**UC Davis Graduate School of Management**

BAX 423 California's Next Top Big Data Scientist
Final Project Report

# SEE through your EARS

---

***Audio-Visual Transformation: Enhancing Inclusivity for the Visually Impaired***

**Submitted by: Group 14 (RATS)**

## Table of Contents

## Executive Summary

The "See through your Ears" project aims to develop a prototype of a product that converts visuals into real-time audio descriptions for individuals with visual disabilities. With an estimated 40-50 million as per WHO blind people worldwide, there is a significant need to enhance inclusivity and independence for this community. With a descriptive sense of the surroundings through audio, the product aims to empower individuals with visual disabilities.

In this report, we evaluate the success of the initiative by identifying key metrics that would help track the value and engagement of this product once it is commercialized. These metrics provide valuable insights into the reception of the product, its impact on users, market penetration, and potential expansion strategies. Analytics play a crucial role in the development and success of this project, enabling data collection, pre-processing, and speech translation.

The product utilizes the COCO (common objects in context) image captioning dataset for training and evaluating the model. The model architecture includes components for image captioning, text-to-speech conversion, and data processing. The report will also provide snippets of the model results to demonstrate the progress made. As this is a prototype, we have included our forward-looking ideas to improve model accuracy and commercialize the product, enhancing its usability and portability.

Through product development, partnerships, and effective marketing, the project has a potential to achieve its objective of enhancing inclusivity and independence for people with visual disabilities. By leveraging analytical techniques and insights gained from data analysis, this prototype can further enhance the accuracy and efficiency of the model, ensuring that the converted visuals are very precisely described in real-time audio, as we plan to scale its usability for a larger target audience.

## Business Objective

Our goal is to bridge the gap between the visual world and those who are visually impaired by developing an innovative product called "See through your Ears." By harnessing the power of deep learning, we aim to convert visuals into real-time audio descriptions, enabling individuals with visual disabilities to gain a descriptive sense of their surroundings.

## Business Initiative

Below are some ideas and actionable steps on comprehensive business initiatives that are aimed at transforming the lives of individuals with visual disabilities. However, this of course needs to be evaluated at a company level that adapts to this solution, and their budget allowance will ultimately drive the implementation forward.

- Product Development and Enhancement

- Conduct market research and user testing to understand the specific needs and preferences of people with visual disabilities.
- Collaborate with assistive technology experts, audiologists, and individuals with visual disabilities to develop and refine the "See through your Ears" product.
- Continuously improve the product based on user feedback, technological advancements, and evolving accessibility standards.

❖ Active Partnerships and Outreach:
- Establish partnerships with organizations, institutions, and NGOs working in the field of visual disabilities to increase awareness and reach a wider audience.
- Collaborate with healthcare professionals, rehabilitation centers, and schools for the visually impaired to promote and distribute the product.
- Organize workshops, seminars, and awareness campaigns to educate individuals with visual disabilities, their families, and caregivers about the benefits and functionalities of the product.

❖ Marketing and Communication:
- Develop a comprehensive marketing strategy targeting individuals with visual disabilities, their families, and relevant stakeholders.
- Utilize online platforms, social media, and assistive technology forums to raise awareness about the product and its features.
- Highlight real-life stories and testimonials from users to demonstrate the positive impact of the product on the lives of people with visual disabilities.

The mentioned three actions will lay the foundation for developing a high-quality product, expanding its reach through partnerships, and effectively promoting it through strategic marketing and communication. By focusing on these areas, the adapting company can maximize the impact of business efforts and achieve the objective of enhancing inclusivity and independence for people with visual disabilities. Once these are done, other business actions to explore could be estimating - Accessibility and Affordability, Continuous Support and User Engagement and Research and Innovation.

## Metrics of Success

To evaluate the success of the initiative and measure the impact of the our product, the business can consider the following key metrics according to our knowledge:

- User Adoption and Engagement: This metric will provide insights into how well the product is being received and utilized by the target audience. Key metrics could be:
    - Number of product downloads or installations (to gauge the level of interest and initial adoption of the product)

    - Active user base and user retention rate (to track the number of active users over time and measuring the rate at which users continue to engage with the product)
    - User feedback and satisfaction ratings (to gathering feedback through surveys, reviews, or ratings to gauge user satisfaction and identify areas for improvement)

- Impact on Users: This metric focuses on assessing the actual impact of the product on the lives of people with visual disabilities. Key metrics could be:
    - User testimonials and success stories (for collecting qualitative feedback and stories from users about how the product has improved their daily lives, independence, and inclusivity)
    - Surveys or interviews measuring improvements (for conducting surveys or interviews to gather quantitative data on the users' ability to navigate their surroundings, understand visual information, and engage with the environment after using the product)
    - Quantitative data on assistive feature usage (for analyzing data on the usage of specific product features, such as audio description functionality, to understand the extent to which the product is meeting users' needs)

- Reach and Awareness: This metric evaluates the reach and visibility of the product, indicating its market penetration and impact on the target audience. Key metrics could be:
    - Website traffic and social media metrics (for monitoring website visits, unique visitors, page views, as well as engagement metrics on social media platforms (followers, likes, comments, shares)
    - Media coverage and press mentions (for tracking the number of media outlets or publications that have covered the product, indicating its visibility in the public domain)
    - Attendance and participation in events (for assessing the level of engagement and participation in workshops, seminars, and awareness campaigns, indicating the effectiveness of outreach efforts)

Lastly, regular data analysis and ongoing evaluation will help guide strategic decision-making and drive continuous improvement.

## Role of Analytics in the Project

Analytics plays a crucial role in the development and success of the "See through your Ears" project. By leveraging an ensemble of analytical techniques and tools, the project extracts valuable insights from the data generated during the image recognition and caption generation processes. These insights help improve the accuracy and efficiency of the model, ensuring that the converted visuals are accurately described in real-time audio. In the subsequent sections, we

discuss in detail the architecture of the analytical framework we used, the steps we implemented and the results validation we obtained. Primarily, using analytics, we are optimizing the text-to-audio transformer, ensuring that the generated audio is clear, concise, and provides an accurate description of the visual content.
Furthermore, analytics can also enable continuous monitoring and evaluation of the product's performance and user feedback. By gauging user interactions, preferences, and feedback, areas of improvement can be identified and data-driven decisions can be made to enhance the user experience and build strategies around usage patterns of the product to develop incremental enhancements on this prototype

## Analytics Methodology

- **Describing the dataset:** We have used the COCO (Common Objects in Context) image captioning dataset which is a widely used dataset in the field of computer vision and natural language processing. It consists of a large collection of images, each accompanied by multiple human-generated captions describing the content of the image. The dataset contains diverse scenes with a wide range of objects and activities, making it suitable for training and evaluating image captioning models.
  Below is a snapshot of the COCO image (image 1 below) captioning dataset which is an existing observational data to train and evaluate the model. The target or outcome variable is the generated caption for a given image, which aims to provide a descriptive sense of the visual content. The explanatory variables or features are the images themselves, which serve as the input for the model (image 2 below).
  To summarize, the entire dataset contains 25.7 GB data which is a combination of train, test and validation datasets. The training set contains 118,287 images, the test set contains 40,670 images while the validation set contains 5,000 images each

Out[3]:

| | image | caption |
|---|---|---|
| 0 | C:/Users/Sripriya Srinivasan/Downloads/coco201... | A clock on a boardwalk near a beach. |
| 1 | C:/Users/Sripriya Srinivasan/Downloads/coco201... | Large air force plane parked out on the runway |
| 2 | C:/Users/Sripriya Srinivasan/Downloads/coco201... | The man sits before a large pizza and many gla... |
| 3 | C:/Users/Sripriya Srinivasan/Downloads/coco201... | a bathroom with a cat walking across the sink |
| 4 | C:/Users/Sripriya Srinivasan/Downloads/coco201... | A man and woman sitting across the table from ... |

Image 1

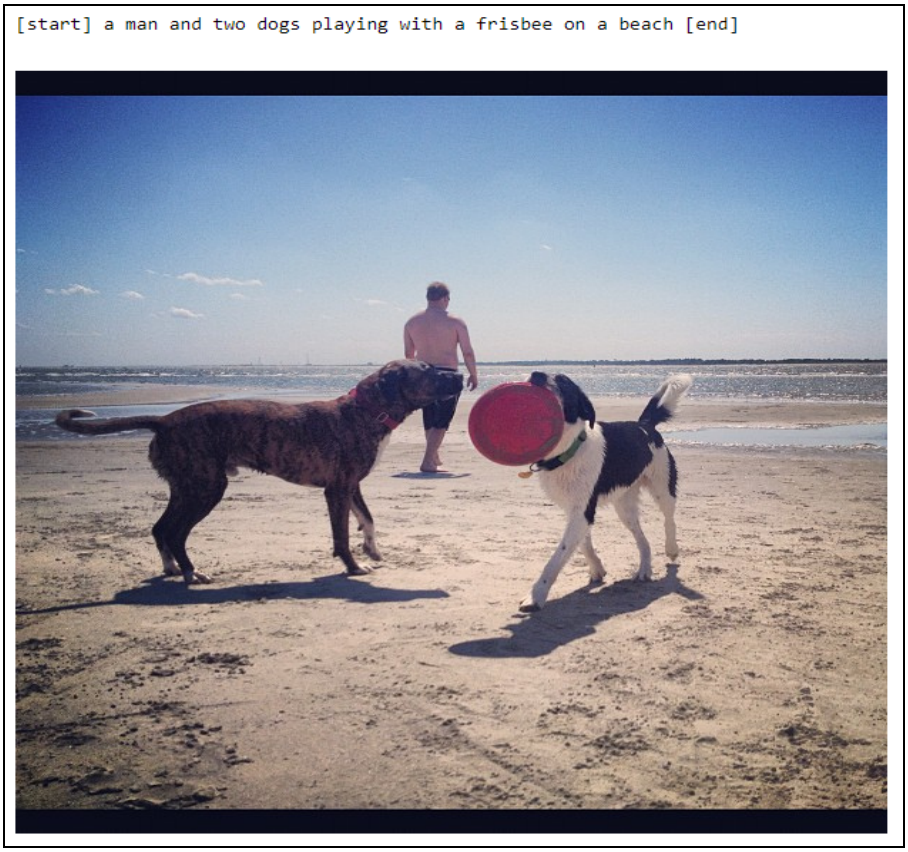



Image 2

- **Data Processing:**
  The input data as seen above is in the form of image url and caption combinations. Before inputting it to the model there are a series of preprocessing steps that are performed. We split the dataset into training and validation sets. It randomly shuffles the image keys and divides them based on a specified split ratio (80% for training and 20% for validation.

  **Caption Processing:**
  - All special characters are removed and the caption is made lower case.
  - Special tokens like `[start]` and `[end]` are added at the start and end of the caption respectively.
  - Finally in order for the model to be able to interpret the caption it is vectorized with a dictionary of vocabulary size of 15,000 and maximum caption length of 40.

  **Image Processing:**
  - The image file is read as a binary string from the image path.
  - It is then decoded as a JPEG image with 3 channels i.e RGB
  - Then the image is resized to 299 x 299
  - Images are scaled and normalized to be input into the model.

- Finally we are also performing image augmentation to enhance the train dataset. This essentially adds rotations and changes the contrast of the images to make the model more robust.

- **Approach:**

  On a high level the process involved from input to output is as follows:
  First, the user has a device that continuously takes pictures of their surroundings and sends it to the model as input.
  Second, the model uses image recognition and detects all the objects in the image and generates a caption.
  Third, further this caption is input to a text to audio transformer that generates the audio which is sent to the earpiece of the user. Overall, the two major components of our project are:

  1. Image Captioning
  2. Text to speech conversion

  **[1] Image Captioning:**

  Once the data is processed and ready to be loaded into the model as per the above mentioned steps, we perform image augmentation, i.e a technique used to artificially increase the size of the training dataset by applying various random transformations to the images. This helps to improve the model's ability to generalize and handle variations in the input data.

- For image captioning, we have used the InceptionV3 model. The architecture of *InceptionV3* consists of multiple layers, including convolutional layers, pooling layers, fully connected layers, and auxiliary classifiers. Here is a brief overview of its architecture:
  - Input Layer: The model takes input images of size 299x299x3.
  - Convolutional Layers: The initial layers of InceptionV3 are convolutional layers that perform feature extraction from the input images. These layers use various filter sizes (1x1, 3x3, 5x5) and apply different types of convolutions (standard, dilated) to capture different levels of detail in the image.
  - Pooling Layers: After some convolutional layers, max pooling is applied to reduce the spatial dimensions of the feature maps while retaining important features.
  - The central element of InceptionV3 is the Inception module, which is intended to effectively capture multi-scale information. Parallel convolutional branches with various filter sizes (1x1, 3x3, and 5x5) and pooling processes make up each Inception module. The output of the Inception module is created by joining the outputs of these branches.

    - Auxiliary Classifiers: To help the model learn more discriminative features, InceptionV3 incorporates auxiliary classifiers at intermediate levels of the network. These support classifiers aid in enhancing gradient flow and regularizing the model during training.
    - Fully Connected Layers: Towards the end of the network, the feature maps are flattened and passed through fully connected layers to perform the final classification.
    - Softmax Activation: The final layer of InceptionV3 uses the softmax activation function to produce probability scores for different classes. The softmax function normalizes the output scores into a probability distribution, where each score represents the probability of the input image belonging to a specific class.
- The TransformerEncoderLayer function represents one layer of the Transformer encoder in the Transformer model architecture. Here is an overview of its architecture:
    - Input: The input to the TransformerEncoderLayer is a sequence of embeddings or feature vectors.
    - Layer Normalization: The input sequence is first passed through a layer normalization operation, which normalizes the values across the feature dimension.
    - Dense Layer: The layer-normalized input is then passed through a dense layer, which applies a linear transformation to the input features and introduces non-linearity using the ReLU activation function.
    - Multi-Head Attention: The output of the dense layer is used as the input to the multi-head attention mechanism. This mechanism consists of multiple parallel attention heads that perform self-attention on the input sequence. Each attention head attends to different parts of the input sequence and generates an attention-weighted representation.
    - Feed-Forward Network: The output of the layer normalization is passed through a feed-forward network (FFN) consisting of two dense layers with a ReLU activation function in between. The FFN introduces non-linearity and applies another linear transformation to the input sequence.
    - Output: The final output of the TransformerEncoderLayer is the output sequence after the feed-forward network and the residual connection. This output represents the encoded representation of the input sequence after one layer of the Transformer encoder.

- The TransformerDecoderLayer function represents one layer of the Transformer decoder in the Transformer model architecture. Here is an overview of its architecture:
    - Input: The input to the TransformerDecoderLayer is a sequence of input embeddings or feature vectors.

- Encoder-Decoder Attention: The output of the layer normalization is passed through another attention mechanism called encoder-decoder attention. This attention mechanism attends to the encoder output, which contains the encoded representation of the image features, and generates an attention-weighted representation of the encoder output based on the input sequence.
- Layer Normalization: Similar to the previous layer, the output of the encoder-decoder attention mechanism is added to the output of the previous layer using a residual connection, and the resulting sum is then normalized using layer normalization.
- Feed-Forward Network: The output of the layer normalization is passed through a feed-forward network (FFN) consisting of two dense layers with a ReLU activation function in between. The FFN introduces non-linearity and applies another linear transformation to the input sequence.
- Output: The final output of the TransformerDecoderLayer is the output sequence after the feed-forward network and the residual connection. This output represents the decoded representation of the input sequence after one layer of the Transformer decoder.

The TransformerDecoderLayer function represents one layer of the Transformer decoder. The number of layers in the Transformer decoder can be determined by stacking multiple instances of the TransformerDecoderLayer. The activation functions used in this architecture are ReLU for the dense layers in the feed-forward network. The self-attention and encoder-decoder attention mechanisms do not use explicit activation functions, but they incorporate the softmax function to compute attention weights.

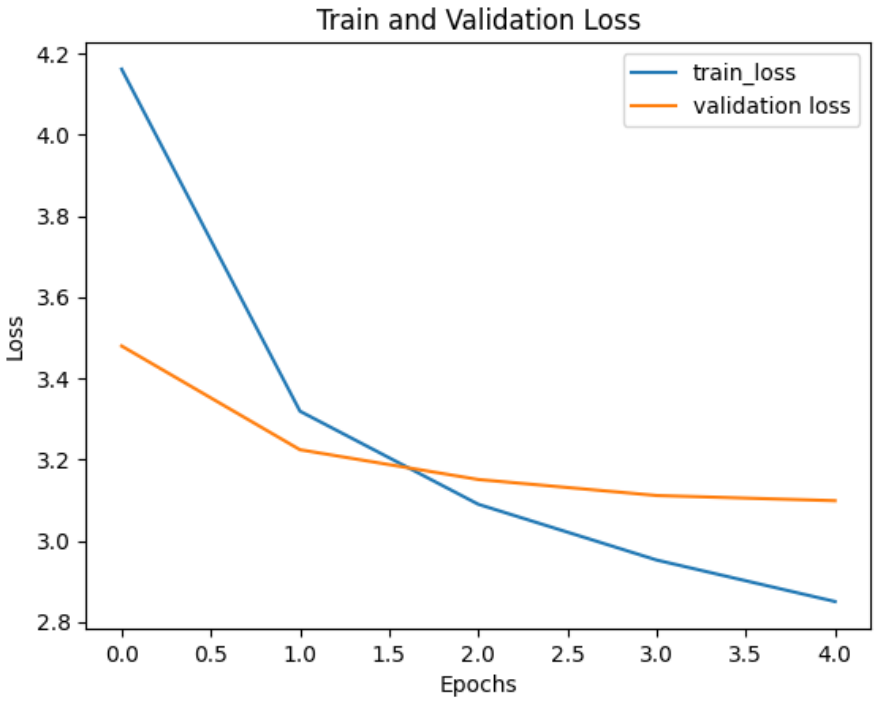
**[2] Text to Speech Conversion:**

GTTS (Google Text-to-Speech) is a Python library tool that enables text-to-speech conversion using the Google Text-to-Speech API. It is used to generate audio files (mp3) from the predicted captions using a simple and straightforward interface. The library utilizes the power of Google's infrastructure to perform the text-to-speech synthesis and generate the corresponding audio output. The two key parameters in this module are:

- Slow (default: False): Specifies whether the generated speech should be slower than normal speed. Set it to True for slower speech.
- Lang (default: 'en'): Specifies the language of the text. You can provide language codes like 'en' for English, 'fr' for French, 'es' for Spanish, etc.

## Results

After training the model on the COCO dataset we observe the following train and validation loss plot. We can see that as the number of epochs increases on the x-axis, the train and validation loss both have a decreasing trend as expected.




Further evaluating the model on the test data we get the following metrics:
**Test Accuracy = 0.427**
**Test Loss = 3.091**

To get a more intuitive evaluation on the model below is an image and corresponding caption that the image captioning model generated.

```
In [57]: img_path = "coder.jpg"

         im1 = Image.open(img_path)

         pred_caption1 = generate_caption('coder.jpg', add_noise=False)
         print('Predicted Caption:', pred_caption1)
         print()
         im1

         Predicted Caption: a man is sitting on a laptop computer
```

Out[57]:

As a second stage of our product, this predicted caption is sent to a text-to-speech converter api that gives a .mp3 with the audio version of the caption generated. (click link to listen)

https://drive.google.com/file/d/1CA8VwFDCbPei7FbrU2et3aacY-pbKlKP/view?usp=sharing

## Model Assessment

As we can observe, the accuracy and performance of the above model is not very impressive. Upon closer inspection and analysis the following reasons have been identified

1. Size of the data used for training:
    - Due to available compute power the model was trained only on 70,000 images with corresponding captions. This was split into 80:20 and used for training and model validation.
    - Also due to limited training data the model was unable to identify all components of an image, for instance in the above picture it can identify a man, laptop and computer - but it fails in framing the caption correctly.

2. Choice of model:
    - In the model above, the image feature extraction is done with a pre-trained CNN model, InceptionV3, these features are input through an encoder followed by a decoder.
3. Absence of fine-tuning, hyper-parameter tuning and regularization techniques.

## Model Improvement - Next steps

1. **Use a more advanced architecture:** The choice of architecture plays a crucial role in the performance of an image captioning model. More advanced architectures, such as attention-based models or Transformer-based models, allow the model to focus on different image regions and capture complex relationships between them. Here, in our model we utilized InceptionV3, but there are other powerful models that we can use. One such model is *Show, Attend, and Tell*, SAT is an attention-based model that combines a convolutional neural network (CNN) for image feature extraction with a recurrent neural network (RNN) for caption generation. It incorporates an attention mechanism that attends to different image regions while generating captions, allowing the model to focus on relevant visual information. Others include *Transformer based models*, they leverage self-attention mechanisms to capture the relationships between image regions and generate captions.

2. **Larger dataset:** Increasing the size of the training dataset can help the model learn more diverse visual features and better generalize to unseen images. The COCO dataset contains 1 caption per image, but there are other datasets like Flickr30k that contain 5 captions for the same image, such datasets could additionally improve the model performance. It will expose the model to different image styles, objects, and contexts, which can improve the accuracy of the captions generated by the model.

3. **Data augmentation:** Data augmentation involves applying various transformations to the training images to create new variations while preserving the semantic meaning. Techniques such as random cropping, rotation, scaling, and flipping help introduce diversity into the training data. By augmenting the dataset, the model is provided with more examples to learn from, making it more robust to different image conditions and improving its accuracy. Though data augmentation has been implemented to enhance the train dataset it could be done on a larger scale to improve the accuracy even further.

4. **Pretraining:** A model can learn general visual qualities that are transferable to the image captioning challenge by pretraining it on a large-scale picture classification task like ImageNet. The model improves its ability to recognize and comprehend visual cues by being trained to identify items in pictures, which can help with the captioning assignment.

5. **Fine-tuning:** After pretraining on a large-scale task, fine-tuning the model on the COCO dataset specifically for the captioning task is essential. During fine-tuning, the model adapts to the specific characteristics and nuances of the COCO dataset. It learns to generate captions that align well with the annotations in the dataset, leading to improved accuracy on this specific task.

6. **Hyperparameter tuning:** Hyperparameters are parameters that regulate how the model learns. The performance of the model can be dramatically affected by experimenting with various settings for hyperparameters like learning rate, batch size, and network depth. One can explore the hyperparameter space methodically with the use of methods like grid search, random search, or Bayesian optimization to identify the ideal setting that will maximize the accuracy of the model.

7. **Ensemble models:** Training multiple models with different architectures or hyperparameter settings and combining their predictions through ensembling techniques can lead to improved accuracy. Each individual model captures different aspects of the data and brings its own strengths. By combining their outputs, one can leverage the diversity and complementarity of the models, resulting in more accurate captions.

8. **Regularization techniques:**Regularization techniques are intended to avoid overfitting, a situation in which the model gets overly focused on the training set and performs badly on unobserved data. By adding randomness and lessening the model's reliance on particular characteristics or patterns, techniques like dropout and weight decay aid in regularizing the model. Improved accuracy on fresh images results from regularization, which pushes the model to acquire more generalizable representations.

# Appendix

**Dataset:**
https://arxiv.org/abs/1405.0312
https://cocodataset.org/#home

**Image Captioning:**
https://paperswithcode.com/dataset/coco-captions
https://www.ijitee.org/wp-content/uploads/papers/v10i3/C83830110321.pdf
https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2

**Speech-to-Text Conversion:**
https://gtts.readthedocs.io/en/latest/