UC San Diego

Explainability Of Multimodal Models

Harin Raja Radha Krishnan
Sai Kaushik Soma
Venkata Harsha Vardhan Gangala

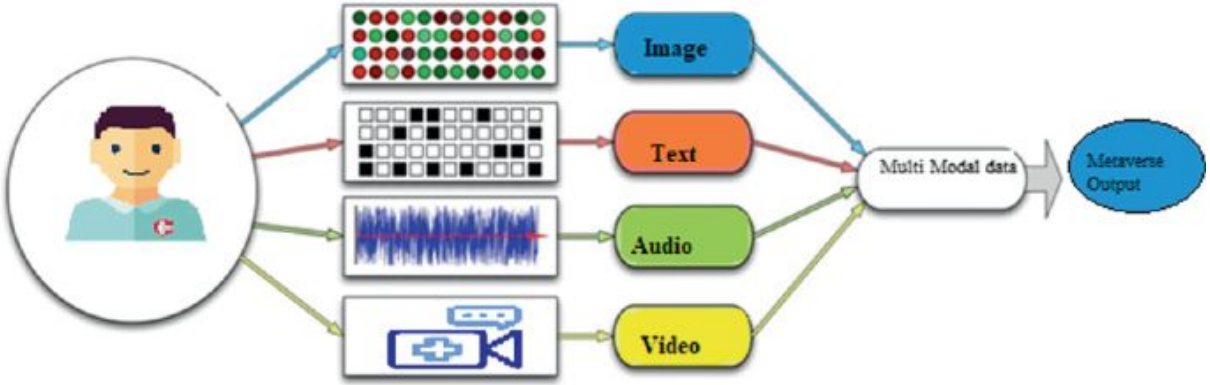
DSC 261 Responsible Data Science | March 12th 2024

Contents

1. Introduction/Motivation
2. Dataset
3. Model Architecture
4. Tabular Model
5. Text Model
6. Visual Model
7. Multi-Modal
8. Results
9. SHAP for Multi-Modal
10. Conclusion

Introduction

- Multimodal refers to the integration of multiple modalities of data within a system or context.
- Multimodal data model allows us to capture a comprehensive view of any scenario, transcending the limitations of relying on a single type of data.
- Methodologies applicable in various sectors: Potential to save lives in healthcare, prevent customer loss in telecom.



Motivation

- Develop a multimodal model integrating text, tabular, and image data and understand how these models make decisions using explainability.
- Tackle inherent challenges: data complexity, diversity, and technical hurdles in data fusion.
- Address the research gap in explaining AI models that blend structured and unstructured data.
- Aim to mirror human decision-making by combining multiple data types for comprehensive insights.
- Drive towards making AI decisions more transparent and actionable, building trust in AI systems.

Pet-Finder Dataset

- We aim to help adoption agencies to operate better, making the process more efficient and effective. If homes can be found for them, many precious lives can be saved — and more happy families created.
- Encompasses various data types:
  - **Categorical Data (Tabular)**: Non-numeric attributes like species, breed, and adoption outcomes.
  - **Numerical Data (Tabular):** Quantitative information such as pet age, cost and weight.
  - **Text Data:** Descriptive narratives about pets
  - **Image Data:** Visual representations offering insights into pet appearance and size.
- The target column with value 1 indicates that pet has been adopted and 0 otherwise.
- The dataset comprises around 15,000 records.

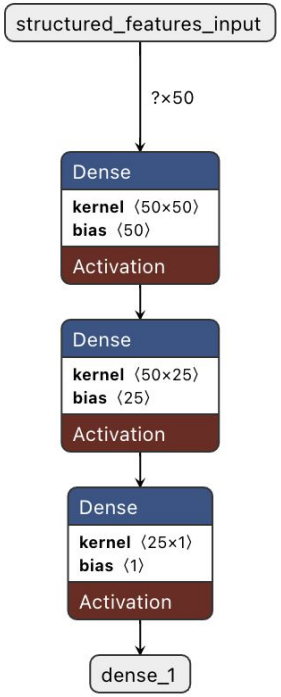
Tabular Model

Model : 3 layer Neural Network

Features Considered: Considered categorical and numerical columns

Hyperparameters:

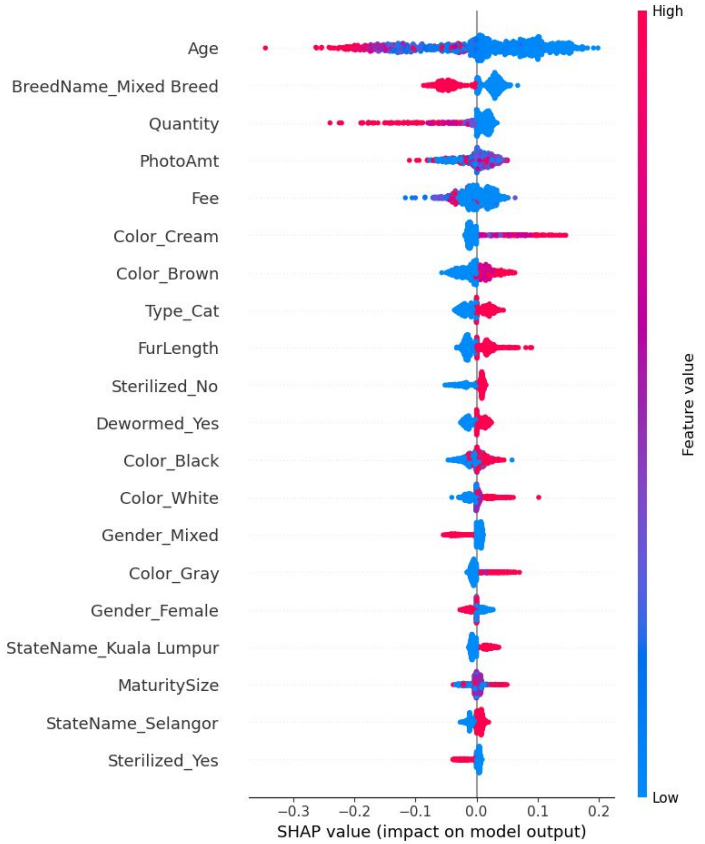
- Epochs: 100
- Learning Rate: 0.1
- Optimizer: Adam
- Loss Function: Binary Crossentropy

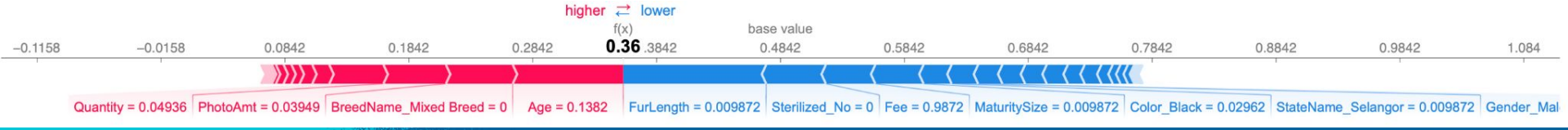


Explaining Tabular Results (SHAP)

.

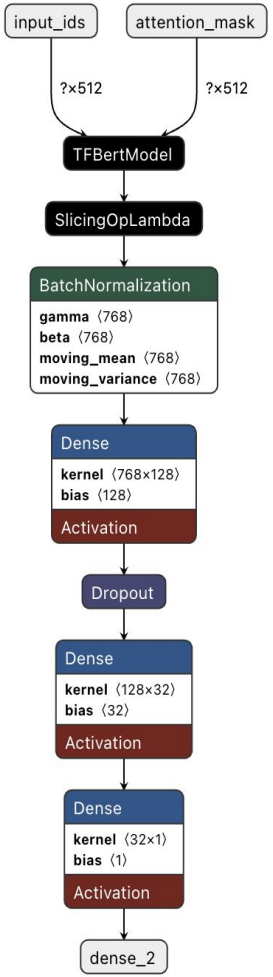
- Red indicates high feature values, blue indicates low.
- Rightward SHAP values suggest faster adoption; leftward, slower.
- Age, breed, color, health, and care status are key influencing factors.
- Positive SHAP values accelerate adoption; negative values decelerate it.
- The baseline represents the model's average prediction for adoption speed.





Text Model

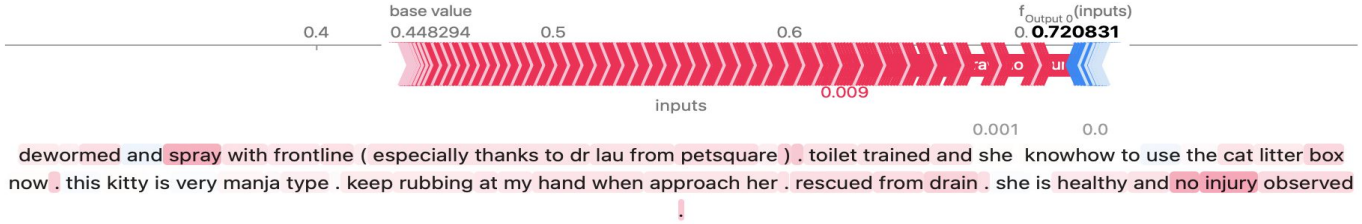
- Model: Finetuning BERT with extra dense layers for binary classification
- Features Considered: Description
- Hyperparameters:
  - Epochs: 25
  - Learning Rate: 1e-3
  - Optimizer: Adam
  - Loss Function: Binary Crossentropy



Explaining Text model Results

- Descriptions highlighting affectionate behavior increase predicted adoption speed.
- Phrases indicating health and care, like "dewormed" and "no injury," positively impact predictions.
- Mention of costs or extra responsibilities for the adopter may negatively influence adoption speed.

Profiles emphasizing a pet's affectionate traits and good health tend to predict quicker adoptions, while references to adopter-incurred costs or responsibilities could potentially delay it.



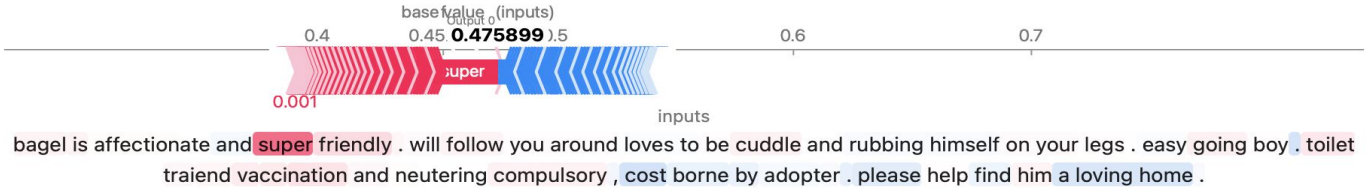
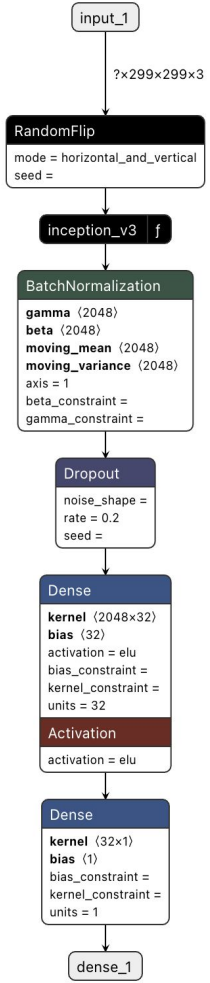


Image Model

- Model: Extending InceptionV3 with dense layers and a binary classifier via transfer learning.
- Features Considered: Pet_img
- Hyperparameters:

  Epochs: 30

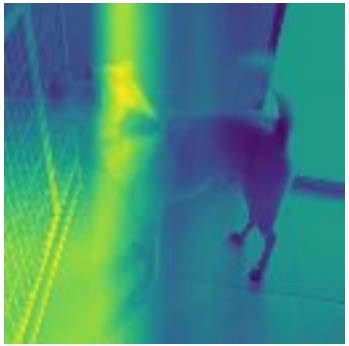
  - Learning Rate: 0.1
  - Optimizer: Adam
  - Loss Function: Binary Crossentropy



Explaining Image Model Results

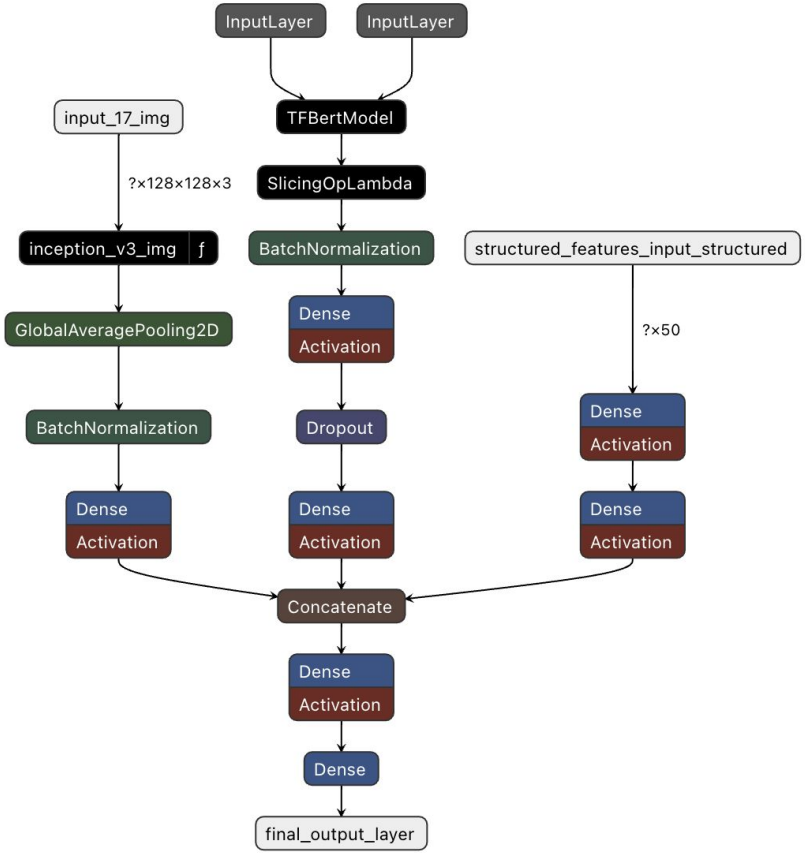
GradCam (Gradient-Class Activation Maps):

- Grad-CAM highlights important regions in images and provides insight into where the model focuses its attention for predictions.
- Despite Grad-CAM highlighting regions the model fails to recognize subjects effectively.
- Limited image availability per pet species hampers the model's learning of specific features.



Multi-Modal Architecture:

- Integrated the 3 fine-tuned models combining different modalities - images, text, and tabular data.
- Each model independently extracts features from its input type.
- The embedding are concatenated into a unified feature vector.
- Additional neural network layers are appended to the concatenated vector.
- Hybrid model further fine-tuned on the complete dataset.
- This method has need further adopted to Bi-Modal architectures - Text + Tabular, and Images + Tabular



UC San Diego

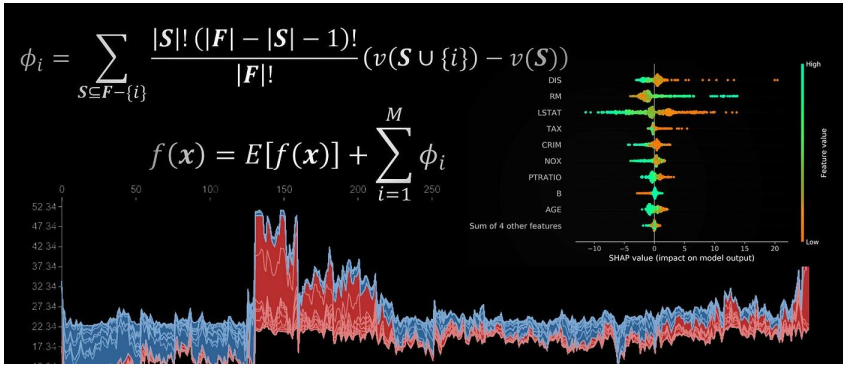
Results

Models	Accuracy	AUC	Recall	Precision
Structured Data (Neural Network)	0.773	0.825	**0.782**	0.775
Text Data (BERT)	0.718	0.842	0.783	0.735
Image Data (Inception V3)	0.552	0.773	0.532	0.582
Hybrid (Image + Structured)	0.615	0.807	0.594	0.635
Hybrid (Text + Structured)	**0.794**	**0.873**	0.768	**0.814**
Hybrid (Text +Image + Structured)	0.652	0.832	0.641	0.679

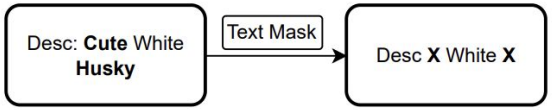
- Using just Text + Tabular features we got the best results
- Problems with Images in the model and limited computation and data

SHAP (Shapley Additive exPlanations)

- Allocates the impact among features in a model based on their contribution to the prediction.
- Determines the marginal contribution of each feature to the prediction output among all features, based on their individual contributions - **Game Theory**.
- SHAP requires a model and a masker to explain the importance of features.
- Maskers manipulate data by hiding parts of features to test impact on model predictions.
- Explainers calculate SHAP values, showing features' impact on prediction deviations.



Text Masker:



- Tokenize text to identify individual components for analysis.
- Apply masking tokens to simulate removal of text segments.
- Assess how each masked text variation influences prediction outcomes.
- As masking is done for all token combinations (complexity - 2^N).

Partition Text Explainer:

- Organizes features into a hierarchy and recursively calculates Shapley values.
- Partitions the text into segments w.r.t to the position or semantic similarity.
- Explanations for each segment are aggregated.

Baseline All-text model (using Text Masker)

- Text-Tabular features are considered.
- Data preparation:
- Each datapoint in Tabular data is converted and concatenated with text into the following format:

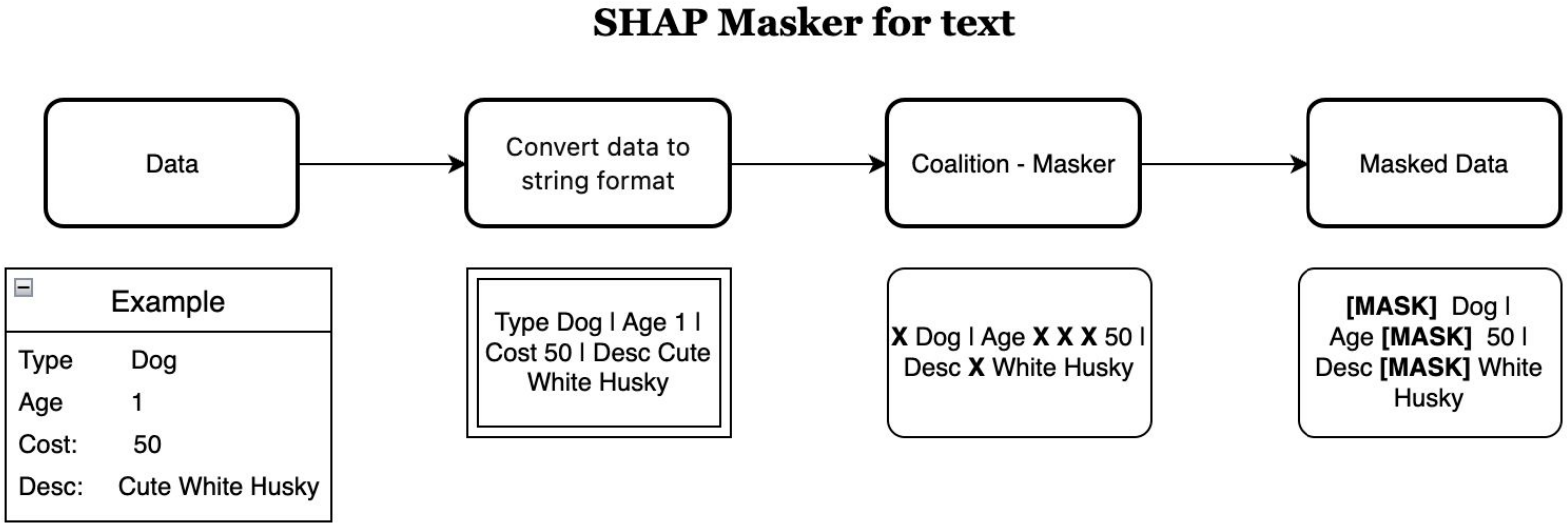
```
"Type Cat | Age 3 | Gender Male | Color1 Black | Color2 White | Color3 NA | MaturitySize 1 | FurLength 1 | Vac
cinated No | Dewormed No | Sterilized No | Health 1 | Quantity 1 | Fee 100 | VideoAmt 0 | PhotoAmt 1.0 | Breed
Name Tabby | StateName Selangor | Description Nibble is a 3+ month old ball of cuteness. He is energetic and p
layful. I rescued a couple of cats a few months ago but could not get them neutered in time as the clinic was
fully scheduled. The result was this little kitty. I do not have enough space and funds to care for more cats
in my household. Looking for responsible people to take over Nibble's care."
```

- BERT model is fine-tuned on the processed data to predict if a pet will be adopted.

Shortcomings of All-Text Masker

- SHAP's Text Masker treats the entire data point as a single text block for masking.
- Incorrect maskings can occur, mixing attribute names and values, leading to ambiguity.
- Ambiguous maskings may result in misleading SHAP explanations, influencing model outcomes improperly.
- To resolve this, distinguishing between attribute names and values during masking is crucial for clear SHAP explanations

Baseline Masker: (All Text)

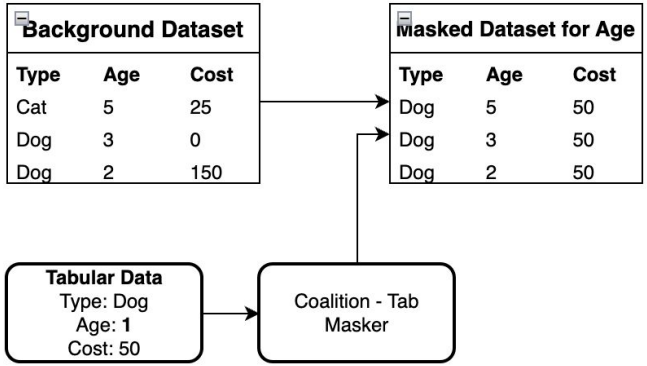


Tabular Masker:

- Mask individual features by replacing them with values from a background dataset to isolate their impact.
- Systematically apply different combinations of masks to the data.
- Evaluate the model's output with and without each feature.
- Complexity - Sample size * Background Dataset Size * 2^N

Tabular Partition Explainer:

- Groups features hierarchically based on correlation values.
- Aggregates effects across partitions to determine overall feature importance.



Joint Masker

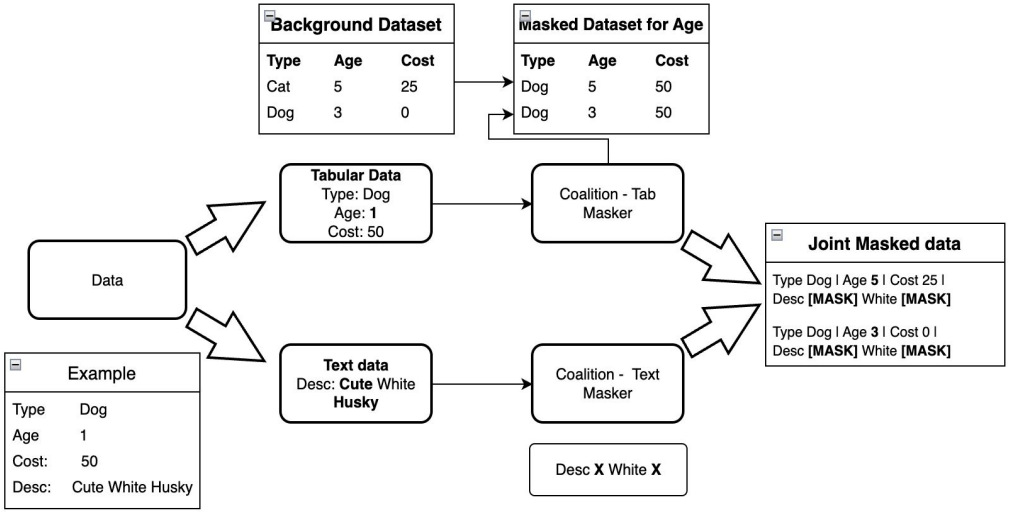
The masker integrates both text and tabular maskers to explain the all-text model.

Tabular masker:

- Background dataset is randomly sampled from the test data (100 samples)
- Background dataset is used to replace the columns where its masked
- The masked data is returned

Text masker:

- Shap's text masker is used to mask the relevant tokens generated by the BERT tokenizer.



Tabular Mask

```
original Data   [48 1 2 1 1 250 0 4.0]
Tabular Masks  [False  True  True False  True  True  True  True]
```

Masked Data	Age	MaturitySize	FurLength	Health	Quantity	Fee	VideoAmt	PhotoAmt
0	5.0	1.0	2.0	1.0	1.0	250.0	0.0	4.0
1	1.0	1.0	2.0	1.0	1.0	250.0	0.0	4.0
2	12.0	1.0	2.0	1.0	1.0	250.0	0.0	4.0
3	2.0	1.0	2.0	1.0	1.0	250.0	0.0	4.0
4	8.0	1.0	2.0	1.0	1.0	250.0	0.0	4.0

Text Mask

```
original Data   ["JoJo is my cute lovely girl. She's so sweet, odorable & manja doggie. Due to my personal reasons, i must let her go to a ne
w good owner & lovely forever home. Hope someone can really take good care, love and being nice to her, could spend more times with her. Anyo
ne interested pls whatsapp or sms me Grace. Thanks for viewing this post and hope to share it."]
Text Masks  [ True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True False
 False False False False False  True  True  True  True  True  True  True
  True False False False False False False False  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True False False False False False False
 False False False False False False False False False False False False
 False False False False False False]
Masked Data  (array(["jojo is my cute lovely girl . she ' s so sweet , odorable & manja doggie . [MASK] i must let her go to a new [MASK] hop
e someone can really take good care , love and being nice to her , could spend more times with her . [MASK]"],
      dtype='<U226'),)
```

Joint Masker

```
original Data
 [96 2 2 1 1 50 0 1.0
 'Found a COCKER SPANIEL two days ago in TAMAN SETAPAK with a leash attached to her. took the dog to a vet where she is currently boarding. F
emale, brown in color appox eight years old with cataract on both eyes - almost fully blind. Please call me at if YOU are willing to give her
a permanent home.']
Entire Mask
 [False  True  True False  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True]
Masked Data
    Age  MaturitySize  FurLength  Health  Quantity   Fee  VideoAmt  PhotoAmt  \
0  5.0           2.0        2.0     1.0       1.0  50.0       0.0       1.0
1  1.0           2.0        2.0     1.0       1.0  50.0       0.0       1.0

                                      Description
0  found a cocker spaniel two days ago in taman s...
1  found a cocker spaniel two days ago in taman s...
```

Proposed Joint Masker

- Joint masker masks tabular and textual features simultaneously
- Categorical data is ordinal encoded for the explainer to find correlations before passing it into the masker, then decoded for model predictions.
- SHAP employs joint masker when explainer called, addresses pitfalls with all-text masker.
- Partition Explainer is used to explain the BERT model trained on the tabular and textual data using Joint masker.

```
cols_to_str_fn = lambda array: " | ".join(
    [f"{col} {str(val)}" for col, val in zip(df, array)])
```

```
"Type Cat | Age 3 | Gender Male | Color1 Black | Color2 White | Color3 NA | MaturitySize 1 | FurLength 1 | Vac
cinated No | Dewormed No | Sterilized No | Health 1 | Quantity 1 | Fee 100 | VideoAmt 0 | PhotoAmt 1.0 | Breed
Name Tabby | StateName Selangor | Description Nibble is a 3+ month old ball of cuteness. He is energetic and p
layful. I rescued a couple of cats a few months ago but could not get them neutered in time as the clinic was
fully scheduled. The result was this little kitty. I do not have enough space and funds to care for more cats
in my household. Looking for responsible people to take over Nibble's care."
```

SHAP Value Generation Process

- Select random instances from the test set for analysis.
- Employ a standard (100-instance) background set from the test set for multimodal analyses.
- Pass the Model and Joint Masker to the SHAP partition explainer.
- SHAP construct masks for each data point, segregating masks for tabular and text features for analysis.
- Determine SHAP values by evaluating each instance's marginal contribution by forming coalition.
- Visualize results with SHAP's text plot feature for interpretation.

Conclusion

- Examined models combining text, tabular, and image data. Observed a decline in performance when incorporating image data due to the heterogeneity of images.
- Integrated text and tabular features to provide comprehensive explanations for model predictions.
- Proposed joint masker for simultaneous masking of tabular and text data, aimed at mitigating potential issues encountered with SHAP's text explainer when applied to text-tabular datasets.
- Enhanced the explainability of multimodal models for better insights into model decisions using pet finder dataset.

Future Work

- Extend our exploration to other explainability tools such as LIME.
- Employ vision explainers such as AsticaVision for image-to-text conversion, enhancing multimodal model interpretation.
- Explore additional data types, like audio or time-series, to enrich multimodal models.
- Aim for improvements in speed and efficiency of explanation mechanisms.

Challenges & Learnings

- Integrating text embeddings with tabular data proved challenging due to dimensional inconsistencies.
- Initial use of Customer Churn Prediction dataset highlighted synthetic data issues, underscoring data quality's importance. Importance of verifying data and label integrity was emphasized.
- Utilized CT-GAN for creating synthetic tabular data from few samples, showcasing its effectiveness.
- Recognized that models might produce illogical explanations despite seemingly accurate outcomes; vigilance needed in varied test distributions.
- Stressed the importance of backing up data, code, and models :)

References

- Burton, James & Al Moubayed, Noura. (2023). SHAP Explanations for Multimodal Text-Tabular Models. 10.21203/rs.3.rs-3405528/v1.
- Lundberg, Scott & Lee, Su-In. (2017). A Unified Approach to Interpreting Model Predictions.
- Enja Kokalj et al. “BERT meets shapley: Extending SHAP explanations to transformer-based classifiers”. In: (2021), pp. 16–21

UC San Diego

Thank you!

Data Preprocessing

- The pet entry with a missing description and no images has been excluded from the dataset (~120 entries).
- Multiple entry columns like color, breed etc have been merged into a primary column
- Importance to categorical features given based based on their order of importance. Ex: Color1 is given more weight in than color 2 and color 3, similarly for breed
- One hot encoded the categorical features
- Imputed data using mode for missing values (~60 entries)

Tabular Masker:

- Applies masks to evaluate feature groups' impact on predictions.
- Iterates over feature coalitions, altering data with background values.
- Background values sampled from data to establish prediction baselines.
- Calculates SHAP values based on changes in model output.
- Partition Explainer groups features hierarchically, efficiently computing SHAP values based on similarities.