Uni-NLX: Unifying Textual Explanations for Vision and Vision-Language Tasks

Fawaz Sammani and Nikos Deligiannis
ETRO Department, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
imec, Kapeldreef 75, B-3001 Leuven, Belgium
fawaz.sammani@vub.be, ndeligia@etrovub.be

Abstract

Natural Language Explanations (NLE) aim at supplementing the prediction of a model with human-friendly natural text. Existing NLE approaches involve training separate models for each downstream task. In this work, we propose Uni-NLX, a unified framework that consolidates all NLE tasks into a single and compact multi-task model using a unified training objective of text generation. Additionally, we introduce two new NLE datasets: 1) ImageNetX, a dataset of 144K samples for explaining ImageNet categories, and 2) VQA-ParaX, a dataset of 123K samples for explaining the task of Visual Question Answering (VQA). Both datasets are derived leveraging large language models (LLMs). By training on the 1M combined NLE samples, our single unified framework is capable of simultaneously performing seven NLE tasks including VQA, visual recognition and visual reasoning tasks with 7× fewer parameters, demonstrating comparable performance to the independent task-specific models in previous approaches, and in certain tasks even outperforming them.[1]

1. Introduction

Moving away from general and high-level explanations such as heatmaps [29, 31, 3, 30], Natural Language Explanations (NLE)[2] [6, 19] offer a detailed, human-friendly textual format explanation. Recently, NLE has been extended to encompass vision and vision-language (VL) tasks [22, 36, 17, 11]. The general pipeline comprises a vision model to encode the image, a task model M_T to generate a prediction for the task at hand (*e.g.,* answer for VQA, class for image classification) and an explainer model M_E which takes the form of a language model to produce an explanation for the prediction via natural text. A subsequent study [26] unifies M_T and M_E into a single compact model that performs both tasks simultaneously by converting all tasks into generative tasks with a single casual language modeling training objective (Figure 1a). This greatly reduces the number of parameters and inference time and associates the reasoning process of M_E to the same answer prediction process in M_T. It also attributes to the fact that explainability techniques are applied on the *same* model responsible for generating the prediction. However, both these approaches require separate finetuning on each NLE task. This results in N separately-parameterized models for N tasks of NLE. Moreover, it requires a separate specialized model to perform each task. In this work, we build upon the work of [26] and consolidate all NLE tasks into a single compact model, dubbed as Uni-NLX (Figure 1b). This unification offers several advantages that previous approaches lack: Firstly, it offers a single model to simultaneously perform all N NLE tasks, thereby requiring $N\times$ less parameters. Secondly, the integration enables mutual learning among all NLE tasks, as they possess similar reasoning capabilities. Lastly, the shared information across diverse tasks enables greater flexibility in answers and explanations (*e.g.,* free-form text generation).

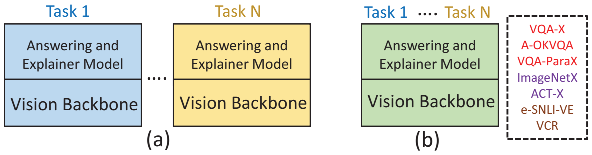


Figure 1. The current SoTA model (a) [26] unifies the answering and explainer models into a single compact model, training separate models for each of the N tasks. Our proposed approach (b) takes a further step by unifying all tasks into a single compact model, resulting in $N\times$ fewer parameters. Our single unified model is capable of simultaneously handling diverse tasks ranging from Visual Question Answering, Visual Recognition and Visual Reasoning.

Furthermore, we propose to leverage knowledge from Large Language Models (LLMs) to obtain two additional NLE datasets: *VQA-ParaX* and *ImageNetX*. VQA-ParaX is

[1] https://github.com/fawazsammani/uni-nlx

[2] it is worth noting that ”explanations” in this context do not refer to explanations of the underlying decision-making process of a model as typical in post-hoc explainability methods, but rather to supplementary information concerning the predicted outcome, incorporated through training

a re-formulation of long-text captioning datasets (*e.g.,* Image Paragraph Captioning [13] or Local Narratives [23]) into question-answer-explanation formats using LLMs in a scalable manner. Moreover, LLMs posses vast knowledge about the world, and can be leveraged to obtain fine-grained, distinctive features and descriptions about different objects. ImageNetX is a dataset encompassing such textual data, which are regarded as explanations for ImageNet [14] categories.

The integration of these two additional datasets with the existing NLE datasets results in a total of 7 NLE datasets, containing approximately 1M (image, text) pairs. The textual component of these pairs comprises the question, answer, and explanation. By training on these pairs, Uni-NLX achieves performance levels comparable to state-of-the-art task-specific NLE models on 4 tasks, while surpassing them on 3 tasks.

2. Related Work

Early works in NLE for vision and vision-language tasks include [10, 22, 15, 36, 17, 11]. They rely on a task model (*e.g.,* UNITER [7]) for multimodal feature extraction and answer prediction, and an explainer model (*e.g.,* GPT-2 [25]) to generate an explanation for the prediction. Most recently, NLX-GPT [26] proposed to unify both these models into a single, compact-sized model (*e.g.,* Distilled-GPT-2) that simultaneously generates and explains an answer using a single casual language modelling objective, while also eliminating the computationally-expensive object-level feature extraction stage [2]. This generative formulation has also proven to be effective in vision-language pretraining methods such as VL-T5 [8], OFA [34] and GIT [33]. Multimodal-CoT [38] builds upon the Chain of Thought Prompting [35] technique and instead generates a rationale (explanation) prior to generating an answer, which serves as a reasoning step for inferring the answer. However, the aforementioned methods require training or finetuning for each task individually, which consequently leads to separately-parameterized models specialized to each task. Different from these methods, our work unifies all tasks into a single compact-sized model, greatly reducing parameters and computational cost.

The authors of [18] perform zero-shot visual classification by measuring the similarity between an image and various distinctive textual features that describe the object in the image. These descriptors are obtained from LLMs. However, this approach relies on a strong retrieval model (*e.g.,* CLIP [24]) and does not have the ability to generate text. Additionally, it is primarily aimed to vision-only tasks. In contrast, our method generates flexible free-form answers and explanations for both vision and vision-language tasks.

3. Method

Following NLX-GPT [26], we formulate the discriminative answer prediction task as a generative text prediction task, along with the explanation. Both the answer and explanation tasks are unified into the model which outputs a single sequence containing the answer and explanation in a textual form. We first describe how we construct additional NLE datasets, and then elaborate on our multi-task unified model.

3.1. Data Synthesis Strategies

We propose to harness the powerful reasoning capabilities of LLMs to formulate two additional NLE datasets: *VQA-ParaX* and *ImageNetX*, in a scalable manner. We utilize GPT-3 [5] with instructional finetuning [20] (ChatGPT) as our LLM.

VQA-ParaX: LLMs posses remarkable ability in reading and re-formulating passages such as summarization and information extraction. The image paragraph captioning dataset [13] contains 19,561 samples and provides detailed descriptions of images which allows the LLM to gain a complete understanding of the image solely through the textual description. Using a LLM, we re-formulate the image paragraph captioning dataset into question-answer-explanation formats. We prompt the LLM with <$\mathtt{I}$, S^i>, where S^i represents the paragraph sample, and $\mathtt{I}$ represents the instruction given to the LLM. For each sample i, we formulate 6 question-answer-explanation triplets, resulting in approximately 123K triplet samples. The instruction $\mathtt{I}$ we use is provided in the supplementary material.

ImageNetX: ImageNet-1K [14] is a dataset used for image classification containing 1K categories. LLMs posses wealth knowledge about the world, which can be harnessed to obtain distinctive features and descriptions of various objects at a granular level. We propose to obtain such textual descriptions from LLM for the ImageNet-1K categories, which are then regarded as explanations for the class category (answer). We prompt the LLM with <$\mathtt{I}$,$\mathtt{c}$>, where $\mathtt{I}$ represents the instruction and $\mathtt{c} \in C$ represents the class category for each of the 1K categories C. We generate 50 descriptions for each class c. In order to account for variations in visual representations of the same textual description within a given class, we assign three distinct training images per description for each class. Consequently, this approach yields a dataset of approximately 141K training samples. The remaining 3K textual descriptions are associated to a single image from the ImageNet validation set, and are divided into validation and test set. The instruction $\mathtt{I}$ we use is provided in the supplementary material.

We provide further analysis, quality assessment and qualitative samples of these two new datasets in the supplementary material.

3.2. Unifying Explanations

To achieve a unified NLE framework across diverse tasks, it is necessary to establish a standardized format of question-answer-explanation. However, certain tasks (*e.g.,* visual recognition) lack inherent questions. To address this, we introduce a consistent question relevant to each task, such as *"What category is this?"* for image recognition, *"What action is this?"* for action recognition, or *"is the following hypothesis true or false?"* for visual entailment. By employing this unified format, all tasks can be formulated using the sequence S: `<question> the answer is <answer> because <explanation>`. The compilation of all available datasets yields a collective corpus of approximately 1M samples. During training, we provide S as input to the model and predict the answer and explanation component of S in an autoregressive manner, utilizing a single causal language modeling training objective with cross-entropy loss. During inference, only the question is fed into the model, which subsequently predicts the answer and explanation using greedy decoding. It is worth noting that the answer can also be provided during inference, in which case the model solely generates the explanation. To allow the model to distinguish between the question, answer and explanation components of S, we utilize three different segment embeddings for each.

4. Experiments

Our unified dataset comprises seven NLE datasets encompassing visual question answering (VQA), vision recognition and visual reasoning tasks. VQA tasks consists of VQA-X [22] (33K samples), A-OKVQA [28] (25K samples) and VQA-ParaX (123K samples). Visual recognition tasks include ACT-X [22] (18K samples) for action recognition and ImageNetX (144K samples) for image classification. Visual reasoning tasks comprises e-SNLI-VE [11] (430K samples) for visual entailment and Visual Commensense Reasoning (VCR) [37] of 192K samples. To establish a fair comparison, our model follows NLX-GPT [26], which uses a distilled version [27] of the GPT-2 transformer language model [5] as the answering and explanation model, and a CLIP visual encoder part [24] as the visual backbone. Our model is trained for a maximum of 20 epochs with a batch size of 64 and a learning rate of 2e-5 which decays linearly to 0.

4.1. Quantitative Results

We evaluate our model quantitatively using automatic natural language generation (NLG) metrics (BLEU [21], METEOR [4], ROUGE-L [16], CIDER [32] and SPICE [1]); all scores are computed with the publicly available code[3]. Following previous works, the evaluation is carried on in two settings: *filtered* and *unfiltered*. In the filtered setting, we only consider the explanations for which the predicted answer is correct. In the unfiltered setting, all explanations are considered, irrespective of whether the

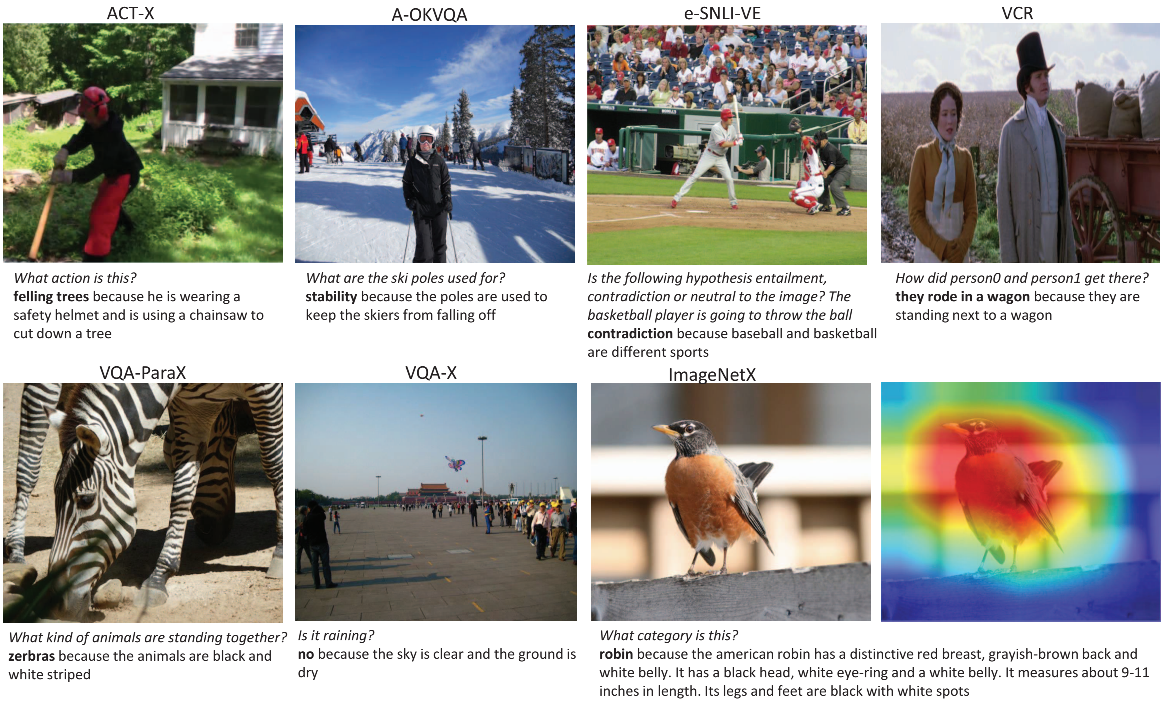
Table 1. Unfiltered Scores for Uni-NLX compared to NLX-GPT [26] on the 7 downstream tasks. Both models are w/o pretraining. B-N, M R, C, S are short for: BLEU-N, METEOR, ROUGE-L, CIDER and SPICE.

	VQA-X							
	B-1	B-2	B-3	B-4	M	R	C	S
NLX-GPT	**59.1**	**43.8**	**32.2**	**23.8**	**20.3**	**47.2**	**89.2**	**18.3**
Uni-NLX	57.9	42.1	30.2	21.7	19.3	45.9	81.1	17.8
	ACT-X							
NLX-GPT	64.4	47.5	34.7	25.6	21.4	48.0	63.5	15.4
Uni-NLX	**65.4**	**49.1**	**36.0**	**26.5**	**22.0**	**48.5**	**67.7**	**16.7**
	e-SNLI-VE							
NLX-GPT	34.3	22.7	15.6	10.9	17.5	31.7	**106.6**	**31.5**
Uni-NLX	**35.3**	**23.6**	**16.5**	**11.8**	**17.8**	**32.2**	106.5	31.3
	VQA-ParaX							
NLX-GPT	**37.1**	**27.0**	**20.4**	**15.5**	**18.5**	**40.9**	**142.6**	31.4
Uni-NLX	35.1	25.7	19.4	14.8	18.2	40.8	139.9	**31.6**
	A-OKVQA							
NLX-GPT	55.0	**39.9**	**29.3**	**20.2**	16.4	**46.2**	**64.4**	15.2
Uni-NLX	**58.2**	39.6	27.6	18.5	**17.1**	44.0	58.1	**16.0**
	ImageNetX							
NLX-GPT	**64.5**	**48.1**	**36.9**	**28.9**	**22.0**	**39.4**	**87.5**	**22.4**
Uni-NLX	62.9	46.3	35.2	27.4	21.4	38.7	82.8	21.3
	VCR							
NLX-GPT	18.5	9.7	5.4	**3.3**	**9.0**	**19.9**	24.2	12.4
Uni-NLX	**18.7**	**9.9**	**5.7**	**3.5**	**9.0**	**19.9**	**24.7**	**12.5**

Table 2. Filtered Scores for Uni-NLX compared to NLX-GPT [26] on the 7 downstream tasks. Both models are w/ pretraining.

	VQA-X							
	B-1	B-2	B-3	B-4	M	R	C	S
NLX-GPT	**64.2**	**49.5**	**37.6**	**28.5**	**23.1**	**51.5**	**110.6**	**22.1**
Uni-NLX	62.1	46.8	34.9	26.0	21.8	48.8	97.8	20.8
	ACT-X							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**71.6**	56.2	43.2	**33.5**	**25.7**	**53.7**	**111.8**	**23.3**
Uni-NLX	71.5	**56.7**	**43.6**	**33.5**	**25.7**	53.5	109.4	22.8
	e-SNLI-VE							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**35.7**	24.0	16.8	11.9	18.1	33.4	114.7	**32.1**
Uni-NLX	35.3	**24.1**	**17.0**	**12.3**	**18.2**	**33.7**	**115.4**	**32.1**
	VQA-ParaX							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**41.9**	**31.5**	**24.7**	**19.9**	**22.3**	**47.2**	**203.7**	41.9
Uni-NLX	41.3	31.2	24.5	19.7	22.0	**47.2**	203.6	**42.1**
	A-OKVQA							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**62.3**	**46.8**	**36.1**	**27.7**	**20.5**	**51.5**	**93.0**	19.3
Uni-NLX	62.1	43.3	30.8	20.8	19.6	48.1	78.1	**19.7**
	ImageNetX							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	69.7	54.1	42.5	33.8	24.7	43.1	107.4	26.1
Uni-NLX	**71.9**	**56.5**	**45.0**	**36.1**	**25.8**	**44.8**	**117.2**	**27.3**
	VCR							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	-	-	-	-	-	-	-	-
Uni-NLX	29.7	23.4	19.9	17.4	17.1	33.6	85.7	23.5

[3] https://github.com/tylin/coco-caption



ACT-X

What action is this?
felling trees because he is wearing a safety helmet and is using a chainsaw to cut down a tree

A-OKVQA

What are the ski poles used for?
stability because the poles are used to keep the skiers from falling off

e-SNLI-VE

Is the following hypothesis entailment, contradiction or neutral to the image? The basketball player is going to throw the ball
contradiction because baseball and basketball are different sports

VCR

How did person0 and person1 get there?
they rode in a wagon because they are standing next to a wagon

VQA-ParaX

What kind of animals are standing together?
zerbras because the animals are black and white striped

VQA-X

Is it raining?
no because the sky is clear and the ground is dry

ImageNetX

What category is this?
robin because the american robin has a distinctive red breast, grayish-brown back and white belly. It has a black head, white eye-ring and a white belly. It measures about 9-11 inches in length. Its legs and feet are black with white spots

Figure 2. Qualitative Examples of Uni-NLX on the 7 NLE tasks. We show the *question*, **answer** and explanation under each image.

predicted answer associated with each explanation is true or false. We utilize the recent state-of-the-art NLX-GPT [26] model as our baseline for evaluating our approach. NLX-GPT also presents results of NLE tasks by fine-tuning a pretrained model on image captioning. In our study, we consider this setting and utilize the pretrained model provided by the official code[4]. Table 1 presents the unfiltered results of Uni-NLX without finetuning the pretrained model, while Table 2 reports the filtered results obtained after finetuning the pretrained model. Additional results on unfiltered results with pretraining and filtered results without pretraining can be found in the supplementary material. In Table 1, Uni-NLX demonstrates superior performance compared to NLX-GPT on ACT-X, e-SNLI-VE, and VCR. Additionally, Uni-NLX achieves performance that is comparable to NLX-GPT across all VQA tasks (VQA-X, VQA-ParaX, and A-OKVQA) and ImageNetX, and surpasses NLX-GPT on certain metrics. Table 2 shows that Uni-NLX outperforms NLX-GPT on e-SNLI-VE and ImageNetX and demonstrates comparable performance to other tasks, and in certain metrics even outperforms them. It is worth noting that NLX-GPT does not present unfiltered results on VCR.

4.2. Qualitative Results

Figure 2 shows qualitative results for each of the seven NLE tasks. As observed, our model generates an answer to the given question and image, supported by a detailed explanation. We discuss limitations such as collapse cases in the supplementary material. For ImageNetX, we additionally show a heatmap visualization obtained from ResNet-18 [9] using Grad-CAM [29]. Compared to heatmap-based visualization techniques which only display high-level, general and entangled features influencing the prediction, Uni-NLX provides detailed and fine-grained explanations describing distinctive and disentangled features influencing the prediction (*e.g.,* red breast, grayish-brown back, black with white spots) in the form of human-friendly text. Furthermore, the attribution maps associated with these distinctive textual attributes have the potential to represent concept activation vectors [12], with the exception that in our case, these concepts are derived automatically from the image rather than obtained manually through annotators. We intend to investigate this avenue in future research.

5. Conclusion

We proposed Uni-NLX, a unified model which simultaneously performs seven NLE tasks. Leveraging a LLM, we also introduced two additional NLE datasets: VQA-Parax for the VQA task, and ImageNetX for the ImageNet recognition task. Experiments demonstrate that Uni-NLX achieves comparable performance to task-specific models in certain tasks, while surpassing them on others. In the future, we plan to investigate adapter models and prompt learning techniques to perform zero-shot NLE.

[4] https://github.com/fawazsammani/nlxgpt

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 3

[2] Peter Anderson, X. He, C. Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 2

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10, 2015. 1

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, 2005. 3

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 2, 3

[6] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In *NeurIPS*, 2018. 1

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, A. E. Kholy, Faisal Ahmed, Zhe Gan, Y. Cheng, and Jing jing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 2

[8] Jaemin Cho, Jie Lei, Haochen Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021. 2

[9] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[10] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *ECCV*, 2016. 2

[11] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. *ArXiv*, abs/2105.03761, 2021. 1, 2, 3

[12] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, 2017. 4

[13] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3345, 2017. 2

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. 2

[15] Qing Li, Qingyi Tao, Shafiq R. Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. *ArXiv*, abs/1803.07464, 2018. 2

[16] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*, 2004. 3

[17] Ana Marasović, Chandra Bhagavatula, Jae Sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *FINDINGS*, 2020. 1, 2

[18] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *International Conference on Learning Representations*, abs/2210.07183, 2023. 2

[19] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions. *ArXiv*, abs/2004.14546, 2020. 1

[20] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, pages 27730–27744, 2022. 2

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2001. 3

[22] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018. 1, 2, 3

[23] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 2

[24] 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. In *ICML*, 2021. 2, 3

[25] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2

[26] Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. Nlx-gpt: A model for natural language explanations in

vision and vision-language tasks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8312–8322, 2022. 1, 2, 3, 4

[27] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. 3

[28] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *European Conference on Computer Vision (ECCV)*, 2022. 3

[29] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2019. 1, 4

[30] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014. 1

[31] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *ArXiv*, abs/1703.01365, 2017. 1

[32] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2014. 3

[33] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *Trans. Mach. Learn. Res.*, 2022, 2022. 2

[34] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, 2022. 2

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022. 2

[36] Jialin Wu and Raymond J. Mooney. Faithful multimodal explanation for visual question answering. *ArXiv*, abs/1809.02805, 2019. 1, 2

[37] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724, 2019. 3

[38] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alexander J. Smola. Multimodal chain-of-thought reasoning in language models. *ArXiv*, abs/2302.00923, 2023. 2