5-2-2023

Improving Multi-Modal Food Detection System with Transfer Learning

Shivani Gowda
Loyola Marymount University

5-2-2023

Improving Multi-Modal Food Detection System with Transfer Learning

Shivani Gowda
Loyola Marymount University, shivanigowdaks@gmail.com

Improving Multi-Modal Food Detection System with Transfer Learning

by

Shivani Gowda

A thesis presented to the

Faculty of the Department of
Computer Science
Loyola Marymount University

In partial fulfillment of the
Requirements for the Degree
Master of Science in Computer Science

May 2, 2023

Improving Multi-Modal Food Detection System with Transfer Learning

Dr. Mandy Korpusik

[redacted]

signature

5/3/2023

Date

Dr. Lei Huang

[redacted]

Signature

5/10/2023

Date

Dr. Junyuan Lin

[redacted]

Signature

5/3/2023

Date

Abstract

Self-assessment of food intake is important for preventing and treating obesity. The current self-assessment methods of food intake are inaccurate and hard to use. In this thesis, we explore ways to improve machine learning (ML) food classification methods which are the core technical problem of food intake self-assessment. We present a food detection system that utilizes a state-of-the art multi-modal architecture called Vision and Language Transformer (ViLT). This architecture combines both food appearance via the image modality, and description via the textual modality to improve the accuracy of food classification. To further enhance the performance, we incorporate other improvements such as curating a branded food item dataset. We apply transfer learning, an ML method that allows reusing a pre-trained model from a related high-resource task as the starting point for a low-resource task such as ours. This approach reduces the cost and time required compared to building a model from scratch. In addition, we compare our approaches with that of Visual Chat GPT, a combination of vision foundation model and a large language model, and find that our approach for food intake assessment is both accurate and cost-effective.

Acknowledgments

I express my gratitude to my thesis advisor, Dr. Mandy Korpusik of the Frank R. Seaver College of Science and Engineering at Loyola Marymount University, for providing me with the opportunity to explore a multimodal food detection system as a continuation of her diet tracking system. Dr. Korpusik consistently guided me in the right direction. I am also grateful to my thesis committee members, Dr. Huang and Dr. Lin, for their insightful ideas and support throughout my thesis work.

I'd like to extend my thanks to the graduate advisor, Dr. Johnson, for his encouragement and support in excelling in my thesis project.

I am grateful to my family for their unwavering support and motivation. My husband, Dr. Thamme Gowda, deserves special recognition for constantly challenging me to improve and strive for excellence. I also want to express my appreciation to Yifan Yu for helping me understand the datasets and base model code.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nutrition is a crucial aspect of our daily lives. It is essential to know the nutrient values of the food we consume to maintain a healthy diet. However, it can be challenging to accurately determine the nutrient content of food. This is where the use of technology can be beneficial. In this thesis, we present a multi-modal model which is based on transformer architecture [Vaswani et al., 2017] that uses both text and image data for food classification.

The use of a multi-modal approach has several advantages. Text data can provide detailed information about the ingredients and preparation methods used in a dish, while image data can provide visual cues about the food's appearance and composition. By combining these two sources of information, our model can make more informed predictions about the food. In the following chapters, we will describe the design and implementation of our multi-modal model in detail. We will also present the results of our experiments and discuss their implications.

We believe that our work has the potential to make a significant contribution to the field of nutrition by providing a new and innovative method for food classification. We hope that our research will inspire further developments in this area and help people make more informed decisions about their diet.

1.1 Motivation

This thesis proposes a multi-modal food detection system with transfer learning [1] that would possibly merge with an existing nutrition application, *Coco Nutritionist* which is available on the Apple App Store [2]. Coco is a personal AI-based digital assistant who can guide you on the path to healthy eating by making diet tracking easy with a conversational calorie counter. Coco nutritionist was developed by members of the Spoken Language Systems Group at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), with the idea of a spoken-language application that would make meal logging even easier. The mobile application provides smart features [3] such as:

- Logging multiple foods at once, by speaking or typing
- Easy barcode scanning
- Timeline of meals
- Comprehensive, personalized list of nutrients
- Daily calorie goal and the streak of consecutive days logged
- Timestamps and setting the time with natural language

The aim of this thesis is to create a food detection system that uses multiple modes of data to track food details easily.

1.2 Scope

This thesis presents a Transformer-based multi-modal diet tracking system for food. The scope of the thesis is limited to demonstrating the feasibility of a transfer learning

[1] *Multi-modal Food Classification in a Diet Tracking System with Spoken and Visual Inputs Shivani Gowda, Yifan (Rosetta) Hu, Mandy Korpusik 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2023), June 2023.*

[2] https://apps.apple.com/us/app/coco-nutritionist/id1173015957

[3] https://korpusik.wixsite.com/coco-nutritionist

approach for developing a food detection system. We begin by describing the methods used for data collection and annotation. In Chapter 3, we discuss the process of data collection. Chapter 4 covers the design and experimentation of our system. In Chapter 5, we present an ablation study to evaluate the impact of different components of our system. Finally, in Chapter 6, we conclude our work and discuss future directions.

Chapter 2

Literature Review

In this chapter, we review some of the background information and relevant approaches for. In Section 2.1, we provide an overview of machine learning (ML) methods applicable to our work, and in Section 2.2, we focus on reviewing diet tracking models.

2.1 Background: Machine Learning Classification

Machine learning (ML), a sub-area of artificial intelligence and computer science, is concerned with the study of methods for the creation of autonomous agents. ML methods learn to perform a chosen task from a given set of examples, with the aim of performing the chosen task reasonably well on unseen data, without explicitly programming the instructions. The score of ML is broad: it varies from automatically classifying emails as spam-or-ham, to automatically driving a car. This thesis focuses on diet-tracking task, which involves both text and image modalities. ML methods has evolved and improved over the time; currently deeplearning based approaches are popular which use artificial neural networks to accomplish the task. While there are several sub areas, we focus on *supervised classification*, which is relevant to the diet-tracking when modeled as food item classification problem. In this classification modeling, given an input x, say a textual description or a photograph of food consumed, the task is to map x to $y \in Y$, a set of labels such as smoothies, salad, muffin, burger, etc.

At a high-level, this has a two steps: firstly, given $x \in \mathbb{R}^m$ where m is the input dimension, we learn a representation, $z \in \mathbb{R}^d$ where d is the hidden dimension, and lastly, we map representation to $P(Y = y|z)$, to discrete probability distribution. Both steps are accomplished and learned end-to-end using artificial neural network, modeled as $P(Y = y|x) = f(Y = y|x; \theta)$, where θ is parameters of model estimated from a set of examples. During the inference time, a class y that has highest probability (single-label) or a set of classes which exceeds a given threshold multi-label are chosen as label.

A key part of this model is learning a meaningful representation $z \in \mathbb{R}^d$ for the given input $x \in R^m$, such that the representation help accomplish the end task (in our case, identifying of a food item in the given input). There are many approaches to learn representations using deeplearning, and historically, the approach varied based on modality of input. For text modality, where input is a sequence of words i.e, a time series, recurrent neural networks such as Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks were popular. For image modality where the input is two or more dimensional i.e. a grid of pixels, convolutional neural networks (CNN) were used. However, in the recent years, Transformers [Vaswani et al., 2017] are the state of the art for representation learning across several modalities.

Transformers: Vaswani et al. [2017] originally proposed Transformers for neural machine translation i.e. a general sequence to sequence mapping problem. However, transformers are later adapted to other task types including sequence classification and image classification. Transformer is a layered architecture, consisting of a stack of identical layers (typically, 6, 12 or 24 layers), as shown in Figure 2-1.[1] Each Transformer Encoder layer has two sublayers, Multi-Head Attention, and Feed Forward sublayers, which are connected through residual connections and Layer Norm (indicated by *Add and Norm* in Figure 2-1).

For simplicity, consider meal dairy as input, i.e. a sequence of words that describe

[1] Transformer model is originally an encoder-decoder network. We focus on encoder only as we do not have usecase for decoder part of model.

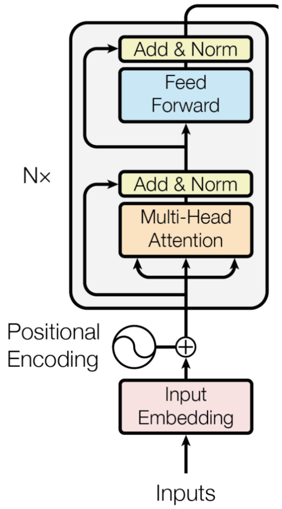


Figure 2-1: Transformer Encoder architecture. Image credit: Vaswani et al. [2017].

food in take. More concretely, $x = x_1, x_2, ...x_m$, where m is the length of sequence. The goal is to assign one or more labels $y \in Y$, the set of food items. Although we need to learn a representation $z \in \mathbb{R}^d$ that captures the meaning of whole sequence x, Transformers, by design, learns contextual representations $z_i \in \mathbb{R}^d$ for each token $\{x_i | i = 1, 2...m\}$. Since each h_i is learned in the context of sequence, a special token inserted at the beginning of input as $x = [CLS], x_1, x_2...x_m$ and its corresponding representation z_0 captures the representation for the whole sequence.

Each z_i is learned using self-attention mechanism, which is described as follows:

$$z_i = \sum_{j=1}^{m} \alpha_{ij}(x_j W^V)$$

Each weight co-efficient, α_{ij} is computed using softmax:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{m} \exp(e_{ik})}$$

Finally, e_{ij} is computed as following:

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d}} \tag{2.1}$$

$W^Q, W^K, W^V \in \mathbb{R}^{d \times d}$ are parameters of attention sub layer. In practice, attention is computed in parallel using batch matrix operations. Transformer employs multiple of such attention mechanism called multi-head attention, which are concatenated.

The feedforward layer is implemented as two layer feedforward network as:

$$FFN(x_i) = max(0, xW_1 + b_1)W_2 + b_2 \tag{2.2}$$

Where W_1, W_2 are parameter matrices and b_1, b_2 are parameter bias vectors.

2.2 Models for Diet-Tracking System

2.2.1 Image Models

The rapid development of neural networks [Bengio, 2009] prompted advancement in the domain of food image classification Coco Nutritionist [Korpusik et al., 2016, 2017a, Korpusik and Glass, 2018, Korpusik et al., 2017b, 2014, 2019]. After the release of the benchmark ImageNet dataset for image classification [Krizhevsky et al., 2012, Zeiler and Fergus, 2014, Simonyan and Zisserman, 2014, Szegedy et al., 2015, He et al., 2016], food image classification models are generally pre-trained on generic ImageNet and fine-tuned on food image datasets (e.g., UEC-Food100, UEC-Food256, or Food-101). One of the earliest works in classifying food images with deep learning was in 2014 with a deep convolutional neural network (CNN)[Kawano and Yanai, 2014]. Later, the authors enhanced their model through transfer learning [Yanai and Kawano, 2015]. Several other CNN-based methods have been explored by other studies concerning food image recognition [Ciocca et al., 2018, Kaur et al., 2019, Farinella et al., 2016, Hnoohom and Yuenyong, 2018], but none multi-modal.

2.2.2 Text Models

For natural language processing (NLP), Transformer-based contextual embedding models such as bidirectional encoder representations from Transformers (BERT)[Devlin et al., 2018] are state-of-the-art. Comprehensive studies have been conducted to analyze the rise of deep learning in text classification[Young et al., 2018, Minaee et al., 2021].

2.2.3 Vision-and-Language Models

Recent Vision-and-Language (V+L) research has been oriented towards pre-training on extensive image-text datasets, observing significant improvements in learning joint modality relationships. Experiments indicate that they achieve notable results in tasks such as visual question answering and text-image generation[Mei et al., 2020]. Starting with ViLBERT [Lu et al., 2019], there is a surge in using Transformers as the main architecture, shifting away from recurrent neural networks (RNNs). Both ViLBERT and LXMERT [Tan and Bansal, 2019] fuse two separate Transformers, one for images and one for text. Later, researchers [Li et al., 2019, Su et al., 2019, Chen et al., 2020] introduced a single-stream Transformer model to better understand joint representations.

2.2.4 Background: Evaluations

Our dataset has many imbalanced classes and is skewed. Most classes fall on the long tail, with only a few food categories having more than 100 examples and the majority having 10 or fewer. To account for this class imbalance as shown in Figure 2-2, reporting accuracy would capture head and ignore long tail hence we report both Macro-f1 and Micro-f1 scores [Gowda et al., 2021]. Micro-f1 assigns equal weight to each instance while Macro-f1 assigns equal weight to each class. Macro-f1 is useful in imbalanced data scenarios like ours.

Consider a test corpus, $\mathrm{T} = \{(\mathrm{x}^{(i)}, y^{(i)}) \mid i = 1, 2, 3...m\}$
where $x^{(i)}$ and $y^{(i)}$ are source and system output respectively.[2]

Let L_x, L_y and $L_{x \cap y}$ where L_x is the *label* of x, L_y is the *label* of $y, L_x \cap L_y$ comman labels in x and $y,$ for each class $c \in L,$

[2] *Equation credit [Gowda et al., 2021]*

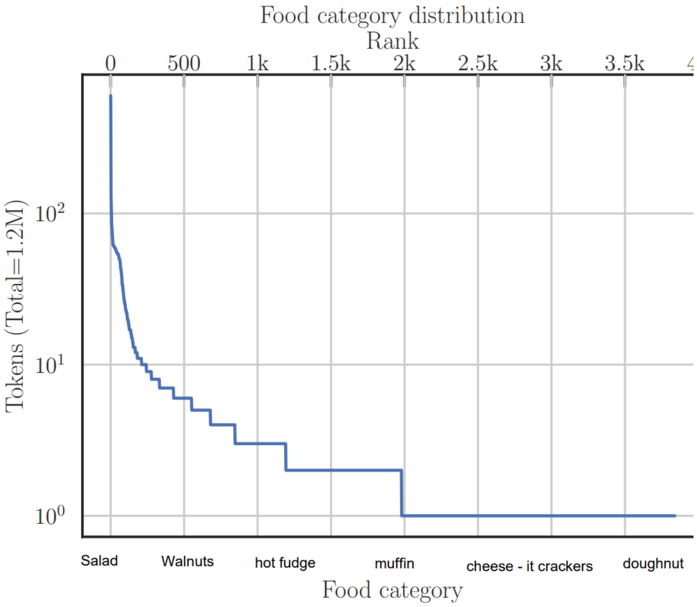


Figure 2-2: Food category distribution. The vertical axis is in log scale. Categories in our dataset are imbalanced as in the real-world scenario.

$$\begin{aligned}\text{PREDS}(c) &= \sum_{i=1}^{m} C(c, x^{(i)})\\ \text{REFS}(c) &= \sum_{i=1}^{m} C(c, y^{(i)})\\ \text{MATCH}(c) &= \sum_{i=1}^{m} \min\{C(c, x^{(i)}), C(c, y^{(i)})\}\end{aligned}$$

where C(c,a) counts the number of tokens of type c in a For each class c ∈ $L_{x\cap y}$, precision (P_c), recall (R_c), and F_β measure ($F_{\beta;c}$) are computed as follows:

$$P_c = \frac{\text{MATCH}(c)}{\text{PREDS}(c)}; \qquad R_c = \frac{\text{MATCH}(c)}{\text{REFS}(c)} \qquad F_{\beta;c} = \frac{(1+\beta^2)P_c \times R_c}{\beta^2 \times P_c + R_c}$$

The macro-average consolidates individual performance by averaging by type, while the micro-average averages by token:

$$\text{MACRO-}F_\beta = \frac{\sum_{c \in L} F_{\beta;c}}{|L|}$$

$$\text{MICRO-}F_\beta = \frac{\sum_{c \in L} f(c) \times F_{\beta;c}}{\sum_{c' \in L} f(c')}$$

where f(c) = REFS(c)+k for smoothing factor k.

2.3 Summary

Even though the aforementioned deep learning architectures are crucial for achieving breakthrough achievements, there is still a lack of real-world deployment of these models. This issue is largely due to the fact that downstream tasks require specific domain expertise. Thus, in our thesis, we collected a food-specific multi-modal dataset and developed a novel architecture to accomplish multi-modality in the real world.

Chapter 3

Food Database

Gathering high-quality training data is a crucial step in the supervised machine learning process, as it directly impacts the performance. In the past, we were able to utilize or build upon existing open-source data sets. However, with the advent of multi-modal learning, the data collection process became more complex, as it requires two modalities of data and thus doubles the time and effort needed. To address this challenge, we have created a new multi-modal dataset by combining text-to-image and image-to-text data sets, as detailed in the next section. The dataset described in 3.1 was obtained from the research of Dr. Korpusik. Originally, it consisted of multi-label text and single-label images.

3.1 Data Collection I: Multi-Label Text and Single Label Images

This part of the dataset includes pairs of food images and user-provided meal descriptions across 696 food categories.

3.1.1 Text-to-Image Data

In the prior work [Korpusik and Glass, 2017, 2019], meal diaries were crowd-sourced from Amazon Mechanical Turk [Korpusik et al., 2014]. Workers were prompted to write down in natural language a description of a particular food item with a property token and a food token. The property token ranges from brands (e.g., McDonald's, Cheerios) to quantity (e.g., half a dozen, a cup). The food token represents a particular food item (e.g., apples, pancakes).

Images were collected for 66 of the 89 most common food classes[1] from Flickr and mapped to the natural language sentences. For the remaining 23 food classes, 100 images per category were used from the Food-101 dataset [Bossard et al., 2014].

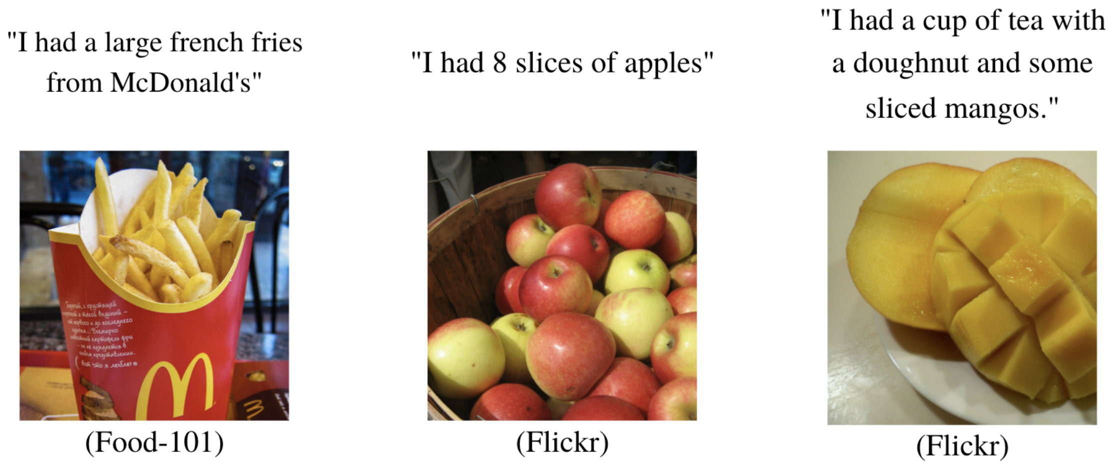

"I had a large french fries from McDonald's"
"I had 8 slices of apples"
"I had a cup of tea with a doughnut and some sliced mangos."
(Food-101)
(Flickr)
(Flickr)

Figure 3-1: Examples of Text-to-Image data

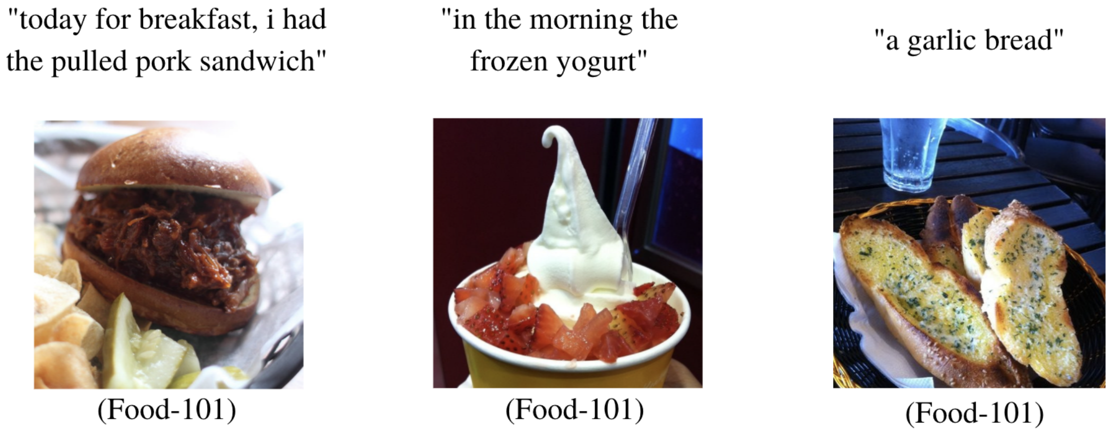

"today for breakfast, i had the pulled pork sandwich"
"in the morning the frozen yogurt"
"a garlic bread"
(Food-101)
(Food-101)
(Food-101)

Figure 3-2: Examples of Image-to-Text data. Image credit: Food-101 and Flickr (Rosetta's research put together the initial multi-modal dataset)

[1] We use the term 'category' interchangeably with terms 'label' and 'class.'

3.1.2 Image-to-Text Data

For the second part of the dataset, 100 images were selected from each of the remaining 78 food categories in the Food-101 dataset. To complete the multi-modal dataset, natural sentences were generated using state-of-the-art automated image captioning with visual attention. However, this approach had two issues: the generated text did not follow the format of nutritional diaries and was not specific enough to be useful. To solve this problem, a custom text template was created based on common phrases found in the natural language dataset.

3.1.3 Merge

In the end, we collected food-specific image-text pairs with images from Flickr or Food-101 and text from our previously collected natural language dataset or a targeted text template. The text may consist of multiple food categories, but the images consisted of only one food. As a result, we used 167 classes for the image modality and 696 total classes for the natural language modality.

3.2 Data collection II: Multi-Label Images

To handle images with multiple labels, we employed the DALL-E 2 API [2]. For sentences that mentioned only a single food label, we retained images from Flickr and Food-101. However, for sentences that mentioned multiple food labels, we generated images using DALL-E 2, as illustrated in Figure 3-3.

3.3 Data Collection III: Branded Dataset

The reason for creating a branded dataset was to solve the problem of our model not being able to accurately identify branded foods due to the lack of branded data in our original dataset. By adding branded items to our dataset, we aimed to enhance the model's ability to recognize and classify branded foods.

[2] https://openai.com/product/dall-e-2

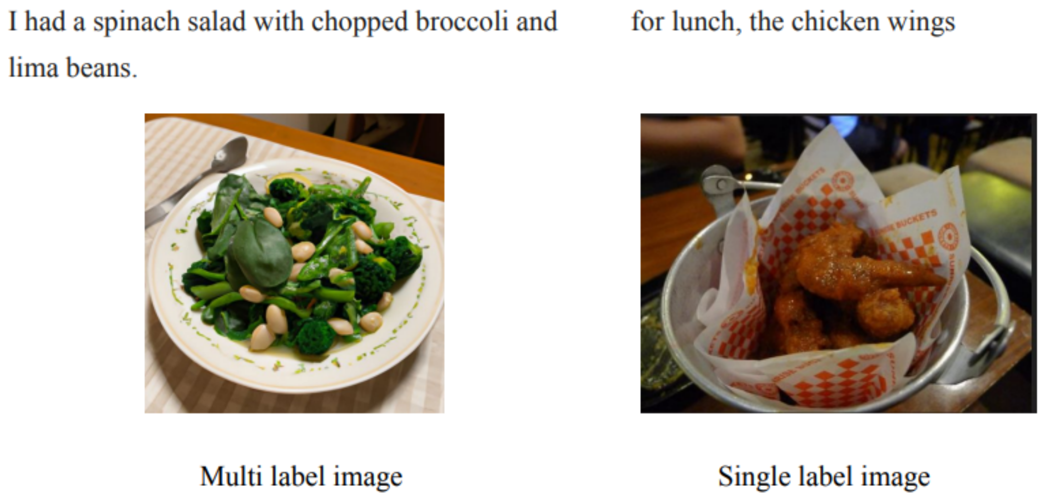

I had a spinach salad with chopped broccoli and lima beans.
for lunch, the chicken wings
Multi label image
Single label image

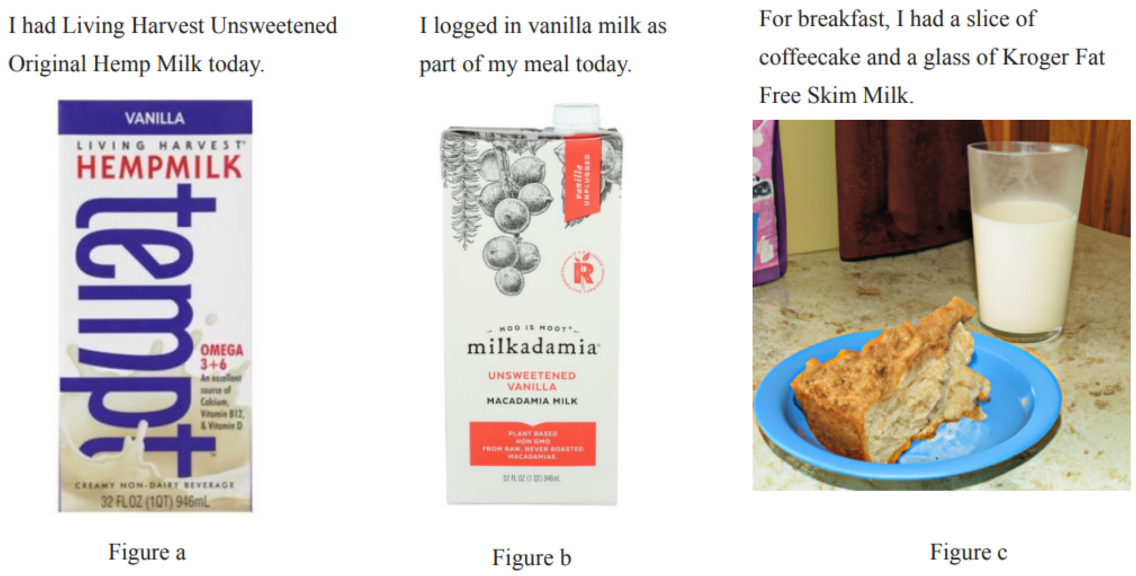
Figure 3-3: Multi-label and single-label images

We gathered images of branded pizza, smoothies, and dairy products from a variety of websites. These images were obtained through web scraping. Different scenarios were considered of how the user would describe the branded data, such as

- The user can share an image and text that features a brand name as shown in Figure (3-4: a).
- The user can share an image featuring a brand name while the text does not contain the branded dataset as shown in Figure (3-4: b).
- In the text, the user can refer to consuming a smoothie from a particular brand, even if the accompanying image does not display the brand name as shown in Figure (3-4: c).

3.3.1 Learning From Both Text and Image

Text generation templates were created as shown in Figure 3-5. Each template was given a single label corresponding to the branded data in the respective image. For instance, in Figure 3-4: a both the image and text contain branded data.



I had Living Harvest Unsweetened Original Hemp Milk today.

I logged in vanilla milk as part of my meal today.

For breakfast, I had a slice of coffeecake and a glass of Kroger Fat Free Skim Milk.

Figure a

Figure b

Figure c

Figure 3-4: Branded data with text and images

Template 1: I drank kroger fat free skim milk today for breakfast.

Template 2: I ate RED BARON Classic Crust Four Cheese Pizza today for lunch.

Template 3: I drank Bolthouse Farms after workout

Figure 3-5: Templates for branded dataset

3.3.2 Learning From Image Only

For text generation templates as shown in Figure 3-6 were created, were the templates were given generalized information of the data, whereas the image contained the branded data information. For example, see Figure 3-4: a where the image alone contains branded data.

Template 1: I drank milk today for breakfast.

Template : I ate pizza today for lunch.

Template 3: I drank smoothie after workout

Figure 3-6: Templates for branded dataset

3.3.3 Learning From Text Only

The DALL·E 2 API does not preform well on the branded dataset, it provides a very generalized image with not much information. To learn from text-only data, the images were generated from the DALL·E 2 API Figure 3-4: c. To make the data much more interesting, for creating text data, we combined branded dataset with other food categories, example: for one branded dataset we added upto 2 (refer to Figure 3-7) other food categories from our dataset. To obtain the best possible combination and more human like we generated the text with the help of GPT-3.5-Turbo refer to Figure 3-8 from OpenAI [3].

Sentence type 1: Today for breakfast, I had two chicken drumsticks and a glass of Private Selection Strawberries and Cream Milk to start my day.

Sentence type 2: Today for breakfast, I had an apple cinnamon toasted English muffin, for lunch I drank cranberry juice, and for dinner, I had a Quest thin crust uncured pepperoni frozen pizza.

Figure 3-7: The labels marked red are branded labels and blue are not branded ones

Prompt: Given all these categories category names give a sentence how a human would log their food using all of these categories

Figure 3-8: Prompt for GPT-3.5 Turbo

Table 3.1: A summary of branded data

Modality	**sentences**	**images**
Branded info. in text and image	418	530
Branded info. only in image	509	
Branded info. only in text	801	801

Table 3.1 provides a summary of the branded dataset. In the following chapters, we will explore how this dataset was used in our experiments and examine the performance of our models on it.

[3] https://platform.openai.com/docs/guides/chat

Chapter 4

Experiments

Our method employs transfer learning to take advantage of existing resources and apply knowledge from one domain to another. Instead of training a deep neural network from scratch, we use pre-trained models and only modified the final layer to fit our specific dataset.

4.1 Model-I: Baseline Vision-and-Language Model without Pre-Trained Weights

Our baseline[1] model is based on CNN and LSTM (i.e., Long Short-Term Memory recurrent neural network) architectures. In this model, concatenatenation of a convolutional neural network (CNN) [Krizhevsky et al., 2012] with an LSTM [Hochreiter and Schmidhuber, 1997] for joint visual and textual classification. Due to the difference in the mixed features of the data, each modality (i.e., image and text) is trained separately. The image input is handled by a CNN, and the text is processed by an LSTM, which are described in the following subsections.

The image features are extracted by the CNN model to output a classification result. The second input is processed by the LSTM model. The classification outputs of the two branches are concatenated together by a final layer to obtain the global

[1] This model/approch was developed by Dr. Korpusik and Rosetta Yifan Hu

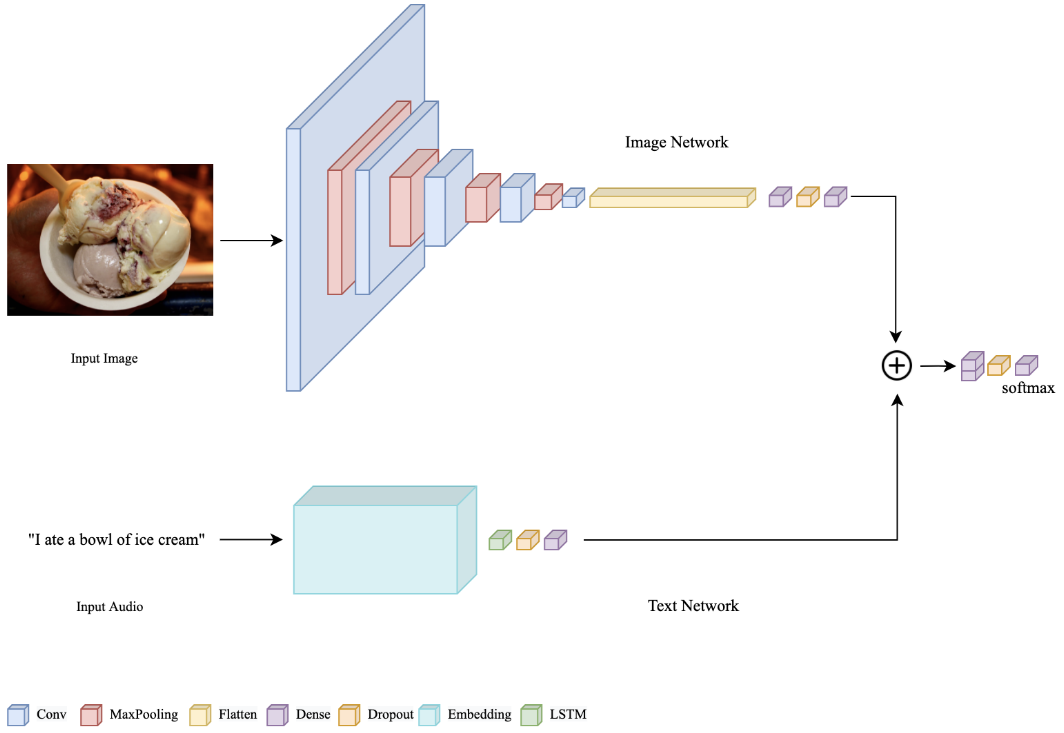


Figure 4-1: CNN+LSTM network (image credit: Dr. Korpusik and Rosetta)

output. The model was trained until convergence, using the Adam optimizer with binary cross-entropy loss and a batch size of eight.

4.2 Model-II: Vision-and-Language Transformer (ViLT) with Pre-Trained Weights

Although the baseline model supports multi-modality, it consists of two drastically different architectures, the CNN and LSTM. Moreover, both the CNN and LSTM have been replaced by Transformers as the state-of-the-art for both text and vision modalities. In this section, we explore Vision-and-Language Transformer (ViLT)[Kim et al., 2021] (see Fig. 4-2), which uses a unified and efficient architecture for both text and image modalities. ViLT builds upon BERT [Devlin et al., 2018], which was the current best model for the language modality, and Vision Transformer (ViT)[Dosovitskiy et al., 2020], which is the current best model for the vision modality. The Transformer

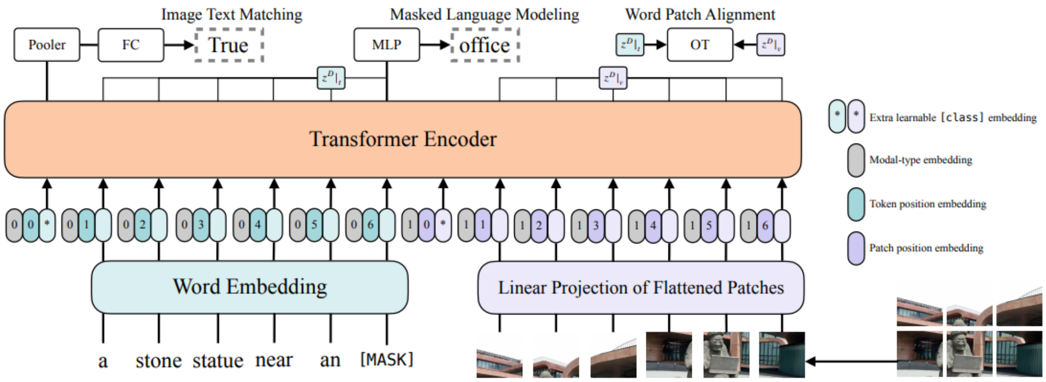


Figure 4-2: ViLT Model overview; image credit: [Kim et al., 2021].

layers in the ViLT model are initialized from ViT, and the language pre-processing pipeline uses

`bert-base-uncased`. Since our food classification task is very similar to the visual question answering task (VQA), we use a ViLT with pre-trained weights, specifically `vilt-b32-mlm`.[2] We further fine-tune the model on our dataset (Section 3.1) with a batch size of 32 until convergence. We use the Adam optimizer with a learning rate of 0.001. During all of our experiments, we kept the sigmoid threshold probabilities within a range of 0.1 to 0.35.

4.3 Experiments

4.3.1 Experiment-I: Multi-Label Food With Singel-Label Images

This section describes the experimental results on the data-set described in Section 3.1. 80% of the data comprises the training set, 10% of the data constitutes the validation set, and the remaining 10% forms the test set. For this experiment we freezed all the layers. As discussed in 2.2.4, we report macro-f1 and micro-f1 scores, Our results in Table 4.1 show that ViLT with pre-trained weights outperforms the CNN+LSTM baseline. In this experiment for ViLT, all the layers were kept fixed and

[2] https://huggingface.co/dandelin/vilt-b32-mlm-itm

the pre-existing weights were used.

Table 4.1: Macro-f1 and Micro-f1 Scores for Multi-Modal Models with single-label images and multi-label text

Model	Macro-f1	Micro-f1
Model-I (CNN+LSTM)	26.3	42.5
Model-II (ViLT with pre-trained weights)	**51.4**	**77.7**

4.3.2 Experiment-II: Multi-Label Food, Multi-Label Images With 696 Categories

Collecting multi-label image data (Section 3.2) was a challenging task due to the lack of reliable sources. However, with the assistance of OpenAI's DALL·E 2 API, we were able to successfully gather the necessary multi-label images. The text data used in this experiment is same as the previous experiment 3.2. In this experiment for ViLT, all the layers were kept fixed and the pre-existing weights were used. The results of our second experiment, which demonstrate the effectiveness of our approach, can be seen in Table 4.2.

Table 4.2: A summary of model with Macro-f1 and Micro-f1 scores described in this work for multi-label images and text with 696 categories

Model	Macro-f1	Micro-f1
ViLT(multi-label text and image with 696 categories)	**67.80**	**89.5**

Compared to the previous experiment 4.2, this experiment has higher macro-f1 and micro-f1 scores. This is expected and we believe it is due to the inclusion of multi-label images.

4.3.3 Experiment-III: Multi-label food, Multi-label images with 3589 categories

In experiment two, the categories were vague. But in this experiment the categories are made more specific. For example, spinach (exp 2) is further divided into New Zealand spinach, spinach salad, cooked spinach, raw spinach, defrosted spinach, etc. In other words, in experiment two, we tried to analyze how the model works on a food category. In experiment three, we understand how the model performs detecting food varieties. Table 4.3 shows the results of experiment three.

In this experiment, we aimed to evaluate the performance of our model as the number of categories increased. To do so, we employed a majority baseline and a random classifier. The last three layers weight were unfreezed for this experiment.

For the majority baseline classifier as shown in Figure 4-3, we looked at two scenarios. In the first scenario, the model predicts the most frequent category (salad) for all sentences. In the second scenario, the model predicts the top five most frequent categories (salad, sandwich, smoothie, spinach, rice) for all sentences.

sentence one: I ate a smoothie with 1 cup canned mixed fruit, 1 cup canned pineapple and 1/2 cup orange juice concentrate.

sentence one output for majority baseline one label classifier(salad): [salad]

sentence one output for majority baseline five label classifier (salad, sandwich, smoothie, spinach, rice): [salad, sandwich, smoothie, spinach, rice]

Figure 4-3: Majority baseline classifier

For the random classifier as shown in Figure 4-4, we looked at the same two scenarios as above, but this time the labels were chosen randomly. In the first scenario, five labels were chosen and in the other scenario, only one label was chosen.

This allowed us to thoroughly analyze the effectiveness of our approach in handling a larger number of categories.

sentence one: I ate a smoothie with 1 cup canned mixed fruit,1 cup canned pineapple and 1/2 cup orange juice concentrate.

sentence one output for random baseline one label classifier: [eggs]

sentence one output for random baseline five label classifier : [Muffins, raw spinach , orange juice, kale, red tomatoes]

Figure 4-4: Random baseline classifier

Table 4.3: A summary of model with Macro-f1 and Micro-f1 scores described in this work with

Model	**Macro-f1**	**Micro-f1**
Majority baseline one label classifier(salad)	0.0158	0.436
Majority baseline five label classifier (salad, sandwich, smoothie, spinach, rice)	0.0380	0.727
Random classifier one label	0.127	0.050
Random classifier five labels	0.051	0.019
ViLT	**38.17**	**68.27**

4.3.4 Adding The Branded dataset

The experiment involved examining how our models will perform with branded dataset refer to Section 3.3. All the layers of ViLT were kept fixed and the pre-existing weights were used.

In this section we combined dataset mentioned in Section 3.2, 3.3.2, 3.3.3 and 3.3.3. Where the model can learn from both text and image (Figure 3-4: a), text (Figure 3-4: c) and image only (Figure 3-4: b).

Table 4.4: A summary of model with Macro-f1 and Micro-f1 scores described in this work

Model	**Macro-f1**	**Micro-f1**
ViLT	**80.78**	**84.83**

4.3.5 Experiment-V: Visual ChatGPT vs ViLT

In the next step, we evaluated the performance of the ViLT model from our previous experiment and the Visual ChatGPT model [Wu et al., 2023]. Due to time constraints, we randomly selected 335 sentences from the test set. While waiting for access to GPT-4, Visual ChatGPT was chosen as the next best option. It combines ChatGPT's conversational abilities with the image understanding and generation capabilities of Visual Foundation Models (VFMs). Instead of creating a new multi-modal model from scratch, Visual ChatGPT uses existing ChatGPT and integrates various VFMs. A Prompt Manager is used to connect ChatGPT and VFMs by providing clear instructions, converting visual information into language format, and managing histories, priorities, and conflicts between different VFMs. On the other hand, GPT-4 is a large multi-modal model that can accept both image and text inputs and produce text outputs. This means that GPT-4 can analyze an image's contents and link that information to a written question.

As indicated in Table 4.5, ViLT demonstrated better performance than visual ChatGPT.

Table 4.5: A summary of model with Macro-f1 and Micro-f1 scores described in this work

Model	**Macro-f1**	**Micro-f1**
ViLT	**85.03**	**86.18**
Visual ChatGPT	54.07	56.19

The ViLT model performed better than Visual ChatGPT on our test sentences. However, *it's worth mentioning that we didn't give Visual ChatGPT access to our 3870 categories.* Instead, we asked it to generate a list of foods from the image and text. This could have negatively impacted Visual ChatGPT's score since the way we describe food can vary.

Chapter 5

Ablation Study

In the previous chapter, ViLT architecture performed better than the CNN+LSTM model. In this section, we conduct an ablation study. We investigate the language and vision modalities, each in isolation.

Language: To understand the diet tracking ability with language only, we fine-tune a BERT [Devlin et al., 2018] contextual embedding model on our downstream sequence classification task of natural language meal diaries. The BERT model is initialized with pre-trained weights from the `bert-base-uncased` model.[1] We fine-tune the model using the Adam optimizer [Kingma and Ba, 2015] (similar to [Devlin et al., 2018]) with weight decay of 0.1 and a learning rate of 0.001. Since our meal diaries' annotations are multi-label, we use binary cross-entropy loss (i.e., 0 or 1 per food).

Vision: To understand the diet tracking ability with images only, we fine-tune a Vision Transformer (ViT) model[Dosovitskiy et al., 2020] on our downstream food image classification task. The ViT model is initialized with pre-trained weights from `ViT_B_16`.[2] We fine-tune the model using the stochastic gradient descent (SGD) optimizer (similar to [Dosovitskiy et al., 2020]), with a learning rate of 0.0001. We use cross entropy loss.

[1] https://huggingface.co/bert-base-uncased

[2] https://pytorch.org/vision/main/models/generated/torchvision.models.vit_b_16.html

5.0.1 Ablation study I: Vision-and-Language Transformer (ViLT) with Pre-Trained Weights

Our first ablation study conducted on dataset with Multi-label sentences and single-label images. Since images in our dataset were annotated with a single-label only, we use cross entropy loss. For a fair comparison we also train another version of ViLT with a single-label objective. Note that our single-label setup has 167 balanced classes, whereas the multi-label setup has 696 imbalanced classes.

The results in Table 5.1 shows that even with single-label images, ViLT achieves significantly higher f1 scores than BERT for multi-label classification and than ViT for single-label classification, demonstrating that both modalities are helpful.

Table 5.1: A summary of model architectures, labels, and Macro-f1 and Micro-f1 scores described in this work

Model	Multi-Label		Single-Label	
	Macro-f1	Micro-f1	Macro-f1	Micro-f1
ViLT	**51.4**	**77.7**	**84.5**	**84.0**
BERT (Text Only)	31.5	56.1		
ViT (Image Only)			76.7	77.5

5.0.2 Ablation study II: Vision-and-Language Transformer (ViLT) with Pre-Trained Weights

Our second ablation study, which was conducted on our final model 4.3.4. The results from Table 5.2 shows that our ViLT models outperformed single modality approaches using either images or text alone.

5.0.3 Conclusion

As shown in Tables 5.1 and 5.2, the ViLT model outperforms both BERT and ViT in terms of f1 scores for food detection. This indicates that incorporating multi-modal information can significantly improve the performance of machine learning models in

Table 5.2: A summary of model with Macro-f1 and Micro-f1 scores described in this work

Model	**Macro-f1**	**Micro-f1**
ViLT	**80.78**	**84.83**
BERT	23.74	46.32
ViT	9.65	24.57

this task. The superior performance of ViLT demonstrates the potential benefits of using multi-modal approaches for food detection and other related tasks.

Chapter 6

Conclusion and Future Work

In our research, we presented a transformer-based method for diet tracking that shows the benefits of using both image and text information.

In our future work, we plan to address class imbalance 2-2 and bias in our dataset by expanding it to include more diverse user input and international cuisines. We added a branded dataset for only 3 categories, and we plan to expand it to other categories as well. We plan to investigate the use of newer models, such as BEIT-3 [Wang et al., 2022], which is a general-purpose multimodal foundation model that achieves state-of-the-art transfer performance on both vision and vision-language tasks. Our goal is to predict not only the food category but also its quantity.

Bibliography

Yoshua Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Cnn-based features for retrieval and classification of food images. *Computer Vision and Image Understanding*, 176:70–77, 2018.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Giovanni Maria Farinella, Dario Allegra, Marco Moltisanti, Filippo Stanco, and Sebastiano Battiato. Retrieval and classification of food images. *Computers in biology and medicine*, 77:23–39, 2016.

Thamme Gowda, Weiqiu You, Constantine Lignos, and Jonathan May. Macro-average: Rare types are important too. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1138–1157, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.90. URL https://aclanthology.org/2021.naacl-main.90.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Narit Hnoohom and Sumeth Yuenyong. Thai fast food image classification using deep learning. In *2018 International ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI-NCON)*, pages 116–119. IEEE, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Parneet Kaur, Karan Sikka, Weijun Wang, Serge Belongie, and Ajay Divakaran. Foodx-251: A dataset for fine-grained food classification. *arXiv preprint arXiv:1907.06167*, 2019.

Yoshiyuki Kawano and Keiji Yanai. Food image recognition with deep convolutional features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 589–593, 2014.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

M. Korpusik and J. Glass. Spoken language understanding for a nutrition dialogue system. *IEEE Transactions on Audio, Speech, and Language Processing*, 25:1450–1461, 2017.

M. Korpusik and J. Glass. Convolutional neural networks and multitask strategies for semantic mapping of natural language input to a structured database. In *Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6174–6178. IEEE, 2018.

M. Korpusik, N. Schmidt, J. Drexler, S. Cyphers, and J. Glass. Data collection and language understanding of food descriptions. *Proceedings of 2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 560–565, 2014.

M. Korpusik, C. Huang, M. Price, and J. Glass. Distributional semantics for understanding spoken meal descriptions. *Proceedings of 2016 IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6070–6074, 2016.

M. Korpusik, Z. Collins, and J. Glass. Semantic mapping of natural language input to database entries via convolutional neural networks. *Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5685–5689, 2017a.

M. Korpusik, Z. Collins, and J. Glass. Character-based embedding models and reranking strategies for understanding natural language meal descriptions. *Proceedings of Interspeech*, pages 3320–3324, 2017b.

Mandy Korpusik and Jim Glass. Deep learning for database mapping and asking clarification questions in dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

Mandy Korpusik, Zoe Liu, and James Glass. A comparison of deep learning methods for language understanding. *Proc. Interspeech 2019*, pages 849–853, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

Tao Mei, Wei Zhang, and Ting Yao. Vision and language: from visual perception to content creation. *APSIPA Transactions on Signal and Information Processing*, 9, 2020.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, 2017.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Keiji Yanai and Yoshiyuki Kawano. Food image recognition using deep convolutional network with pre-training and fine-tuning. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2015.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.