Multi-Modal Food Classification in a Diet Tracking System with Spoken and Visual Inputs

Shivani Gowda*
Computer Science Department
Loyola Marymount University
Los Angeles, CA, USA
sks@lion.lmu.edu

Yifan Hu*
Computer Science Department
Loyola Marymount University
Los Angeles, CA, USA
yhu9@lion.lmu.edu

Mandy Korpusik
Computer Science Department
Loyola Marymount University
Los Angeles, CA, USA
Mandy.Korpusik@lmu.edu

***Abstract*—In this paper, we present multi-modal approaches to diet tracking. As health and well-being become increasingly important, mobile applications for diet tracking attract much interest. However, these applications often require users to log their meals based on relatively unreliable memory recall, thereby underestimating nutritional intake and, thus, undermining the efforts of nutrition tracking. To accurately record dietary intake, there is an increasing need for image computational methods. We investigated multi-modal transfer learning approaches on a novel, food-specific image-text dataset, specifically a Vision-and-Language Transformer that achieves a held-out test set Micro-F1 score of 77.70% and Macro-F1 score of 51.43% for 696 food categories. We aim to give other researchers new insight into the process of developing domain-specific, multi-modal deep learning models with small datasets.**

***Index Terms*—Transfer Learning, Convolutional Neural Network, Long Short-Term Memory, Vision-and-Language, Transformer**

I. Introduction

The worldwide concern of obesity dominates many headlines [1]. Thus, in recent years, there has been an increased interest in tracking diets [2], [3]. We previously built a mobile application prototype, *Coco Nutritionist*, [4]–[9] that lets users record food intake with natural language and accurately estimates calories and other nutrients consumed. Unbiased estimation of daily nutritional intake is crucial for maintaining a healthy lifestyle. The current methods of dietary assessment only rely on self-reports, which often leads to underestimation. Thus, it is necessary to supplement memory recall with automatic food recognizers. Such a computer-aided dietary intake analysis system could detect food items in photos, along with quantity estimates, and translate those to nutritional values.

In recent years, the rapid development of the computer vision field greatly boosted the accuracy and robustness of models across various domains. As deep learning models became increasingly popular, various architectures and methods have been utilized, and the Transformer [10] has emerged as state-of-the-art. The emphasis on performance has facilitated an endless collection of optimization and training methods. Even though these novel approaches are beneficial to the field as a whole, they also pose the dilemma for researchers to select a combination that is suitable for their specific dataset. In this work, we report our process of arriving at one of the most optimal models for the diet tracking task.

The contributions of this work are as follows:

1) We present a multi-modal dataset for diet tracking. Our dataset consists of two modalities: meal diaries as the language modality, and food images as the vision modality. Our dataset contains over 16,600 images paired with meal diaries and covers a label space of over 696 food categories (Section III).
2) We present models for diet tracking and demonstrate how each modality contributes to the final performance.
3) Our experiments and analysis show that the multi-modal Transformer with pre-trained weights achieves state-of-the-art performance (Sections IV-C and VI).

II. Related Work

We present related work in the following subsections:

A. Image and Text Classification for Diet Tracking

The rapid development of neural networks [11] prompted advancement in the domain of food image classification. After the release of the benchmark ImageNet dataset for image classification [12]–[16], 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) [17]. Later, the authors enhanced their model through transfer learning [18]. Several other CNN-based methods have been explored by other studies concerning food image recognition [19]–[22], but none multi-modal.

For natural language processing (NLP), Transformer-based contextual embedding models such as bidirectional encoder representations from Transformers (BERT) [23] are state-of-the-art. Comprehensive studies have been conducted to analyze the rise of deep learning in text classification [24], [25].

B. 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

*Equal contributions.

results in tasks such as visual question answering and text-image generation [26]. Starting with ViLBERT [27], there is a surge in using Transformers as the main architecture, shifting away from recurrent neural networks (RNNs). Both ViLBERT and LXMERT [28] fuse two separate Transformers, one for images and one for text. Later, researchers [29]–[31] introduced a single-stream Transformer model to better understand joint representations.

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 paper, we collected a food-specific multi-modal dataset and developed a novel architecture to accomplish multi-modality in the real world.

III. Data Collection

Collecting training data is one of the important aspects of supervised machine learning. It is well-known that quality data is crucial to model performance. Previously, we were able to directly use or expand upon an open-source dataset. However, with multi-modal learning, the model requires two modalities of data, doubling the time and effort required for collection. In the new multi-modal dataset we constructed, we approached the problem by combining two sets of data: text-to-image and image-to-text which are described in detail in the following section. The dataset maps an image containing food to a user's natural language meal description. There is a total of 16,600 images from 696 food categories.

Fig. 1. Examples of Text-to-Image data

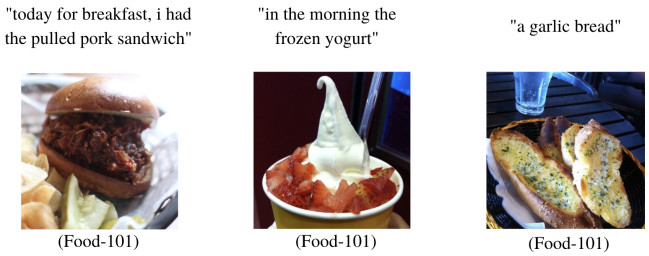

Fig. 2. Examples of Image-to-Text data

A. Text-to-Image Data

In our prior work [32], [33], meal diaries were crowd-sourced from Amazon Mechanical Turk [8]. 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).

At the beginning of the data collection process, we attempted to use novel zero-shot text-to-image generators (e.g, Deep-Daze [34]) to gather more images. However, the results proved to be unsatisfactory for real food image classification. Many of our natural language meal descriptions include multiple food categories,[1] but finding images with the exact combination of food categories mentioned in the diary is hard, so we used images with one of the food categories in the natural language meal description (see Figure 1). We have 696 food categories from which we picked the 89 most common classes within our assembled meal diary dataset. For 66 of the 89 classes, we collected food images from Flickr's dataset and mapped them to the natural language sentences. For the remaining 23 food classes, we used 100 images per category from the training set of the Food-101 dataset [35].

B. Image-to-Text Data

For the second part of the dataset, we selected 100 images each from the remaining 78 food categories in the Food-101 dataset. In total, we had 167 food categories (i.e., 89 classes in text-to-image and 78 classes in image-to-text). Due to the lack of natural language food dietary data, we needed to generate natural sentences in order to complete the multi-modal dataset.

During our first attempt, we employed state-of-the-art automated image captioning with visual attention [36]. There were two issues with this approach. First and most importantly, the generated text does not follow the format of nutritional diaries with a quantity and food token. Often, the generated text only describes the visual aspects of the image, which renders it insufficient for a diet tracking system. Second, similar to many pre-trained models, the description is not specific enough to the food categories to be useful. Due to the fact that each image is a close-up of the food, the generated captions are so similar that even if we replaced the food token with a more distinct category, the image caption would remain inadequate.

To solve the first problem, we attempted to generate a diary, using a text generator [37] trained on our set of existing dietary entries. However, we realized that this form of generation is too general in the category of food to be valuable as training data. Thus, we decided to generate a diary based on a custom text template composed of the most common phrases found in the natural language dataset (see Figure 2).

C. Merge

In the end, we collected 16,600 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 were consist of only one food. As a result, we used 167 classes for the image modality and 696 total classes for the natural language modality.

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

IV. Models

Our approach is based on transfer learning which is a technique to leverage existing resources and transfer knowledge learned from a related domain to another. Rather than building a deep neural network from scratch, we utilized the weights from models that are already trained on large-scale datasets and only replaced the last layer to adapt to our specific dataset.

A. Model-I: *Baseline Vision-and-Language Model without Pre-Trained Weights*

Our first conceptually simple approach for the vision-and-language pipeline is based on CNN and LSTM (i.e., Long Short-Term mMemory recurrent neural network) architectures. The motivation behind this approach is to develop a single neural network that is able to effectively employ information from two modalities and analyze the results without pre-trained weights. In this model, we concatenated a convolutional neural network (CNN) [12] with an LSTM [39] 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.

1) CNN: For this network, we use 2x32, 2x64, and 1x128 dimension layers with a 3x3 kernel and rectified linear unit (ReLU) activation function. In our model architecture, we use max-pooling to extract the maximum value with a 2x2 filter. Each convolution layer, except for the last one, is followed by a max-pooling layer to decrease the dimensionality of the convolution output vector. After several convolution and pooling layers, the feature map is flattened and fed into a fully-connected layer. To reduce overfitting, we added a dropout layer before feeding the CNN output vector into a softmax activation layer to generate a probability distribution.

2) LSTM: For our LSTM model, we first fed the input text into an embedding layer, followed by an LSTM (128 dimensions) and a dropout layer. This layer is then followed by a linear layer of 696 neurons. This process allows the network to assign a domain-specific food category to each natural language sentence.

3) Concatenation: Our proposed model consists of the two separate input layers, with each one followed by a hidden layer, and then merged by a concatenation layer. The merge is followed by a dense layer, a dropout layer, and a final output classification layer with a softmax activation function.

The image features are extracted by the CNN model outlined in section IV-A1 to output a classification result. The second input is processed by the LSTM model described in section IV-A2. The classification outputs of the two branches are concatenated together by a final layer to obtain the global output. We used the Adam optimizer with binary cross-entropy loss, a batch size of eight, and trained until convergence.

B. Model-VII: *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 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) [38] (see Fig. 3), which uses a unified and efficient architecture for both text and image modalities. ViLT builds upon BERT [23], which is the current best model for the language modality, and Vision Transformer (ViT) [40], which is the current best model for the vision modality. The Transformer 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 III) with a batch size of 32 until convergence. We use the Adam optimizer with a learning rate of 0.001.

C. Experiments

This section describes the experimental results on the dataset described in Section III. 80% of the data comprises the training set, 10% of the data constitutes the validation set, and the remaining 10% forms the test set. The process described above demonstrates the difficulty in collecting a quality multi-modal dataset for a specific downstream task. As shown in Figure 4, our dataset has class imbalance. We see that only a small fraction of food categories have more than 100 examples, and the majority of them have 10 or fewer examples. Due to the class imbalance, we report both Macro-F1 and Micro-F1 scores [41]. While the Micro-F1 metric assigns equal weight to each instance, the Macro-F1 metric assigns equal weights to each class. Macro-F1 is a useful metric in data imbalance scenarios like ours. Our results in Table I show that ViLT with pre-trained weights outperforms the CNN+LSTM baseline.

TABLE I
Macro-f1 and Micro-f1 Scores for Multi-Modal Models

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

V. Ablation of Vision-and-Language Transformers

The ViLT architecture is 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 [23] contextual embedding model on our downstream sequence classification task of natural language meal diaries. The BERT model is initialized

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

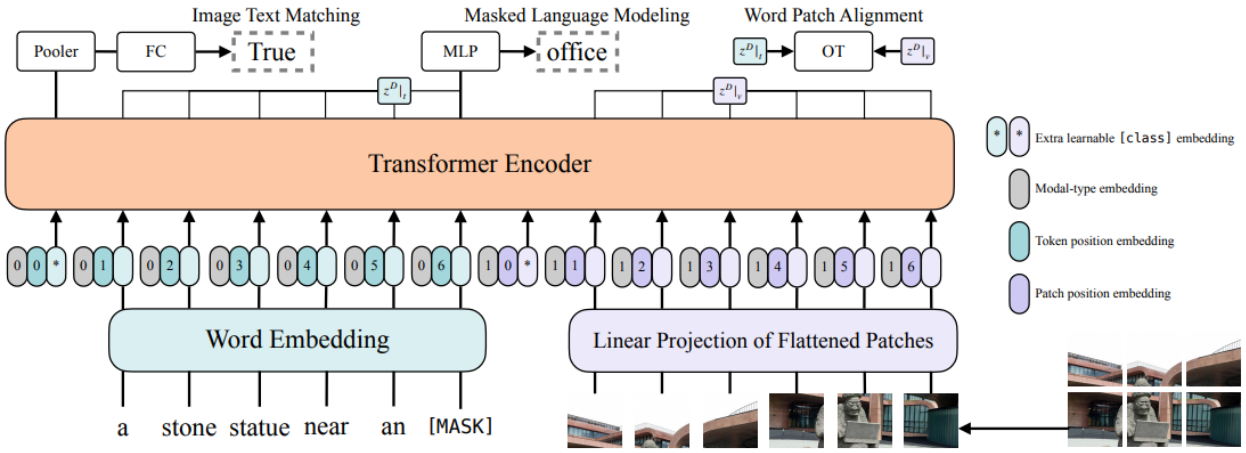

Fig. 3. ViLT Model overview; image credit: [38].

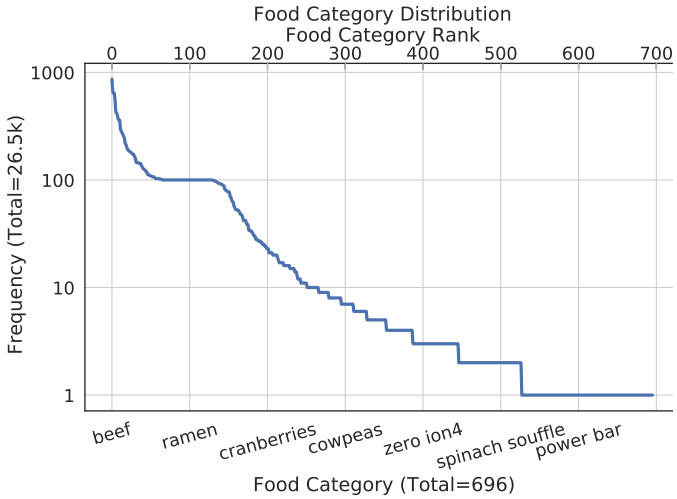

Fig. 4. Food category distribution. The vertical axis is in log scale. Categories in our dataset are imbalanced as in the real-world scenario.

with pre-trained weights from the `bert-base-uncased` model.[3] We fine-tune the model using the Adam optimizer [42] (similar to [23]) 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 [40] on our downstream food image classification task. The ViT model is initialized with pre-trained weights from `ViT_B_16`.[4] We fine-tune the model using the stochastic gradient descent (SGD) optimizer (similar to [40]), with a learning rate of 0.0001. Since images in our dataset are annotated with a single label only (see Section III-A), 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. As shown in Table I, 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.

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

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

TABLE II
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

VI. Conclusion and Future Work

In this paper, we have proposed two related models that attempt to achieve food classification in the domain of diet tracking. We have achieved significant results with our ViLT model trained on a custom dataset of 16,600 image-text pairs.

In future work, we will continue to expand our dataset to be more inclusive and reflective of user input and also handle the imbalance in the classes. The current dataset also suffers from bias due to the limited dataset. For example, we rely on the USDA food database, which is primarily American food, but aim to expand to international cuisines in the future. In addition, the two modalities are not always reflective of each other. Some text descriptions include more food items than the image, while images might not necessarily reflect the quantity described in the text due to the fact that we combined inputs from different sources. In the future, we will not only predict the food category, but also the quantity.

References

[1] "Obesity and overweight," 2021.

[2] Megan E Rollo, Elroy J Aguiar, Rebecca L Williams, Katie Wynne, Michelle Kriss, Robin Callister, and Clare E Collins, "ehealth technologies to support nutrition and physical activity behaviors in diabetes self-management," *Diabetes, metabolic syndrome and obesity: targets and therapy*, vol. 9, pp. 381, 2016.

[3] Rosalind Fallaize, Rodrigo Zenun Franco, Jennifer Pasang, Faustina Hwang, and Julie A Lovegrove, "Popular nutrition-related mobile apps: an agreement assessment against a uk reference method," *JMIR mHealth and uHealth*, vol. 7, no. 2, pp. e9838, 2019.

[4] 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)*, pp. 6070–6074, 2016.

[5] 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)*, pp. 5685–5689, 2017.

[6] 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)*. IEEE, 2018, pp. 6174–6178.

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

[8] 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)*, pp. 560–565, 2014.

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

[10] 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)*, 2017, pp. 5998–6008.

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

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

[13] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

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

[15] 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*, 2015, pp. 1–9.

[16] 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*, 2016, pp. 770–778.

[17] 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*, 2014, pp. 589–593.

[18] 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)*. IEEE, 2015, pp. 1–6.

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

[20] 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.

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

[22] 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)*. IEEE, 2018, pp. 116–119.

[23] 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.

[24] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[25] 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)*, vol. 54, no. 3, pp. 1–40, 2021.

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

[27] 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.

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

[29] 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.

[30] 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.

[31] 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*. Springer, 2020, pp. 104–120.

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

[33] 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.

[34] "Deep-daze," 2021.

[35] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool, "Food-101–mining discriminative components with random forests," in *European conference on computer vision*. Springer, 2014, pp. 446–461.

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[37] Max Woolf, "textgenrnn," 2020.

[38] Wonjae Kim, Bokyung Son, and Ildoo Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.

[39] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] 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.

[41] 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*, Online, June 2021, pp. 1138–1157, Association for Computational Linguistics.

[42] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.