IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Harmonizing Multimodal Tasks: High-Efficiency Visual Encoding for Understanding and Generation

M.Santhoshkumar[1], V.Divya[2]

[1]Research Scholar, School of Computing Science, Vels Institute of Science, Technology & Advanced Studies, Pallavaram, Chennai, India.

[2]Assistant Professor, Department of Computer Applications (UG), School of Computing Science, Vels Institute of Science, Technology & Advanced Studies, Pallavaram, Chennai, India.

Abstract—Unified multimodal models usually rely on a single visual encoder to handle both understanding and generation. However, this shared approach forces a compromise between the distinct levels of information granularity each task requires, often degrading multimodal understanding. To resolve this bottleneck, we present **Janus**, an autoregressive framework that **decouples visual encoding into separate pathways** while maintaining a single, unified transformer for processing. By untangling these visual roles, Janus eliminates task conflict and introduces unparalleled flexibility, allowing understanding and generation components to utilize their optimal encoding methods independently. Empirical evaluations demonstrate that Janus outpaces existing unified models and matches or exceeds the performance of specialized, task-specific architectures. Its simplicity and flexibility position Janus as a foundational architecture for next-generation unified multimodal systems.

1. Introduction

In recent years, multimodal large models have made significant advancements in both under-standing and generation domains. In the field of multimodal understanding, researchers follow the design of LLaVA by using a vision encoder as a bridge to enable large language models (LLMs) to understand images. In the field of visual generation, diffusion-based ap-proaches have seen notable success. More recently, some works have explored autoregressive methods for vision generation, achieving performance comparable to diffusion models. To build more powerful and generalist multimodal models, researchers have sought to combine multimodal understanding and generation tasks. For instance, some studies have attempted to connect multimodal understanding models with pretrained dif-fusion models. For example, Emu uses the output of the LLM as a condition for a pretrained diffusion model, and then relies on the diffusion model to generate images. However, strictly speaking, this approach cannot be considered a truly unified model, because the visual generation functionality is handled by the external diffusion model, while the multimodal LLM itself lacks the capability to directly generate images.

The output of understanding task not only involves extracting information from images but also involves complex semantic reasoning. Therefore, the granularity of the vision encoder's representation tends to mainly focus on high-dimensional semantic representation. By contrast, in visual generation tasks, the main focus is on generating local details and maintaining global consistency in the image.

The representation in this context necessitates a low-dimensional encoding that is capable of fine-grained spatial structure and textural detail expression. Unifying the representations of these two tasks within the same space will lead to conflicts and trade-offs. Consequently, existing unified models for multimodal understanding and generation often compromise on multi-modal understanding performance, falling markedly short of the state-of-the-arts multimodal understanding models. We explore this issue further in the ablation study.

To the best of our knowledge, we are the first to highlight the importance of decoupling visual encoding within the unified multimodal understanding and generation framework. Our experimental results show that Janus surpasses existing unified models with comparable parameter sizes on both multimodal understanding and generation benchmarks, achieving state-of-the-art results.

2. Related Work

2.1. Visual Generation

Visual generation is a rapidly evolving field that combines concepts from natural language processing with advancements in transformer architectures. Autoregressive models, influenced by the success in language processing, leverage transformers to predict sequences of discrete visual tokens (codebook IDs). These models tokenize visual data and employ a prediction approach similar to GPT-style techniques. Additionally, masked prediction models draw upon BERT-style masking methods, predicting masked sections of visual inputs to improve synthesis efficiency, and have been adapted for video generation. Concurrently, continuous diffusion models have showcased impressive capabilities in visual generation, complementing discrete methods by approaching generation through a probabilistic lens.

2.2. Multimodal Understanding

Multimodal large language models (MLLMs) integrate both text and images. By leveraging pretrained LLMs, MLLMs demonstrate a robust ability to understand and process multimodal information. Recent advancements have explored extending MLLMs with pretrained diffusion models to facilitate image generation. These methods fall under the category of tool utilization, where diffusion models are used to generate images based on the conditions output by the MLLM, while the MLLM itself does not have the ability to directly perform visual generation. Moreover, the generative ability of the entire system is often constrained by the external diffusion model, making its performance inferior to directly using the diffusion model on its own.

2.3. Unified Multimodal Understanding and Generation

Unified multimodal understanding and generation models are considered powerful for fa-cilitating seamless reasoning and generation across different modalities. Traditional approaches in these models typically use a single visual representation for both understanding.

3. Janus: A Simple, Unified and Flexible Multimodal Framework

3.1. Architecture

For pure text understanding, multimodal understanding, and visual generation, we apply independent encoding methods to convert the raw inputs into features, which are then processed by an unified autoregressive transformer. Specifically, for text understanding, we use the built-in tokenizer of the LLM to convert the text into discrete IDs and obtain the feature representations corresponding to each ID. For multimodal understanding, we use the SigLIP encoder to extract high-dimensional semantic features from images. These features are flattened from a 2-D grid into a 1-D sequence, and an understanding adaptor is used to map these image features into the input space of the LLM. For

visual generation tasks, we use the VQ tokenizer from to convert images into discrete IDs. After the ID sequence is flattened into 1-D, we use a generation adaptor to map the codebook embeddings corresponding to each ID into the input space of the LLM. We then concatenate these feature sequences to form a multimodal feature sequence, which is subsequently fed into the LLM for processing. The built-in prediction head of the LLM is utilized for text predictions in both the pure text understanding and multimodal understanding tasks, while a randomly initialized prediction head is used for image predictions in the visual generation task. The entire model adheres to an autoregressive framework without the need for specially designed.

3.2. Training Procedure

Stage I: Training Adaptors and Image Head. The main goal of this stage is to create a conceptual connection between visual and linguistic elements within the embedding space, enabling the LLM to understand the entities shown in images and have preliminary visual generation ability. We keep the visual encoders and the LLM frozen during this stage, allowing only the trainable parameters within the understanding adaptor, generation adaptor and image head to be updated.

Stage II: Unified Pretraining. In this stage, we perform unified pretraining with multimodal corpus to enable Janus to learn both multimodal understanding and generation. We unfreeze the LLM and utilize all types of training data: pure text data, multimodal understanding data, and visual generation data. Inspired by Pixart, we begin by conducting simple visual generation training using ImageNet-1k to help the model grasp basic pixel dependencies. Subsequently, we enhance the model's open-domain visual generation capability with general text-to-image data.

Stage III: Supervised Fine-tuning. During this stage, we fine-tune the pretrained model with instruction tuning data to enhance its instruction-following and dialogue capabilities. We fine-tune all parameters except the generation encoder. We focus on supervising the answers while masking system and user prompts. To ensure Janus's proficiency in both multimodal understanding and generation, we don't fine-tune separate models for a certain task. Instead, we use a blend of pure text dialogue data, multimodal understanding data and visual generation data, ensuring versatility across various scenarios.

3.3. Inference

During inference, our model adopts a next-token prediction approach. For pure text under-standing and multimodal understanding, we follow the standard practice of sampling tokens sequentially from the predicted distribution.

3.4. Possible Extensions

It is important to note that our design, which features separate encoders for understanding and generation, is straightforward and easy to extend.

Multimodal Understanding. (1) For the multimodal understanding component, a stronger vision encoder can be chosen without worrying about whether the encoder is capable of handling vision generation tasks, such as EVA-CLIP, InternViT, etc. (2) To handle high-resolution images, dynamic high-resolution techniques can be used. This allows the model to scale to any resolution, without performing positional embedding interpolation for ViTs. Tokens can be further compressed to save computational cost, for instance, using pixel shuffle operation.

Visual Generation. (1) For visual generation, finer-grained encoders can be chosen in order to preserve more image details after encoding, such as MoVQGan. (2) Loss functions specifically designed for visual generation can be employed, such as diffusion loss. (3) A combination of AR (causal attention) and parallel (bidirectional attention) methods can be used in the visual generation process to reduce accumulated errors during visual generation.

Support for Additional Modalities. The straightforward architecture of Janus allows for easy integration with additional encoders, accommodating various modalities such as 3D point cloud, tactile, and EEG. This gives Janus the potential to become a more powerful multimodal generalist model.

4. Experiments

In this section, we present a series of comprehensive experiments designed to assess the perfor-mance of our method across a range of visual understanding and generation tasks. We begin by detailing our experimental setup, which includes the model architecture, training datasets, and evaluation benchmarks. Next, we report the performance of Janus, followed by a comparison with other state-of-the-art models on various benchmarks for multimodal understanding and generation. We also conduct extensive ablation studies to verify the effectiveness of the proposed method. Lastly, we provide some qualitative results.

4.1. Data Setup

In this section, we provide details of the pretraining and supervised finetuning datasets.

Stage I. We use a dataset that includes 1.25 million image-text paired captions from ShareGPT4V for multimodal understanding and approximately 1.2 million samples from ImageNet-1k for visual generation. The ShareGPT4V data is formatted as "<image><text>". The ImageNet data is ~~organized into a text-to-image data format using the category names: "<category_name><ima~~ge>". Here, the "<>" symbols represent placeholders.

Stage II. We organize the data into the following categories. (1) Text-only data. We use pre-training text copus from DeepSeek-LLM. (2) Interleaved image-text data. We use Wiki-How and WIT dataset. (3) Image caption data. We use images from. Among them, we employ open-source multimodal model to re-caption images in. The image caption data is formatted into question-answer pairs, for ex-ample, "<image>Describe the image in detail.<caption>". (4) Table and chart data. We use corresponding table and chart data from DeepSeek-VL. The data is formatted as "<question><answer>". (5) Visual generation data. We utilize image-caption pairs from various datasets including, along with 2M in-house data. For images from, we filter based on aesthetic scores and image sizes, resulting in 20% remaining. During training, we randomly use only the first sentence of a caption with a 25% probability to encourage the model to develop strong generation capabilities for short descriptions. ImageNet samples are presented only during the first 120K training steps, while images from other datasets appear in the later 60K steps. This approach helps the model first learn basic pixel dependencies before progressing to more complex scene understanding, as suggested by. The visual generation data is provided in the format: "<caption><image>".

Stage III. For text understanding, we use data from. For multimodal understanding, we use instruct tuning data from. For visual generation, we use image-text pairs from (a subset of that in stage II) and 4M in-house data. We utilize the following format for instruction tuning:"User:<Input Message> \n Assistant: <Response>". For multi-turn dialogues, we repeat this format to structure the data.

4.2. Comparison with State-of-the-arts

Multimodal Understanding Performance. We compare the proposed method with state-of-the-art unified models and understanding-only models in Table 2. Janus achieves the overall best results among models of similar scale. Specifically, compared to the previous best unified model, Show-o [86], we achieve performance improvements of 41% (949 → 1338) and 30% (48.7 → 59.1) on the MME and GQA datasets, respectively. This can be attributed to Janus decoupling the visual encoding for multimodal understanding and generation, mitigating the conflict between these two tasks. When compared to models with significantly larger sizes, Janus remains highly competitive. For instance, Janus outperforms LLaVA-v1.5 (7B) on several datasets, including POPE, MMbench, SEED Bench, and MM-Vet.

Visual Generation Performance. We report visual generation performance on GenEval, COCO-30K and MJHQ-30K benchmarks. As shown in Table 3, our Janus obtains 61% overall accuracy on GenEval, which outperforms the previous best unified model Show-o (53%) and some popular generation-only methods, e.g., SDXL (55%) and DALL-E 2 (52%). This demonstrates that our approach has better instruction-following capabilities. As shown in Table 4, Janus achieves FIDs of 8.53 and 10.10 on the COCO-30K and MJHQ-30K benchmarks, respectively, surpassing unified models Show-o and LWM, and demonstrating competitive performance compared to some well-known generation-only methods. This demonstrates that the images generated by Janus have good quality and highlights its potential in visual generation.

4.3. Qualitative Results

Visualizations of Visual Generation provides qualitative comparisons between our model, diffusion-based models like SDXL, and the autoregressive model LlamaGen. The results show that our model demonstrates superior instruction-following capabilities in visual generation, accurately capturing most of details in the user's prompt. This indicates the potential of the unified model in the realm of visual generation. More visualizations can be found.

Multimodal Understanding on MEME Images. Janus accurately interprets the text caption and captures the emotion conveyed in the meme. In contrast, both Chameleon and Show-o struggle with accurately recognizing the text in the image. Additionally, Chameleon fails to identify objects in the meme, while Show-o misinterprets the dog's color. These examples highlight that the decoupled vision encoder significantly enhances Janus's fine-grained multimodal understanding ability compared to the shared encoder used by Chameleon and Show-o. More multimodal understanding exmples can be found.

5. Conclusion

In this paper, we introduced Janus, a simple, unified and extensible multimodal understanding and generation model. The core idea of Janus is to decouple visual encoding for multimodal understanding and generation, which could alleviate the conflict arising from the differing demands that understanding and generation place on the visual encoder. Extensive experi-ments have demonstrated the effectiveness and leading performance of Janus. It is also worth noting that Janus is flexible and easy to extend. In addition to having significant potential for improvement in both multimodal understanding and generation, Janus is also easily extendable to incorporate more input modalities. The above advantages suggest that Janus may serve as an inspiration for the development of the next generation of multimodal general-purpose models.

References

[1] Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://www.anthropic. com, 2024.

[2] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A fron-tier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966, 2023.

[3] Y. Bai, X. Wang, Y.-p. Cao, Y. Ge, C. Yuan, and Y. Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. arXiv preprint arXiv:2306.16934, 2023.

[4] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954, 2024.

[5] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, et al. Pixart-
alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv

preprint arXiv:2310.00426, 2023.

[6] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin. Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793, 2023.

[7] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821, 2024.

[8] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24185–24198, 2024.

[9] X. Chu, L. Qiao, X. Lin, S. Xu, Y. Yang, Y. Hu, F. Wei, X. Zhang, B. Zhang, X. Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. arXiv preprint arXiv:2312.16886, 2023.