Optimizing Waste Classification through Large-Scale Language Models in Deep Learning

Jianchun Qi, Minh Nguyen, Wei Qi Yan
Department of Computer and Information Sciences
Auckland University of Technology, New Zealand

ABSTRACT

Effective waste management is vital for resource conservation, environmental protection, and sustainable human progress. Traditional classification methods struggle to keep pace with the diverse and complex nature of waste materials, this leads to inefficiencies in sorting and recycling processes. To address these challenges, we introduce a novel multimodal approach that harnesses the capabilities of large language models (LLMs) to enrich waste classification with semantic insights. Our method generates descriptive prompts specifically tailored to the waste imagery, which are then used to infer semantic attributes relevant to classification tasks. The descriptions facilitate a transformative converter architecture that bridges textual and visual domains, enabling our model to interpret waste images with enhanced precision. We present the first multimodal waste classification model that leverages the LLM-generated textual descriptors alongside visual features. Extensive testing shows that our approach outperforms existing models, which achieves a top accuracy improvement of 62.20%. A comprehensive suite of ablation studies further underscores the method's efficacy and resilience, confirming its potential to advance waste management by integrating the complementary strengths of both image and textual data. This dual-dimension assistance not only elevates classification accuracy but also underscores the broader implications of AI's role in environmental technologies.

Keywords: LLM, Context information, Image processing, Waste classification

INTRODUCTION

Waste management is an increasingly important topic. Inadequate waste management mechanisms pose a great challenge to the protection of the ecological environment, the improvement of public health, and the safeguarding of human health. For example, the current open waste dumps, this waste management method is prone to produce hazardous chemicals (Mohanraj, Senthilkumar, Chandrasekar & Arulmozhi, 2023), pollute the soil and water, and damage the ecosystem, in addition to becoming a breeding ground for pathogens, which can easily lead to the spread of infectious diseases (Amasuomo & Baird, 2016) (Ferronato & Torretta, 2019). Proper waste disposal practices hold significant implications for ecological sustainability, resource efficiency, and public health enhancement. At present, waste classification is a crucial step in waste management, aiding in resource recycling and minimizing resource depletion. It contributes to reducing the reliance on waste incineration and landfills, thereby lessening pollution and safeguarding ecosystems (Gundupalli, Hait & Thakur, 2017). However, traditional methods of waste classification, which are typically semi-manual or semi-automatic, struggle to keep pace with the increasing volumes of waste, often resulting in inefficient sorting and adverse health effects on workers. Consequently, there is a pressing need to incorporate more sophisticated technologies, such as artificial intelligence, into waste management. The integration of advanced AI-driven classification techniques can lead to more effective, efficient, and health-conscious waste management practices. This, in turn, supports economic growth and environmental protection, steering us toward the sustainable coexistence of humanity and nature (Qiu et al., 2022) (Shi, Tan, Wang & Wang, 2021).

Although deep learning models for waste classification are constantly being improved and have obtained significant classification and detection results, there is still room for improvement. A slew of waste classification models, such as the optimized DenseNet121 and ResNet-10 using fusion schemes, have waste classification accuracies as high as over 85% (Ahmad, Khan & Al-Fuqaha, 2020) (Mao, Chen, Wang & Lin, 2021). However, the datasets they use are only simple recyclable waste categories, such as glass, cardboard, plastic, paper, and metal, which cannot measure the real waste classification application scenarios. According to the waste classification standard, waste should be classified into four categories, namely, recyclable waste, wet waste, dry waste, and hazardous waste. While the ETHSeg model groups the four categories of waste based on X-rays, the classification accuracy of small objects in waste remains low (Qiu et al., 2022). Thus, the lack of a waste dataset, the low accuracy of small object waste classification, and the intensive manual annotation effort due to the wide variety of waste categories are the important challenges faced by artificial intelligence in waste classification tasks.

Artificial Intelligence (AI), particularly large-scale language models, has shown remarkable promise in various applications, including image classification tasks. The cutting-edge GPT-4, for instance, can perform image classification, text-to-image conversion, and image-to-text translation. Its superior generalization and zero-shot learning abilities enable the processing of complex datasets with high efficiency. Building on this potential, our book chapter introduces a pioneering approach to waste classification by harnessing the semantic capabilities of large language models. We utilize MiniGPT-4 to generate textual descriptions of waste images, then input into the pre-trained language model RoBERTa (Liu et al., 2019) (Zhu, Chen, Shen, Li & Elhoseiny, 2023). Concurrently, we process the waste images directly through the Swin Transformer model (Liu et al., 2021). The culmination of our efforts is the novel Image-Text Aware Adaptive Attention mechanism, which integrates outputs from these dual pathways.

Despite these advancements, large language models are not without their challenges in practical scenarios. They are composed of multi-layer neural networks with hundreds of millions of parameters, necessitating substantial computational resources and extended training periods. This complexity results in considerable training expenses. Furthermore, an increase in model parameters can complicate the model's interpretability and elevate its complexity (Singla, 2023).

Therefore, to solve this problem and avoid manually collecting image description information from a large language model, we introduced MiniGPT-4 into our model through an API interface, aiming to leverage the rich semantics of the large language model in a simplified way to create an efficient and highly accurate waste classification model. Our description-driven approach to image classification shows promise, particularly when image data is scarce, making it well-suited for the task of waste classification. Furthermore, the quality of datasets is foundational to the training and validation of machine learning models. The diversity, balance, and consistency within datasets are crucial to guarantee high efficiency and strong generalization capabilities in model training. To this end, we have curated our own dataset, named WasteNet, to contribute a wider variety of waste images to the domain of waste classification. Our contributions are listed as follows:

- By combining with a large language model, the rich semantic information of MiniGPT-4 is converted into waste image data, which improves the accuracy of waste classification. And the effectiveness of our model has been verified through many experiments;

- We created an in-depth waste dataset, WasteNet, which contains four categories of waste classification: Recyclable waste, hazardous waste, wet waste, and dry waste. In the dataset, the status of waste is variable, and there are patterns such as obscured and folded. The dataset conforms to the waste classification standards and is closer to the real waste classification scenarios, which can provide more diversified and richer data for the waste classification task;
- An attention mechanism suitable for our model, Image-Text Aware Adaptive Attention, was designed to integrate image and text features to provide the model with the necessary nonlinear and decision-making capabilities.

RELATED WORK

Waste detection

Waste detection is becoming popular. Using deep learning for waste classification has a few advantages, such as scalability, high accuracy, and convenience (Altikat, Gulbe & Altikat, 2022) (Huang, He, Xu & Huang, 2020) (Adedeji & Wang, 2019) (Zhou et al., 2022). The waste classification models based on deep neural networks have been continuously proposed. MobileNetV2 was trained for waste classification with an accuracy of 82.92% (Yong, Ma, Sun & Du, 2023). After that, the EnCNN-UPMWS model generated by combining CNN with an unequal precision measurement weighting strategy improves the waste classification accuracy to 92.85% by the two key points of weight coefficients and predicted probability vectors (Zheng & Gu, 2021). Other than the two models based on traditional CNN structure, the up-and-coming Transformer model is also applied for the waste classification task. Compared to CNN, the advantage of the self-attention mechanism, which is not limited by local interactions, allows Vision Transformer to achieve 96.98% in waste classification (Huang, Lei, Jiao & Zhong, 2021). All three models for waste classification models achieved more than 80% accuracy.

However, the dataset has a simple background, fewer types of waste data, and a lack of stacked waste states. Even if the algorithms were improved, the real form of waste (such as deformed waste) was less considered, which may make the classification results uneven. Therefore, we collected our own waste dataset and tried to apply MiniGPT-4 to the waste classification task, making full use of the emergent ability of the large language model to improve the waste classification, and contribute to the environmental protection and the sustainable development of humans and nature.

Large language model

Large language model (LLM) aims to generally refer to deep learning models with many parameters and complex structures, which can contain millions or even hundreds of millions of parameters and are able to handle large-scale data, learn more complex features, and have more powerful generalization ability and higher accuracy (Kasneci et al., 2023) (Liu et al., 2023) (Thirunavukarasu et al., 2023) (Waisberg et al., 2023). At present, large language models are gradually becoming the mainstream development direction in the field of deep learning. The large-scale multilingual and multimodal machine translation model, SeamlessM4T, which can perform speech translation between up to 100 languages (Barrault, 2023). Afterwards, a large language model that can role-play animated characters was proposed, called ChatHaruhi (Li et al., 2023). In addition to the domains of speech translation and role-playing chat robots, large language models also have excellent performance in writing code, modifying code bugs, and textual question and answer, such as the highly regarded ChatGPT model (OpenAI, 2023).

The large language models were explored in specific fields and the usage of large language models for intelligent change in healthcare are of much interest. The ability to write postoperative patient discharge summaries and recognize images of patient lesions using GPT-4 was demonstrated to have the potential to aid medical innovation (OpenAI, 2023) (Waisberg et al., 2023). Subsequently, the application of GPT-4 in biomedical engineering has been explored, which has demonstrated excellent performance in the areas of medical devices, bioinformatics, and medical imaging (Cheng et al., 2023). Finally, large language model can also be applied to healthcare, such as providing users with healthcare-related information support for weight loss and mental health (Egli, 2023). It is thus conjectured that GPT-4 has unlimited potential to help other domains.

LLMs combined with vision tasks

Large language models combined with vision tasks also gradually being developed. Continuous improvements in large language models have evolved the functionality from processing text to processing visual images, bringing significant benefits to many text-to-image interaction tasks. Visual ChatGPT connects a range of visual foundation models into ChatGPT, which enables users to interact with ChatGPT in the form of text and images (Chen, Guo, Yi, Li & Elhoseiny, 2022). It also provides complex visual instructions that allow multiple models to work together. Visual ChatGPT can also understand and respond to both text-based and vision-based inputs, while reducing the barriers to accessing text-to-image models. Then, Google proposed the multimodal visual language model PaLM-E, which has 562 billion parameters (Driess et al., 2023). Based on the language model, PaLM-E performs continuous observation, e.g., receives image or sensor data and encodes it into a series of vectors of the same size as the language token. In this way, PaLM-E can continue to understand sensory information in the same way it processes language. The success of these models validates the future possibilities of such multimodal models.

METHODOLOGY

The structure of our framework

Deep learning models, such as CNN, have performed well on computer vision tasks over the past years (Krizhevsky, Sutskever & Hinton, 2012). It is able to learn features directly from image data. This means that the model can predict the output directly from the input data without any human intervention at intermediate steps in the training process. Recently, it has been confirmed that large language models have excellent performance (Chen, Guo, Yi, Li & Elhoseiny, 2022) (Driess et al., 2023) (OpenAI, 2023) (OpenAI, 2023), therefore, we speculate that applying large language models to image classification models by introducing multimodal information so that the image classification model is not limited to learning features only from the image data, which will bring about a performance improvement. Our model framework is shown in Figure 1.

Firstly, we input our waste image to MiniGPT-4, which can generate a detailed description of the image. At this point, we give the prompt "Describe these images for waste classification" that is designed to fit the prompt more closely to our task, thus increasing the accuracy of waste classification. In this step, we take the approach of describing the images by introducing the MiniGPT-4 API. The purpose of this method is to save resource allocation without affecting the classification performance. Hence, we did not choose to adopt the better-performing GPT-4 model since GPT-4 has not opened the API function for the image-to-text module. Later, we feed all the image descriptions generated by using MiniGPT-4 to the RoBERTa model, which converts the text descriptions into high-dimensional embedding vectors that provide semantic classification decisions for the images based on their descriptions (Liu et al., 2019).

Simultaneously, we also send the image data to Swin Transformer for direct image classification. Then, we add a new attention mechanism layer, Image-Text Aware Adaptive Attention, as a way of fusing two different feature representations of the visual features generated by using Swin Transformer and the text descriptive features of the image generated by MiniGPT-4. Image-Text Aware Adaptive Attention serves to dynamically assign weights based on the input data to improve the flexibility of the model, which further improves the accuracy of the model.

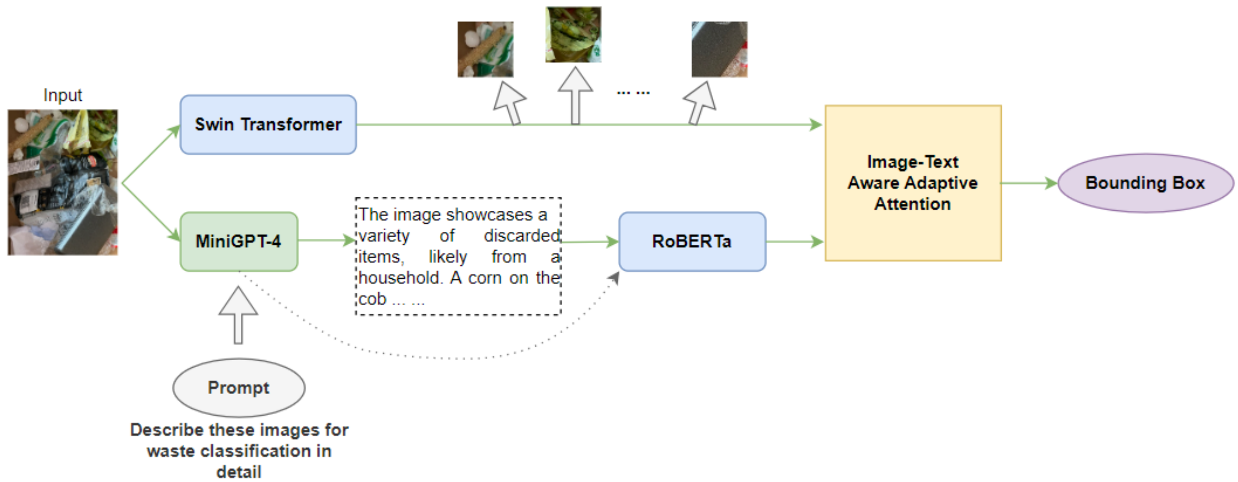

Figure 1. The framework

Image descriptions generated by using large language model

Our proposed multimodal waste classification model can integrate textual descriptions and image processing content for waste classification. In the entire model, generating image descriptions through MiniGPT-4 is one of the most important aspects. First, the waste image and the prompt related to this image are inputted into the model to get detailed description information of the waste image. During this process, even if the same image is input into MiniGPT-4, the image description generated is different every time.

Therefore, we conjectured that different descriptions generated for the same image would have multiple effects on the model training results. Thus, we explored the effects of different lengths of image descriptions and different prompts of input on the model results in our ablation experiments. The specific experimental details are shown in Section 5. Based on the experimental results, we finally chose to define the prompt as "Describe these images for waste classification in detail". Moreover, we selected the description information of the image as shown in Figure 2. As can be seen, "description1" gives key information about the waste in the image. Whereas "description2" lacks information about some objects such as corn cobs and plastic bags. Therefore, if we choose to define the prompt as "Describe these images in detail", the accuracy of the model may be affected.

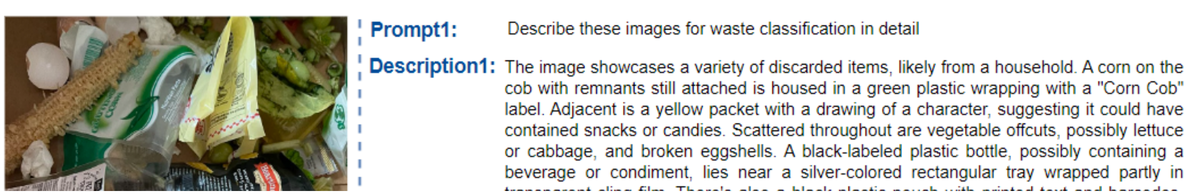

Figure 1. The structure of our model.

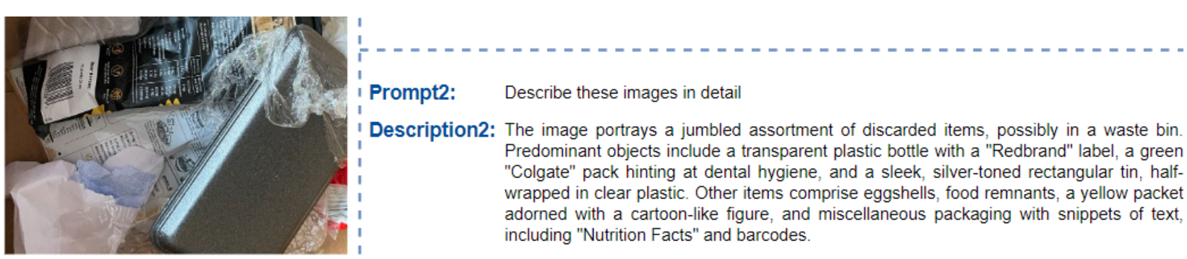

Figure 2. The image descriptions generated by using MiniGPT-4 with various prompts.

Image-text aware adaptive attention

Since our model leverages multimodal capabilities to integrate both image and textual information, it is essential to effectively fuse the image features with the text features derived from the model. This enables the model to consider information from both modalities when making classification decisions. An effective method for this fusion is the use of an attention mechanism (Guo et al., 2022) (Lieskovská, Jakubec, Jarina & Chmulík, 2021) (Yan et al., 2019) (Zhu, Cheng, Zhang, Lin & Dai, 2019). In this study, we implement an Image-Text Aware Adaptive Attention technique, which enriches sequence representations by amalgamating global and local contextual information. Notably, this approach is able to dynamically allocate weights based on the characteristics of the input data, enhancing the model's flexibility to handle various tasks and inputs, and consequently improving its accuracy.

For illustration, taking into consideration of Figure 2, which uses the prompt "Describe these images for waste classification in detail". Despite this detailed description, a number of image elements, such as the corn in the upper left corner, may not be captured in full detail. Yet, if this corn is categorized as wet waste in the image classification, the Image-Text Aware Adaptive Attention mechanism can adjust the weights to refine the classification accuracy for such items. It is anticipated that this mechanism will not only benefit items like the corn, which may be overlooked by the image classification model, but also detailed in the text by the large language model. Therefore, the deployment of a dynamically adaptive attention mechanism can significantly enhance the model's accuracy.

This attention mechanism has three modules, the first one is dynamic weight calculation module, which is responsible for dynamically calculating the weights according to the characteristics of the input data, considering the output feature matrix of Swin Transformer as *X* and the output feature matrix of RoBERTa as *Y*. To better utilize the characteristics of the two lines, W_1 is defined as the weight of *X*, W_2 is defined as the weight of *Y*. Then Eq. 1, and Eq. 2 show the dynamic weight calculation function as follows:

$$W_1 = \frac{e^{\frac{2}{n}\sum_{i=1}^{n} x_i} - 1}{e^{\frac{2}{n}\sum_{i=1}^{n} x_i} + 1} \tag{1}$$

$$W_2 = \frac{e^{\frac{2}{m}\sum_{j=1}^{m}(Y_i - mean(Y))^2} - 1}{e^{\frac{2}{m}\sum_{j=1}^{n}(Y_j - mean(Y))^2} + 1} \tag{2}$$

where *n* and *m* represent the number of elements in the *X* and *Y*, respectively, *i* is the index of the anchor in the image.

The second is the weight normalization module, which ensures that the sum of the weights is 1.0, so that the relative importance of the two lines can be intuitively explained, which is shown in Eq. 3 and Eq. 4.

$$\widehat{W_1} = \frac{\exp(W_1)}{\exp(W_1) + \exp(W_2)} \tag{3}$$

$$\widehat{W_2} = 1 - \widehat{W_1} \tag{4}$$

Thus, the final module is designed for feature fusion, which is responsible for integrating the representation of image and text information for classification. It is written as Eq. 5.

$$R = \widehat{W_1} \times X + \widehat{W_2} \times Y \tag{5}$$

Indeed, integrating textual information with image features could be approached through a loss function; however, for the task of waste classification, the attention mechanism presents a more apt solution. For instance, when considering the contrastive distance loss function (Cheng & Wang, 2019), the objective is to bring features of similar samples closer in the feature space, while distancing those of dissimilar samples. For the model propounded in this book chapter, if the contrastive distance loss function is used, although the image features and text features can be more aligned, a more accurate sample matching strategy is needed to determine which sample features should be labeled similarly. Once the sample feature selection strategy is inappropriate, it will have a negative impact on model training (Qi & Su, 2017). Our task is waste classification, for a given image and text description, other waste images and text information need to be selected as positive or negative samples. How to select the appropriate images and text information is a great challenge, and it also adds time cost for model training. On the contrary, for the attention mechanism it can help the model to focus more accurately on the key information in the image and text descriptions, thereby improving the classification accuracy (Niu, Zhong & Yu, 2021). Therefore, we choose to adopt the attention mechanism as a method to integrate image and text information.

RESULT ANALYSIS

In this chapter, the dataset is called WasteNet, the waste images in the dataset are all taken by ourselves. WasteNet contains four categories of waste (classified according to waste classification standards), namely recyclable waste, hazardous waste, dry waste, and wet waste, with a total of 3,326 waste images. We employed the LabelMe annotation tool to manually annotate images. Each image contains multiple types of waste, so the number of labeled wastes in the dataset exceeds 3,326. Some waste data images are shown in Figure 3. In Figure 3, our dataset details the stacking phenomenon between waste objects, and some waste is also deformed, which is consistent with real waste classification scenarios. Figure 3 also illustrates the description of the image produced by MiniGPT-4.

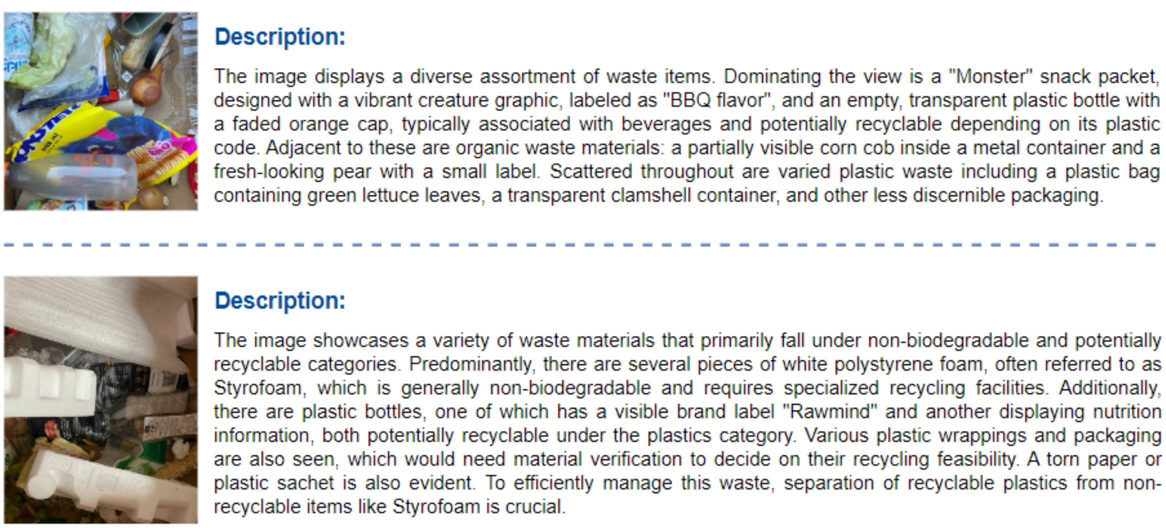

Figure 3. Examples of image descriptions generated by using MiniGPT-4 in WasteNet.

Utilizing the WasteNet dataset, we evaluated our proposed method against other leading-edge image classification models, primarily focusing on Average Precision (AP) as our primary metric for comparison. The outcomes in Table 1 demonstrate that our integrated approach—combining MiniGPT-4 with Swin Transformer—achieves an AP value of 62.20%, outperforming all other models listed (Liu et al., 2021). To illustrate this, utilizing the Swin Transformer alone for waste image classification yields an AP of 60.70%, which is 1.50% lower than that of our combined method. The Vision Transformer model and the ResNet model follow suit, recording AP values of 59.10% and 55.30%, respectively, both trailing our method

(Dosovitskiy et al., 2020) (He, Zhang, Ren & Sun, 2016). Other models, such as ConvNeXt, DenseNet, and EfficientNet, show average precision values of 55.20%, 51.20%, and 50.50%, in that order (Liu et al., 2022) (Huang, Liu, Van Der Maaten & Weinberger, 2017) (Tan & Le, 2019). The VGG model, with the lowest AP at 49.90%, is positioned at the end of the ranking (Simonyan & Zisserman, 2014). Additionally, Figure 4 presents the loss values associated with our model. These experimental results show that our method is effective and is able to significantly improve the accuracy of waste classification. It can also maintain a high level of accuracy and reliability when dealing with diverse and complex waste datasets. These experimental results show that our method is effective and significantly improve the accuracy of waste classification. It can also maintain a high level of accuracy and reliability when dealing with diverse and complex waste datasets.

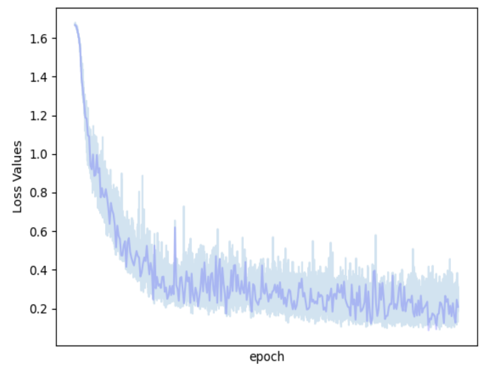

Figure 4. The Loss values of the model.

Table 1. Comparisons of AP values with different models

Models	AP
Ours	62.20
Swin Transformer	60.70
Vision Transformer	59.10
ResNet	55.30
ConvNeXt	55.20
DenseNet	51.20
EfficientNet	50.50
VGG	49.90

ABLATION STUDIES

Attention analysis of large language models and pre-trained language models

To further validate the effectiveness of MiniGPT-4, we conducted ablation experiments on different large language models, as shown in Table 2. The value of AP of our model reaches 62.20%, which is the best effect. Later, we kept the model structure unchanged and replaced MiniGPT-4 with other large language models, such as Blip and Clip (Li, Li, Xiong & Hoi, 2022) (Radford et al., 2021). Their AP values are very close to the AP values of our model, but still lower than our model by 0.90% and 2.10%, respectively. Furthermore, LLaVA is also tested and obtained AP values of 56.70% (Liu, Li, Wu & Lee, 2023). Finally, the model with the lowest AP value was Otter (Li et al., 2023).

Afterwards, we experimented with the selection of pre-trained language models as well, the results are shown in Table 3. If RoBERTa is replaced by BERT, XLNet, and ELECTRA, AP values of 60.10%, 56.90%, and 54.20% are obtained, respectively, which are lower than those of our model (Devlin, Chang, Lee & Toutanova, 2019) (Clark, Luong, Le & Manning, 2020) (Yang et al., 2019). We conjecture that in the context of our task, pre-trained language models are required to more accurately understand the text of image descriptions generated by large language models, while RoBERTa and BERT perform well on the task of processing and understanding the text, and can easily be used for new classification tasks. The experiments verify that RoBERTa is the best choice. In summary, MiniGPT-4 has the potential to lead the innovation in waste classification, and we will continue to explore the application of large language models, such as GPT-4, to waste classification in our future work.

Table 2. Comparisons of AP values with different large language models

Models	AP
Ours	62.20
Blip	61.30
Clip	60.10
LLaVA	56.70
Otter	53.30

Table 3. Comparisons of AP values with different pre-trained language models

Models	AP
Ours	62.20
MiniGPT-4 + Swin Transformer +BERT	60.10
MiniGPT-4 + Swin Transformer +XLNet	56.90
MiniGPT-4 + Swin Transformer +ELECTRA	54.20

Analysis of different prompts

In this chapter, we chose "Describe these images for waste classification in detail" as the prompt to input MiniGPT-4. To verify whether this prompt is suitable for our waste classification task, we also tested another prompt, "Describe these images in detail", and the results can be seen in Table 4. If we utilize "Describe these images in detail" in the prompt, the AP value will be reduced by 61.30%. This may be relevant to our dataset and classification task. From Figure 2, we also intuitively see that the descriptions obtained by inputting the two prompts are very different. The description given by the "Describe these images in detail" lacks object information, such as corn cobs and plastic bags, which may have a negative impact on the accuracy of this model.

Table 4. Comparisons of AP values with various prompts

Prompts	APs
Describe these images in detail	61.30
Describe these images for waste classification in detail	62.20

Analysis of prompts of different lengths

We also expanded our investigation to include the effect of prompt length on model performance. While limiting the description of a waste image to a single sentence, we found that essential information was often omitted— including specific items like lettuce, Kiwifruit, and cans. This omission can lead to a reduction in model accuracy. As indicated in Table 5, the Average Precision (AP) decreases by approximately 4.50% when using a single-sentence prompt compared to the more detailed prompt, "Describe these images for waste classification in detail."

Table 5. Comparisons of AP values of prompts with different lengths

Prompts	APs
Describe these images for waste in one sentence	57.70
Describe these images for waste classification in detail	62.20

Analysis of different strategies

Finally, we performed ablation experiments on the attention mechanism. If we utilize Adaptive Channel-wise Attention (Li, Liu, Zhang & Cheng, 2020) instead of Image-Text Aware Adaptive Attention, the resulting AP value is 58.50%. On the contrary, if we do not apply the attention mechanism strategy but employ the Contrastive Distance Loss function (Ding et al., 2023), its AP value is the lowest, only 54.60%. We conclude from Table 6 that the model performs best when we use Image-Text Aware Adaptive Attention. Figure 5 shows the AP change trend of these three strategies. This attention mechanism can dynamically adjust weights based on the properties of the input data rather than relying solely on fixed learning parameters. This provides greater flexibility to the model. However, it may be an oversimplification. Therefore, we will explore more complex weight calculation strategies in the future.

Table 6. Comparisons of AP values with different strategies

Prompts	APs
Image-Text Aware Adaptive Attention	62.20
Adaptive Channel-wise Attention	58.50
Contrastive Distance Loss	54.60

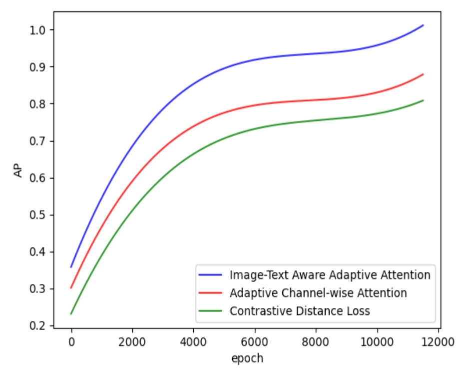

Figure 5. The AP values with different strategies.

CONCLUSION

In conclusion, our chapter presents the development of a multimodal waste classification model that leverages a large language model to transform semantic features of waste imagery into richer image data, demonstrating stronger generalization capabilities and enhanced accuracy. Employing an Image-Text Aware Adaptive Attention mechanism, our approach innovatively fuses disparate image features, utilizing a dynamic, adaptive weighting strategy based on input properties. This confers a more nuanced weight computation within the model, bolstering its generalization prowess.

Additionally, we have constructed a more expansive and realistic waste dataset, substantially augmenting the resources available for waste classification tasks. Our model has achieved a mean Average Precision (mAP) of up to 62.20%, evidencing an increase in classification accuracy while simultaneously reducing the computational resource demand. This marks a step forward in the fusion of waste management and artificial intelligence, offering a scalable solution that can benefit academic and practical applications alike.

Nevertheless, our work is not without limitations. The current waste dataset contains 3,326 images—a figure we aim to increase in future work to create a more comprehensive collection. The potential for more precise prompts for the large language model also remains an area for further investigation. Moreover, the development of a more sophisticated weight calculation strategy for the Image-Text Aware Adaptive Attention mechanism could enhance model performance. Continuing this research will allow us to refine and extend the capabilities of AI in waste classification, pushing the boundaries of environmental technology.

REFERENCES

Adedeji, O., & Wang, Z. (2019). Intelligent waste classification system using deep learning convolutional neural network. Procedia Manufacturing, 35 (pp.607-612).

Ahmad, K., Khan, K., & Al-Fuqaha, A. (2020). Intelligent fusion of deep features for improved waste classification. IEEE Access, 8 (pp.96495-96504).

Altikat, A. A. A. G. S., Gulbe, A., & Altikat, S. (2022). Intelligent solid waste classification using deep convolutional neural networks. International Journal of Environmental Science and Technology (pp.1-8).

Amasuomo, E., & Baird, J. (2016). The concept of waste and waste management. J. Mgmt. & Sustainability, 6 (pp.88).

Barrault, L., Chung, Y. A., Meglioli, M. C., Dale, D., Dong, N., Duquenne, P. A., ... & Wang, S. (2023). SeamlessM4T: Massively multilingual & multimodal machine translation. arXiv preprint arXiv:2308.11596.

Chen, J., Guo, H., Yi, K., Li, B., & Elhoseiny, M. (2022). Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18030-18040).

Cheng, K., Guo, Q., He, Y., Lu, Y., Gu, S., & Wu, H. (2023). Exploring the potential of GPT-4 in biomedical engineering: the dawn of a new era. Annals of Biomedical Engineering, 51(8) (pp.1645-1653).

Cheng, Y., & Wang, H. (2019). A modified contrastive loss method for face recognition. Pattern Recognition Letters, 125 (pp.785-790).

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In International Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171-4186).

Ding, N., Tang, Y., Fu, Z., Xu, C., Han, K., & Wang, Y. (2023). GPT4IMAGE: Can large pre-trained models help vision models on perception tasks?. arXiv preprint arXiv:2306.00693.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Wahid, A., ... & Florence, P. (2023). PaLM-E: An embodied multimodal language model. International Conference on Machine Learning Article No.: 340 (pp.8469 – 8488)

Egli, A. (2023). ChatGPT, GPT-4, and other large language models: the next revolution for clinical microbiology? Clinical Infectious Diseases, 77(9) (pp.1322-1328).

Ferronato, N., & Torretta, V. (2019). Waste mismanagement in developing countries: A review of global issues. International journal of environmental research and public health, 16(6) (1060).

Gundupalli, S. P., Hait, S., & Thakur, A. (2017). Multi-material classification of dry recyclables from municipal solid waste based on thermal imaging. Waste Management, 70 (pp.13-21).

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., ... & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. Computational Visual Media, 8(3) (pp.331-368).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

Huang, G. L., He, J., Xu, Z., & Huang, G. (2020). A combination model based on transfer learning for waste classification. Concurrency and Computation: Practice and Experience, 32(19), e5751.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700-4708).

Huang, K., Lei, H., Jiao, Z., & Zhong, Z. (2021). Recycling waste classification using vision transformer on portable device. Sustainability, 13(21) (pp.11572).

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103 (pp.102274).

Kaza, S., Yao, L., Bhada-Tata, P., & Van Woerden, F. (2018). What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050. World Bank Publications.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25.

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., ... & Liu, Z. (2023). MIMIC-IT: Multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425.

Li, C., Leng, Z., Yan, C., Shen, J., Wang, H., Mi, W., ... & Sun, H. (2023). Chatharuhi: Reviving anime character in reality via large language model. arXiv preprint arXiv:2308.09597.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning (pp. 12888-12900)

Li, W., Liu, K., Zhang, L., & Cheng, F. (2020). Object detection based on an adaptive attention mechanism. Scientific Reports, 10(1), 11307.

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. Electronics, 10(10) (pp.1163).

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in Neural Information Processing Systems, 36 (pp.34892-34916).

Liu, J., Pan, C., Yan, W. (2022) Litter detection from digital images using deep learning. Springer Nature Computer Science.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-radiology, 1(2) (pp,100017).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBRTTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).

Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11976-11986).

Mao, W. L., Chen, W. C., Wang, C. T., & Lin, Y. H. (2021). Recycling waste classification using optimized convolutional neural network. Resources, Conservation and Recycling, 164 (pp.105132).

Mohanraj, C., Senthilkumar, T., Chandrasekar, M., & Arulmozhi, M. (2023). Conversion of waste plastics into sustainable fuel. In Waste to Profit (pp. 41-52). CRC Press.

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. Neurocomputing, 452 (pp.48-62).

Qi, C., & Su, F. (2017). Contrastive-center loss for deep neural networks. In IEEE International Conference on Image Processing (ICIP) (pp. 2851-2855).

Qi, J., Nguyen, M., Yan, W. (2022) Waste classification from digital images using ConvNeXt. PSIVT 2022

Qi, J., Nguyen, M., Yan, W. (2022) Small visual object detection in smart waste classification using Transformers with deep learning. IVCNZ'22

Qi, J., Nguyen, M., Yan, W. (2024) NUNI-Waste: Novel semi-supervised semantic segmentation waste classification with non-uniform data augmentation. Springer Multimedia Tools and Applications.

Qi, J., Nguyen, M., Yan, W. (2024) CISO: Co-iteration semi-supervised learning for visual object detection. Multimedia Tools and Applications

Qi, J., Nguyen, M., Yan, W. (2024) Waste classification using improved deep learning method. Optimization, Machine Learning, and Fuzzy Logic: Theory, Algorithms, and Applications. IGI Global

Qi, J. (2024) Deep Learning Methods for Waste Classification. PhD thesis, Auckland University of Technology, New Zealand.

Qiu, L., Xiong, Z., Wang, X., Liu, K., Li, Y., Chen, G., ... & Cui, S. (2022). ETHSeg: An amodel instance segmentation network and a real-world dataset for x-ray waste inspection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2283-2292).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763).

Shi, C., Tan, C., Wang, T., & Wang, L. (2021). A waste classification method based on a multilayer hybrid convolution neural network. Applied Sciences, 11(18), 8572.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Singla, A. (2023). Evaluating ChatGPT and GPT-4 for visual programming. In ACM Conference on International Computing Education Research-Volume 2 (pp. 14-15).

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105-6114).

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8) (pp.1930-1940).

Waisberg, E., Ong, J., Masalkhi, M., Kamran, S. A., Zaman, N., Sarker, P., ... & Tavakkoli, A. (2023). GPT-4: A new era of artificial intelligence in medicine. Irish Journal of Medical Science (1971-), 192(6) (pp.3197-3200).

Wang, L., Yan, W. (2021) Tree leaves detection based on deep learning. International Symposium on Geometry and Vision.

Yan, C., Tu, Y., Wang, X., Zhang, Y., Hao, X., Zhang, Y., & Dai, Q. (2019). STAT: Spatial-temporal attention mechanism for video captioning. IEEE Transactions on Multimedia, 22(1) (pp.229-241).

Yan, W. (2019) Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics. Springer Nature.

Yan, W. (2023) Computational Methods for Deep Learning: Theory, Algorithms, and Implementations. Springer Nature.

Yan, W. (2025) Robotic Vision: From Deep Learning to Autonomous Systems. Springer

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, 32.

Yong, L., Ma, L., Sun, D., & Du, L. (2023). Application of MobileNetV2 to waste classification. PLOS One, 18(3), e0282336.

Zheng, H., & Gu, Y. (2021). EnCNN-UPMWS: Waste classification by a CNN ensemble using the UPM weighting strategy. Electronics, 10(4) (pp.427).

Zhou, H., Yu, X., Alhaskawi, A., Dong, Y., Wang, Z., Jin, Q., ... & Lu, H. (2022). A deep learning approach for medical waste classification. Scientific Reports, 12(1) (pp.2159).

Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592.

Zhu, X., Cheng, D., Zhang, Z., Lin, S., & Dai, J. (2019). An empirical study of spatial attention mechanisms in deep networks. In IEEE/CVF International Conference on Computer Vision (pp. 6688-6697).