National College of Ireland

Deep Learning and Natural Language Processing Approach for Real Estate Property Description Generation

MSc Research Project
Data Analytics

Sai SriMaha Vishnu Valluri
Student ID: X19208758

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sai SriMaha Vishnu Valluri
Student ID:	X19208758
Programme:	Data Analytics
Year:	2020/21
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	16/08/2021
Project Title:	Deep Learning and Natural Language Processing Approach for Real Estate Property Description Generation
Word Count:	7741
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sai SriMaha Vishnu Valluri
Date:	23rd September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Deep Learning and Natural Language Processing Approach for Real Estate Property Description Generation

Sai SriMaha Vishnu Valluri
X19208758

Abstract

Real estate as an area of business has evolved over time. The development of online real estate marketing has contributed significantly to this improvement. The improvement of online real estate platforms along with increasing interest of individuals to handle their own sales has presented new opportunities for introducing innovative solutions into this field. In this study, exploration of the use of deep learning architectures such as convolutional neural networks and LSTM networks has been done in an attempt to create an image captioning and image tagging architecture for online real estate platforms. As a part of this, experiments have been done to understand and compare the effectiveness of attention based image captioning architectures as compared to traditional image captioning methods. The experiments have been evaluated on the basis of BLEU scores, accuracy and loss metrics. The image captioning model has showed a BLEU-1 score of 0.5456 while the accuracy of the image classification model was noted to be 76.86%

1 Introduction

The real estate market has traditionally depended extensively on real estate agents and a lot of footwork put in by the agents. Homeowners and prospective house buyers, for the most part, had only one point of common contact which was through an agent. This led to real estate markets being very local in the way they functioned and consequently it made certain situations quite difficult. The biggest issue faced in this process was that of distance decay1. This was the case of the American real estate market until 1999. In 1999, the advancements in internet technology led to the advent of internet real estate platforms.

According to the National Association of Realtors (NAR)[1], 84% of homebuyers in the United States of America currently begin the search for a new house online. In the context of online real estate listings, it is well known that creating a listing for a property would involve creating good property descriptions. Property descriptions are necessary for getting the attention of a buyer and giving proper information regarding a certain property. Although there is an abundance of data and advancements in technology, the process of creating image descriptions and tagging areas of the property on these

[1] https://www.nar.realtor/research-and-statistics/research-reports/highlights-from-the-profile-of-home-buyers-and-sellers

platforms are mostly to be done manually. The aim of this research is to explore possible methods that would help in tackling this issue with the use of deep learning techniques such as image classification and image captioning.

1.1 Research Question

RQ : "Can the application of computer vision algorithms and natural language processing techniques help in automating the process of image tagging real estate property images and property description generation?"

1.2 Objectives

Table 1 below shows the objectives that have been planned and achieved.

Table 1: Table of objectives

Sno	Objective	Description	Evaluation
1	Literature review	Identify and critically evaluate peer-reviewed literature in the areas of neural network-based computer vision techniques, natural language processing techniques, in the context of the real estate sector or work from other fields that could apply to the case at hand.	Critical review and comment of peer reviewed literature.
2	Data identification and pre-processing	Identify appropriate data for image captioning training that would be beneficial for image classification and content recognition tasks. Identifying housing images datasets for image classification.	Critical analysis and judgement of data quality based on need.
2.1	Image resizing	Resizing images to accommodate the requirements of the models that will be used.	
2.2	Feature extraction	Identify feature extraction techniques that are beneficial from the image captioning and image classification point of view.	

3	Identifying appropriate image processing techniques	Identifying computer vision techniques that are ideal for image classification and image captioning on the basis of the data that has been used and literature review.	
3.1	Image classification using InceptionV3	Training InceptionV3 model on Housing images dataset after splitting the dataset into training and testing segments.	
3.2	Evaluation of Image Classification model (Inception V3)	Testing the model on the testing segment of the Housing Image Dataset. Results of the testing phase are recorded and evaluated objectively.	Evaluation has been done on the basis of accuracy, sensitivity and specificity.
4	Identifying Image captioning techniques	Identifying image processing models and text generation techniques on the basis of data used and literature review.	
4.1	Image feature extraction for image captioning	Extracted features using InceptionV3 and VGG16 models separately for experimentation.	
4.2	Image captioning using LSTM and InceptionV3	Carried out image captioning using image features obtained from the InceptionV3 image encoder and generated text using an LSTM decoder. Results and evaluation metrics were recorded	
4.3	Image Captioning using LSTM with InceptionV3 and VGG16 Places365 CNN	Carried out image captioning using image features obtained from the InceptionV3 and VGG16 Places365 image encoder and generated text using an LSTM decoder. Results and evaluation metrics were recorded	
5	Evaluation of LSTM approach with InceptionV3	Evaluation metrics obtained and stored for further analysis	Evaluation metrics used are BLEU Scores

6	Evaluation of LSTM and VGG16 Places365 combined method	Evaluation metrics obtained and stored for further analysis	Evaluation metrics used are BLEU scores
7	Comparison of image captioning results	Quantitative evaluation metrics of InceptionV3 and VGG16 based approaches were compared and discussed.	Evaluation metrics used are BLEU scores

The contribution of the work done here can be understood from two perspectives. The first would be in the context of deep learning. The intention here is to understand and exhibit the effectiveness of transfer learning and gauge the capability of this concept as a potential solution to commercial issues through the medium of the real estate domain. This is done using image classification and image captioning as tools to carry out the task. The choice of real estate as a domain in this specific context has been made as this is a field where there is a good amount of data but this data is still unavailable for general use to develop innovative solutions. Additionally, the experiments performed here would help in understanding potentially good approach methodologies for image captioning and image classification tasks.

In the context of the field of real estate, the benefiting stakeholders would be the companies that host online listings. Sellers and buyers would also benefit from this system as it makes the process of posting a housing listing much easier from the seller's perspective while possibly aligning all listings to a certain established standard of quality. This in turn could help in increase the usage of a certain specific online listing platform as opposed to others. Additionally, the techniques used here would enable the concept of using image classification and image captioning as a standalone tool in the context of online real estate marketing.

The technical report is structured in the following way. Section 1 provides a brief introduction into the objectives, motivation and contribution of this research. Section 2 is a review of work done by researchers in areas related to the work done in this research. Section 3 is a detailed explanation of the methodology approach used to achieve desired goals and objectives. This is followed by Section 4 which would look into the design specification. Section 5 gives a brief explanation of model implementation and training followed by Section 6 in which results of evaluation are discussed. Section 7 concludes the research work and discusses possible improvements and use-case scenarios in the future.

2 Related Work

This section investigates work done in the fields of image classification, content recognition, image captioning, and text generation. This section is divided into sub-sections in the order mentioned. Literature review will be done with the aim to understand, discuss, reason and critique relevant methods. Design and implementation decisions were taken on the basis of the review of related work.

2.1 A Review of Image Classification and Content Recognition

The idea of employing image processing and analysis in the field of real estate has been gaining popularity in the recent years. The field itself provides various opportunities to apply image processing concepts. A large section of the work done in this area is focused mostly on price estimation and evaluation of properties. The approach taken by researchers falls into one of two categories i.e., creating a neural network specific for the task, or using the concept of transfer learning to carry out the required tasks. In this section both approaches will be analysed critically.

Poursaeed et al. (2018) have attempted to quantify the aesthetic quality of real estate properties to aid in the process of property evaluation. This has been done by the group by first establishing a metric that helps in objectively describing the "luxury level" of a property. The training data used for this has been obtained from the Houzz, Places and Google Images dataset. The group opted the DenseNet network to carry out image classification. DenseNet model is a fair choice of model as it helps in diminishing the vanishing gradient problem that is an issue. The DenseNet model is a "pre-made" convolutional network. The architecture is already established and can be leveraged for specific tasks. This is a useful approach considering the network has already been established and tested. This helps reducing the time and complexity involved in building a network while more focus can be put on fine-tuning the model to achieve the intended goals.

You et al. (2017) have also worked on formulating a solution for the task of assessing property prices based on the visual features. Along with this, location of the property and the properties around it have also been considered to evaluate the pricing. The task of image classification and content recognition in this case is carried out by the pre-trained GoogleNet. As mentioned by Szegedy et al. (2015) in their hallmark research, GoogleNet is a convolutional neural network consisting of 22 layers. The use of this network and networks based on this architecture is efficient because of the fact that it has been designed to function in a way that the network could optimize internal resource utilization.

As opposed to the mentioned techniques for image classification, Bappy et al. (2017) have approached the problem of image classification by using LSTM networks instead of the traditional CNN approach. The use of an LSTM network for image processing tasks has its benefits as well. CNN networks in reality need high input dimensionality in order to function efficiently. In a case where the network has to be made from scratch, it would require large amounts of data in order to work as intended. This can be avoided by using an LSTM network as the input requirements are not that high and this helps in reducing computational complexity as well. While the choice of the type of neural network depends on the task and the availability of resources, using convolutional neural networks specifically helps in dealing with the process of feature extraction. The versatility and capabilities of convolutional neural networks allow a wide range of methods to handle feature extraction. Lu et al. (2014) have carried out feature extraction from images by creating a double column neural network through which RGB inputs are used as opposed to taking the features of the images alone. Additionally, the use of a double column system aided them to carry out training from both a global and local view. This in itself is another way of learning aesthetic-related features from images.

2.2 A Critical Review of Attention Based Models

Image captioning methods rely heavily on attention based techniques to gather features from images. Image captioning techniques traditionally use a single image feature extraction mechanism which creates a set of global image features. This is not beneficial because it leads to irrelevant features being collected (Xu et al.; 2015). One issue that this leads to is the creation of captions that are not descriptive or useful in many cases. One way to tackle this is to develop on the attention mechanism of the encoder-decoder architecture. One key contribution to this solution was presented by Dang et al. (2019). The work done here was to create an attention mechanism with the use of two pretrained CNN networks. This enabled the image captioning to be done on the basis of image features provided by a multi-feature image representation. On the basis of the results presented, it can be seen that this method as compared to traditional attention mechanisms has been more successful. Evaluation and comparision of performance has been done using BLEU scores. Hossain et al. (2019) have exhibited the impact of attention based image captioning mechanisms as well. According to the work done here, it is stated and proved that traditional methods of image feature processing is wasteful in the sense that the information collected cannot contribute to the quality of image captions. The concept of using global features alone is not sufficient and this has been proven. The solution provided here involves the use of DenseNet (Huang et al.; 2017) weights. This helps in creating a multi level set of features. In the work done by Peng et al. (2019), it can be seen that the efficient use of attention mechanisms could impact the quality of image captions generated. The use of two CNNs for image feature extraction enabled the creation of local and global features.

2.3 A Critical Review of Image Captioning

Image captioning in deep learning is the process of training a model to understand the contents of an image and describe it in a natural human language. This by itself is a fairly complicated task as multiple types of data are required to be considered. The advancements made in deep learning techniques and collaborative efforts made by researchers in different contexts has made this task possible. Although, the application of image processing in the real estate domain has primarily been in property evaluation, work done using image captioning in other domains is very helpful in understanding appropriate methods to carry out the intended tasks.

Puscasiu et al. (2020), proposed a method to carry out the task of image captioning effectively. The work done by this group involves the creation of a CNN-RNN based encoder-decoder architecture. Using the InceptionV3 CNN network, initial image captioning is done. The output of this network is used as the input to an RNN based on a GRU and attention layer. This method would involve the generation of a RNN network tailored to carry out this task in specific. While being beneficial it is heavily dependant on the availability of data to be able to reach desired quality in the outputs. While availability of data is subject to vary, the advantage of this approach would be the option to control the size of the caption and the language style that the model would pick. Additionally, creating an entire network from scratch is proven to be computationally expensive and resource intensive.

Going further into tackling issues regarding lack of data, Zhao et al. (2020) have proposed a novel approach to the image captioning problem when there is a lack of image-description pair data. This method would involve the establishment of a cross-domain

training architecture. This led to the use of a training dataset with image-description pairs (MSCOCO) for training while applying the trained model onto images without these attributes (Oxford-102). The researchers here used the ResNet50 CNN network for training purposes while a GRU network was established for the text generation task. Typical CNN networks could also be used for image captioning as they are well capable to carry out such tasks, but this may not be efficient considering the fact that this would lead to a lack of control on the direction in which it would go. While this deals with the case of not having sufficient training data, the availability of data could lead to highly efficient and productive systems. This was highlighted by Zou et al. (2020). The group has worked on a novel idea which is to identify image aesthetic and not just the content of the images. This was followed by describing the images using image-captioning concepts. The aesthetic of the image would have to be quantified in order to be measurable and this has been done by training the CNN to estimate lighting, composition, focus, and depth of field. The image captioning section is handled with the help of bi-directional LSTM networks consisting of two layers. Essentially, the use of separate decoders for each aesthetic element leads to the creation of highly descriptive captions. The use of an LSTM network instead of an RNN also benefits the outcomes as LSTMs have better control as compared to RNNs. The use of LSTM networks is seen in certain other contexts as well. In the work presented by Liu et al. (2019) , it is seen that a CNN-LSTM encoder-decoder architecture has been established. The intended purpose was to make use of visual paraphrases in order to create descriptive and diverse captions. In such a scenario, as seen earlier, the most commonly adopted way is to have a CNN for image content extraction, while two LSTM networks are employed to generate target text. This separation of responsibilities helps in each network carrying out its own tasks efficiently.

In all the cases mentioned above, generated text has been evaluated on the basis of BLUE, CIDEr and SPICE scores as these are the standard metrics of machine generated text evaluation. Additional metrics that could be looked at for a better understanding are METEOR and ROGUE-L values as done by Zhao et al. (2020).

On the basis of reviewing related work, it is understood that image classification can be achieved successfully with the use of pretrained neural networks given that there is good training data. The area of image captioning provides a lot more methods in which efficient results can be achieved. In view of the objective, that is, to achieve detailed descriptions, attention based mechanisms have been seen to be efficient. Text generation can also be efficiently achieved using LSTM or RNN networks. The advantages that these networks provide is in terms of control on the direction in which the content would be going to, that is, keeping the context in check, having the option to manipulate the size of the entire description that would be generated and in general control over the style of the language based on the domain the generated text would be used in.

2.4 Identified Gaps and Conclusion

Based on looking at the work done by researchers in similar fields, there is now some backing to the idea and motivation. Image classification in most cases has been a task that required a certain level of accuracy and preciseness. Upon looking at research done regarding this, the solution to this becomes quite clear. On the basis of reviewed literature, it becomes clear that CNNs are capable of doing a good job at image classification. The time and amount of data used for training and fine tuning a model is a key influential factor in deciding the outcome of the model.

In the case of image captioning, It was noted that both LSTM and CNN models can carry out image processing tasks but looking at the amount of work done using CNNs, quality of work and the ability to replicate the work, CNNs, especially pretrained CNNs are likely solutions for the image processing part of the work that has been done here. This is a decision made taking the computational constraints, data related constraints and required precision of work into account.

As for text generation, the use of pretrained image captioning models would definitely help in carrying out image captioning. While this would streamline the solution into the application of a single model to carry out both the tasks this may be an issue in terms of the quality of work. One of the issues in using a pretrained image captioning network would be the issue of not having enough control on how the output should turn out. This would not be a problem in cases where image-caption pair training data is available. Flow of control in data has been one key factor upon which the basis of LSTM networks lies on. Using this as a potential solution for descriptive image captioning while combining with content specific image feature extraction could be beneficial in establishing descriptive and specific image captioning.

3 Image Classification and Captioning Methodology

Data mining research methodology can be done in multiple ways. Some examples of this are the KDD or CRISP-DM methodologies. In this case, the CRISP-DM methodology has been adopted. Figure 1 gives a brief understanding of the CRISP-DM architecture which has been modified. Following this a detailed explanation of each process has been provided.

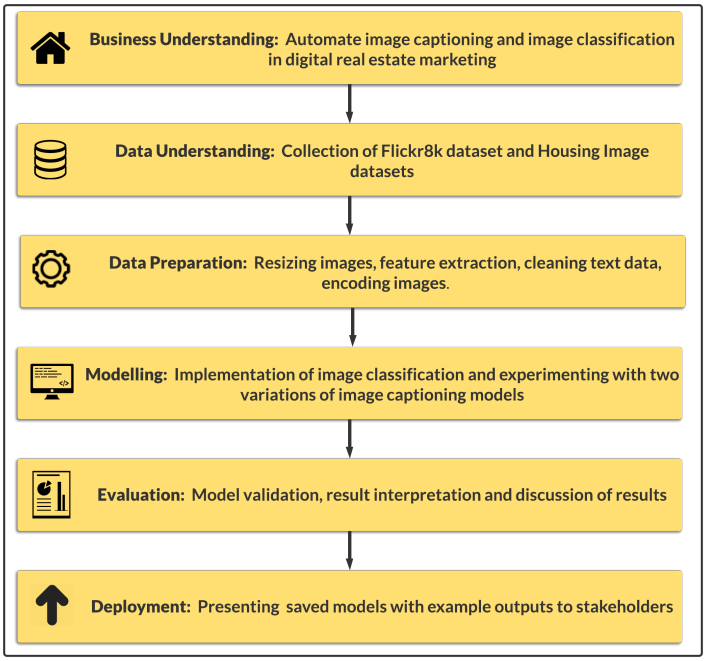

Figure 1: Modified CRISP-DM methodology

3.1 Business Understanding

The commercial focus of this research is to investigate potential data based solutions for automating the task of generating image descriptions in real estate listing websites. This is necessary as detailed descriptions of real estate properties are proven to be influential on the amount of interest customers would show in that property. Moreover, there is a need to be descriptive and concise with the language being used in these platforms to drive in profitable interactions from customers.

3.2 Data Understanding

For training purposes, the Flickr8k dataset was used. This is a benchmark dataset that is used for image captioning tasks and has been established in the literature review as well. The dataset consists of a variety of images collected from Flickr with each image having 5 ground truth captions. For image classification, the dataset being used is a dataset consisting of housing images from the online real estate platform Zillow.

3.3 Data Preparation

Data preparation is the first step of any machine learning process. The data available first needs to be transformed into forms that would be compatible with the models being used or in accordance with computational restrictions. In this case, the initial task is to carry out image processing.

3.3.1 Image Processing

Image processing for image captioning requires the images present in the dataset to be stored into an array first. This task is followed by normalising the images and compressing them in accordance to the requirements of the InceptionV3 model and VGG16 Places CNN model as these two models were be used for the experiments that were done here. Next, feature extraction needs to be done. Feature extraction is a type of dimensionality reduction (Kumar and Bhatia; 2014) . In this process, the pixel data of the images are stored in such a way that the stored data would be useful to the model to carry out any tasks further. One of the more efficient methods to execute feature extraction would be to make use of pretrained neural networks as done by Peng et al. (2019). Pretrained neural networks are effective in cases where time, computational resources and available data are limited Singla et al. (2019). In this research, a similar approach has been taken. The features are extracted from the images using a version of the VGG16 network. In specific, the VGG16 Places CNN model (Zhou et al.; 2017). This is a VGG16 model that has been trained on the Places365 Dataset (Zhou et al.; 2017). The Places365 Dataset is a dataset that comprises of images of various places and locations i.e., areas inside or outside houses, public indoor locations, outdoor locations and many more. The data is labelled into 6 classes. This is beneficial in the case of this research as the model is trained to extract features with respect to the Places365 dataset. This would help in gathering data relating to the image scene content which is necessary in order to caption images in a more descriptive manner in order to capture the entirety of scene features. The features would essentially be the data collected from the model from the first layer till the last fully connected layer.. Features extracted in this stage have been stored as pickle files for future use and modelling purposes.

For image classification, Housing Images dataset (Poursaeed et al.; 2018) images are initially stored in a zip file format. This dataset consists of 116,000 images belonging to 7 classes i.e., Exterior, Interior, Bedroom. Living Room, Dining room, Bathroom and Kitchen. This dataset has then been split into training, testing and validation sets in the ratio of 80%, 10% and 10% respectively. Additionally this dataset has been sampled to a smaller quantity of data for processing purposes.

3.3.2 Text Processing

Text data cannot be processed as sentences that exist in natural language by machines. It must be processed into forms that are machine comprehensive. For this purpose, the following steps need to be taken to process text data before carrying out any modelling tasks on text.

- **Removing Punctuation:** Removal of punctuation is necessary due to multiple reasons. Existence of punctuation increases complexity. Additionally, punctuation are not necessary in this specific case as the focus is on single line descriptions.
- **Converting text to lowercase:** Converting all the text in the dataset into lower case helps in maintaining uniformity. This is also necessary as there may be a risk of the machine comprehending words with upper case letters as different compared to the same words without upper case letters. This would increase complexity as well if not dealt with.
- **Adding starting and ending tags:** It is necessary to indicate the start and end of a sentence with tokens such as 'startseq' and 'endseq'. This would help the user in enabling the machine to understand the start and end of a sentence. This helps in text generation as well to understand the sequence in which the text "prediction" would work.
- **Tokenization:** Sentences that have been "cleaned" in the earlier steps are broken into words in order to form a dictionary. This would consist of the entirety of the vocabulary used in the sentences.
- **Vectorization:** The words that are stored in the tokenization stage are further converted into numerical forms called vectors and can be used to learn word sequences by the machine. This is the form that is finally usable by the machine and this would lead to the beginning of training.

3.4 Modelling

In the previous steps, information that is relevant to the model's learning methods has been collected from the training data through the process of text pre-processing and image feature extraction. The information obtained through this would be in the from of image scene features and tokenized sequential text dictionaries. This will be the basis on which the models would be trained. The CNN part of the model would encode the image features and provide that as an output. The encoded features consist of contextual information present in the images. This is followed by using the encoded data as an input to the text model. The text model consists of an embedding layer which is used to learn the vector representations of the text captions that have been created in the earlier steps.

An LSTM layer is used before the final layer with softmax activation is used to predict the possible words that could occur in the sequence with reference to the images.

The image data for image classification model will undergo image resizing and feature extraction as a part of modelling for the inceptionV3 network. These features are then used to train the final layer of the network to carry out classification.

3.5 Evaluation

traditional methods of analysing accuracy and related metrics is not applicable in this specific context as text generation relies on a lot more factors than just accuracy alone. Taking this into account, evaluation metrics such as Bilingual Evaluation Understudy (BLEU) (Papineni et al.; 2002) has been used. These metrics serve as the benchmark metrics for machine learning tasks involving text generation. The BLEU metric is known to have its flaws but it is still a standard metric for quality assurance of text generation. The metric evaluates generated text on a scale of 0 to 1. It evaluates text on the basis of how well generated text represents the sequences and patterns found in the text on the basis of which learning is done. This makes the BLEU metric a quantified version of understanding how close machine generated text is to human generated text.

Evaluation of image classification model has been done by understanding the model performance with reference to how well the model training is carried out. This leads to the usage of metrics such as change in validation loss and validation accuracy as training progresses. These metrics for both sets of tasks along with qualitative evaluation have been used for evaluation.

4 Image Processing Design Specification

In this section, the project design process is explained with the help of an architecture diagram shown in Figure 2.

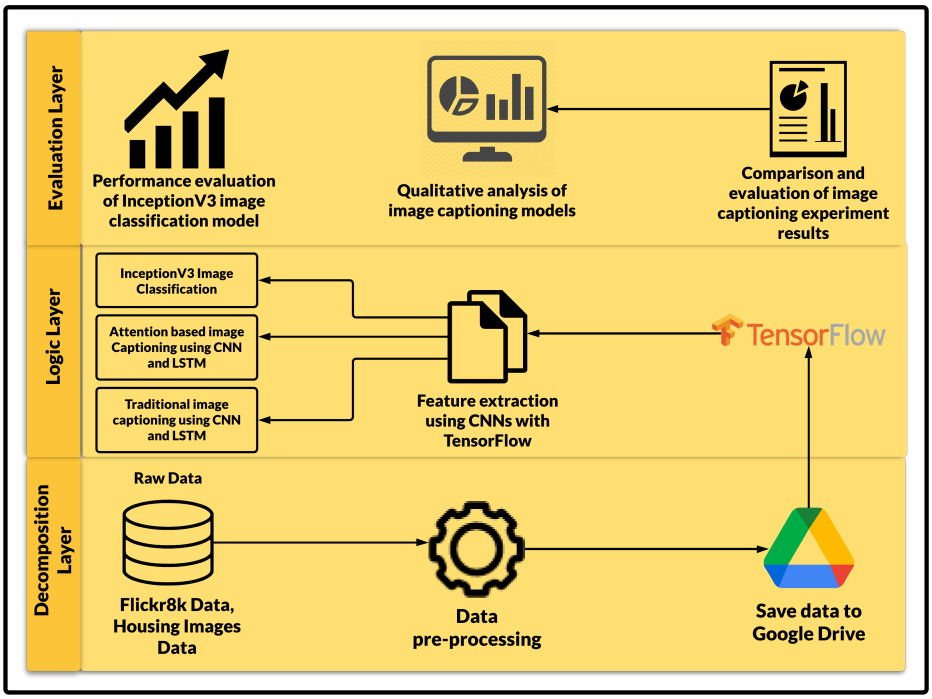

Figure 2: Real estate image processing architecture

The design architecture consists of three layers. The first layer is the Decomposition Layer. This is where data is collected, pre-processed and stored for future use. The next layer is the Logic Layer. In this layer feature extraction is done followed by executing the proposed experiments. The last layer is the Evaluation Layer. In this layer, image captioning experiment results are analysed and compared both quantitatively and qualitatively. Image classification model performance is also analysed as a part of this layer.

The underlying models that enable the chosen approaches to work are examined next. Tasks relating to image processing were carried out using convolutional neural networks while text based tasks were executed using LSTM models.

4.1 Convolutional Neural Networks

Convolutional neural networks are a class of neural networks in deep learning that are well known for their effectiveness in the field of image processing and image feature extraction (Chollet; 2017). CNNs are also efficient for transfer learning purposes. In terms of image classification, CNNs can leverage the concept of transfer learning to speed up the entire process. As explained by Yim et al. (2015), image classification using CNNs using transfer learning can be done by utilizing the layered architecture. Different layers through the network analyse and process different levels or dimensions of information. Each layer moves this information forward as the input to the next layers. The final layer of a CNN handles the task of classification. When using transfer learning approaches. All the weights in the network are already stored after training on large datasets. In this case, the InceptionV3 network has been used and this has been trained on the ImageNet dataset (Deng et al.; 2009). All the layers, except the final layer of the model will be

used. The final layer is trained on data upon which classification should be done. This enables domain specific classification, depending on the training data being used.

For image captioning, CNNs play an integral role. One of the main aspects of image captioning is to understand the content of the image that is being fed into a model. This process is solely dealt with by CNNs. The way this is done is through feature extraction. High level features are initially extracted from the channels of the image. Following this, the high-level features are reduced in size spatially by pooling to identify features that would best represent the content of the image. These feature vectors are further processed by the final connected layer. This entire process essentially enables the model to "see" and recognise content and group similar content into prediction classes. Using the concept of transfer learning in this scenario is helpful as the models used here i.e., InceptionV3 and VGG16 are trained on large amounts of data which helps in extracting distinct features that would be beneficial for content recognition on new data as well.

4.2 Long Short-Term Memory Networks(LSTM)

Long short-term memory(LSTM) networks are a class of recurrent neural networks(RNN) that are efficient with text generation and text processing tasks. One of the main reasons for using this over the RNN is the ability of this type of networks to overcome issues or vanishing and exploding gradients faced by RNN networks(Chollet; 2017). Along with this LSTM networks control the flow of information efficiently which make these networks good for sequential tasks like text generation.

5 Implementation

In this section, the process of implementing the proposed approach for this research is explained in detail.

5.1 Setup

A major requirement in order to carry out the intended objectives successfully is to have an environment capable of running deep neural networks efficiently and quickly. For this purpose a capable graphics processing unit (GPU) is required. GPUs are faster, easy to use and more efficient than standard CPUs that are present in computers but are also expensive. To overcome this issue, the entire research project implementation has been carried out using the Google Colab Environment. Google Colab is a cloud based python IDE based on Jupyter Notebook. It provides access to efficient and powerful GPUs such as the Nvidia K80, P100 or the T4. GPUs are essential to run processes on the TensorFlow package.

The host machine that was used was a Dell Inspiron with an i7 processor, 12GB of RAM and an Nvidia 230MX GPU. Jupyter Notebook was also used to carry out certain parts of the pre-processing locally.

5.2 Data Identification and handling

The Housing Image Dataset used for image captioning was initially downloaded into Google Drive as a zip file and then extracted through Google Colab. Custom python functions were defined to download and extract and split the dataset into training, testing

and validation data. This step is followed by sampling the dataset randomly to a smaller quantity of data i.e., 7000 images for training and 2100 images for testing. Python functions were also defined to carry out this pre-processing task. This smaller dataset has been used for image classification task. Figure 3 shows a sample of one image from the 7 classes each.

Figure 3: Sample images from House Image Dataset

The next step of pre-processing involved processing the Flickr8k data which is used for image captioning. Custom functions were defined to process this data. This step would involve identifying the text files containing image data and parsing the data to the images. This would help in referring to the data in further steps as it was needed. Following this, image descriptions were pre-processed. Figure 4 shows a sample image along with its 5 provided captions.

The first step of doing this is to clean the text, this would involve breaking down sentences into words, converting all the words into lower case forms and removing punctuations. This step is followed by resizing images of both datasets in order to fit image

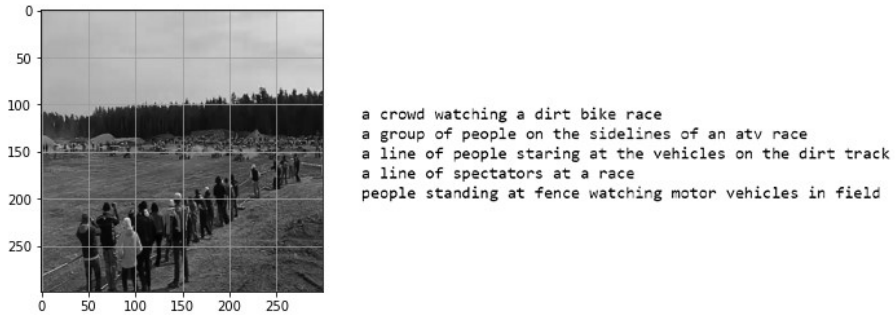

Figure 4: Sample image with training captions from Flickr8k Dataset

requirements. In case of the InceptionV3 model, images are resized to 299x299 in order to satisfy the requirements. Processed files are stored as pickle files for future use

Feature extraction is carried out after pre-processing images. Feature extraction is also handled by the InceptionV3 model and the VGG16 model. The features extracted are stored as pickle files and used for training in the steps of the research project.

5.3 Model Implementation

Two models are built in total in order to carry out image classification and image captioning.

5.3.1 Image Captioning

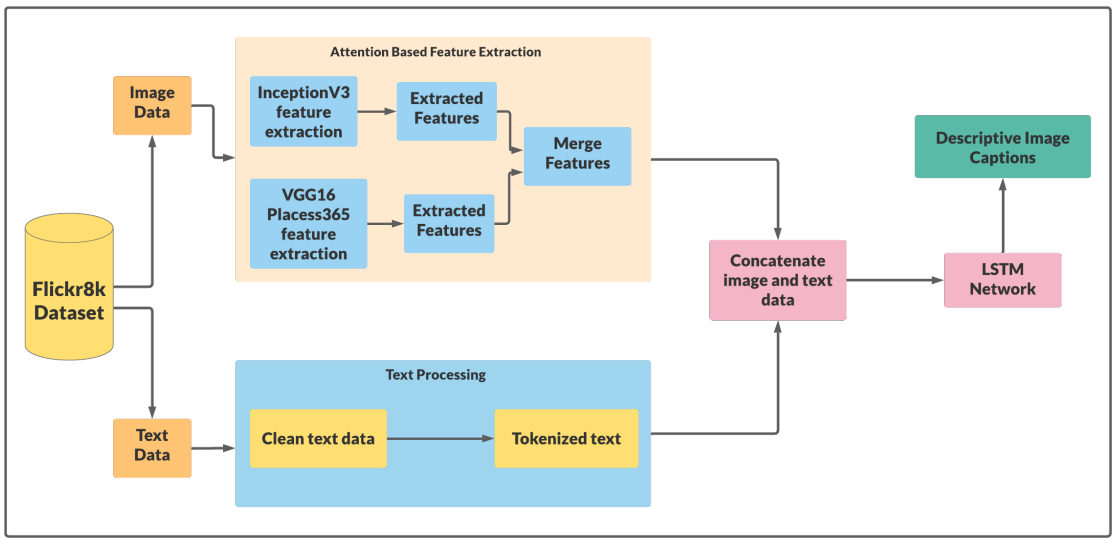

Figure 5: Image Captioning methodology

Figure 5 explains the image captioning methodology in terms of flow of data and processes. The InceptionV3 model is first built for processing image features. The first

layer of the model processes the image features. The input shape for this layer is 2048. The image features after being processed are passed on to a dense layer which reduces the dimensionality of the features to 300 to match with the text embeddings. Following this is a repeat vector function which repeats this process for the maximum length of the text data. The same process is repeated with the VGG16 Places365 CNN. Features extracted from InceptionV3 and Places365 CNN are merged together to form one set of extracted features.

Regarding text generation, LSTM models have been implemented. The text data available in the Flickr dataset have been converted to vector representations on the basis of word sequences. The model learns the sequences through the vector representation to be able to generate captions. This level of learning is carried out by an embedding layer present after the input layer. Finally, the features from the CNN models and processed text were concatenated and passed to an LSTM model with 256 units. The final layer is a dense layer and uses softmax activation to predict the likelihood of the next word in the sequence and passes it appropriately.

5.3.2 Image Classification

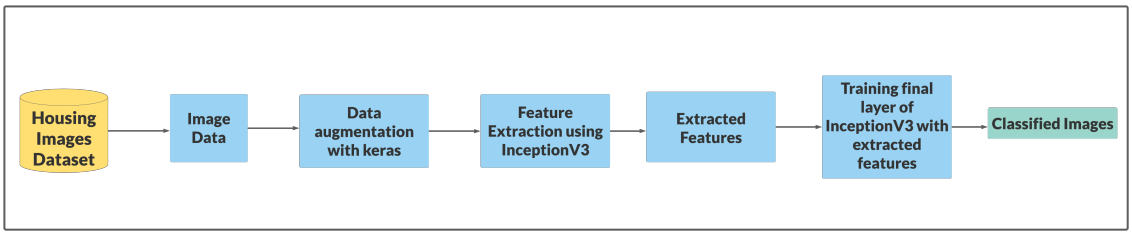

Figure 6: Image classification methodology

Figure 6 depicts the flow of data and processes involved in the image classification model. InceptionV3 model was built for the purpose of image classification. The model would require creating a data generator function which helps in augmenting the images to create more features which would enhance the quality of training. Images are processed in batches of 118. These batches of images are passed to the final layer for training. This process is carried on for 10 epochs first and its performance has been evaluated. The same process is run for 20 epochs next to analyse the improvement in performance. Finally, classification is tested qualitatively by testing on examples from external data and test data.

5.4 Model Training

5.4.1 Image Captioning

Image captioning models are trained on the Flickr8k dataset. This dataset is split into 6000 images for training and 1000 images for testing. Data generators are used in this case as well to load images into the model for training. The images have been loaded in batches of 300 to avoid and RAM related storage issues on Google Colab. Similar to the image classification model, 'categorical_crossentropy' was used as the loss function. In case of backpropagation optimization, 'RMSProp' was used instead of relu. The choice of

activation function is based on literature review according to which RMSProp is a good choice for tasks language related neural network tasks. The model was run for 15 epochs and evaluated according to the metrics mentioned in the previous sections.

5.4.2 Image Classification

The model for image classification is trained using the house image dataset. The dataset is split into three parts proportionally as 80% for training and 20% for testing. Image generator functions are used to augment images as the model runs. This enables a certain level of robustness in the model performance. A data generator was used to load the images in batches of 118 images to avoid running out of RAM storage space. The model is loaded using the 'categorical_crossentropy' as the loss function and 'relu' as the backpropagation optimizer. The model was run for 20 epochs and weights were saved as each epoch completed a run.

6 Evaluation and Results of Developed Classification and Captioning Models

In this section, the results of the tasks carried out and quantitative comparison of experiments is done. Results of image classification are first examined and discussed. This is followed by examining the results of the image captioning experiments. Two experiments have been carried out in this case. The first is based on executing the image captioning model using the InceptionV3 CNN for feature extraction. The second experiment is to use the VGG16 Places CNN for image feature extraction and evaluate performances of both approaches. Evaluation for image captioning is done on the basis of BLEU scores while image classification performance has been analysed using loss and accuracy metrics of the model.

6.1 Image Classification

Image classification is carried out 3 times. The model was run on a different epoch setting each time to be able to find the optimal number of epochs for efficient image classification. The plot of accuracy and validation are shown below in Figure 7. Table 2 shows the actual values of accuracy and loss for better understanding.

Table 2: Image classification results

Epochs	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
10	1.1997	0.6566	1.1245	0.6967
15	0.8282	0.7586	0.7876	0.7686
20	0.6126	0.8287	0.6367	0.8105

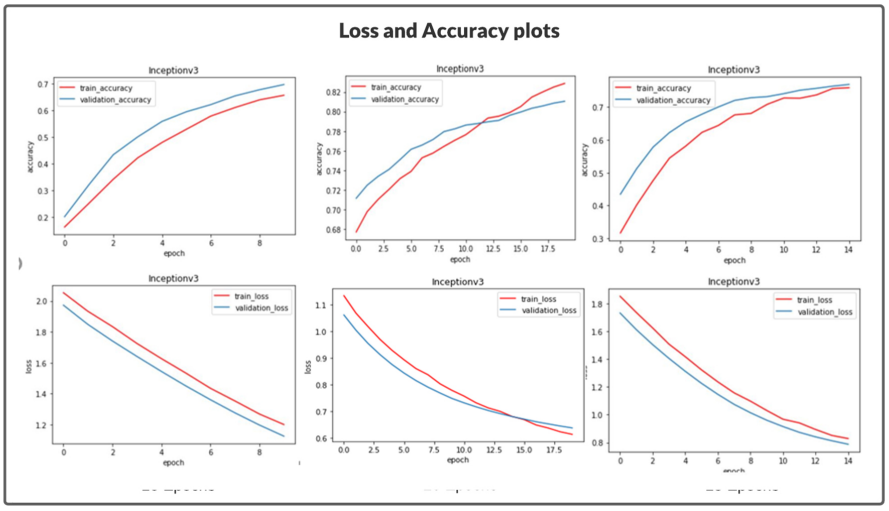

Figure 7: Loss and accuracy plots

6.2 Experiment 1: InceptionV3 Based Feature Extraction for Image Captioning

The first experiment is conducted using the InceptionV3 CNN for global scene feature extraction. This helps provide a baseset of evaluation scores which would help in understanding the effectiveness of the proposed alternate solution. Along with this, an LSTM netowrk is used for image caption generation. The test has been done at 12 and 15 epochs. In Table 3, results for the first experiment have been presented.

Table 3: InceptionV3 based image captioning results

Epochs	BLEU-1	BLEU-2	BLEU-3	BLEU-4
10	0.5322	0.000153	0.2502	0.1312
15	0.5072	0.000145	0.2420	0.1272

It can be seen that the model achieves a BLEU-1 score of 0.53 in the first test with 12 epochs. This has been seen to reduce to 0.50 when the model is run for 15 epochs.

6.3 Experiment 2: VGG16 Places CNN Based Feature Extraction for Image Captioning

The second experiment is similar to what had been done in the first experiment. The aim of doing this experiment is to investigate if the VGG16 Places CNN paired with inceptionV3 would be a better option for an image encoder in order to be able to produce descriptive image captions.Table 4 shows the results of this experiment.

Table 4: InceptionV3 and VGG16 Places365 CNN based image classification results

Epochs	BLEU-1	BLEU-2	BLEU-3	BLEU-4
10	0.5456	0.000199	0.2556	0.1330
15	0.4695	0.000078	0.2178	0.1117

Based on the results it can be seen that the BLEU-1 score drops significantly upon increasing epochs. This indicates overfitting in the model.

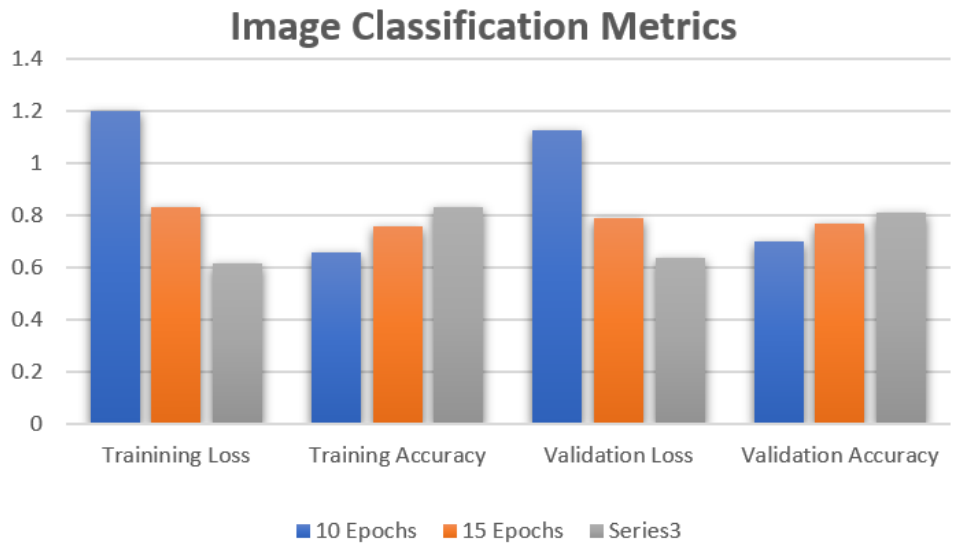

Figure 8: Image classification results

The main issue here is that the validation loss is higher than the training loss. Although the difference is not significantly too high, this could still cause issues of overfitting. The model would not work too effectively when presented with new data. Plotting the loss and accuracy for training and testing data helped in understanding the ideal scenario to obtain a well trained model. At 15 epochs, the model accuracy was reported at 0.7686 which is higher than at 10 epochs. The validation loss was also lesser than the training loss which is a good sign. Validation accuracy has also been seen to be higher than training accuracy.

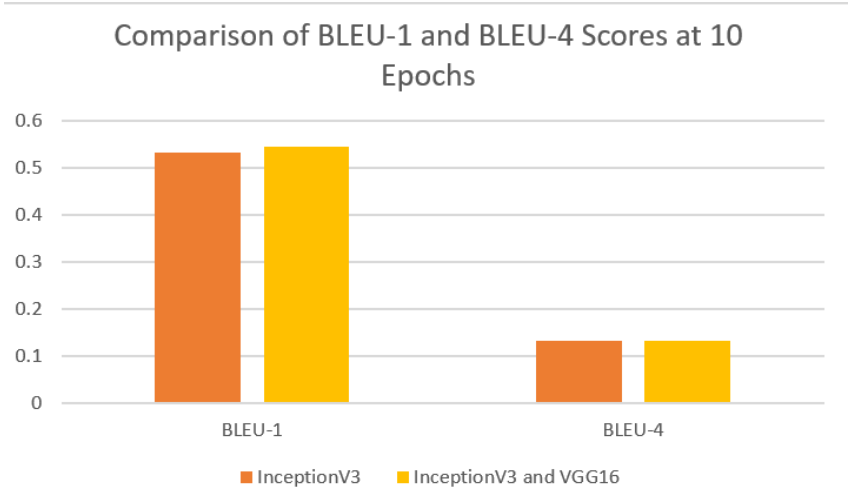

Figure 9: Comparison of BLEU-1 anf BLEU4 scores

In regards to image captioning, it can be seen from Figure 9 that the results of BLEU-1 scores of an image captioning model based on attention has a better performance than traditionally established image captioning models. Generally BLEU-1 and BLEU-4 scores are representative of how well the generated text could represent a defined corpus. The

models of both experiments have been run for 10 and 15 epochs. In both cases, it was noted that the BLEU-1 scores drop significantly on increasing the number of epochs indicating overfitting. This was the case with both experiments. Hence analysis has been narrowed down to looking at the results of 10 epochs. The BLEU-1 score of the attention based model was recorded to be 0.5456 while that of the traditional model was seen at 0.5322. BLEU- 4 scores have also showed similar trends. Improving epochs could improve performance but only to a certain extent. The scores achieved here are satisfactory in the view of this research.

6.4.2 Qualitative Analysis

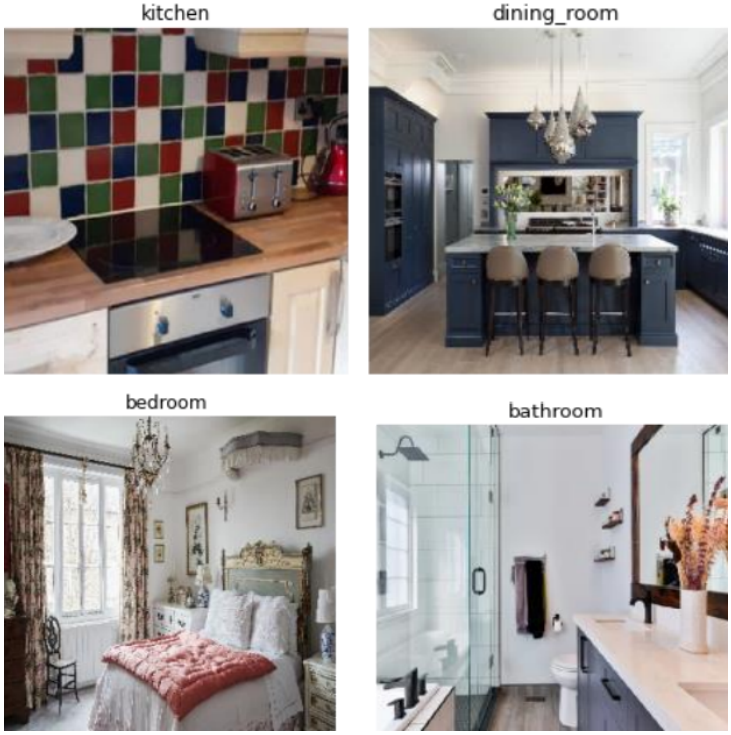

Figure 10: Outputs of Image classification

Qualitatively analysing the results helped in understanding the model performances better.In the case of image classification, models running at 10 and 20 epochs have wrongly classified certain images while most of the images presented to the model run at 15 epochs was classified correctly.

In the case of image captioning models implemented, Good BLEU scores are only representative of how well the model would be able to generate text that is semantically correct. It does not take the content of the information into account. As seen in Figure 11, The captions produced by the InceptionV3 based approach is given as "A boy in a blue shirt and a white shirt is playing soccer". This sentence is not too correct semantically. It does show that the right objects in the images are being recognised. Compared to the combined approach, the output states the caption as "A boy in a red shirt is playing soccer in a field". This statement captures the objects in the image accurately and also takes the surroundings of the subject in the image into account. Additionally, semantic structure has also improved.

Figure 11: Outputs of Image captioning

7 Conclusion and Future Work

The objective of this research was to accomplish two goals. One was to find a way to tag images according to the areas of a house or real estate property. The other objective was to explore and experiment with possible image captioning techniques to create a method of descriptive image captioning that could potentially help in real estate listing websites. The underlying theme of the work done here is to leverage transfer learning techniques. Image classification has been achieved with the use of the InceptionV3 CNN. The model has achieved a validation accuracy of 81.05%. This answers one part of the problem which is tagging images according to the room of a house that it shows. The second part of the objective is to create a descriptive image captioning method. Upon reviewing related work, the possible solution has been narrowed down to attention based mechanisms. Experiments were done to understand the effectiveness of attention based mechanisms in comparison with traditional image captioning mechanisms. On the basis of these experiments, it was noted that the attention based mechanisms have a qualitatively better performance in the context of the objectives. The model has been able to capture the features of not just the subject, but also the surroundings and background features of the image accurately. These approaches to image captioning and image classification provide a fitting solution to the problem of automating the process of image captioning and image tagging with classification and could be beneficial as a product in the field of real estate marketing.

The future aspects of this research could lead to improvement in image captioning techniques. Advancements in text generation could help image captioning approaches significantly. Deep learning techniques like transformers are already efficient in text generation. Coupling this with the ability to process image features could help in the

future. In the context of the real estate field, increasing availability of data could help in improving the quality of deep learning based solutions to such problems. Finally, in reference to the work done here, if put together in the form of a single pipeline architecture along with a user friendly GUI could lead to the creation of a tool that could be used over a wide range of online platforms.

References

Bappy, J. H., Barr, J. R., Srinivasan, N. and Roy-Chowdhury, A. K. (2017). Real estate image classification, *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp. 373–381.

Chollet, F. (2017). *Deep learning with Python*, Simon and Schuster.

Dang, T. X., Oh, A., Na, I.-S. and Kim, S.-H. (2019). The role of attention mechanism and multi-feature in image captioning, *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*, pp. 170–174.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database, *2009 IEEE conference on computer vision and pattern recognition*, Ieee, pp. 248–255.

Hodosh, M., Young, P. and Hockenmaier, J. (n.d.). Flickr8k dataset.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F. and Laga, H. (2019). A comprehensive survey of deep learning for image captioning, *ACM Computing Surveys (CsUR)* **51**(6): 1–36.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K. Q. (2017). Densely connected convolutional networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Kumar, G. and Bhatia, P. (2014). A detailed review of feature extraction in image processing systems.

Liu, L., Tang, J., Wan, X. and Guo, Z. (2019). Generating diverse and descriptive image captions using visual paraphrases, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4240–4249.

Lu, X., Lin, Z., Jin, H., Yang, J. and Wang, J. Z. (2014). Rapid: Rating pictorial aesthetics using deep learning, *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 457–466.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation, *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318.

Peng, Y., Liu, X., Wang, W., Zhao, X. and Wei, M. (2019). Image caption model of double lstm with scene factors, *Image and Vision Computing* **86**: 38–44.

Poursaeed, O., Matera, T. and Belongie, S. (2018). Vision-based real estate price estimation, *Machine Vision and Applications* **29**(4): 667–676.

Puscasiu, A., Fanca, A., Gota, D.-I. and Valean, H. (2020). Automated image captioning, *2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)*, IEEE, pp. 1–6.

Singla, A., Bertino, E. and Verma, D. (2019). Overcoming the lack of labeled data: Training intrusion detection models using transfer learning, *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, IEEE, pp. 69–74.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015). Going deeper with convolutions, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention, *International conference on machine learning*, PMLR, pp. 2048–2057.

Yim, J., Ju, J., Jung, H. and Kim, J. (2015). Image classification using convolutional neural networks with multi-stage feature, *Robot Intelligence Technology and Applications 3*, Springer, pp. 587–594.

You, Q., Pang, R., Cao, L. and Luo, J. (2017). Image-based appraisal of real estate properties, *IEEE transactions on multimedia* **19**(12): 2751–2759.

Zhao, W., Wu, X. and Luo, J. (2020). Cross-domain image captioning via cross-modal retrieval and model adaptation, *IEEE Transactions on Image Processing* **30**: 1180–1192.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. and Torralba, A. (2017). Places: A 10 million image database for scene recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Zou, X., Lin, C., Zhang, Y. and Zhao, Q. (2020). To be an artist: Automatic generation on food image aesthetic captioning, *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, pp. 779–786.