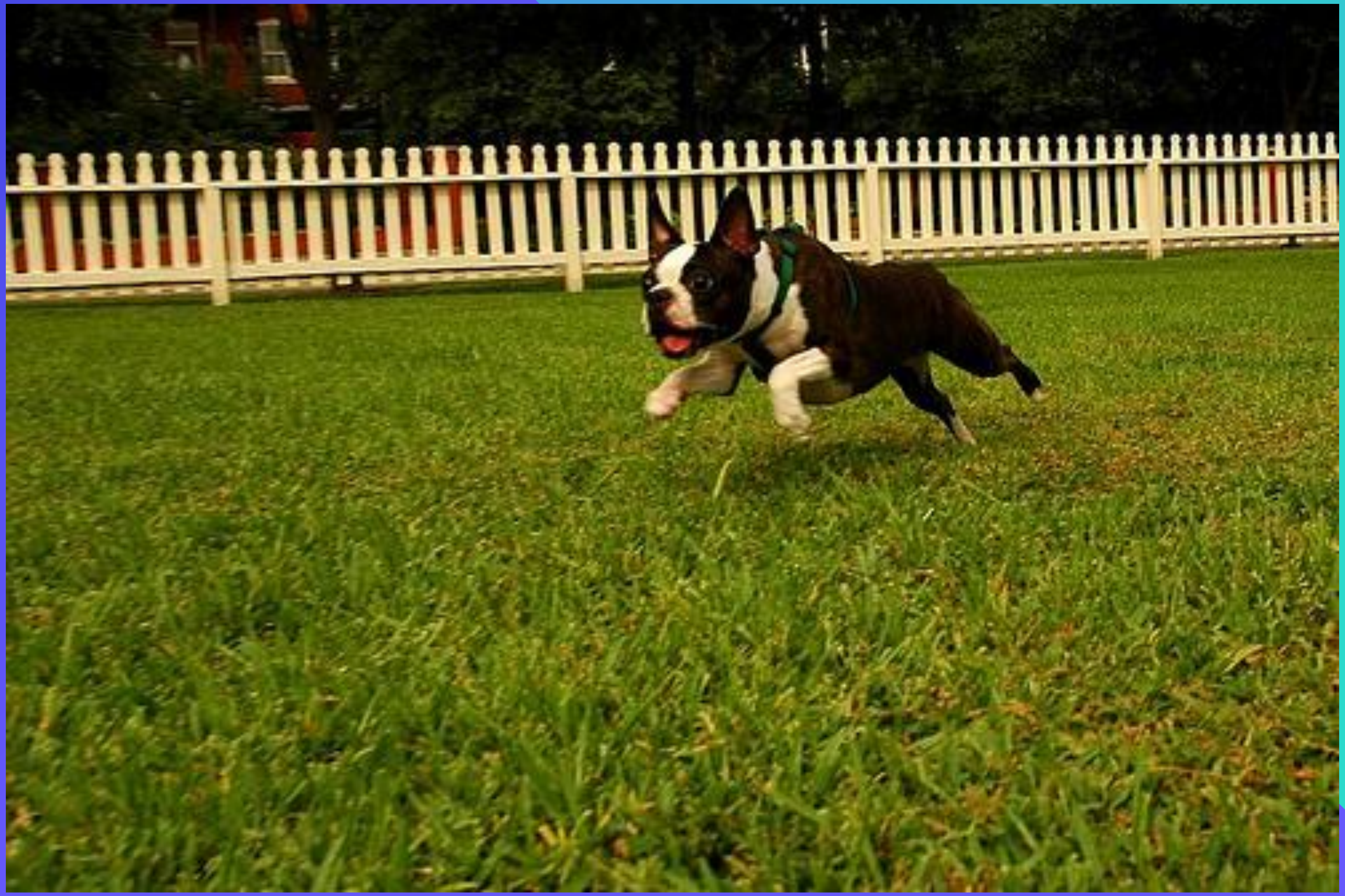
COMS 4995: AML

Image Captioning

Group 25



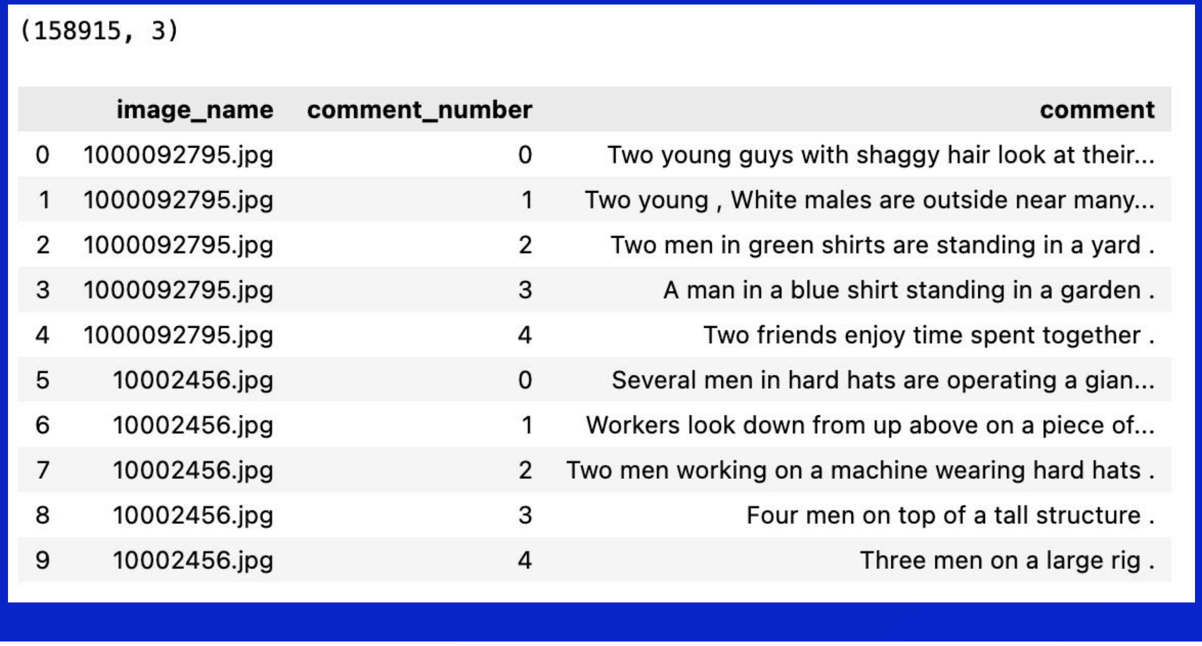
1009434119.jpg

A black and white dog is running in a grassy garden surrounded by a white fence

Data Exploration - Flickr Dataset

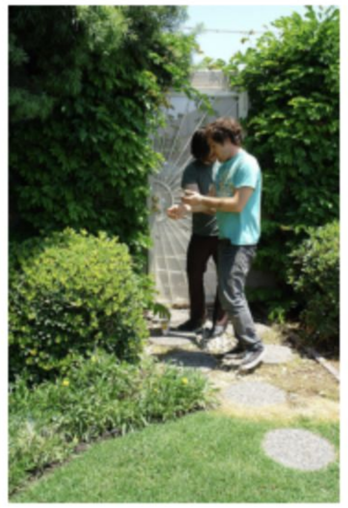
The Flickr Dataset consists two types of data: images and captions

- There are 31783 images in total
- Each image has 5 captions given by different people, so the resultant dataset consists 158915 rows



(158915, 3)

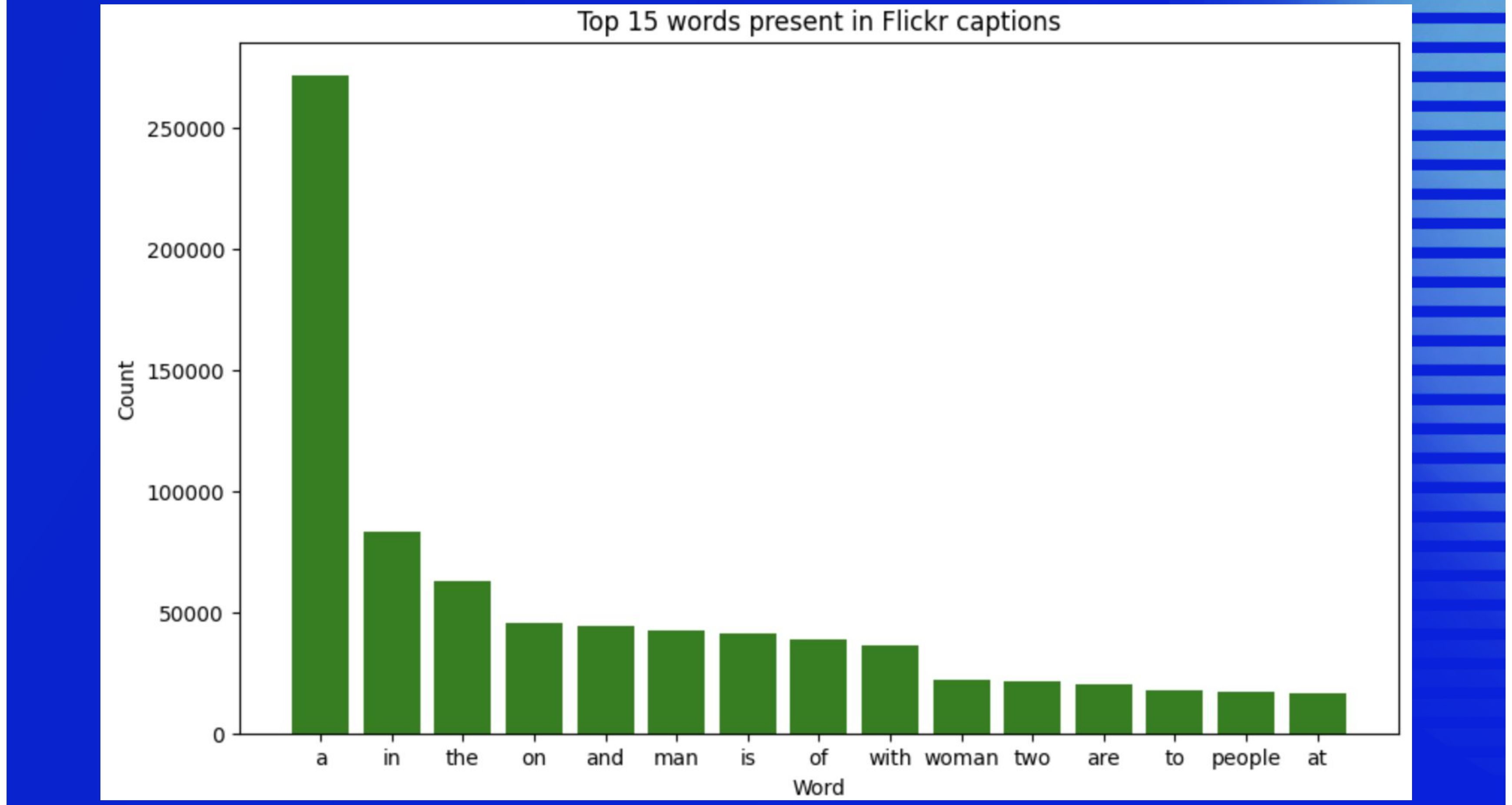
	image_name	comment_number	comment
0	1000092795.jpg	0	Two young guys with shaggy hair look at their...
1	1000092795.jpg	1	Two young , White males are outside near many...
2	1000092795.jpg	2	Two men in green shirts are standing in a yard .
3	1000092795.jpg	3	A man in a blue shirt standing in a garden .
4	1000092795.jpg	4	Two friends enjoy time spent together .
5	10002456.jpg	0	Several men in hard hats are operating a gian...
6	10002456.jpg	1	Workers look down from up above on a piece of...
7	10002456.jpg	2	Two men working on a machine wearing hard hats .
8	10002456.jpg	3	Four men on top of a tall structure .
9	10002456.jpg	4	Three men on a large rig .



Caption 1: Two young guys with shaggy hair look at their hands while hanging out in the yard .
Caption 2: Two young , White males are outside near many bushes .
Caption 3: Two men in green shirts are standing in a yard .
Caption 4: A man in a blue shirt standing in a garden .
Caption 5: Two friends enjoy time spent together .

Word Count Distribution:

- The 15 most common words are: a, in, the, on, and, man, is, of, with, woman, two, are, to , people, at



Top 15 words present in Flickr captions

- The 10 least common words are: solicitor, ther, stumbled, unfortunate, acoustics, runaway, aghast, portraying, mostlyempty, topdown

Data Exploration - Subset Dataset

Subset of Flickr

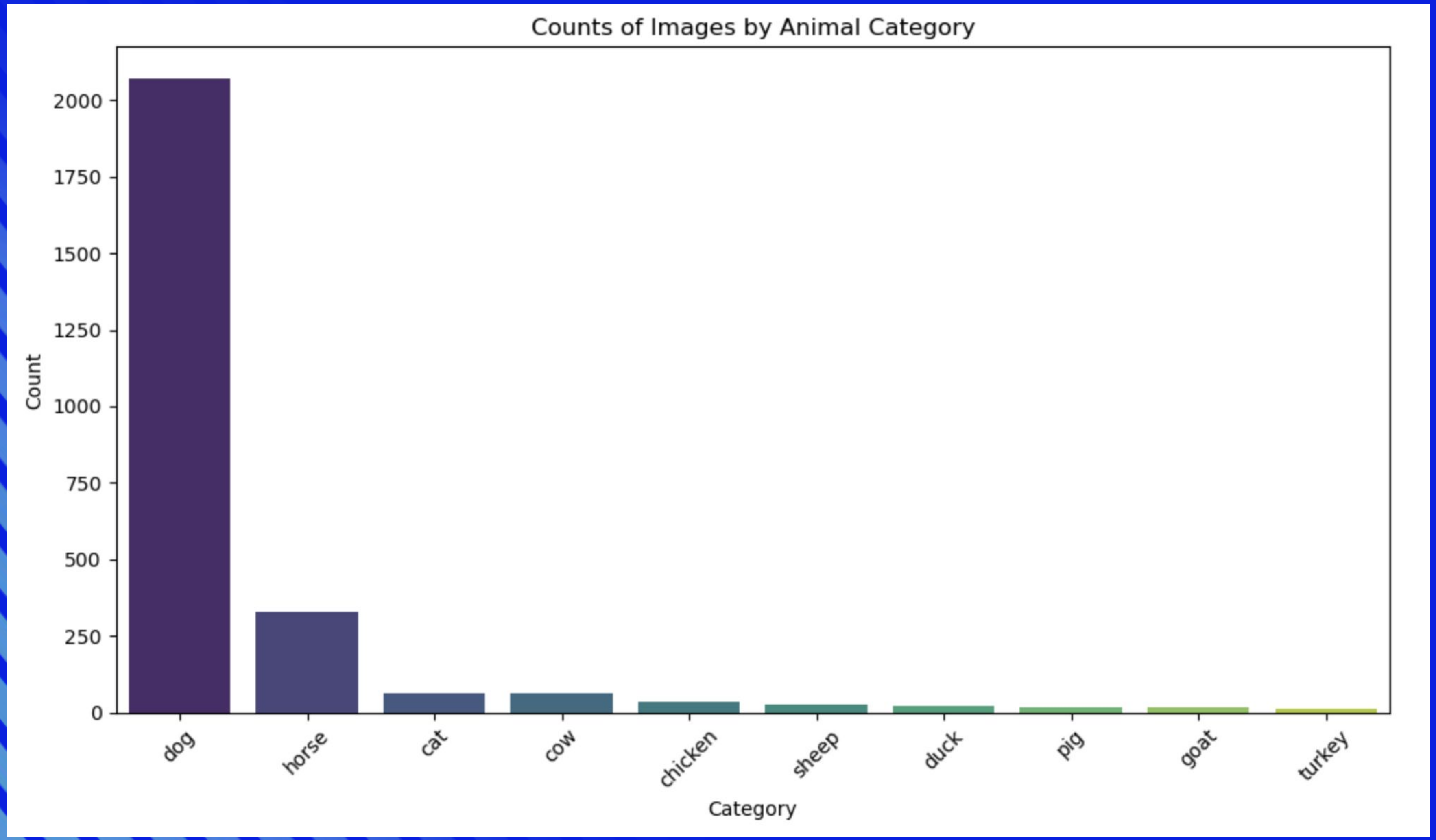
- Our project involves captioning images containing animals that could be useful for visually impaired farm owners or pet owners to monitor their pets remotely.
- After filtering out these images, the size of the animal dataset consists 2644 images and 9936 captions.

Categories of Animals

- On analysing the dataset, we aim to caption images containing the following domesticated animals:

Dog, Horse, Cat, Cow, Sheep, Chicken, Duck, Goat, Pig, Turkey

Distribution of Animal Categories



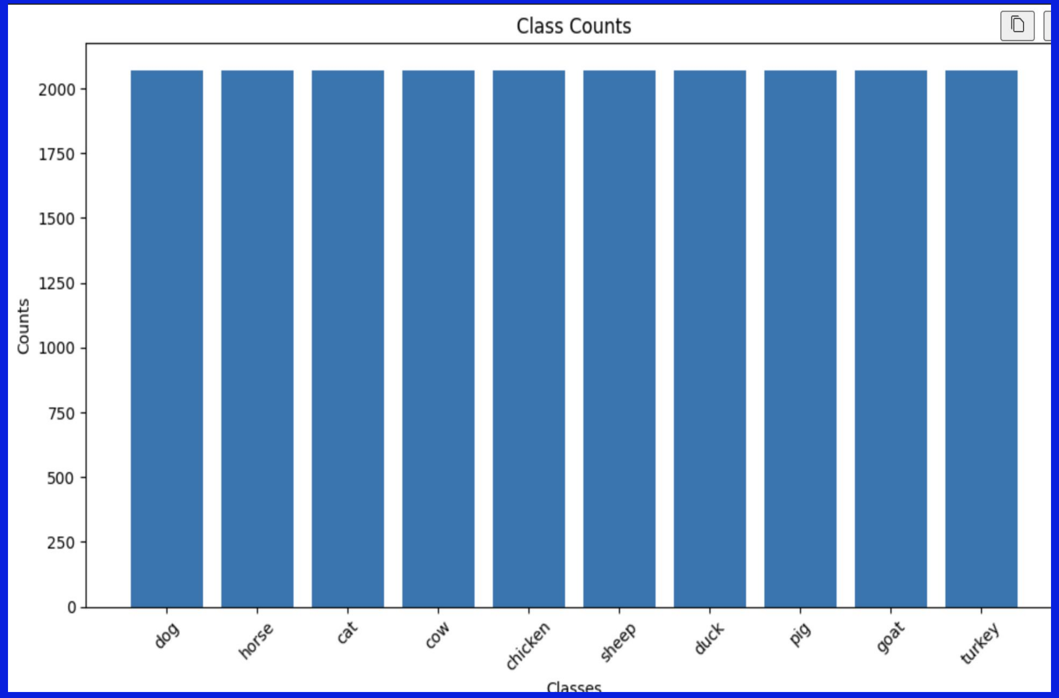
- We observe a huge imbalance in the images with respect to the categories of animals it captions.
- The 'dog' category has the highest count, with over 2000 images, indicating it's the most represented category in this dataset.
- This will be addressed in the sampling slide, so that we train our model on a balanced dataset.

Cleaning and Sampling

Cleaning the Dataset:

- **Relevance Check**: We are focusing only on the animals in the given dataset. Hence, we removed images that do not contain animals and kept those that contain animals.
- **Quality Check**: We checked the captions for spelling and grammatical errors
- **Normalization:** We normalized the vocabulary in the captions. This includes making all text lowercase, removing punctuation and special characters.
- **Eliminating Duplicates**: Finally, we eliminated duplicate images or captions to ensure the diversity of the dataset so that our model does not overfit to repeated examples

Sampling the Dataset:



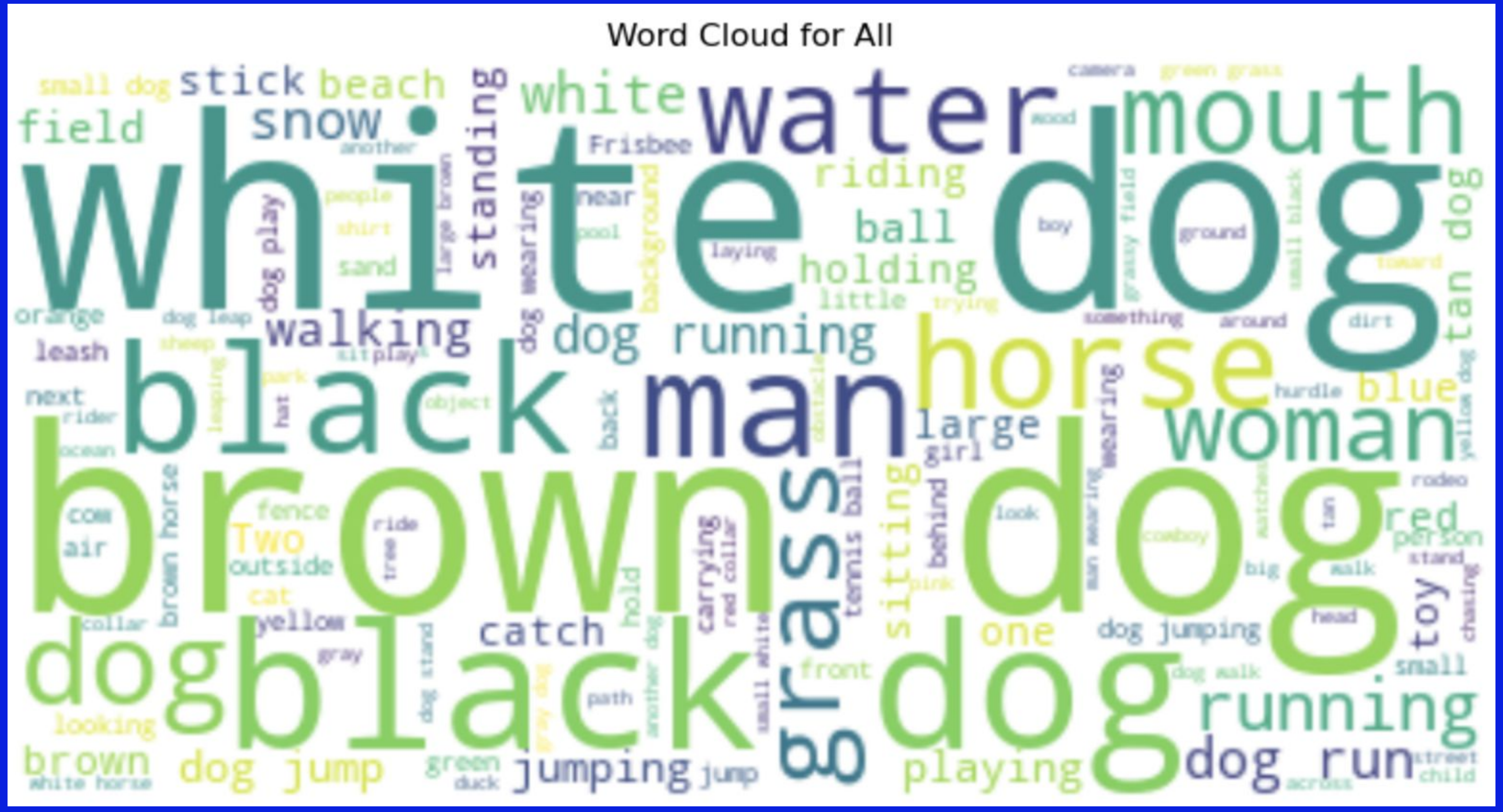
- **Diverse Representation**: We ensured that there is a diverse representation of animals to avoid model bias toward the most common animals
- **Oversampling/SMOTE**: The imbalance in the categories is a problem, so we used SMOTE to create synthetic representation of the minority classes - Sheep, Chicken, Duck, Goat, Pig, Turkey. All the classes now have a count of 2071, as seen in the graph.
- **Stratified Splitting for Cross Validation**: We used Stratified Splitting to tackle the imbalanced dataset to perform cross fold validation.
- **Caption Length Consideration**: We also considered the length of captions, as overly long/short captions could introduce additional challenges for the model.

Data Exploration - Word Clouds

We are interested in exploring some of the common words appear in each animal category, hence, we created word clouds across the entire dataset and for the majority class “dog” and the minority class “turkey”:

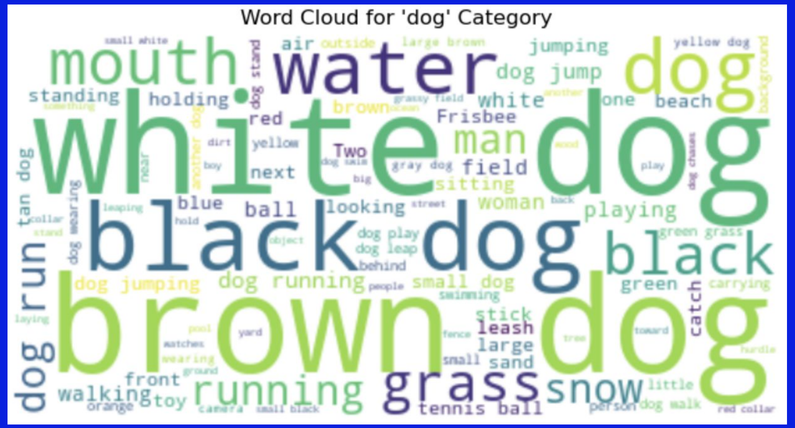
Whole Dataset

- The most frequently occurred words across the entire dataset are “white”, “dog”, “brown”, “black”, “man”, “grass”, etc.
- The prominence of "dog" in the cloud hints at the dataset include a lot of dog images, with a strong emphasis on activities, colors, and settings.


Word Cloud for All

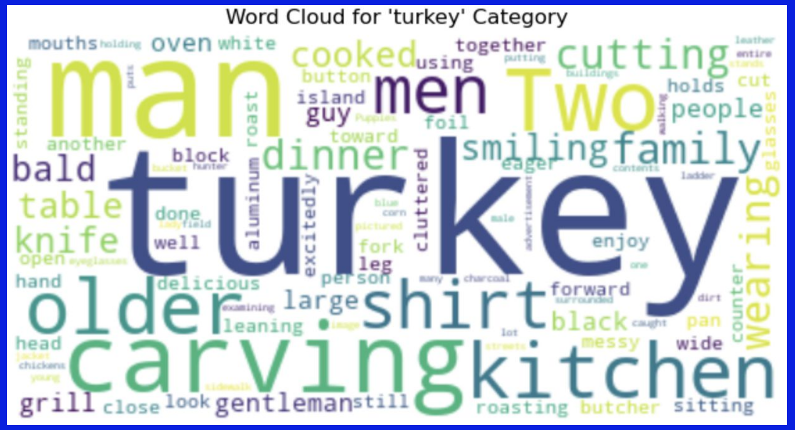
Majority: Dog

- The prominence of action words like "running" along with references to colors and objects such as "ball" suggests that this category includes descriptions of dogs in various activities and environments.


Word Cloud for 'dog' Category

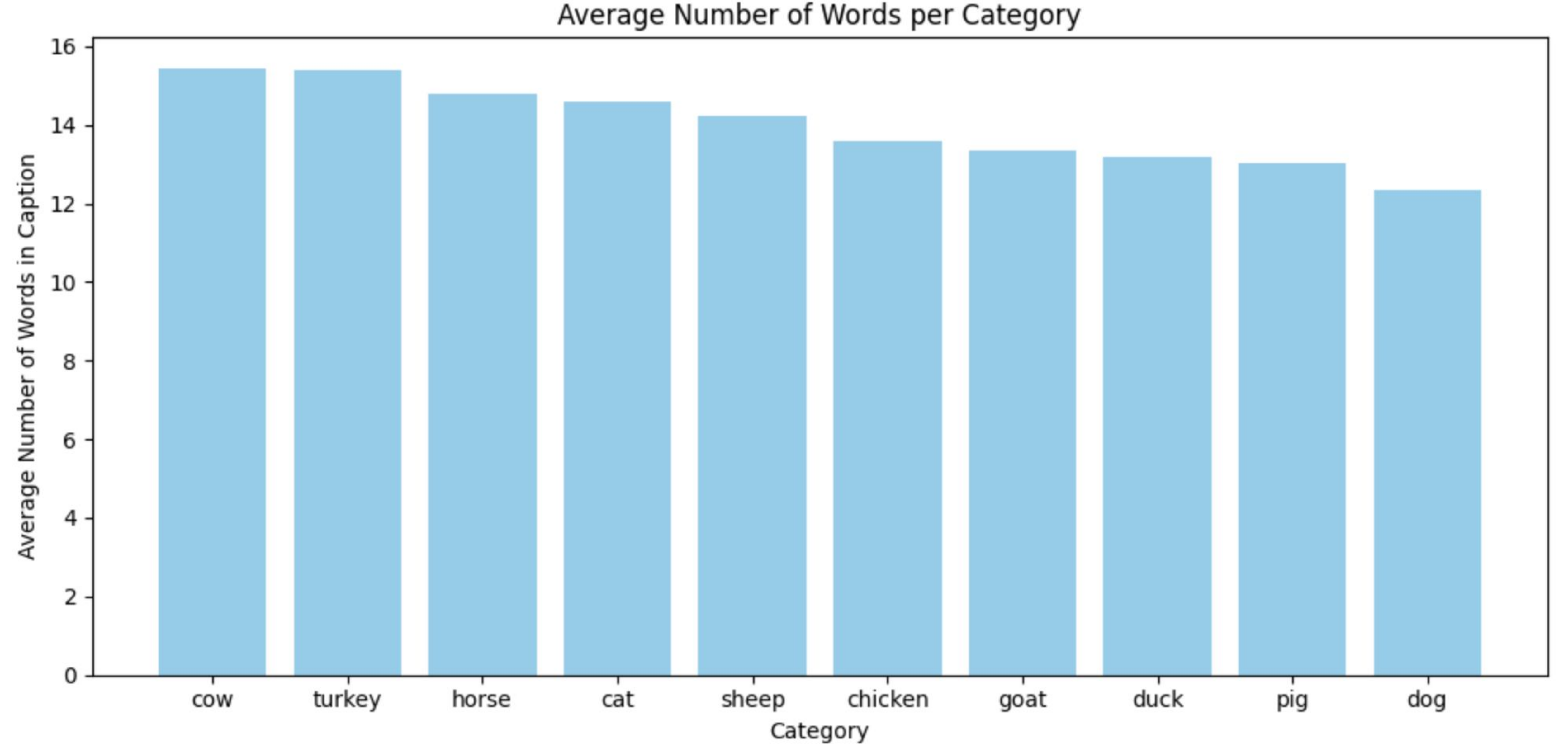
Minority: Turkey

- The presence of words like "knife" suggests that this category include descriptions of cooking or family gatherings centered around meals.


Word Cloud for 'turkey' Category

Data Exploration - Avg Num of Words

We observe the average number of words present in a caption to be most for the Cow (~15) category and least for the most frequently appearing category which is Dog (~13). This tells us information on how the sentence structure is (not very descriptive in nature), and leads us to the idea of using an RNN (such as LSTM) in our NLP model to capture the dependencies between these 10-15 words present.

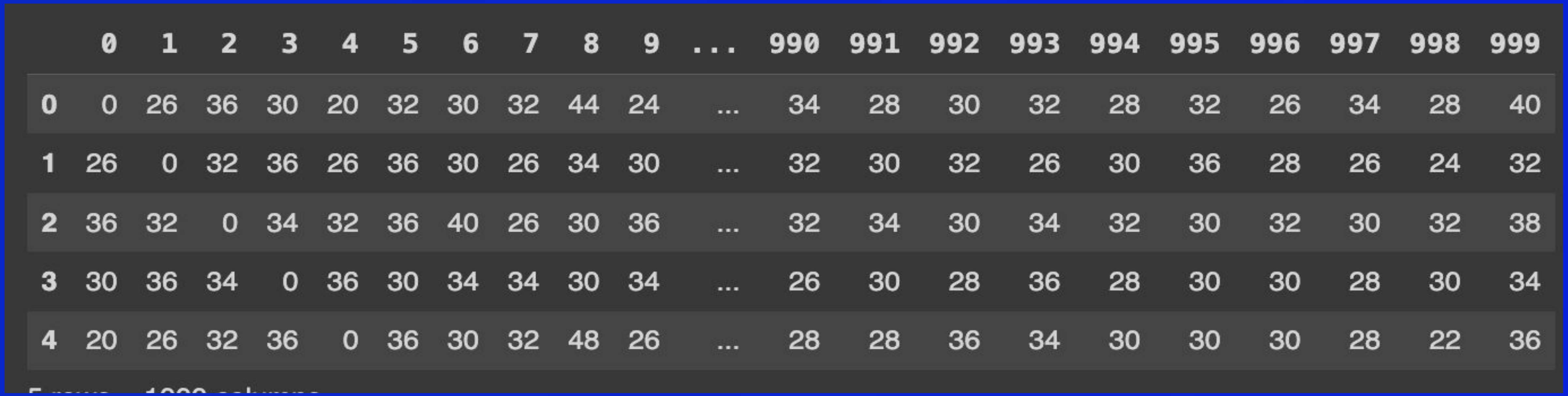


Data Exploration - Hash Value Analysis

Perceptual hashing algorithm can be used to calculate phash values for images in the dataset. Phash values help us analyse similar images and captions in the dataset. The hamming distance between phash values of two images determines the similarity between them. Image Hash Matrix of the phash values can help determine most similar images

Hashing Matrix:

- Hash matrix is a similarity matrix displaying the hamming distance between every image in the dataset. Lesser the value, greater the similarity

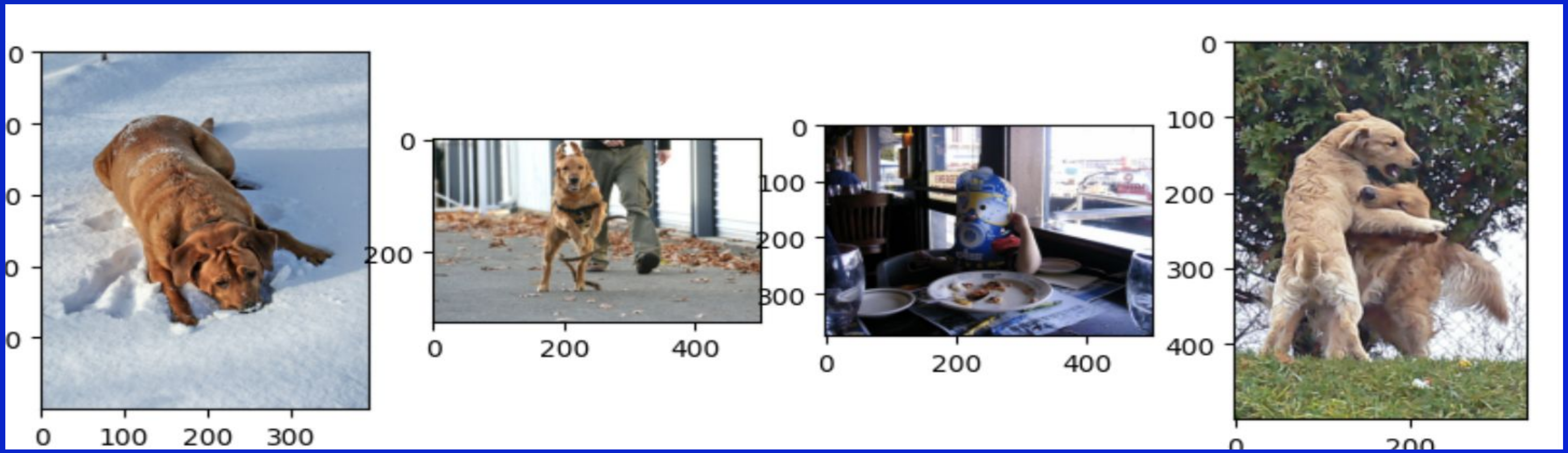


	0	1	2	3	4	5	6	7	8	9	...	990	991	992	993	994	995	996	997	998	999
0	0	26	36	30	20	32	30	32	44	24	...	34	28	30	32	28	32	26	34	28	40
1	26	0	32	36	26	36	30	26	34	30	...	32	30	32	26	30	36	28	26	24	32
2	36	32	0	34	32	36	40	26	30	36	...	32	34	30	34	32	30	32	30	32	38
3	30	36	34	0	36	30	34	34	30	34	...	26	30	28	36	28	30	30	28	30	34
4	20	26	32	36	0	36	30	32	48	26	...	28	28	36	34	30	30	30	28	22	36

5 rows × 1000 columns

Using the hash matrix for a random image we have picked the top few most similar images and their captions:

Similar images



Captions for images with similar hash values

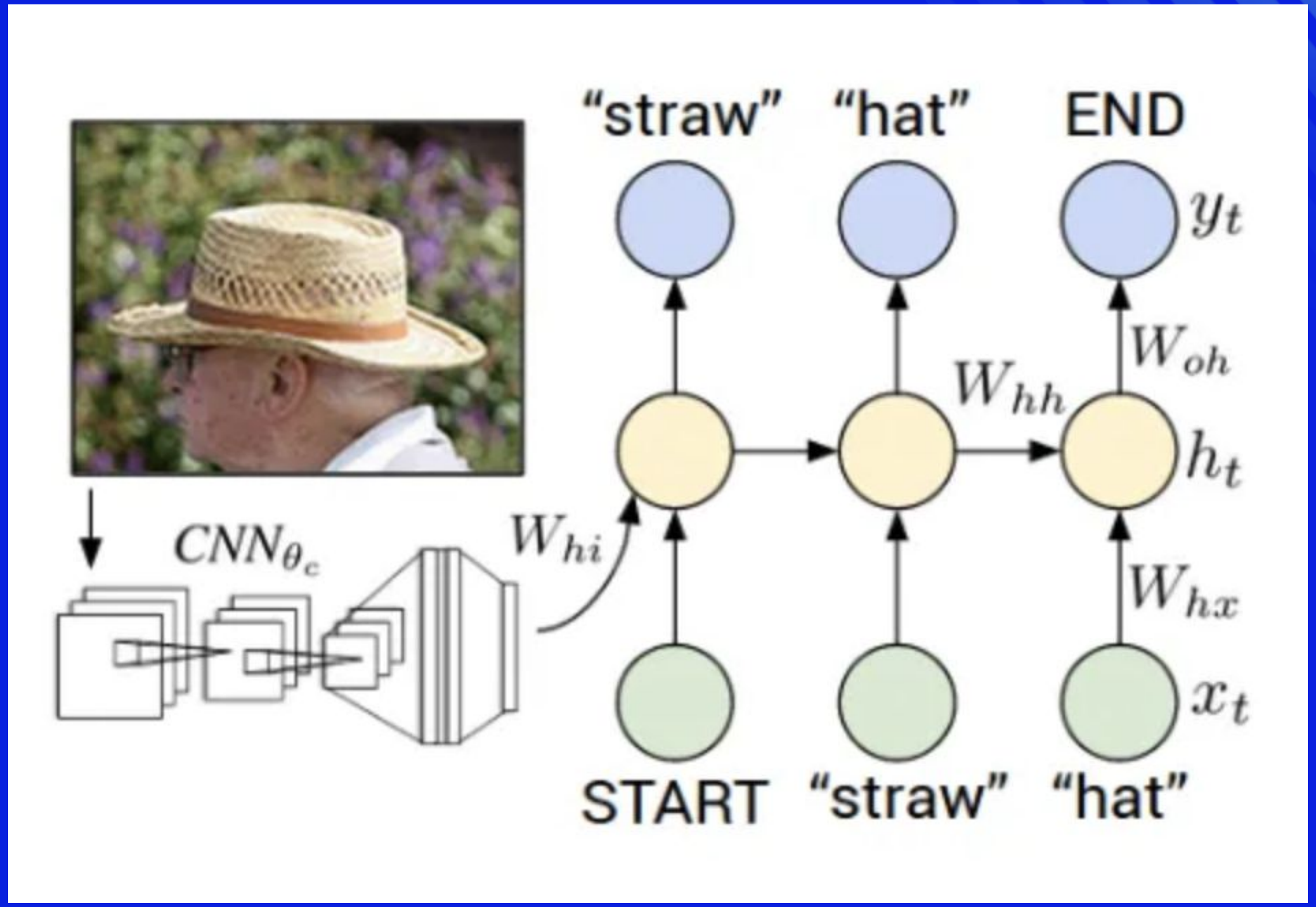
- Leftmost Image: A fluffy little dog running through the snow
- Rightmost Image: Dogs fight in grass over toy

Proposed Machine Learning Techniques

Image Captioning = Natural Language Processing + Computer Vision

(1) CNN + RNN

- CNN: Extracts the features from the image
  - ○ Involves learning features from raw pixels using transformations at every layer
- RNN: Generates a caption from the extracted information from the image
  - ○ eg. Recurring NN that is Long Short-Term Memory, capable of working in sequence prediction tasks
- Can be thought of as Encoder (CNN) + Decoder (RNN), where the last hidden state of CNN is connected to the decoder



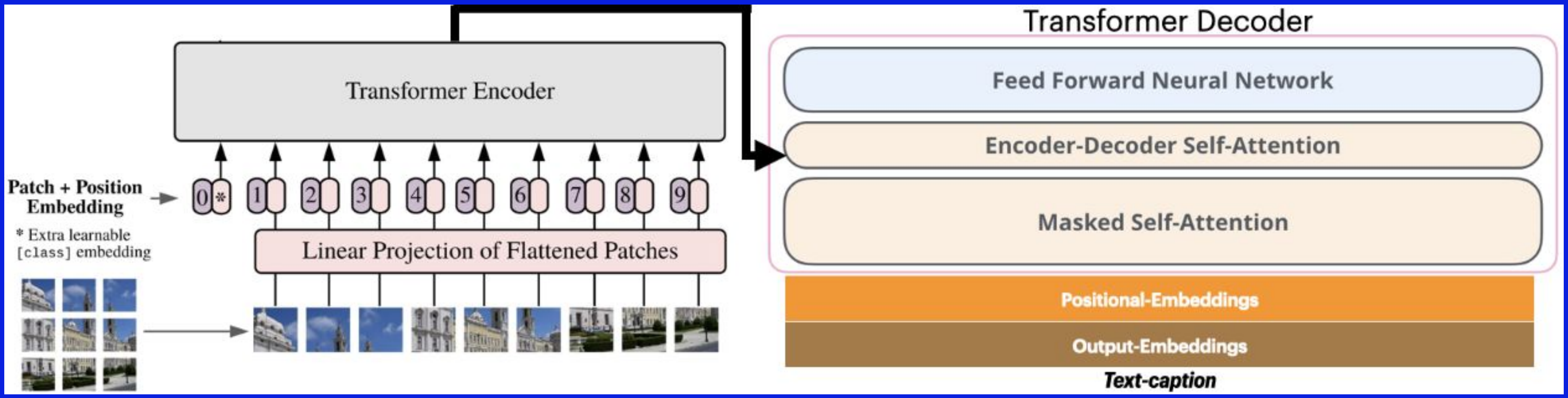
Ref: https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2

Proposed Machine Learning Techniques

Image Captioning = Natural Language Processing + Computer Vision

(2) Using Transformers

- We could use a pre-trained Transformer-based Vision model (eg. ViT) as the encoder that encodes the image and a pre-trained language model as the decoder which is an autoregressive model that generates the caption (eg. gpt, BERT)


Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Transformer Decoder
Feed Forward Neural Network
Encoder-Decoder Self-Attention
Masked Self-Attention
Positional-Embeddings
Output-Embeddings
Text-caption

Ref: https://towardsdatascience.https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers//image-captioning-in-dee-learning-9cd23fb4d8d2

Thank you!