Probing Multimodal Large Language Models for Global and Local Semantic Representations

Mingxu Tao[1]†, Quzhe Huang[1]†, Kun Xu[2], Liwei Chen[2], Yansong Feng[1]✉, Dongyan Zhao[1]
[1]Peking University [2]Kuaishou Technology
{thomastao, huangquzhe, zhaodongyan}@pku.edu.cn
syxu828@gmail.com chenliwei03@kuaishou.com
✉fengyansong@pku.edu.cn

Abstract

The advancement of Multimodal Large Language Models (MLLMs) has greatly accelerated the development of applications in understanding integrated texts and images. Recent works leverage image-caption datasets to train MLLMs, achieving state-of-the-art performance on image-to-text tasks. However, there are few studies exploring which layers of MLLMs make the most effort to the global image information, which plays vital roles in multimodal comprehension and generation. In this study, we find that the intermediate layers of models can encode more global semantic information, whose representation vectors perform better on visual-language entailment tasks, rather than the topmost layers. We further probe models regarding local semantic representations through object recognition tasks. We find that the topmost layers may excessively focus on local information, leading to a diminished ability to encode global information. Our code and data are released via https://github.com/kobayashikanna01/probing_MLLM_rep.

Keywords: Probing Study, Interpretability, Multimodal Large Language Model

1. Introduction

Recently, Large Language Models (LLMs) have achieved remarkable advancements in various natural language processing applications (Touvron et al., 2023; OpenAI, 2024), owing to pre-training on massive text corpus. It becomes a popular topic nowadays to transfer the powerful capacity of LLMs to Multimodal Large Language Models (MLLMs) through image-caption corpus (Alayrac et al., 2022; Li et al., 2023). These MLLMs show an impressive ability to handle multimodal tasks, including Image Captioning (IC, Plummer et al., 2015) and Visual Question Answering (VQA, Goyal et al., 2017). However, existing research has predominantly focused on the ability of MLLMs to generate single tokens one by one, while lacking investigations about how their representation vectors can encode global multimodal information. In generation tasks like IC and VQA, when predicting the next token, the models may only need to focus on a local part of the image and a subsequence of the text to handle the task. But in tasks like image-text retrieval (Xie et al., 2019), the MLLMs should aim to encode the global semantic information of the entire image and text, when predicting whether they have correlation.

In this work, we focus on understanding and uncovering how the global and local semantic information is encoded in the decoder-only MLLMs. To track the representing ability of each layer in MLLMs, we use probing study, a popular tool to investigate model interpretability (Tenney et al., 2019; Jawahar et al., 2019). Previous probing studies of pure-text language models have explored the representing ability of models in various levels, from local to global semantics (Liu et al., 2019; Talmor et al., 2020). However, to the best of our knowledge, existing works about vision-language models sorely focus on investigating the ability to represent local semantic information, for instance, from a lexical perspective (Dahlgren Lindström et al., 2020). We also note that previous works (Ma et al., 2022; Dai et al., 2023b) mainly study the encoder-only or encoder-decoder models, with less than 1B parameters, such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2022). Our work investigates the representing ability of decoder-only MLLMs, from both global and local perspectives, thus bridging a gap in prior works.

Our main contributions in this paper are:

(1) We design an image-text entailment task to probe MLLM's ability to encode global cross-modal information and design a pair of prompts for object recognition to study local representation.

(2) We find that, when encoding global information, it is the intermediate layers rather than the topmost layers that perform the best.

(3) Through the probing study of local representations, we find the topmost layers may excessively focus on local information, leading to a diminished ability to encode global information.

(4) To the best of our knowledge, we are the first to find and discuss the potential shortcomings

†This work was done when Mingxu Tao and Quzhe Huang were interns at Kuaishou Technology.

LREC-COLING 2024, pages 13050–13056
20-25 May, 2024.

of decoder-only MLLMs in representing global semantic information. We hope our findings could encourage the community to explore ways to improve the pre-training process of MLLMs, and even to improve the architecture designs of MLLMs.

2. Related Works

Exploiting the local and global semantic representations is commonly employed in the processing of image data (Bian et al., 2017; Lv et al., 2019; Chen et al., 2021; Zhao and Zhou, 2022). By adjusting the receptive field size of CNN layers, the model can capture information at various granularities (Simonyan and Zisserman, 2015; Dai et al., 2023a). However, in the MLLMs, the structure of each Transformer layer can usually be similar or the same to others. Therefore, we wonder how MLLMs represent the local and global information, especially when the inputs are sequences of visual tokens but not matrices of pixels.

Previous studies (Chi et al., 2020; Vulić et al., 2020) in pre-trained language models (PLMs, i.e., BERT) employ probing tasks to investigate which layer in the model make the most effort to encode lexical, syntactic, or semantic information. These works reveal that the lower layers in BERT can encode lexical information, while the upper ones tend to encode syntactic and semantic information. In this work, we follow previous works and employ multimodal probing tasks to study the granularity of information represented by each layer in decoder-only MLLMs.

We also note that there are other methods available for studying the representing mechanisms of LLMs. For example, Sajjad et al. (2023) propose removing specific layers of PLMs to investigate their effects by comparing the performance of the models before and after removal. Previous works (Kovaleva et al., 2019; Rogers et al., 2021) also use neuron-wise examinations and visualization methods to provide detailed analyses. Although these methods are mainly implemented on encoder-only PLMs, we believe that they may provide insights for future research on the interpretability of MLLMs.

3. Global Multimodal Representation

We first aim to investigate how each layer can encode the global cross-modal semantic information. Motivated by natural language inference, where the alignment between global meanings of two sentences plays a vital role (MacCartney et al., 2008; Tay et al., 2018), we design an image-text entailment task whose goal is to decide whether a caption can entail a given image or not.

We thus build a dataset based on MS COCO (Lin et al., 2014), which contains more than 200K labeled images and five captions for each image. Formally, we denote the images as $\mathcal{M} = \{m_i | i = 0, 1, \cdots\}$ and the five human-written caption texts of m_i as $\mathcal{T}_i = \{t_{i,k} | k = 0, \cdots, 4\}$.

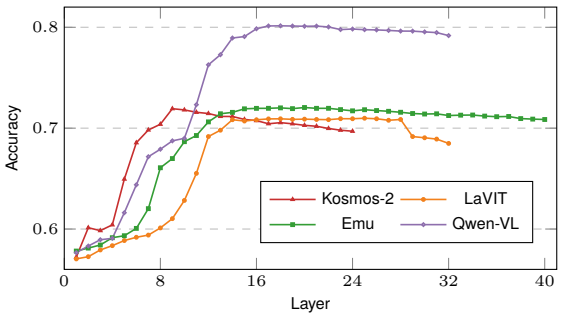


Figure 1: Performance on the image-text entailment task when using the representations at each layer.

We define this image-text semantic entailment task as a binary classification task. For each image $m_i \in \mathcal{M}$, we select all its captions $\mathcal{T}_i$ to construct positive image-text pairs. Furthermore, we also use the image m_i and captions sampled from $\bigcup_{j \neq i} \mathcal{T}_j$ to form negative examples. For each positive pair $\langle m_i, t_{i,k} \rangle$, we randomly sample 5,000 captions from $\bigcup_{j \neq i} \mathcal{T}_j$. To ensure a balanced number of positive and negative examples, we select one negative caption that has the highest similarity[1] with $t_{i,k}$ as the negative sample.

Following previous probing studies (Hupkes et al., 2018; Jawahar et al., 2019), we freeze all parameters of the multimodal large language model (MLLM). We use the following prompt to combine the image and caption pairs as input: `[Image]` *This image describes "*`[Caption]`*". Is it right? Answer:*. We then extract the hidden-state features generated by each layer of MLLM, and take the vectors corresponding to the last tokens as representations of the whole inputs. For the L-th layer, whose feature vector can be denoted as $\mathcal{H}_L$, we train a binary classifier $f_L : \mathcal{H}_L \mapsto \{0, 1\}$. In this paper, we employ single-layer linear classifiers for experiments and use Adam (Kingma and Ba, 2015) as the optimizer.

We examine several popular decoder-only MLLMs, including Kosmos-2 (Peng et al., 2023), LaVIT (Jin et al., 2023), Emu(Sun et al., 2023) and Qwen-VL (Bai et al., 2023), with parameter scales ranging from 7B to 14B. We employ *accuracy* to measure the extent to which the representation vectors can encode information for the image-text entailment task. Experimental results are illustrated in Figure 1. From the results, we find represen-

[1]The similarity is calculated by model `all-mpnet-base-v2` (https://huggingface.co/sentence-transformers/all-mpnet-base-v2).

tation vectors of the topmost layers do not yield optimal performance. For instance, in Kosmos-2, a model consisting of 24 Transformer layers, we find that its representation vectors generated by the 9th layer demonstrate the best performance in the image-text entailment task. Similarly, the 23rd layer in LaVIT, the 20th layer in Emu, and the 18th layer in Qwen-VL achieve the best performance, all of which are not the topmost layers in their corresponding models. As the depth of layers increases to the topmost, all models' abilities to encode global multimodal information exhibit a diminishing trend.

In light of previous research, which demonstrates the upper layers in BERT can possess the strongest ability to represent global semantic information (Jawahar et al., 2019; Koto et al., 2021), we intuitively hypothesize that the same phenomenon may appear in MLLMs. However, our experimental results display a deviation from the conclusions of prior works on encoder-only PLMs.

Revisiting the pre-training process of decoder-only MLLMs, we find there is a gap between their pre-training objective and the ability to encode global semantic information. Since models learn how to generate the sequence token by token, the representation vectors encoded by upper layers may inherently focus more on information related to the local token which will be generated next, rather than all the context tokens. For MLLMs that have been pre-trained but without being fine-tuned on downstream tasks, their predicted tokens in zero-shot scenarios may not always perform well in addressing complex tasks that need global information. Hence, focusing on encoding the local semantic features of such tokens does not contribute to addressing the image-text entailment task. This may be the reason why representation vectors of intermediate layers outperform the upper layers.

4. Local Multimodal Representation

To investigate whether the upper layers encode more local information about the token to be generated than the lower layers, we employ the MS COCO dataset again and conduct an object recognition task. MS COCO comprises annotations for 80 distinct categories. For an image m_i, its annotated object category list can be denoted as $\{\mathcal{O}_k|k \in \mathbb{I}\}$, where $\mathbb{I} \subseteq \{0, 1, \cdots, 79\}$ is an indicator set denoting the categories of objects present in image m_i. In this work, we regard the recognition task for different object categories as 80 separate binary classification tasks, in which the model needs to predict whether the input image contains a specific type of object.

We first study whether the feature vectors encoded by each layer can be used to predict a specific object category, when we provide sufficient

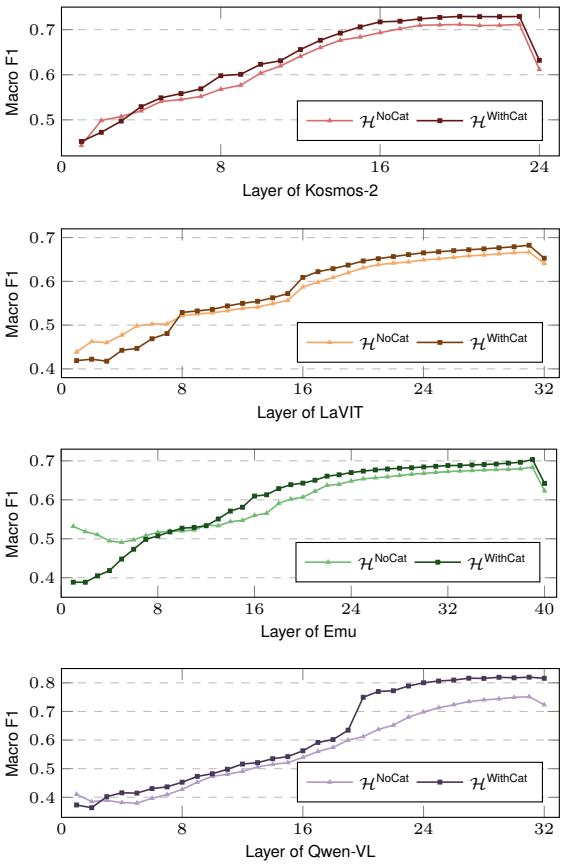


Figure 2: Performance on the object recognition task when using the representations at each layer of different MLLMs.

cues of other categories. It is intuitive that if there are n types of objects in an image and we provide $(n-1)$ categories of them in the text input, a well pre-trained MLLM should then output the n-th remaining category. Thus, we extract the vision-language features by following prompt: `[Image]` *This image contains the following types of objects:* `[Obj_1], [Obj_2], ..., [Obj_n-1],`. Similar to the entailment task, we also freeze the parameters of MLLMs and collect representation vectors of the last token as features of the whole input sequence. We denote the representation vectors of layer L as $\mathcal{H}_L^{\text{WithCat}}$. It is important to note that, to prevent data leakage, when training and evaluating the probing model for the category c, the input object list for image m_i should be $\{\mathcal{O}_k|k \in \mathbb{I} \wedge k \neq c\}$. To mitigate the potential impact arising from the order of input object categories, we shuffle the object lists during both training and evaluation.

In the probing study, as we look at higher layers, the improvement of model performance might be attributed to the expansion of the model's parameter scale, resulting in enhanced representing ability. To eliminate the influence of scale expansion on performance, we formulate another set of experiments to perform object recognition without any

Prompt without Categories				Prompt with Categories			
Pos. Set		Neg. Set		Pos. Set		Neg. Set	
Token	Freq.	Token	Freq.	Token	Freq.	Token	Freq.
A	.9662	A	.9532	man	.1063	and	.1250
a	.0079	a	.0129	people	.1017	grass	.0416
black	.0063	zebra	.0102	woman	.0636	building	.0281
two	.0035	Gir	.0059	and	.0514	street	.0279
tennis	.0022	two	.0044	tennis	.0431	mirror	.0229
snowboarder	.0020	elephant	.0017	baseball	.0373	tree	.0227
an	.0018	an	.0009	person	.0294	water	.0187
baseball	.0015	brown	.0009	beach	.0242	window	.0170
Two	.0013	Two	.0007	boy	.0238	animal	.0161
skateboard	.0013	bananas	.0007	skateboard	.0213	plate	.0157
OTHERS	.0059	OTHERS	.0087	OTHERS	.4979	OTHERS	.6643

Table 1: Frequency of the top 10 frequently generated tokens.

category cues, serving as the baseline. We use the following prompt: `[Image]` *This image contains the following types of objects:*. The vector set of layer L extracted by this prompt is denoted as $\mathcal{H}_L^{\text{NoCat}}$.

Similar to the entailment task, we also examine the four large-scale multimodal models, including Kosmos-2, LaVIT, Emu, and Qwen-VL. Due to the imbalance in the ratio of positive to negative examples in the object recognition tasks, we employ the Macro Average F1 score across all categories as our evaluation metric. The experimental results are illustrated in Figure 2. From the results, we first find that, across the upper layers of all four MLLMs, the probing models trained on $\mathcal{H}^{\text{WithCat}}$ outperform the ones trained on $\mathcal{H}^{\text{NoCat}}$. However, in the lowermost layers, providing several categories as input can hurt the probing model's performance. The results probably indicate that the upper layers, those closer to the token probability prediction layer, tend to encode more local features of the tokens to be decoded, rather than global semantic information.

To further validate the hypothesis, it is necessary to examine whether given a subset of object categories as input, the model indeed produces tokens that are relevant to the remaining categories present in the image. We take the `person` category and the Kosmos-2 model as a case study. We randomly select 10,000 examples from the test set, with 5,409 of them containing the `person` objects (positive examples), while the remaining images do not (negative examples). We employ the two prompts to extract $\mathcal{H}^{\text{NoCat}}$ and $\mathcal{H}^{\text{WithCat}}$ as input, capturing the first new tokens generated by Kosmos-2. We then examine the frequency distributions of the generated tokens on positive and negative examples separately. In Table 1, we list the statistical results for the top 10 most frequently generated tokens for each setting.

We can find when employing the prompt that includes all object categories except `person`, there is a significant difference in the distributions of the first tokens generated by the model for positive and negative examples. We note that, in the case of positive examples, **5 out of the top 10** most frequently generated tokens have meanings associated with *person* (red-colored), while all of the 10 tokens of negative examples lack semantic relevance to *person*.

Nevertheless, when using the prompt without providing any categories, the model generates *"A"* or *"a"* with a frequency exceeding 96%, both in the positive and negative example sets. We also find among the top 10 frequent tokens, several of them convey meanings corresponding to the object categories other than `person`, such as *tennis*, *snowboarder*, *zebra*, and other tokens that are colored by blue. It indicates that the representation vectors may randomly encode one category of the objects appearing in the image.

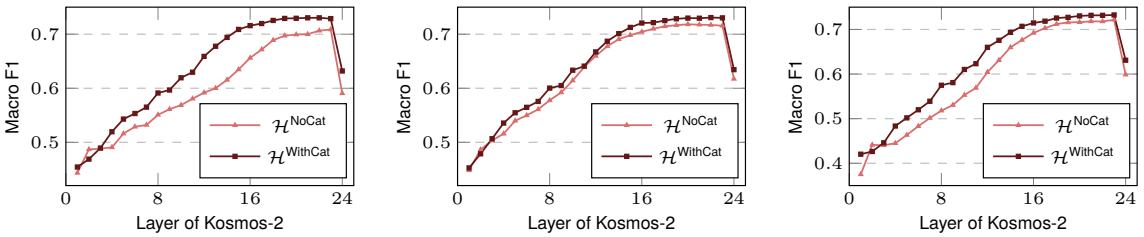
By comparing the results of positive and negative sets generated by the two prompts separately, we can infer that the topmost layer of a MLLM can be effective in representing the local semantic features of the token to be decoded.

Furthermore, we note the model's performance continuously improves from the lowermost to the second-to-last layer, while it significantly declines in the topmost layer. We revisit the frequency distributions of the first generated tokens. We can find there is an overlap in the tokens generated by the model for positive and negative examples, such as *"A"*, *"a"*, *"and"*, *"skateboard"*, and etc. These overlapping tokens may indicate the model produces indistinguishable representation vectors, which negatively affect the performance of probing models. We conjecture that, compared to preceding layers, representation vectors in the topmost layer of a MLLM may lose certain global semantic information but shift their focus towards specific tokens to be outputted, although these tokens may not have relevant meanings to *person*. This could also be the reason why the intermediate layers, rather than the topmost layers, perform better in the image-text entailment tasks.

5. Results of More Prompts

Through a pair of cross-modal prompts, we find in the decoder-only MLLMs, the deficiency of upper layers in encoding global semantic information may arise from that such layers focus excessively on the local information of one token. In order to examine the robustness of our findings, we also conduct experiments with different prompts to probe the model's ability to perform object recognition tasks. The employed prompts are listed in Table 2. These prompts use diverse forms of expression and possess varying lengths.

We replace the prompts in Section 4 with them and implement experiments based on the same settings. We take Kosmos-2 as an instance, and the results based on each prompt are shown in Figure 3. Comparing the results in Figure 2 and Figure 3, we find despite using different prompts,



(a) Results of Prompt 1. (b) Results of Prompt 2. (c) Results of Prompt 3.

Figure 3: Performance on object recognition of the representations extracted with various prompts.

Prompt 1: `[Image]` *What types of objects are there here? Please list them:* `[Obj_1], [Obj_2], ..., [Obj_k],`
Prompt 2: `[Image]` *Objects in this picture are:* `[Obj_1], [Obj_2], ..., [Obj_k],`
Prompt 3: `[Image]` *There can be several types of objects in this image, including up to eighty kinds of objects. These objects can be any color, including red, green, blue, orange, yellow, purple, pink, and etc. Some of these objects can be very huge, while others can be very small. In the meantime, there are also many objects which can be overlapping with others. Please look carefully at the image for any detailed information. Now, you can write which type of objects you can find in the image:* `[Obj_1], [Obj_2], ..., [Obj_k],`

Table 2: Variant prompts to extract representation vectors for object recognition.

the Kosmos-2 model performs consistently across all groups of experiments. It indicates that our findings can be prompt-agnostic.

6. Conclusion

In this paper, we investigate how the decoder-only MLLMs represent the global and local cross-modal semantic information, through prompt-based probing study. We experiment with four open-source models, extracting representation vectors using various prompts. Our findings remarkably remain consistent across diverse models and prompts, which indicates **the upper layers in MLLMs focus too much on** the semantic features of **the next token** to be generated. It may result in a loss of global information in the upper layers. Our findings shed light on understanding the potential mechanism of MLLMs to represent global and local features. We hope this paper can inspire our community to delve into more effective pre-training mechanisms for MLLMs.

7. Acknowledgement

This work is supported in part by NSFC (62161160339) and Kuaishou. We would like to thank the anonymous reviewers for their helpful comments and suggestions.

8. Bibliographical References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *CoRR*, abs/2308.12966.

Xiaoyong Bian, Chen Chen, Long Tian, and Qian Du. 2017. Fusing local and global features for high-resolution scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(6):2889–2901.

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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mohan Chen, Xinxuan Zhao, Bingfei Fu, Li Zhang, and Xiangyang Xue. 2021. Rethinking local and global feature representation for semantic seg-

mentation. In *The 32nd British Machine Vision Conference*.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Adam Dahlgren Lindström, Johanna Björklund, Suna Bensch, and Frank Drewes. 2020. Probing multimodal embeddings for linguistic properties: the visual-semantic case. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 730–744, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2023a. R-fcn: Object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023b. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2136–2148, Dubrovnik, Croatia. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *J. Artif. Int. Res.*, 61(1):907–926.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. 2023. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *CoRR*, abs/2309.04669.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yafei Lv, Xiaohan Zhang, Wei Xiong, Yaqi Cui, and Mi Cai. 2019. An end-to-end local-global-fusion feature extraction network for remote sensing image scene classification. *Remote Sensing*, 11(24).

Zheng Ma, Shi Zong, Mianzhi Pan, Jianbing Zhang, Shujian Huang, Xinyu Dai, and Jiajun Chen.

2022. Probing cross-modal semantics alignment capability from the textual perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5739–5749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 802–811, Honolulu, Hawaii. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *CoRR*, abs/2303.08774.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *CoRR*, abs/2306.14824.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2023. On the effect of dropping layers of pre-trained transformer models. *Comput. Speech Lang.*, 77(C).

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *CoRR*, abs/2307.05222.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-On What Language Model Pre-training Captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1565–1575, Brussels, Belgium. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

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

Yuchi Zhao and Yuhao Zhou. 2022. Fuse local and global semantics in representation learning. *CoRR*, abs/2202.13837.