored.

9.2 Commonsense Generation Examples

In Figure 10 and Figure 11, we show qualitative examples of commonsense generation on images from the Sherlock dataset.

9.3 Topic Modeling Evaluation

We evaluate the quality of topic models based on topic coherence (topic keywords must shame some level of semantic relatedness) and topic segregation, which measures the lexical and semantic overlap between topics. Note that a higher value indicates a better performance in all of the metrics mentioned below.

Normalized Pointwise Mutual Information (NPMI) (Lau et al., 2014) is one of the most well-known automatic coherence metric. It measures how much more likely the most representative terms of a topic co-occur than if they were independent. NPMI returns a high score when the top N words that describe a topic, summed over all pairs w_i and w_j, have high joint probability $P(w_j, w_i)$ compared to their marginal probability (Hoyle et al., 2021). The range of NPMI is between [-1,1], whereas a higher value indicates a more coherent topic. Usually, NPMI is calculated by using a sliding window of 10 words to identify co-occurrences.

C_v (Röder et al., 2015b) uses a variation of NPMI to calculate the coherence over a sliding window with size 110. It calculates the co-occurrence of a word of a given topic against all words of the same topic. It ranges between [0,1], where a higher value suggests more coherent topics.

External word embeddings topic coherence (WECO) (Ding et al., 2018) provides an additional measure of how similar the words in a topic are. It is based on word embeddings (Mikolov et al., 2013). First, it is computed as the average pairwise cosine similarity of the word embeddings of the top 10 words in a topic. Then, the overall average of those values for all the topics is reported.

Topic diversity (TD) (Dieng et al., 2020) is computed as the percentage of unique words in the top 25 words of all topics. Its range is between [0,1]. A value near zero suggests redundant topics, while a value near one suggests more different topics.

Inversed Rank-Biased Overlap (I-RBO) (Bianchi et al., 2021a) evaluates how diverse the topics generated by a single model are. When comparing topics, the rank of each term matters (Terragni et al., 2021); it is not the same if topics share words at high ranks as if they do at low ranks. I-RBO is the reciprocal of the standard RBO (Webber et al., 2010), and it is computed by considering the top 10 words of topics. In this metric, two topics that share some of the keywords, although at different rankings, are penalized less than two topics that share the same keywords at the highest ranks. It ranges between [0,1].

9.4 Additional topic modeling results

We also run our experiments considering different numbers of topics (i.e., 50, and 75). Table 3 and Table 4 show the result of different document representations into our multimodal topic modeling algorithm with 50 and 75 topics, respectively.

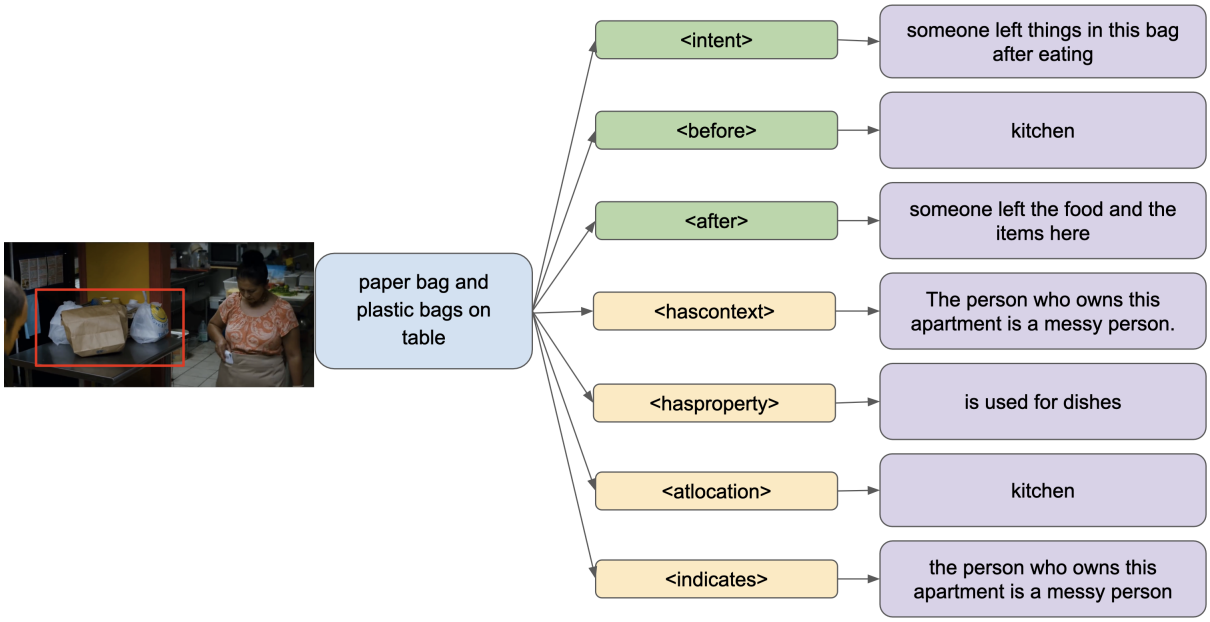


Figure 10: Example of commonsense inference generation on all seven relations that VisualCOMET+ supports. The three older VisualCOMET relations are in green, and the four new relations are in yellow. We generate reasonable inferences given the text cue and the image, but also show that our model is not perfect, and may produce repetitive (less diverse) inferences across relations.

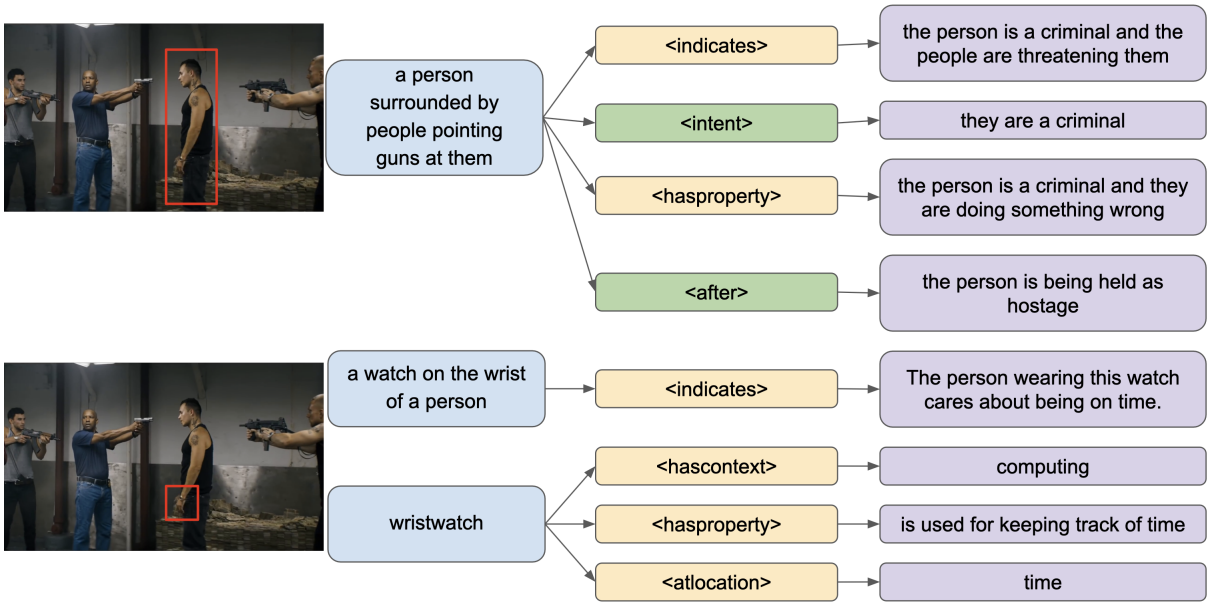


Figure 11: Example of commonsense inference generation on the same image, but with different ROI and text cues. We show that the model adapts well to varying cues.

Documents embeddings	**Coherence**			**Diversity**	
	NPMI	C_v	**WECO**	**TD**	**IRBO**
Text	**-0.03**	**0.38**	0.21	0.31	0.93
Text-Image	-0.04	**0.38**	**0.22**	**0.41**	**0.97**
Text- Inferences	-0.04	**0.38**	**0.22**	0.40	**0.97**
Text-Image-Inferences	**-0.03**	**0.38**	0.21	0.35	0.95

Table 3: Comparison of topics' coherence and topics' diversity between document representations. Each result averaged over 11 runs. We compute all the metrics for 50 topics

Documents embeddings	**Coherence**			**Diversity**	
	NPMI	C_v	**WECO**	**TD**	**IRBO**
Text	-0.04	**0.38**	0.20	0.22	**0.93**
Text-Image	**-0.03**	0.37	**0.21**	0.22	0.92
Text- Inferences	-0.04	**0.38**	0.20	0.19	0.91
Text-Image-Inferences	**-0.03**	0.37	**0.21**	**0.24**	**0.93**

Table 4: Comparison of topics' coherence and topics' diversity between document representations. Each result averaged over 11 runs. We compute all the metrics for 75 topics