Medical Visual Textual Entailment for Numerical Understanding of Vision-and-Language Models

Hitomi Yanaka[*], Yuta Nakamura[*], Yuki Chida, Tomoya Kurosawa
the University of Tokyo
{hyanaka,yuki.chida,kurosawa-tomoya}@is.s.u-tokyo.ac.jp
yutanakamura-tky@umin.ac.jp

Abstract

Assessing the capacity of numerical understanding of vision-and-language models over images and texts is crucial for real vision-and-language applications, such as systems for automated medical image analysis. We provide a visual reasoning dataset focusing on numerical understanding in the medical domain. The experiments using our dataset show that current vision-and-language models fail to perform numerical inference in the medical domain. However, the data augmentation with only a small amount of our dataset improves the model performance, while maintaining the performance in the general domain.

1 Introduction

Vision-and-language models have made great progress on complex tasks, going beyond image recognition and towards reasoning over images and texts (Antol et al., 2015; Xie et al., 2019; Suhr et al., 2019). Following the success of pre-trained language models (Devlin et al., 2019, inter alia), recent advances in vision-and-language models have been made by the introduction of large-scale pre-training (Li et al., 2019; Kim et al., 2021; Singh et al., 2022). However, as with pre-trained language models, it is unclear what information pre-trained vision-and-language models learn and use in their predictions, and what their limitations are.

While a large body of research (Naik et al., 2018; Rozen et al., 2019; Ravichander et al., 2019; Richardson et al., 2020) has provided challenging reasoning tasks to probe the reasoning ability of pre-trained language models, such work has been more limited for vision-and-language models. Furthermore, previous visual reasoning datasets are usually provided by the general domain of images, and analysis across different domains is desirable.

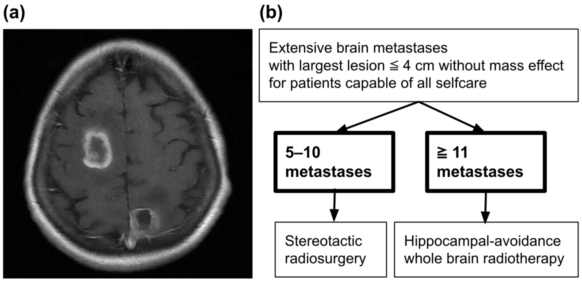
(a)
(b)
Extensive brain metastases with largest lesion ≦ 4 cm without mass effect for patients capable of all selfcare
5–10 metastases
≧ 11 metastases
Stereotactic radiosurgery
Hippocampal-avoidance whole brain radiotherapy

Figure 1: The practical example of the need for visual reasoning in the medical domain. (a) A magnetic resonance imaging (MRI) image showing two brain metastases[1]. (b) Treatment strategy depending on the lesion number of brain metastases (modified from Gondi et al. (2022), not for medical use).

Our focus is to investigate whether current vision-and-language models have the ability to infer numerical relationships between images and texts in the medical domain, which is crucial for real vision-and-language applications such as systems for automated medical image analysis. Consider the example of images and textual descriptions in a medical article presented in Figure 1. The lesion number affects the treatment strategy for diseases such as brain metastasis. If systems can automatically judge whether the lesion number in given images matches that in arbitrary query texts, they can support medical decision-making. Recently, a vision-and-language model focusing on the medical domain (Delbrouck et al., 2022) has begun to be provided but is not yet fully developed.

With this motivation, we provide a visual reasoning dataset focusing on numerical inference in the medical domain by adding annotations to the previous medical image and caption dataset MedICaT (Subramanian et al., 2020). We call our dataset MedVTE, which will be publicly available at https://github.com/ynklab/MedVTE. Using MedVTE, we investigate the extent to which current pre-trained vision-and-language models have the ability of numerical

[*] Equal Contribution.

[1] https://radiopaedia.org/cases/haemorrhagic-intracranial-metastases-from-breast-cancer

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 8–18
July 14, 2023 ©2023 Association for Computational Linguistics

understanding on visual reasoning tasks across images and texts in the medical domain. The experiments show that current models have much room to perform numerical inference in the medical domain.

2 Background

2.1 Vision-and-language understanding

Regarding standard vision-and-language understanding tasks, SNLI-VE (Xie et al., 2019) is a large general domain dataset for the Visual Textual Entailment (VTE) task. The dataset consists of image-sentence pairs annotated with a three-class label (*entailment*, *contradiction*, or *neutral*), indicating whether a premise image entails a hypothesis sentence. There have been studies investigating the counting ability of vision-and-language models on visual question-answering tasks and object detection tasks (Chattopadhyay et al., 2017; Zhang et al., 2018; Song and Qiu, 2018; Trott et al., 2018; Acharya et al., 2019; Parcalabescu et al., 2021). However, since previous studies only use datasets in the general domain, it is unclear the extent to which models can maintain the ability to understand numerical expressions in the medical domain.

For visual reasoning in the medical domain, Li et al. (2020) have compared the performance of four pre-trained vision-and-language models and traditional CNN-RNN models on two datasets of thoracic findings classification tasks: MIMIC-CXR (Johnson et al., 2019) and OpenI datasets. The results showed that the pre-trained models outperformed the traditional models. Our VTE dataset gives a fine-grained analysis of the capacity of the pre-trained vision-and-language models for numerical understanding in the medical domain.

2.2 Clinical NLP

Clinical NLP is one of the practical fields of NLP, and various reasoning tasks in the medical domain have been provided. For sentence-level language understanding tasks, emrQA (Pampari et al., 2018) is a large-scale QA dataset on electronic medical records, and MedSTS (Wang et al., 2020) is a resource for Semantic Textual Similarity (STS) tasks in the medical domain. The most related dataset to ours is MedNLI (Romanov and Shivade, 2018), a physician-annotated Natural Language Inference (NLI) dataset with premises extracted from clinical notes. However, a recent study has reported annotation artifacts in MedNLI (Herlihy and Rudinger, 2021). To avoid such undesired artifacts, we cover a variety of numerical expressions.

3 MedVTE Datasets

We introduce MedVTE, visual reasoning datasets in the medical domain involving numerical expressions. MedVTE is composed of pairs of medical images, captions, and three-class entailment labels (*entailment*, *contradiction*, or *neutral*). MedVTE focuses on the relationship between the number of lesions, such as cancer in an image and the numerical expression in a text.

We created MedVTE by selecting examples involving numerical expressions from MedICaT dataset (Subramanian et al., 2020). MedICaT contains 217,060 figure-caption pairs in medical articles, whose captions sometimes refer to the number of the depicted lesions (e.g., tumors or nodules). The selection is conducted by one medical expert.

3.1 Premise–hypothesis collection

In MedVTE, a premise is a MedICaT figure, and a hypothesis is one complete sentence containing one or more lesion numbers. We created 409 examples for the MedVTE dataset in total.

Step 1. Cleaning We removed 58 MedICaT figure–caption duplicate pairs. We also mitigated occasional errors in MedICaT captions, such as missing letters or interrupted sentences. Some MedICaT captions are provided in two versions, the one by the MedICaT authors and the other from the S2ORC dataset (Lo et al., 2020). In such cases, we always chose the longer one to avoid including incomplete sentences.

Step 2. Figure collection We collected MedICaT figure–caption pairs whose captions include lesion numbers in a rule-based approach. We assigned Penn Treebank part-of-speech (POS) tags (Marcus et al., 1993) to all captions. We then applied a spaCy rule-based matcher to accept only captions having a numeral followed by a noun suggesting lesions. This step left us 431 figure–caption pairs. See Appendix A for details.

Step 3. Hypothesis collection Every MedVTE hypothesis is a single sentence including one or more lesion numbers. We collected hypotheses by splitting captions into sentences and selecting sentences containing at least one lesion number.

Sentence selection was performed in a rule-based approach as in Step 2 followed by manual

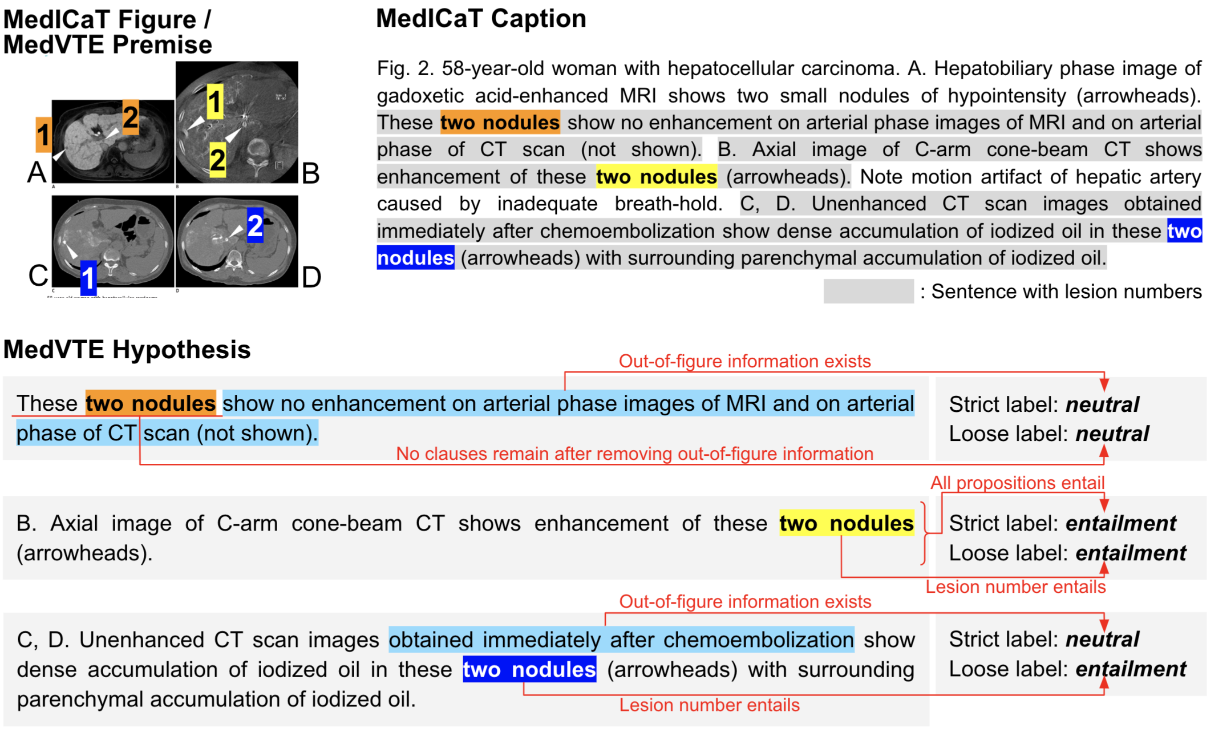
MedICaT Figure / MedVTE Premise
A
B
C
D
1
2
1
2
1
2
MedICaT Caption
Fig. 2. 58-year-old woman with hepatocellular carcinoma. A. Hepatobiliary phase image of gadoxetic acid-enhanced MRI shows two small nodules of hypointensity (arrowheads). These two nodules show no enhancement on arterial phase images of MRI and on arterial phase of CT scan (not shown). B. Axial image of C-arm cone-beam CT shows enhancement of these two nodules (arrowheads). Note motion artifact of hepatic artery caused by inadequate breath-hold. C, D. Unenhanced CT scan images obtained immediately after chemoembolization show dense accumulation of iodized oil in these two nodules (arrowheads) with surrounding parenchymal accumulation of iodized oil.
: Sentence with lesion numbers
MedVTE Hypothesis
These two nodules show no enhancement on arterial phase images of MRI and on arterial phase of CT scan (not shown).
Out-of-figure information exists
No clauses remain after removing out-of-figure information
Strict label: neutral
Loose label: neutral
B. Axial image of C-arm cone-beam CT shows enhancement of these two nodules (arrowheads).
All propositions entail
Lesion number entails
Strict label: entailment
Loose label: entailment
C, D. Unenhanced CT scan images obtained immediately after chemoembolization show dense accumulation of iodized oil in these two nodules (arrowheads) with surrounding parenchymal accumulation of iodized oil.
Out-of-figure information exists
Lesion number entails
Strict label: neutral
Loose label: entailment

Figure 2: MedVTE examples. Premises are MedICaT figures and hypotheses are MedICaT caption sentences containing numerical expressions of lesions. For each hypothesis, the strict label considers all information, and the loose label is only determined by comparing lesion numbers. Corresponding lesion numbers are colored in orange, yellow, and dark blue. Light blue spans indicate out-of-figure information, which is beyond the figure's scope and deemed unverifiable by the medical expert based on the figure alone.

reviews. In manual reviews, we removed erroneous lesion numbers where integers do not actually count lesions, such as cell line names *Walker 256 tumor*. We also excluded invalid premise figure-hypothesis sentence pairs meeting the below criteria:

- the figure file contains multiple article figures
- the hypothesis is not a single sentence
- the hypothesis does not make sense due to ungrammaticality.

When multiple hypothesis sentences corresponded to a single premise figure, we treated each premise figure-hypothesis sentence pair as an independent sample. We obtained 409 premise-hypothesis pairs for 373 premise figures, where 430 lesion numbers appear in total.

3.2 Labeling

We assigned two types of entailment labels, *strict labels* and *loose labels*, to premise-hypothesis pairs on MedVTE.

Models	Train \ Test	SNLI-VE	MVTEl	MVTEs
ViLT	SNLI-VE	0.757	0.243	0.290
	+MVTEl	0.757	0.443	0.366
	+MVTEs	0.745	0.371	0.416
FLAVA	SNLI-VE	0.790	0.236	0.281
	+MVTEl	0.791	0.428	0.356
	+MVTEs	0.791	0.355	0.408

Table 1: F1-macro scores of each baseline model and dataset. MVTEl and MVTEs indicate MedVTE annotated with loose labels and strict labels, respectively. +MVTEl indicates SNLI-VE mixed with MVTEl.

Strict labels follow the common practice of annotating visual reasoning datasets to compare all the information, not only numerical one but also medical background knowledge, of a premise figure and a hypothesis sentence. However, we found that the considerable number of strict labels became *neutral* under given images because out-of-figure information in hypothesis sentences (i.e., information that is not acquired from images), such as "this image was obtained six months after surgery," is necessary to judge their labels as *entailment*.

To realize separate assessments of the numerical reasoning abilities of models under only given images, we add loose label annotations rather than editing hypothesis sentences. Loose labels only compare numerical information of a premise figure and a hypothesis sentence. This approach provides an option to focus on numerical reasoning abilities with loose labels, or to fully measure medical reasoning abilities with strict labels, which requires expert knowledge to recognize out-of-figure information.

The following is the definition of loose labels. Details are available in Appendix C.

- *entailment*: All lesion numbers are consistent with the premise figure
- *contradiction*: One or more lesion numbers are smaller than those depicted in the premise figure
- *neutral*: Either of the following is satisfied: (i) one or more lesion numbers are larger than those depicted in the premise figure although the others are consistent, (ii) the number of lesion numbers cannot be determined only from the premise figure, or (iii) no clauses remain after removing out-of-figure information from the hypothesis.

Figure 2 shows MedVTE examples. In the top and middle examples, their loose labels are the same as their strict labels. In the bottom example, its loose label is different from its strict label with the consideration of out-of-figure information. The distribution of loose labels in MedVTE is $(\textit{entailment}, \textit{neutral}, \textit{contradiction}) = (310, 95, 4)$, and that of strict labels is $(\textit{entailment}, \textit{neutral}, \textit{contradiction}) = (208, 197, 4)$.

4 Experiments and Analysis

4.1 Experimental setup

Models Vision-and-language models are categorized into three broad types based on their encoding style, fusion encoder, dual encoder, and a combination of both. We used two vision-and-language models for our experiments: a Vision-and-Language Transformer model (ViLT) (Kim et al., 2021) and a Foundational Language And Vision Alignment model (FLAVA) (Singh et al., 2022). ViLT is a fusion-encoder style model which has 112M parameters. FLAVA is a fusion-encoder plus dual-encoder style model which has 243M parameters. See details of pre-training datasets for each model in Appendix D.

Training For baseline models, we use vision-and-language models fine-tuned with the training set of SNLI-VE. We split the MedVTE dataset as train:test=306:103 and evaluate the performance of the models on the MedVTE test set. To investigate whether a small portion of additional training data in the medical domain contributes to knowledge transfer for visual reasoning, we evaluate models fine-tuned with the SNLI-VE training set mixed with the MedVTE training set. We fine-tune the models for three epochs for each dataset and use F1-macro scores for evaluation metrics. Details on the hyperparameters can be found in Appendix D.

4.2 Baseline results

Table 1 shows baseline results. While both ViLT and FLAVA models trained with SNLI-VE achieved around 75% on in-domain SNLI-VE, their performance was very low on MedVTE.

When we evaluated models trained with SNLI-VE mixed with a subset of MedVTE, the performance on MedVTE was improved while maintaining the performance on SNLI-VE. However, the overall performance on MedVTE was still lower than 50%. This indicates that numerical inference in the medical domain is challenging for vision-and-language models even when they train with a subset of MedVTE. Regarding the difference between loose labels and strict labels with a subset of MedVTE, the performance improvement on MedVTE strict labels was lower than that on loose labels. This suggests that the ability to use out-of-figure information is difficult to obtain from the data augmentation.

5 Conclusion

We created the visual reasoning dataset MedVTE, focusing on numerical understanding in the medical domain. The experiments using MedVTE showed that current vision-and-language models struggled with performing numerical inference in the medical domain. However, the data augmentation with only a small amount of our MedVTE dataset improved the model performance, while maintaining the performance in the general domain. In future work, we increase the size of our MedVTE dataset and make further analysis of vision-and-language models to investigate the extent to which the size of a fine-tuning dataset affects the performance of numerical inference in the medical domain. Improving automated numerical vision-and-language understanding in the medical domain could aid therapeutic decision-making that depends on lesion numbers.

6 Limitation

Since hypothesis sentences were created and labeled by medical experts, the size of our current dataset is small. In particular, the number of examples of contradiction is small because the hypothesis sentences were created based on captions to efficiently construct our dataset. However, we can increase the number of examples of contradiction by rewriting phrases in the hypothesis sentences. The claim of this study is that we can relatively efficiently create a VTE dataset in the medical domain from the existing image caption dataset, and can empirically demonstrate the challenges of current vision-and-language models on the VTE dataset. Although increasing the data size is an important next step, it is beyond the scope of this paper.

Acknowledgements

We thank the three anonymous reviewers for their helpful comments and suggestions, which improved this paper. This work was supported by JST, PRESTO grant number JPMJPR21C8, Japan.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8076–8084.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra, and Devi Parikh. 2017. Counting everyday objects in everyday scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022. ViLMedic: a framework for research at the intersection of vision and language in medical AI. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. RedCaps: Web-curated image-text data created by the people, for the people. In *Proceedings of the NeurIPS Datasets and Benchmarks*.

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

Vinai Gondi, Glenn S. Bauman, Lisa Bradfield, Stuart H. Burri, Alvin R. Cabrera, Danielle A. Cunningham, Bree R Eaton, Jona A. Hattangadi-Gluth, Michelle M. Kim, Rupesh R. Kotecha, Lianne Kraemer, Jing Li, Seema Nagpal, Chad G Rusthoven, John H. Suh, Wolfgang A. Tomé, Tony J.C. Wang, Alexandra S. Zimmer, Mateo Ziu, and Paul D. Brown. 2022. Radiation therapy for brain metastases: An astro clinical practice guideline. *Practical radiation oncology*, 12(4):265–282.

Christine Herlihy and Rachel Rudinger. 2021. MedNLI is not immune: Natural language inference artifacts in the clinical domain. 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 1020–1027, Online. Association for Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(317):2052–4463.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 5583–5594.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings* *of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li an Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yikuan Li, Hanyin Wang, and Yuan Luo. 2020. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1999–2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2Text: Describing images using 1 million captioned photographs. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 1143–1151, Red Hook, NY, USA. Curran Associates Inc.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, Groningen, Netherlands (Online). Association for Computational Linguistics.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 647–664.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8713–8721.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Ohad Rozen, Vered Shwartz, Roee Aharoni, and Ido Dagan. 2019. Diversify your datasets: Analyzing generalization via controlled variance in adversarial datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 196–205, Hong Kong, China. Association for Computational Linguistics.

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

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629.

Zichen Song and Qiang Qiu. 2018. Learn to classify and count: A unified framework for object classification and counting. In *Proceedings of the 2018 International Conference on Image and Graphics*

Processing, page 110–114. Association for Computing Machinery.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 818–827, Vancouver, Canada. Association for Computational Linguistics.

Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. MedICaT: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120, Online. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl S. Ni, Douglas N. Poland, Damian Borth, and Li-Jia Li. 2015. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59:64–73.

Alexander Trott, Caiming Xiong, and Richard Socher. 2018. Interpretable counting for visual question answering. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. 2018. Learning to count objects in natural images for visual question answering. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

A Sample selection rules

This section explains detailed MedVTE sample selection rules.

We employed rule-based approaches to select figure-caption pairs from the MedICaT dataset so that all sampled captions refer to the number of lesions.

We selected sentences in the MedICaT captions containing LESION-NUMBER-EXPRESSIONs. We defined a LESION-NUMBER-EXPRESSION as any token subsequence of a single sentence of a caption that satisfies all of the following *Rules 1* to *3*:

- *Definition 1.* COMPARATIVE is a string whose lowercase form is either *at least*, *at most*, *more than*, or *less than*.

- *Definition 2.* NUMBER is a single token whose Penn Treebank part-of-speech (POS) tag (Marcus et al., 1993) is CD (cardinal number).

- *Definition 3.* LESION-NOUN is a single token whose POS tag is either NN (noun, singular or mass) or NNS (noun, plural).

- *Rule 1.* A LESION-NUMBER-EXPRESSION must be a concatenation of COMPARATIVE, NUMBER, and LESION-NOUN in this order, or a concatenation of NUMBER and LESION-NOUN in this order.

- *Rule 2.* The lemma of LESION-NOUN must be either *cancer*, *lesion*, *mass*, *metastasis*, *nodule*, or *tumor*.

- *Rule 3.* A LESION-NUMBER-EXPRESSION must not appear immediately after a token whose lowercase form is either *figure, figures, fig, figs, patient, case, day, sample, type, category, group, grade, level, stage, rads, pirads, birads, cin, score, likert, c, t, l, s, segment, gs, suv, +, +1, +2, +3, +4, +5, mm, cm, mm2, cm2, mm3,* or *cm3*.

In our implementation, we first assigned POS tags to all MedICaT captions using Berkeley Neural Parser (Stern et al., 2017; Kitaev and Klein, 2018; Kitaev et al., 2019). We then built a spaCy rule-based matcher and applied it to all parsing results.

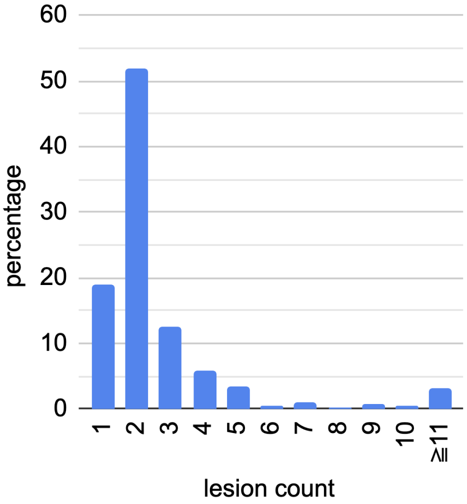

Figure 3: Distribution of the quantity of 424 of the 430 lesion numbers in the MedVTE hypotheses. Note that the remaining six lesion numbers are excluded because they appear immediately after a comparative expression such as “at least” or “more than.”

B Dataset statistics

Of the 409 MedVTE premise-hypothesis pairs, 300 (73.3%) have radiological premise figures, twelve (2.9%) have scopic premise figures, and the remaining have other various types of premise figures including histopathological images.

MedVTE contains 430 lesion numbers in total because three of the 409 hypotheses (0.7%) contain three lesion numbers, fifteen hypotheses (3.7%) contain two lesion numbers, while the remaining 391 hypotheses (95.6%) contain one lesion number.

Six of the 430 lesion numbers (1.4%) include comparative expressions, four of which are associated with “at least” and the others are accompanied by “more than.” Figure 3 shows the distribution of the remaining 424 lesion numbers. The most frequent lesion number is two, occurring 223 times in the dataset (52.6%). 398 lesion numbers (92.6%) are between one and five, and fourteen lesion numbers (3.3%) are greater than ten.

C Details of labeling

C.1 Loose labels

Each MedVTE premise image consists of one or more subfigures that are often excerpts of a vast series of radiological, pathological, or endoscopic images. Therefore, it must be considered that the premise image may not reflect the entire patient and may contain only a subset of the lesions that are actually present, or conversely, the same lesion

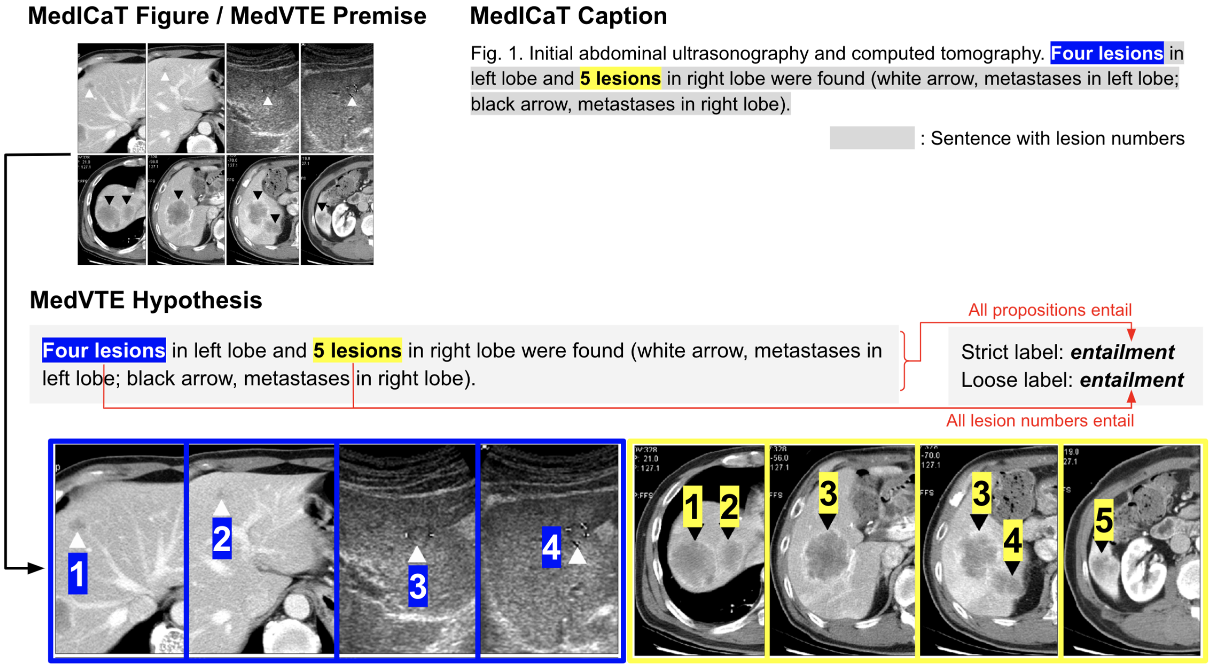
MedICaT Figure / MedVTE Premise
MedICaT Caption
MedVTE Hypothesis
: Sentence with lesion numbers
All propositions entail
All lesion numbers entail
1
2
3
4
1
2
3
3
4
5

Fig. 1. Initial abdominal ultrasonography and computed tomography. Four lesions in left lobe and 5 lesions in right lobe were found (white arrow, metastases in left lobe; black arrow, metastases in right lobe).

Four lesions in left lobe and 5 lesions in right lobe were found (white arrow, metastases in left lobe; black arrow, metastases in right lobe).

Strict label: ***entailment***
Loose label: ***entailment***

Figure 4: Another example of MedVTE. The four subfigures outlined in yellow apparently have *six* lesions. However, the medical expert has determined that the yellow subfigures demonstrate *five* lesions and assigned *entailment* label because it is explainable that the lesion numbered "3" repeatedly appears in the second and third subfigures at the different levels.

may repeatedly appear across multiple subfigures as in Figure 4. This phenomenon is prevalent not only in the medical articles from which MedVTE originates but also in the real-world clinical practice that we target for application.

We regard each hypothesis as a set of propositions. For each proposition addressing the lesion number in the hypothesis sentence, the following procedure was employed to determine the veracity or falsity.

(a) If the medical expert determines that the quantities are equal in the premise figure and the hypothesis sentence, the proposition is supported.

(b) When the lesion number in the hypothesis sentence apparently exceeds that in the premise figure, the medical expert is requested to carefully review the premise figure and determine if the gap can be explained by the following reason:

- The original caption is correct, but the medical expert initially missed some lesions due to subtle image findings.

If so, the hypothesis is supported. Otherwise, the loose label is *neutral* because it is impossible to judge which of the following is happening:

- The original caption is correct, but the premise figure does not show all the lesions
- The original caption has overcounted the lesions.

(c) When the lesion number in the hypothesis sentence appears to be smaller than the premise figure, the medical expert is asked to examine the premise figure again and determine which of the following is the most convincing:

- The original caption is correct, but the medical expert initially overcounted the lesions due to equivocal image findings
- The original caption is correct, but the medical expert initially overcounted the lesions due to the same lesion repeatedly appearing across multiple subfigures
- The original caption has undercounted the lesions.

In the first or second case, the hypothesis is supported. In the last case, the loose label is *contradiction*.

C.2 Strict labels

When a hypothesis contains propositions that cannot be judged true or false from the premise image alone, we consider it out-of-figure information. The following are examples of propositions that we regard as out-of-figure information:

- Mention to other figures than the premise figure (e.g., “show no enhancement on arterial phase images of MRI and on the arterial phase of CT scan (not shown)”)
- Numerical values for elapsed time, such as days, months, or years (e.g., “Axial contrast-enhanced CT *six weeks* pre-RF ablation (a) demonstrates two lesions”)
- Specific lesion size numbers (e.g., “The two nodules were *1.2 cm* in diameter”).

If the hypothesis sentence includes out-of-figure information, we set the strict label to *neutral* regardless of the loose label. Otherwise, the strict label is the same as the loose label.

D Model details

ViLT is pre-trained on MSCOCO (Lin et al., 2014)+VG (Krishna et al., 2017)+CC (Sharma et al., 2018)+SBU (Ordonez et al., 2011). FLAVA is pre-trained on filtered YFCC100M (Thomee et al., 2015)+CC12M (Changpinyo et al., 2021)+WIT (Srinivasan et al., 2021)+Red-Caps (Desai et al., 2021)+LN (Pont-Tuset et al., 2020)+MSCOCO+VG+CC+SBU.

We basically adopted models and parameters implemented in transformers[2]. We attached a 2-layer classifier head ourselves for FLAVA since there was no model implementation for classification tasks in the library. Table 2 and Table 3 show hyperparameters in ViLT and FLAVA models, respectively.

Hyperparameter	Value
Encoder	
hidden size	768
number of heads	12
number of layers	12
intermediate size	3072
dropout probability	0
patch size	32×32
input image size	384×640
Classifier Head	
hidden size	768
Others	
text vocabulary size	30522
Training	
epochs	3
gradient accumulation steps	3
per device batch size	48
learning rate	5e-05
AdamW weight decay	0
AdamW β_1	0.9
AdamW β_2	0.999

Table 2: Hyperparameters in ViLT

[2] https://huggingface.co/docs/transformers/v4.20.1/en/index

Hyperparameter	Value
Image Encoder	
hidden size	768
number of heads	12
intermediate size	3072
number of layers	12
dropout probability	0
patch size	16×16
input image size	224×224
Text Encoder	
hidden size	768
number of heads	12
intermediate size	3072
number of layers	12
dropout probability	0
Multimodal Encoder	
hidden size	768
number of heads	12
intermediate size	3072
number of layers	6
dropout probability	0
Classifier Head	
hidden size	1536
Others	
text vocabulary size	30522
image dVAE codebook size	8192
Training	
epochs	3
gradient accumulation steps	3
per device batch size	24
learning rate	1e-05
learning rate schedule	linear
warmup updates	2000
AdamW weight decay	1e-02
AdamW β_1	0.9
AdamW β_2	0.999

Table 3: Hyperparameters in FLAVA