Exploring Visual Embedding Spaces Induced by Vision Transformers for Online Auto Parts Marketplaces

Cameron Armijo, Pablo Rivas

Department of Computer Science
Baylor University, Texas, USA
{Cameron_Armijo1, Pablo_Rivas}@Baylor.edu

Abstract

This study examines the capabilities of the Vision Transformer (ViT) model in generating visual embeddings for images of auto parts sourced from online marketplaces, such as Craigslist and OfferUp. By focusing exclusively on single-modality data, the analysis evaluates ViT's potential for detecting patterns indicative of illicit activities. The workflow involves extracting high-dimensional embeddings from images, applying dimensionality reduction techniques like Uniform Manifold Approximation and Projection (UMAP) to visualize the embedding space, and using K-Means clustering to categorize similar items. Representative posts nearest to each cluster centroid provide insights into the composition and characteristics of the clusters. While the results highlight the strengths of ViT in isolating visual patterns, challenges such as overlapping clusters and outliers underscore the limitations of single-modal approaches in this domain. This work contributes to understanding the role of Vision Transformers in analyzing online marketplaces and offers a foundation for future advancements in detecting fraudulent or illegal activities.

Introduction

The transformer architecture, originally developed for natural language processing (NLP), has proven highly successful in a variety of tasks, from language translation (Vaswani et al. 2023) to text generation (Radford and Narasimhan 2018). Its core strength lies in the self-attention mechanism, which enables the model to capture relationships between all parts of an input sequence simultaneously. Building on the success of transformers, this architecture has been adapted to computer vision through Vision Transformers. These models take the same transformer principles—such as self-attention and tokenization—and apply them to image data by dividing images into fixed-size patches that are treated as input tokens. Vision Transformers have reached state-of-the-art benchmarks in computer vision, comparable to and even exceeding CNNs (Dosovitskiy et al. 2021). The adaptability of transformers to visual data demonstrates their versatility and effectiveness, opening new avenues for image analysis tasks such as object detection with models like DETR (Carion et al. 2020; Shehzadi et al. 2023) and image synthesis with models such as DCGAN (Radford, Metz, and Chintala 2016; Goodfellow et al. 2014). These models are derivatives of the original Vision Transformer architecture, each tailored for their specific tasks.

This study focuses exclusively on image classification tasks using ViT-Base, a single-modal deep learning model based on the Vision Transformer architecture. Specifically, we fine-tune ViT-Base on a dataset comprising images from online auto parts listings to evaluate its computer vision capabilities. Due to the nature of these online listings, only image data is considered (single-modality), while textual data such as captions from these posts is excluded. Figure 1 illustrates an example image from the dataset, highlighting the typical content of these listings.

While previous works have demonstrated the superiority of multimodal approaches (Hamara and Rivas 2024; Rashid and Rivas 2024) using models like ImageBind and OpenFlamingo, this study specifically evaluates a single-modal approach. This analysis is significant in understanding the limitations and potential of Vision Transformers when contextual data is unavailable or omitted. The contributions of this research are as follows:

- We evaluate ViT-Base on a large-scale, real-world dataset of online auto parts listings, focusing solely on visual data to isolate the model's ability to capture patterns.

Figure 1: Example of an interior view from an online auto parts listing, showcasing a car's seating and dashboard. This image represents the type of visual data used for clustering and analysis in this study.

Proceedings of the AAAI 2025 Workshop on AI for Social Impact: Bridging Innovations in Finance, Social Media, and Crime Prevention.

- By clustering and analyzing image embeddings, we demonstrate ViT's effectiveness in grouping visually similar items while also identifying challenges, such as overlapping clusters and outliers.
- Our findings highlight the limitations of single-modal models in the absence of contextual information, offering a direct comparison to the performance of multimodal approaches.
- We propose directions for enhancing single-modal models and emphasize the role they can play in scenarios where multimodal inputs are unavailable or impractical.

This work offers a critical perspective on the capabilities and constraints of Vision Transformers, laying a foundation for future advancements in single-modal and hybrid approaches.

Related Work

The application of transformer architectures in computer vision has recently gained significant attention. Initially designed for natural language processing tasks, the Transformer model (Vaswani et al. 2023) demonstrated exceptional capabilities in learning relationships between sequential data, leading to its adaptation for image processing tasks. Vision Transformer (Dosovitskiy et al. 2021) represents a significant shift in image classification approaches, outperforming traditional Convolutional Neural Networks (CNNs) such as ResNet (He et al. 2016) in various benchmarks, particularly with large-scale datasets.

The success of ViT can be attributed to its self-attention mechanism, which provides a global receptive field, unlike the local receptive fields inherent in CNNs. Dosovitskiy et al. (2021) demonstrated that ViT can capture long-range dependencies across image regions, making it suitable for diverse computer vision tasks. However, ViT's dependence on large datasets for pre-training has been a known limitation, prompting studies like Touvron et al. (2021) to introduce hybrid approaches, integrating CNNs and Transformers for improved performance on smaller datasets.

Beyond image classification, several studies have adapted the transformer architecture to other computer vision tasks. Carion et al. (2020) proposed DETR (DEtection TRansformer), which applies the transformer architecture to object detection, while Radford, Metz, and Chintala (2016) leveraged transformer architectures for generative models, illustrating their versatility. Using self-attention for object detection and generative tasks further demonstrates the potential of transformer-based models in capturing intricate patterns.

Recent works have also explored multimodal approaches to tackle the challenges in computer vision tasks, particularly in understanding complex datasets like those found in online marketplaces. Multimodal models like ImageBind (Girdhar et al. 2023) and OpenFlamingo (Alayrac et al. 2023) have demonstrated the effectiveness of combining text and image data, which often provides richer context and improves the interpretability of results. For instance, Hamara and Rivas (2024) used a multimodal model for analyzing car part listings from online marketplaces, achieving higher clustering accuracy than single-modal approaches. These results highlight the limitations of single-modal models like ViT when applied to data that could benefit from context.

In the domain of combating illicit activities in online marketplaces, machine learning models have shown promise. Rashid and Rivas (2024) utilized multimodal transformers to detect counterfeit products by integrating visual and textual cues, achieving superior performance compared to image-only models. Such studies underscore the advantage of using multimodal data for tasks requiring contextual understanding, which can be crucial in distinguishing between legitimate and illicit listings. Our work diverges from these approaches by focusing solely on the visual component, evaluating the effectiveness of ViT in detecting patterns in a single-modality context. This provides insight into the capabilities and limitations of visual-only analysis in addressing issues like the sale of stolen car parts.

Despite the promising results of multimodal approaches, single-modality models have their advantages, particularly in scenarios where only one type of data is available or where computational resources are limited. Wu et al. (2020) argued that simplifying the input modality can reduce computational complexity and yield effective representations if the model is sufficiently trained. Thus, our study aims to contribute to this area by investigating how well a single-modality ViT model can classify and cluster car parts from online listings, identifying the strengths and areas for improvement.

Our work builds on the foundational advancements in Vision Transformers and contributes to the body of knowledge by applying ViT to a practical, real-world problem involving the analysis of car part listings. The following section discusses the ViT architecture.

Overview of ViT-Base Model

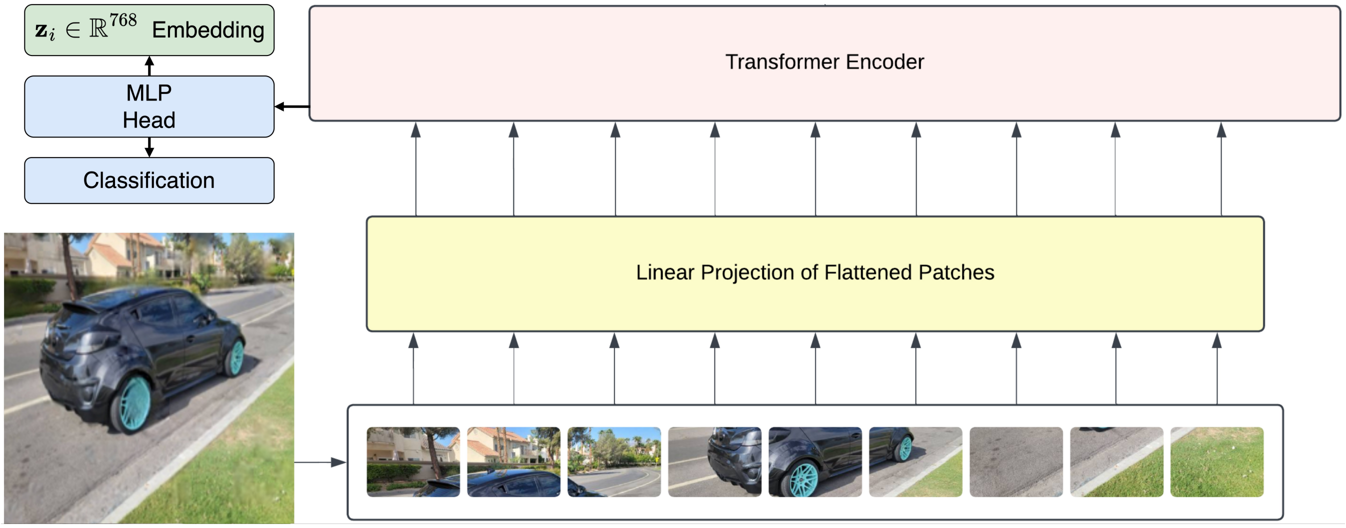
This section provides an overview of the ViT-Base model, an implementation of the Vision Transformer architecture specifically designed for image classification. Key aspects of the model architecture, mechanisms, and training are depicted in Figure 2 and summarized next to highlight its capabilities for image classification.

Architecture

A ViT draws directly from the transformer architecture (Vaswani et al. 2023), adapting its self-attention mechanism for image processing tasks. Unlike CNNs, which capture spatial hierarchies through convolutional layers (Cordonnier, Loukas, and Jaggi 2020), a ViT processes images as sequences of tokens. These tokens are derived from image patches and retain high-level spatial information through embeddings.

For an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and number of color channels respectively, the image is divided into non-overlapping patches of size $P \times P$. Each patch is flattened into a vector and linearly projected into an embedding space, forming a sequence of patch embeddings:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1 \mathbf{E}; \ldots; \mathbf{x}_p^N \mathbf{E}] + \mathbf{E}_{\text{pos}},$$



$\mathbf{z}_i \in \mathbb{R}^{768}$ Embedding
MLP Head
Classification
Transformer Encoder
Linear Projection of Flattened Patches

Figure 2: Vision Transformer architecture, illustrating the process of dividing an input image into patches, applying a linear projection, and processing through a transformer encoder. The final output is classified using an MLP head. Adapted from (Dosovitskiy et al. 2021; Vaswani et al. 2023).

where $\mathbf{x}_{\text{class}}$ is a learnable class embedding, $\mathbf{E}$ is the patch embedding matrix, and $\mathbf{E}_{\text{pos}}$ is the positional embedding. The resulting sequence is fed into a stack of transformer layers, each consisting of multi-headed self-attention and MLP blocks, normalized with LayerNorm and connected via residual connections (Dosovitskiy et al. 2021).

Patch Embedding

A critical innovation in ViT is its patch embedding mechanism. Instead of processing raw pixel data, ViT divides each image into patches, typically of size 16 × 16 pixels. Each patch $\mathbf{x}_p$ is flattened into a vector and passed through a linear projection layer:

$$\mathbf{x}_p\mathbf{E}, \quad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},$$

where D is the dimension of the embedding space. This transformation retains essential visual features, enabling the model to process spatial information effectively.

Positional Encoding

Images, unlike text, lack an inherent order. To encode spatial information, ViT introduces positional embeddings, $\mathbf{E}_{\text{pos}}$, which are added to the patch embeddings. These positional encodings can be either learnable or fixed and ensure that spatial relationships between patches are preserved, providing the transformer with context about each patch's location within the image.

Self-Attention

The self-attention mechanism is central to ViT's ability to capture global dependencies across the image. For a sequence of input embeddings $\mathbf{z} \in \mathbb{R}^{N \times D}$, self-attention computes pairwise relationships using query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) matrices:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D}}\right)\mathbf{V}.$$

Multi-headed self-attention extends this by employing multiple attention mechanisms in parallel, allowing the model to focus on different aspects of the input. This mechanism enables ViT to capture both local and global relationships, unlike CNNs, which are limited by their receptive field (Cordonnier, Loukas, and Jaggi 2020).

Training

Training ViT from scratch demands substantial data and computational resources due to its reliance on self-attention, which scales quadratically with the number of patches. Pre-training on large-scale datasets like ImageNet-21k provides a strong initialization, allowing fine-tuning on smaller, domain-specific datasets. In this study, ViT is fine-tuned on a car part image classification dataset, where the pre-trained classification head is replaced by a task-specific feedforward layer:

$$\mathbf{y} = \text{softmax}(\mathbf{W}\mathbf{z}_L),$$

where $\mathbf{z}_L$ is the output of the final transformer layer, and $\mathbf{W} \in \mathbb{R}^{D \times K}$ maps the latent space to the K output classes.

By leveraging pre-training, ViT achieves robust performance even with limited task-specific data, highlighting its flexibility and effectiveness in single-modality image classification tasks.

Methodology

This study investigates the use of Vision Transformers for clustering images of auto parts collected from online consumer-to-consumer marketplaces. The methodology depicted in Figure 3 involves dataset acquisition, embedding

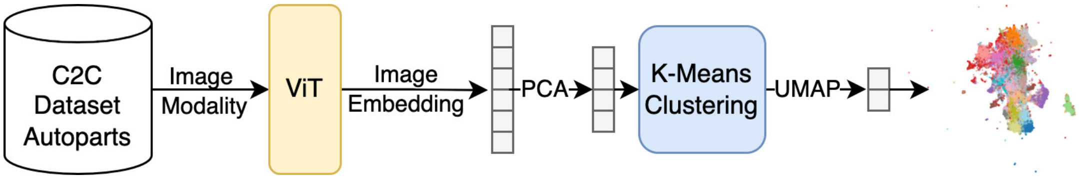


Figure 3: In the proposed methodology, the input data are images that are embedded with a ViT and then analyzed in search of cluster information.

extraction, clustering, and visualization. A detailed overview of each step is presented below.

Dataset

The dataset used in this study was derived from two popular consumer-to-consumer (C2C) platforms, Craigslist and OfferUp. Posts containing car parts were identified and scraped using automated tools. From OfferUp, a total of 650,654 posts were collected, comprising approximately 500GB of images. Similarly, 637,679 posts were obtained from Craigslist, amounting to 50GB of image data. Each post often contained multiple images, providing a diverse and multimodal dataset. The complete data collection methodology, including filtering criteria, scraping procedures, and dataset structure, is detailed in our prior work (Hamara and Rivas 2024).

For this analysis, we selected a random sample of 85,000 images from the full collection, focusing exclusively on the visual component. The text data, while valuable for providing context, was intentionally excluded to evaluate the effectiveness of single-modal Vision Transformers in capturing visual patterns. This subset was processed into embeddings for clustering and visualization.

As clustering is an unsupervised task, this study has no traditional train/val/test split. Instead, we performed a deduplication step to ensure no duplicate images were included in the dataset, preventing potential biases in cluster formation. The dataset was used purely for exploratory analysis, with no supervised learning involved. This ensures that the clustering results reflect the inherent structure in the data rather than being influenced by a label.

The multimodal nature of the dataset, including both text and images, highlights its potential for future research involving multimodal models. However, due to privacy concerns, the dataset will remain confidential.

Embeddings

The pre-trained ViT-Base model, with its 12 transformer layers and 768-dimensional output space, serves as the feature extractor. For each image $\mathbf{x}_i$, the model generates a fixed-size embedding vector $\mathbf{z}_i \in \mathbb{R}^{768}$:

$$\mathbf{z}_i = \text{ViT}(\mathbf{x}_i),$$

where $\mathbf{z}_i$ encapsulates key visual features of $\mathbf{x}_i$. These embeddings are high-dimensional representations designed to capture semantic and structural information within the images. The embeddings were subsequently normalized and stored for downstream tasks, including dimensionality reduction and clustering.

Clustering

To analyze and interpret the extracted embeddings, we applied clustering methods to group similar images. Given the high dimensionality of the embeddings, we used UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction (McInnes, Healy, and Melville 2020), projecting the data into a lower-dimensional space (2D) for visualization. UMAP preserves both local and global data structures, making it ideal for embedding analysis.

For the clustering process, k-means was used due to its simplicity and effectiveness in partitioning data into non-overlapping clusters. We experimented with reduced embedding dimensions of 16, 32, 64, and 128. For each configuration, we evaluated clustering quality using three metrics:

- **Silhouette Score** (Rousseeuw 1987), which measures how similar an object is to its cluster compared to other clusters.
- **Calinski-Harabasz Index (C-H)** (Caliński and Harabasz 1974), which evaluates inter-cluster variance.
- **Davies-Bouldin Index (D-B)** (Davies and Bouldin 1979), which quantifies the average similarity between clusters.

The k-means objective function minimizes the within-cluster variance:

$$J = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mu_i\|^2,$$

where C_i is the set of points in cluster i, and μ_i is the centroid of cluster i. After experimentation, the optimal value of k was determined to be 20, balancing intra-cluster cohesion and inter-cluster separation.

While K-Means is a widely used clustering algorithm due to its simplicity and efficiency, it has inherent limitations. Specifically, K-Means assumes that clusters are spherical and equally sized, which may not align with the underlying structure of the data. Additionally, K-Means struggles with irregularly shaped clusters and is sensitive to the initial cluster centroids. Alternative clustering methods, such as DBSCAN (Ester et al. 1996) or hierarchical clustering (Murtagh and Contreras 2012), could provide a better fit for complex structures in the data. Future work could explore these methods to assess whether they yield improved clustering quality for this dataset.

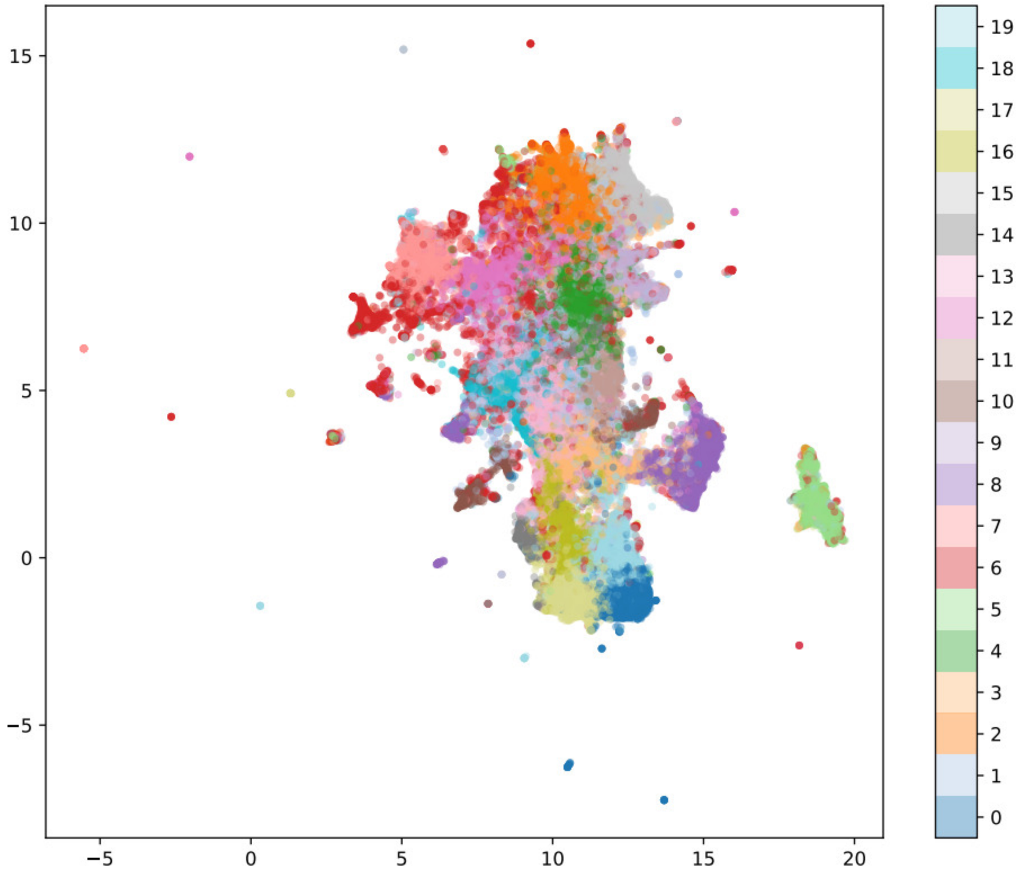


Figure 4: UMAP visualization of embeddings reduced to 64 dimensions, illustrating the clustering of images from online auto parts listings. Each color represents a distinct cluster identified using K-Means, revealing patterns and relationships within the dataset.

Results of Clustering Dimensions

Our experiments revealed that using 64 dimensions provided the best overall performance. Table 1 summarizes the clustering metrics across different embedding dimensions.

Table 1: Reduced embedding dimensions along with corresponding index scores. Calinski-Harabasz and Davies-Bouldin are abbreviated as C-H and D-B, respectively.

Dim.	**Silhouette**	**C-H**	**D-B**
16	.0126	925.5	4.412
32	.0134	937.5	4.274
64	.0152	942.4	4.164
128	.0151	944.3	4.253

While 128 dimensions slightly outperformed 64 in the Calinski-Harabasz Index, 64 dimensions offered the best trade-off between computational efficiency and cluster purity. The low Silhouette Scores suggest some overlap between clusters, attributed to outliers in the dataset, such as images containing mixed content (e.g., both powertrains and vehicle exteriors). This limitation highlights the challenges of single-modal approaches, especially when applied to multimodal datasets.

Interpretation

The clustering results indicate that ViT's embeddings can group visually similar images effectively, with a clear differentiation between major categories of auto parts. However, the presence of outliers emphasizes the importance of context, which a single-modal approach may miss. This finding underscores the potential advantages of multimodal models but also demonstrates the capability of ViT for image-only analysis in domains where additional modalities are unavailable or impractical.

The following section will analyze the findings to evaluate their implications for auto parts classification and clustering.

Results and Discussion

Our analysis revealed clusters that exhibit distinct patterns and characteristics, as shown in the UMAP visualization in Figure 4. These clusters highlight coherent groupings within the dataset, capturing similarities in image features while also exposing outliers that may require further investigation. The visualization demonstrates that the use of UMAP effectively reduced the high-dimensional embedding space to two dimensions, preserving important structural relationships within the data.

Cluster Analysis

Using k-means clustering with $k = 20$, we identified distinct groups of images corresponding to specific auto part categories. To validate these clusters, we employed k-Nearest Neighbors (KNN), locating the ten posts nearest to each cluster centroid based on Euclidean distance. The alignment of these nearest images with their respective centroids, as displayed in Figures 5, 6, and 7, supports the validity of our clustering approach.

The distinctiveness of some clusters was evident in their thematic consistency. For example:

- **Cluster 0**: Primarily contained full vehicle exteriors, including sedans, SUVs, and trucks.
- **Cluster 1**: Dominated by individual exterior components, such as mirrors, bumpers, and grilles.
- **Cluster 2**: Grouped powertrain elements, including engines, transmissions, and drivetrain components.
- **Cluster 3**: Captured body panels like doors, trunks, and hoods, as illustrated in Figure 8a and Figure 8b.
- **Cluster 4**: Focused on towing accessories, such as trailer hitches and tow bars.

The clustering performance metrics from Table 1 corroborated these observations. For 64 dimensions, the Calinski-Harabasz Index (942.4) indicated strong inter-cluster separation, while the Davies-Bouldin Index (4.164) suggested moderate intra-cluster cohesion. However, the low silhouette score (0.015) revealed some degree of cluster overlap, which is reflected in the presence of outliers.

The results indicate that some clusters exhibit significant

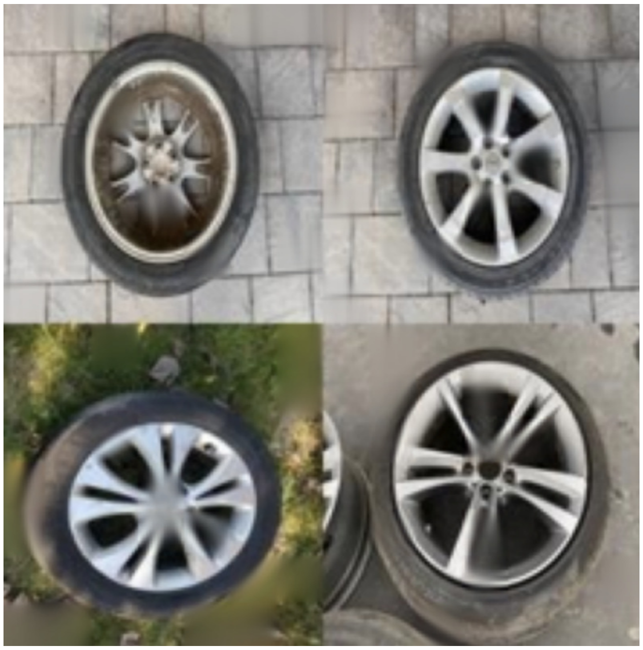


Figure 5: Representative images from posts located near a cluster centroid that appears to represent images of objects that look like *wheels*.

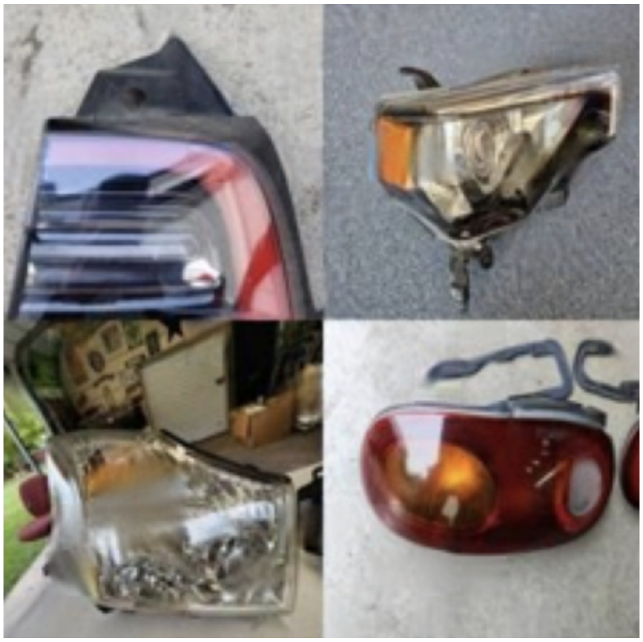


Figure 6: Representative images from posts located near a cluster centroid that appears to represent images of objects that look like *lights*.

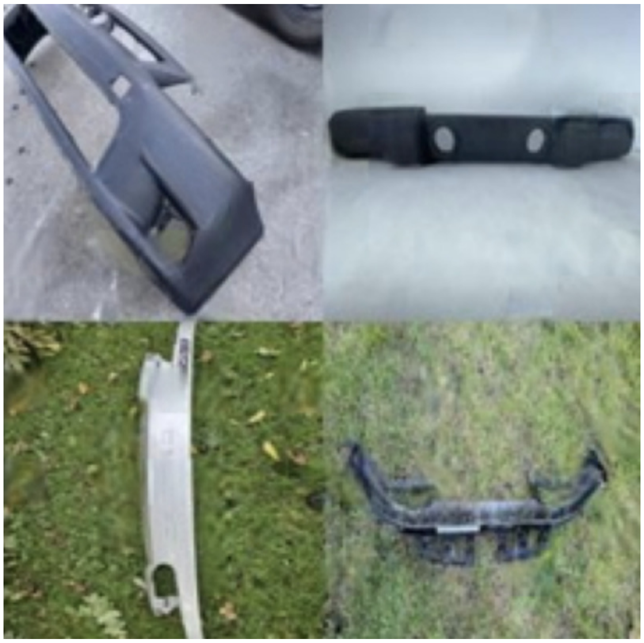
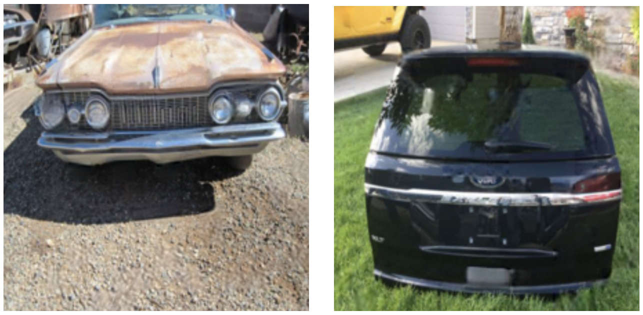


Figure 7: Representative images from posts located near a cluster centroid that appears to represent images of objects that look like *bumpers*.



(a) Hood of a vehicle, representing an image grouped in cluster 3.

(b) Trunk of a vehicle, representing another image grouped in cluster 3.

Figure 8: Examples of images from cluster 3, showcasing body panels. These images highlight the visual consistency of cluster 3 in grouping related components, such as vehicle hoods and trunks.

overlap, which may be partially attributed to the limitations of K-Means. Since K-Means assumes spherical clusters, it may struggle to capture more complex relationships between auto-part images. For example, DBSCAN, which groups data points based on density, could potentially handle cases where clusters have irregular shapes or varying densities. Likewise, hierarchical clustering may offer a more flexible way to identify nested structures in the data. Future work should explore these methods to determine their effectiveness in improving clustering performance for online marketplace images.

Since this study employs an unsupervised clustering approach, no separate train/val/test split exists. Instead, the dataset was processed as a whole, with a de-duplication step ensuring no duplicate samples were present. This allows the clustering results to emerge naturally from the data's inherent structure rather than being influenced by training procedures. Future work could explore whether supervised approaches or hybrid methods incorporating labeled data might further refine these clusters.

Outliers and Limitations

The UMAP visualization also revealed a number of outliers that did not align well with any cluster. These outliers often included images with mixed or ambiguous features, such as a single image displaying both powertrain components and exterior parts. This highlights a key limitation of using a single-modal approach: while ViT is effective at capturing visual patterns, it lacks the contextual understanding that multimodal models can provide by integrating textual descriptions or other metadata.

Additionally, the low silhouette score suggests that some clusters may overlap, particularly in cases where visual similarities exist between different auto parts. For instance, mirrors and body panels may share reflective surfaces or geometric shapes that lead to misclassification. Addressing this issue would likely require incorporating additional data modalities or refining the embedding process to better differentiate these subtle features.

Implications for Single-Modality Models

Despite the noted limitations, the results underscore the capability of ViT to classify and group auto part images based purely on visual data. This finding is particularly relevant for applications where only image data is available or where multimodal approaches are infeasible due to computational or privacy constraints.

The ability of ViT to effectively group visually distinct categories, such as powertrain components versus exterior panels, demonstrates its potential for tasks involving large-scale, single-modality datasets. These results also highlight the utility of dimensionality reduction techniques like UMAP in improving interpretability and revealing structural relationships in high-dimensional data.

While the ViT-based approach successfully groups visually similar auto-part images, the presence of overlapping clusters and outliers suggests that additional context could improve clustering quality. A multimodal approach, such as the one explored in (Hamara and Rivas 2024), leverages both images and textual descriptions from online marketplace listings to enhance representation learning. By incorporating textual context, models like ImageBind (Girdhar et al. 2023) can better differentiate between visually similar objects that serve different functions (e.g., two identical-looking car parts belonging to different vehicle models).

Future work should explore other multimodal Vision-Language Models (VLMs) that integrate textual metadata with image embeddings to refine clustering performance. This could provide richer semantic groupings and mitigate some of the cluster overlap observed in the single-modal setting.

Big Picture Perspective

The findings of this study contribute to the broader understanding of single-modal architectures in computer vision. While multimodal approaches often outperform single-modal models by leveraging complementary information, this work demonstrates that a well-optimized ViT-based pipeline can achieve meaningful results in specific domains. For instance, the ability to detect patterns in C2C auto part listings could support applications in crime prevention and fraud detection.

Moreover, the challenges identified, such as cluster overlap and the handling of outliers, highlight areas for further research. Enhancing single-modal models through techniques like self-supervised learning or advanced embedding strategies could mitigate some of these issues. Future work could also explore how these methods compare to multimodal models, providing insights into the trade-offs between simplicity and performance in practical applications.

Conclusions

This analysis demonstrates that while ViTs perform reasonably well in clustering auto parts listings, the dataset poses challenges for a single-modal approach. Our study diverged from previous studies that employed multimodal techniques

(Hamara and Rivas 2024), which leveraged the fusion of visual and textual data to capture richer contextual information. For instance, the multimodal approach achieved a silhouette score of 0.3819, significantly outperforming the 0.015 score from our single-modal ViT-based approach. This disparity underscores the advantages of multimodal models like ImageBind, which are better equipped to handle datasets containing both images and textual metadata.

Despite these limitations, our results highlight ViT's ability to isolate visual patterns in a dataset where text and images are often complementary. ViT was able to group listings into reasonably coherent clusters; however, cluster overlap and the presence of outliers revealed key shortcomings. One major limitation stems from the exclusion of textual data, such as captions, which often provide crucial contextual information about the images. Without this data, the model struggled to distinguish between visually similar components that serve different functions. Furthermore, many listings included images of multiple vehicle components within a single post, contributing to less distinct clusters and complicating the clustering process.

The findings of this study suggest several areas for improvement. One promising direction is to fine-tune the ViT model with a domain-specific pre-training dataset, allowing it to better capture the nuances of auto part imagery. Another potential enhancement involves experimenting with hybrid models incorporating textual embeddings while maintaining a primary focus on visual data. Additionally, post-processing techniques such as outlier detection and filtering could be developed to improve the clarity and coherence of clusters.

Ultimately, the results of this study serve as a stepping stone toward developing robust tools for analyzing online marketplaces. By improving clustering performance and integrating contextual information, this approach could play a critical role in detecting patterns of illicit activity, such as identifying stolen auto parts or fraudulent listings. Continued efforts in refining single-modal and multimodal methods will help bridge the gap between theoretical advancements and practical applications in this domain.

Acknowledgments

This work was funded in part by the National Science Foundation under grants CNS-2210091 and CNS-2136961, and by the U.S. Department of Education under grant P116Z230151.

References

Alayrac, J.-B.; et al. 2023. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Caliński, T.; and Harabasz, J. 1974. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1): 1–27.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. arXiv:2005.12872.

Cordonnier, J.-B.; Loukas, A.; and Jaggi, M. 2020. On the Relationship between Self-Attention and Convolutional Layers. arXiv:1911.03584.

Davies, D. L.; and Bouldin, D. W. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2): 224–227.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 226–231.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. arXiv:1406.2661.

Hamara, A.; and Rivas, P. 2024. From Latent to Engine Manifolds: Analyzing ImageBind's Multimodal Embedding Space. arXiv:2409.10528.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 770–778.

McInnes, L.; Healy, J.; and Melville, J. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.

Murtagh, F.; and Contreras, P. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1): 86–97.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434.

Radford, A.; and Narasimhan, K. 2018. Improving Language Understanding by Generative Pre-Training.

Rashid, M. B.; and Rivas, P. 2024. AI Safety in Practice: Enhancing Adversarial Robustness in Multimodal Image Captioning. arXiv:2407.21174.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

Shehzadi, T.; Hashmi, K. A.; Stricker, D.; and Afzal, M. Z. 2023. Object Detection with Transformers: A Review. arXiv:2306.04670.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, 10347–10357.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.

Wu, B.; Xu, C.; Dai, X.; Wan, A.; Zhang, P.; Yan, Z.; Tomizuka, M.; Gonzalez, J.; Keutzer, K.; and Vajda, P. 2020. Visual Transformers: Token-based Image Representation and Processing for Computer Vision. arXiv:2006.03677.