Behaviour Analysis in Retail store using Image Captioning Techniques

[a] *Deepa k* [b] *HariPrithika* [c] *Smirthika R* [d] *Tamilselvi* [e] *Yugesh C R*
[a,b,c,d,e] *Deparment of computer Application-2024*

Abstract

In this paper, we explore the application ofencoder-decoder architectures for generating captions in retail settings. We focus on extracting meaningful captions from videoframes to describe various retail scenes. The model utilizes a pre-trained CNN (ResNet50) and attention mechanisms to enhance the qualityand context relevance of generated captions.The performance of our system is evaluated on real-world retail video datasets, showcasing theeffectiveness of deep learning models in automating descriptive tasks.

I.Introduction

The rapid advancements in deep learning haveenabled models to understand and describe visual content effectively.

In retail, automating the process of generating captions for video footage can provide valuableinsights into customer behaviour and store layout optimization. In this work, we apply encoder-decoder architectures to extract video frames and generate meaningful captions that describe the retail environment.

II.Why behavior analysis of in-store consumer behavior is important

For behavior analysis the objective is the discovery of laws and principles governing behavior through experimentation. Skinner (1953, 1976) stressed that behavior is of scientic interest in its own right, not as a sign or symbol of something else. In line with this, several marketing scholars have called for more in-store experiments in retailing. Grewal and Levy (2007) suggest, based on their review of articles published in Journal of Retailing over 2002–7, that measure- ment of "actual behavior" represents a new avenue for further research in consumer behavior. They say, "more work is needed that focus on measuring actual behavior . . . which track and observe actual movement or perhaps determine actualusage and consumption" (Grewal & Levy, 2007, p. 450). Levy et al. (2004) urge researchers to conduct more real-life feld experiments to compare for alternative pricing strategies in retailing.

Another epistemological focus for behavior analysts has been the avoidance of formal devel-opment and testing of theories, or deductive theorizing. Marketing scientists, however, tend to emphasize deductive theory testing. However, one of the cornerstones of marketing is that the marketing mix is made up of elements such as product, price, place and promotion, classes of stimuli that can and often are explicitly used to influence consumer choice. The function of these marketing factors is dependent on consumers' environment and experienced consequences, but this process is not very well understood. The various elements that make up the marketing mix are mainly used as criteria for what is important in marketing strategy. In line with this, Davenport et al. (2011) recommend retailers to think of every offer (in-store offers, coupons etc.) as "a test", and as such to collect and use their customer data in a more sophis-ticated way to determine the effectiveness of various promotional efforts on consumer choice behavior.

Finally, Shankar et al. (2011) claim that controlled experiments are needed to test the effectiveness of different aisle placement and shelf positions, as well to understand the usage situation and effectiveness of new technologies and in-store promotional instruments (such as

in-store TV, shelf-talkers, and shopping carts). They say, "the model of how shopper marketing works is still a black box.This calls for effective ways to study shoppers in their 'natural habitats' compared to florescent-lit 'lab' environments. That is, more field studies are needed to supplement lab studies and validate the results from the lab studies" . Descriptive consumer behavior analytic research and findings can be used to criticize armchair theorizing in marketing; when data deviates from theory or when research becomes too focused on the model instead of the true subject matter, consumer behavior.

The degree of decision-making in the store also suggests there is a considerable upside in doing more in-store experiments. According to POPAI (Point-of-Purchase Advertising International) data, more than 70% of the brand decisions are made after the shopper enters the store (Liljenwall, 2004). Furthermore, many shopping trips take place without a shopping list or any planning from home

(Thomas & Garland, 2004). Many consumers therefore use a store's environment and its shelves as cues for what to buy. This suggests that retailers could benefit greatly from an active retailing approach (see Sorensen, 2009) grounded in insight and intel-ligence derived from behavior analysis of in-store consumer behavior. As retailers continue to invest in their own private labels, shopper behavior insight will also be important to grow these product lines through active retailing.

III.The importance of in-store applied behavior analysis

Research in behavior analysis has produced many useful applications in terms of methods to predict or control behavior. Applied behavior analysis is now used electively in many important and diverse areas such as developmental disabilities, problem behavior, education and organiza-tions. On the other hand, the field has "gotten stuck" in developmental disabilities. For example, 60% of data-based articles published in the Journal of Applied Behavior Analysis, the field's agship periodical, from 2001 to 2005 were in this area of research (Woods et al., 2006) and this trend does not seem to be changing. Applied behavior analysis is therefore, unfortunately, not as relevant to society in general as many analysts in the field would like, but the potential is vast. Today, many countries and markets have shown signs of an economic downturn, which has led to more competition among retailers and subsequently a lower turnover and margins. Overstoring is thus a Challenge in more and more markets (Grewal et al., 2007), which means a disproportional increase in the number of retailers in relation to the growth in the population. With declining growth from new customers entering the store, further growth can only be achieved if existing customers buy more, or start buying more quality brands (growing the share of wallet – see e.g. Nitzberg, 2009). That is to use in-store applied behavior analysis to extract more surplus from consumers once they are in the store, for instance boosting sales by more elective aisle and display management strategies (Bezawada et al., 2009). This prompts retailers to focus on in-store merchandising and promotion, which again requires deep understanding of in-store shopping behavior.

IV.Methodology

A.Video Frame Extraction

We first process the video data by extractingframes using OpenCV. Each frame is then converted into pixels and stored as images forfurther processing.

B.Preprocessing

Preprocessing is a critical step in ensuring that the input data is in a format suitable for training the encoder-decoder model.

This step includes several stages: **resizing images**, **normalizing pixel values**, and applying **data augmentation** toimprove model generalization

1.Resizing Frames

The extracted frames come in various dimensionsdepending on the video resolution. To ensure consistency, all frames are resized to a fixed dimension of **224x224 pixels**, which matches the input size expected by the pre-trained **ResNet50** model used for feature extraction.

2.Normalization

After resizing, the pixel values of each frame are normalized. The **ResNet50** model requires input values to be scaled between -1 and 1, which is achieved by subtracting the mean pixel values (based on ImageNet dataset statistics) and dividingby the standard deviation.

3.Data Augmentation

To make the model robust to variations in the data, **data augmentation** techniques areapplied. These techniques help the modelgeneralize better to unseen data by artificially increasing the diversity of the training dataset. The following augmentation techniques are employed:

- **Horizontal flipping:** Randomlyflipping the frames horizontally.
- **Rotation:** Randomly rotating the framesby up to 20 degrees.
- **Zooming:** Randomly zooming into theframes by up to 10%.
- **Brightness adjustment:** Randomly changing the brightness of the frames.

1.Splitting Dataset

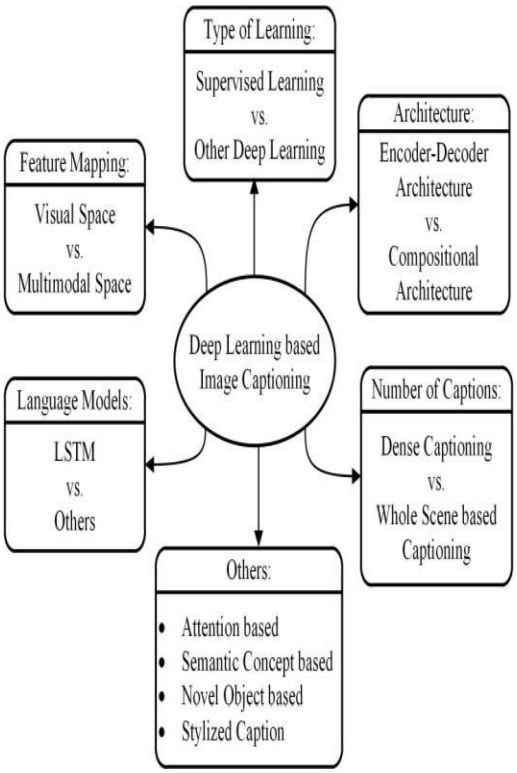
After preprocessing, the dataset is split into **training**, **validation**, and **test** sets. We use an 80-10-10 split to ensure the model has sufficient data for training while allowing us to evaluate performance on unseen data. Thissplit ensures that the model can generalize to new retail video frames and perform well in real-world applications.

A. Encoder-Decoder Model

The encoder-decoder architecture is widely used for sequence generation tasks such as image captioning, where the goal is to translate visual features into coherent text. This architecture consists of two main components: the **encoder**, which processes the input data(image frames), and the **decoder**, which generates descriptive captions.

Encoder:

The **ResNet50** model, pre-trained on ImageNet, is employed as the encoder to extract feature representations from each image (frame) in the video. ResNet50 is a Convolutional Neural



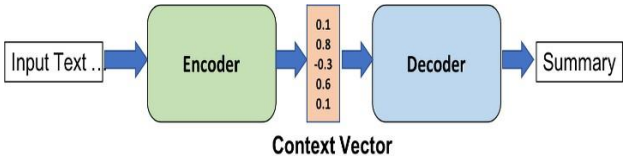
Network (CNN) that is well-suited for feature extraction due to its deep architecture and ability to capture hierarchical patterns in images. The outputof the ResNet50 model is a 2048-dimensional feature vector that encapsulates key information about the frame, such as objects, colors, and spatial structures. This feature vector serves as the contextfor the decoder to generate a caption.

Decoder:

The **Long Short-Term Memory (LSTM)**network is utilized as the decoder to translate the image features into natural language captions. LSTMs are a type of Recurrent Neural Network (RNN) that excels at handling sequential data, making them ideal for text generation tasks.

The LSTM takes the feature vector produced by the encoder and generates a sequence of words, forming a descriptive sentence.

Each word in the caption is generated based on the previous words and the context provided by the encoder



Model Training

We train the model using a combination of cross-entropy loss and an Adam optimizer. The dataset is split into training, validation, and test sets. To prevent overfitting, we apply early stopping based on validation loss

Results and Discussion

__Caption Quality

We evaluate the captions generated by our model using the BLEU (Bilingual Evaluation Understudy) score, which measures the similarity between the generated caption and the ground truth. In addition, we use humanevaluation to assess the contextual relevanceand fluency of the captions.

Conclusion and Future Work

This paper demonstrates the feasibility of usingencoder-decoder architectures for captiongeneration in retail environments. The integration of attention mechanisms significantly improves the contextual relevanceof the generated captions. Future work could explore the use of Transformer-based models for further performance gains and investigate real-time applications

REFERENCES:

1. https://www.researchgate.net/publication/329037107_Image_Captioning_Based_on_Deep_Neural_Networks
2. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00571-w
3. http://papers.neurips.cc/paper/9293-image-captioning-transforming-objects-into-words.pdf

4. https://www.sciencedirect.com/science/a rticle/pii/S2590123023002347
5. https://arxiv.org/pdf/2203.01594
6. https://www.ijert.org/research/image - caption-generating-deep-learning-model-IJERTV10IS090120.pdf
7. https://www.irjmets.com/uploadedfiles/ paper/issue_6_june_2022/26209/final/fin _irjmets1655531403.pdf