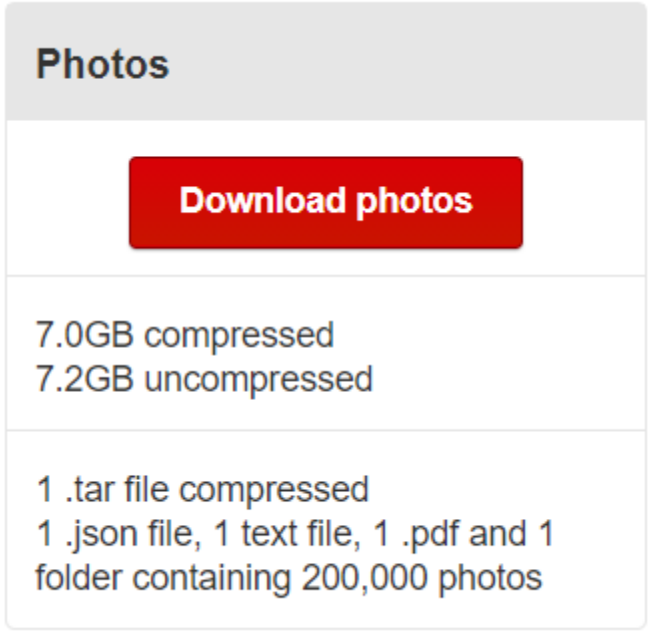
Food Image Captioning in Yelp Dataset

Introduction

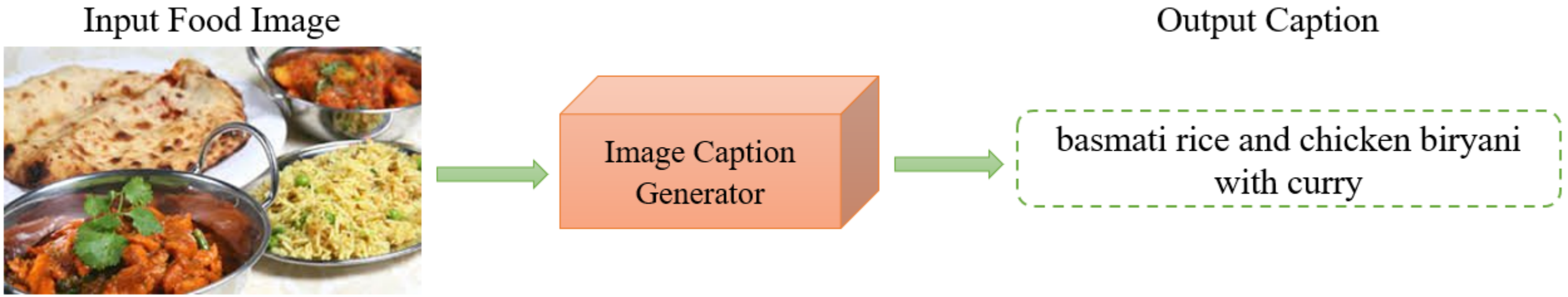
Dataset For Food Image Captioning: For data acquisition we have used the subset of yelp data set which consist of image and the data related to those images.Dataset has 200k images along with the image description in json file.



Problem Statement

Problem Statement:

To build a model that can describe the content of the food image in the form of text given an input image of food.



- Generate food caption consisting of food name probably the various items of food image.
- Different images with different context like single object, multiple objects or background of the image.

Understanding of the Data

Key Features of Dataset

- Dataset has 200K images along with the json file.
- Json file has following attributes

  caption : String, the photo caption, if any.

  photo_id : String, 22 character unique photo id.

  business_id : String, 22 character business id maps to business in business.json.

  label : String, the category the photo belongs to.

- Size of json file: 200000 rows x 5 columns.
- Missing values:

  caption : 107850

  photo_id : 0

  business_id : 0

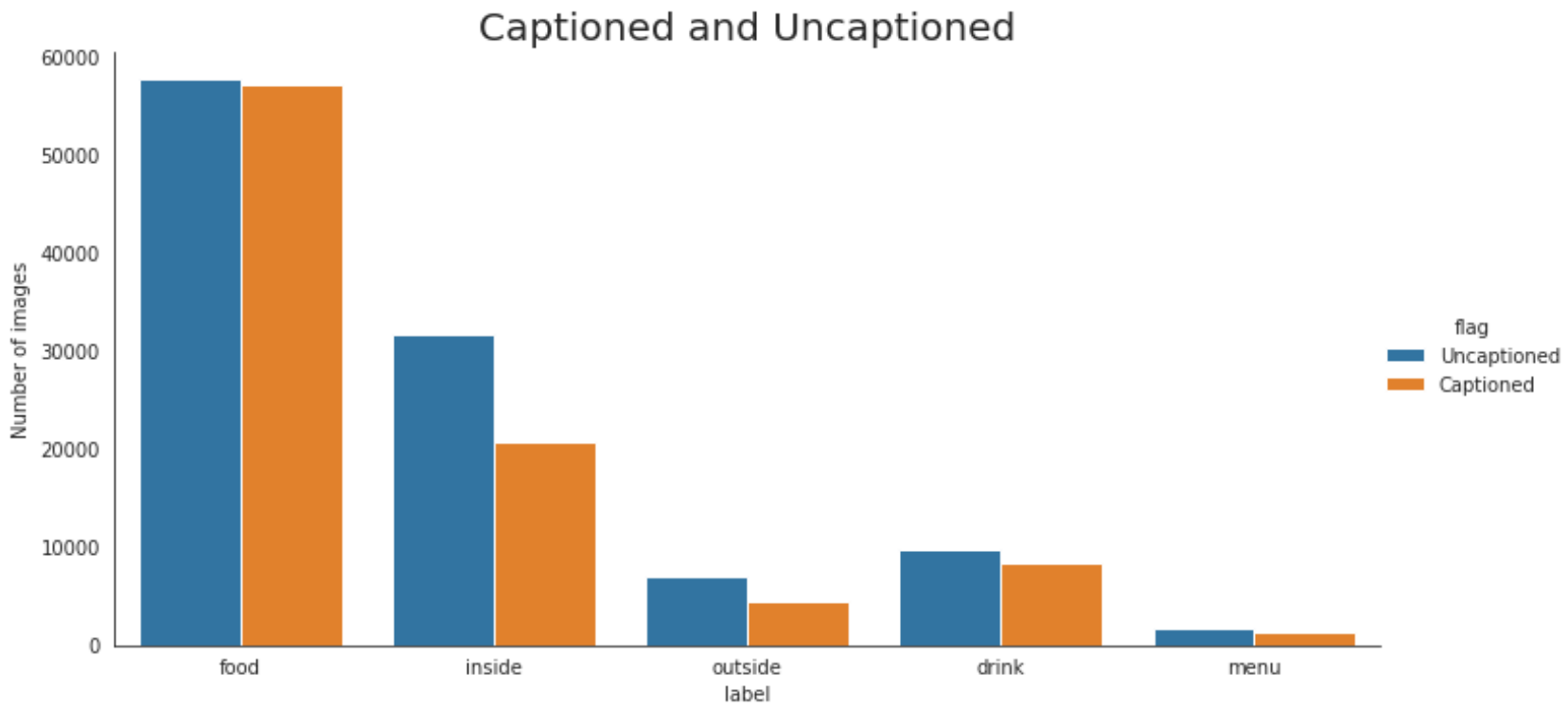
  label : 0

Data set has 92150 observations with caption

Data set has 107850 observations without caption.

Understanding of Data

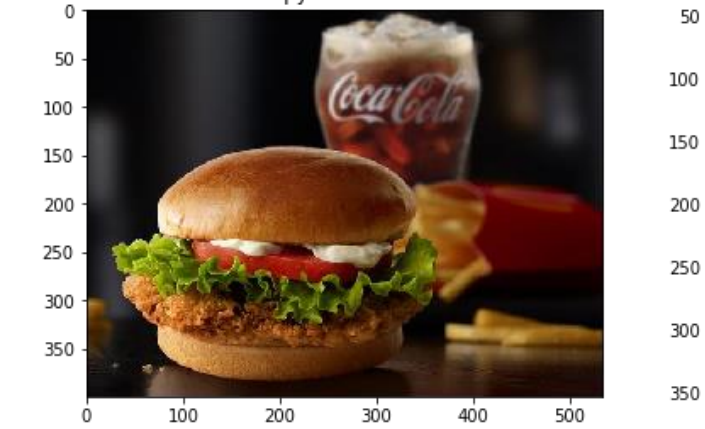
Data distribution (captioned vs uncaptioned)

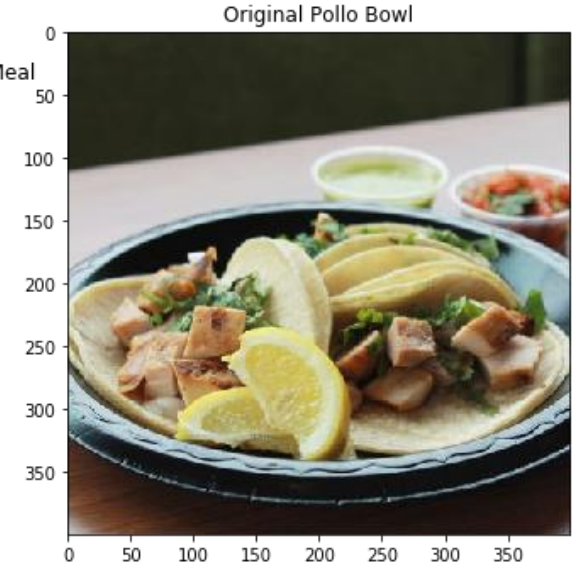


- Captioned food images : 57,151
- Uncaptioned food images: 57,723

Understanding of Data

View of Data: Food images along with caption





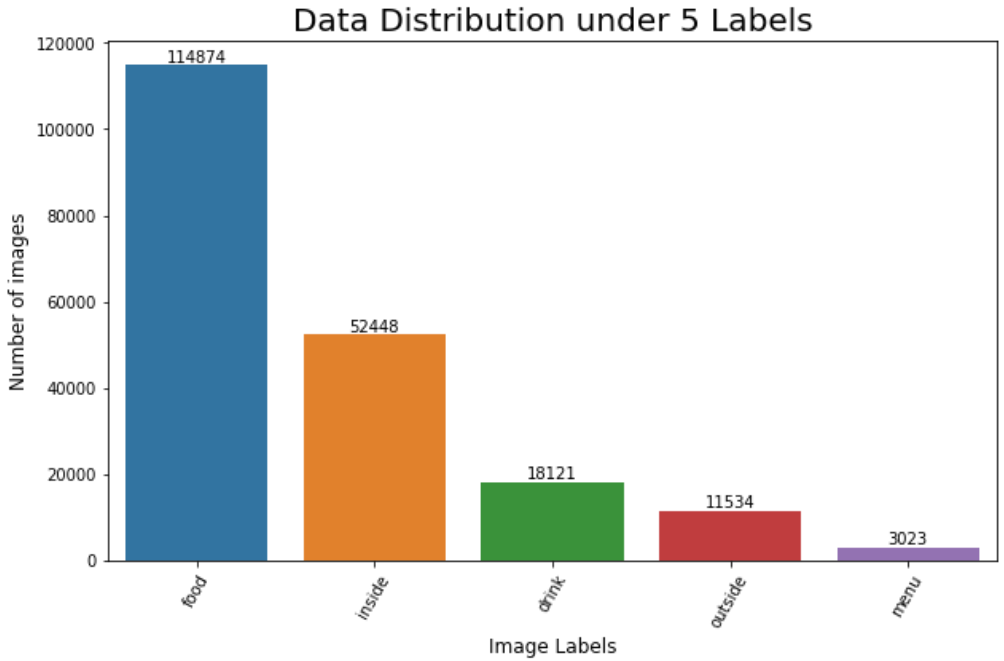
Understanding of Important Attributes

Photo Id : To fetch the image along with caption.

Caption : To check whether image is provided with caption or not.

Label : To categorize images under 5 labels.

{‘food’, ‘drink’, ’inside’, ’outside’, ‘menu’}



Required: Images with caption which belong to food category.

Literature Survey

A Survey on Food Computing – Dataset [14]

For Image Captioning

Name	Data Type	Task
MS COCO	Image + 5 Captions	Image Captioning
Flickr30K	Image + Caption	Image Captioning

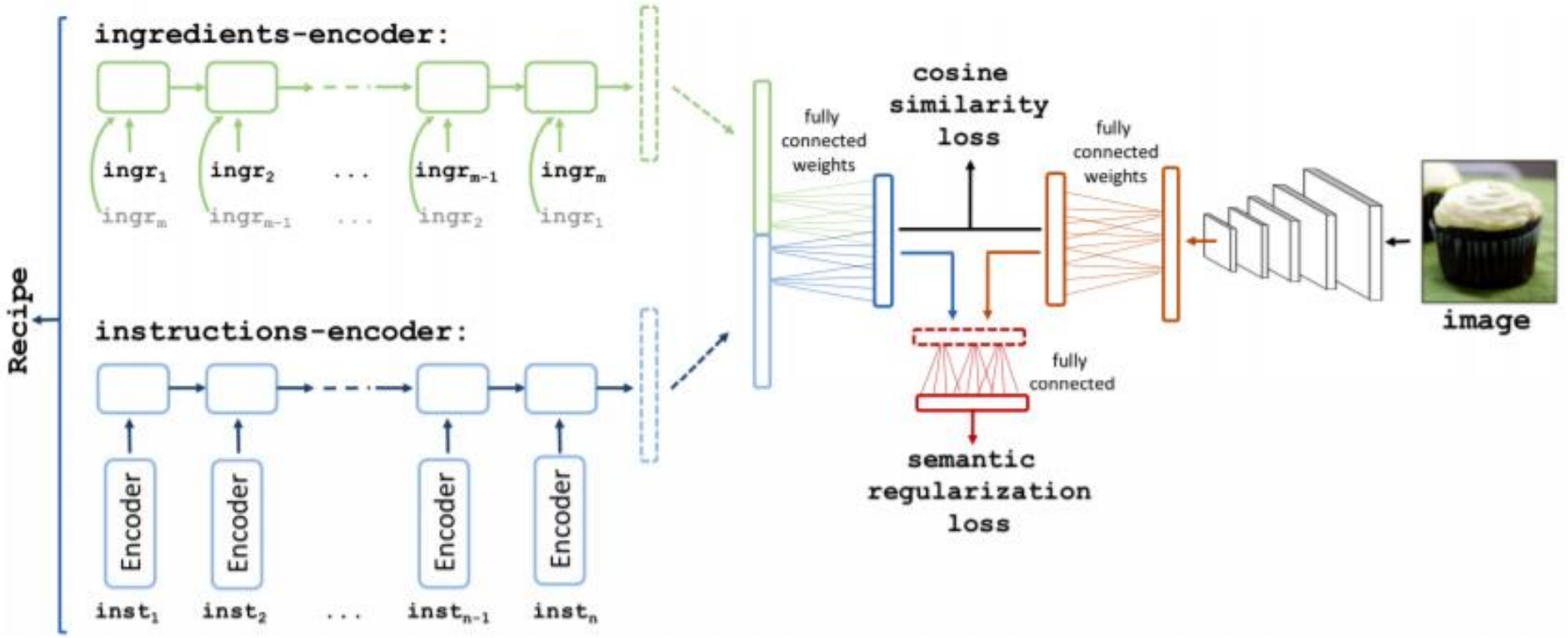
For Food Domain

Name	Data Type	Task
Food101	Image + Text	Cross Modal Retrieval
Recipe1M	Image + Text	Cross Modal Retrieval
Yummly28k	Image + Text	Cross Modal Retrieval
Yelp	Image+ Caption	Image Captioning

Literature Survey(In-Domain)

Learning Cross-modal Embeddings for Cooking Recipes and Food Images(2017)[5]

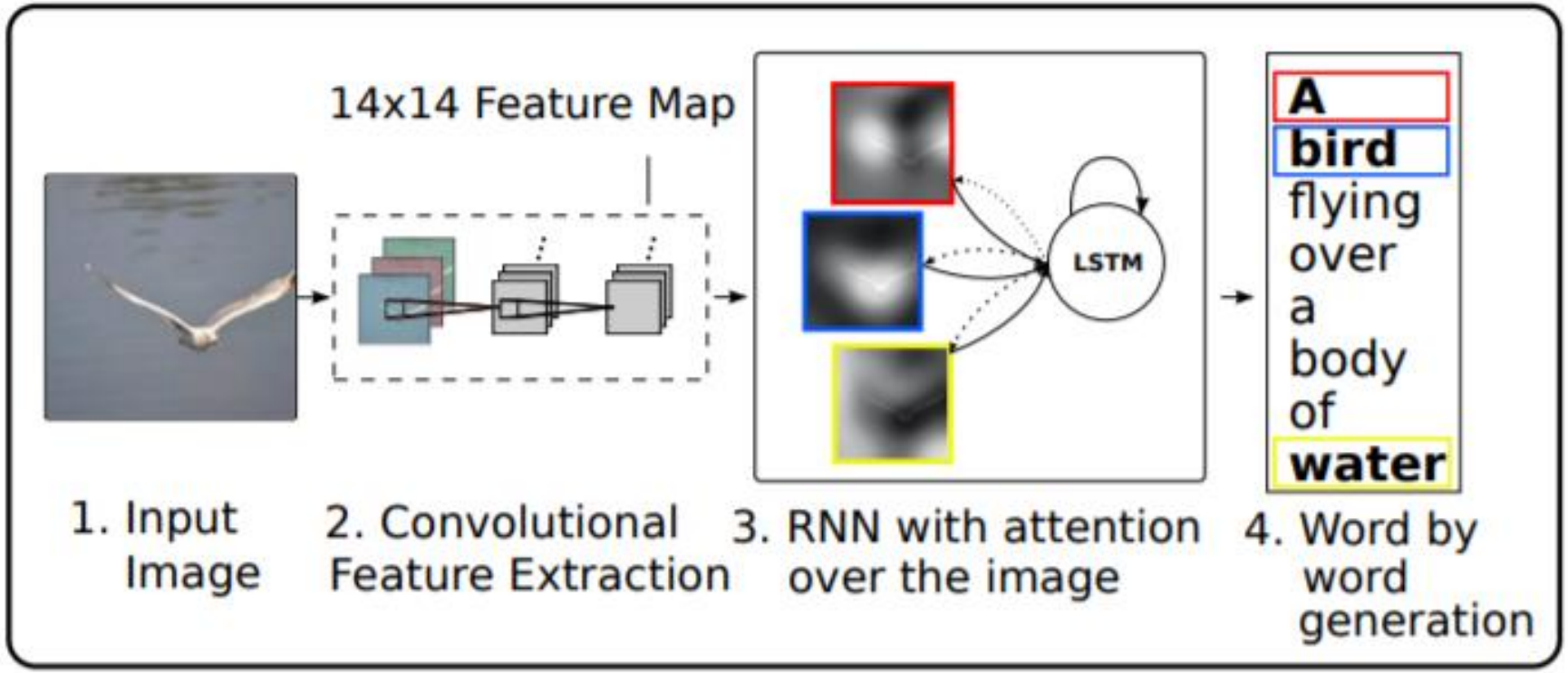
Joint neural embedding model with semantic regularization. Our model learns a joint embedding space for food images and cooking recipes.



Literature Survey(Out-Domain)

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (ICML 2015)[2]

Attention Based Image Captioning



Reference:

"Show, Attend and Tell" by Xu et al. ICML 2015.

Proposed Methodology

Caption

Caption Pre-processing

Input Food Image

CNN Encoder

LSTM Decoder

Intermediate Caption

chicken biryani with curry

Multi Label Classification

Intermediate Label

basmati rice with masala chicken

Natural Language Generator

Generated Caption

Basmati rice and masala chicken biryani with curry

Reference:

- Baig, Muhammad & Shah, Mian & Wajahat, Muhammad & Zafar, Nauman & Arif, Omar. (2018). "Image Caption Generator with Novel Object Injection". 1-8. 10.1109/DICTA.2018.8615810.

Pre-processing (On caption)

Pre-processing on Caption using NLP:

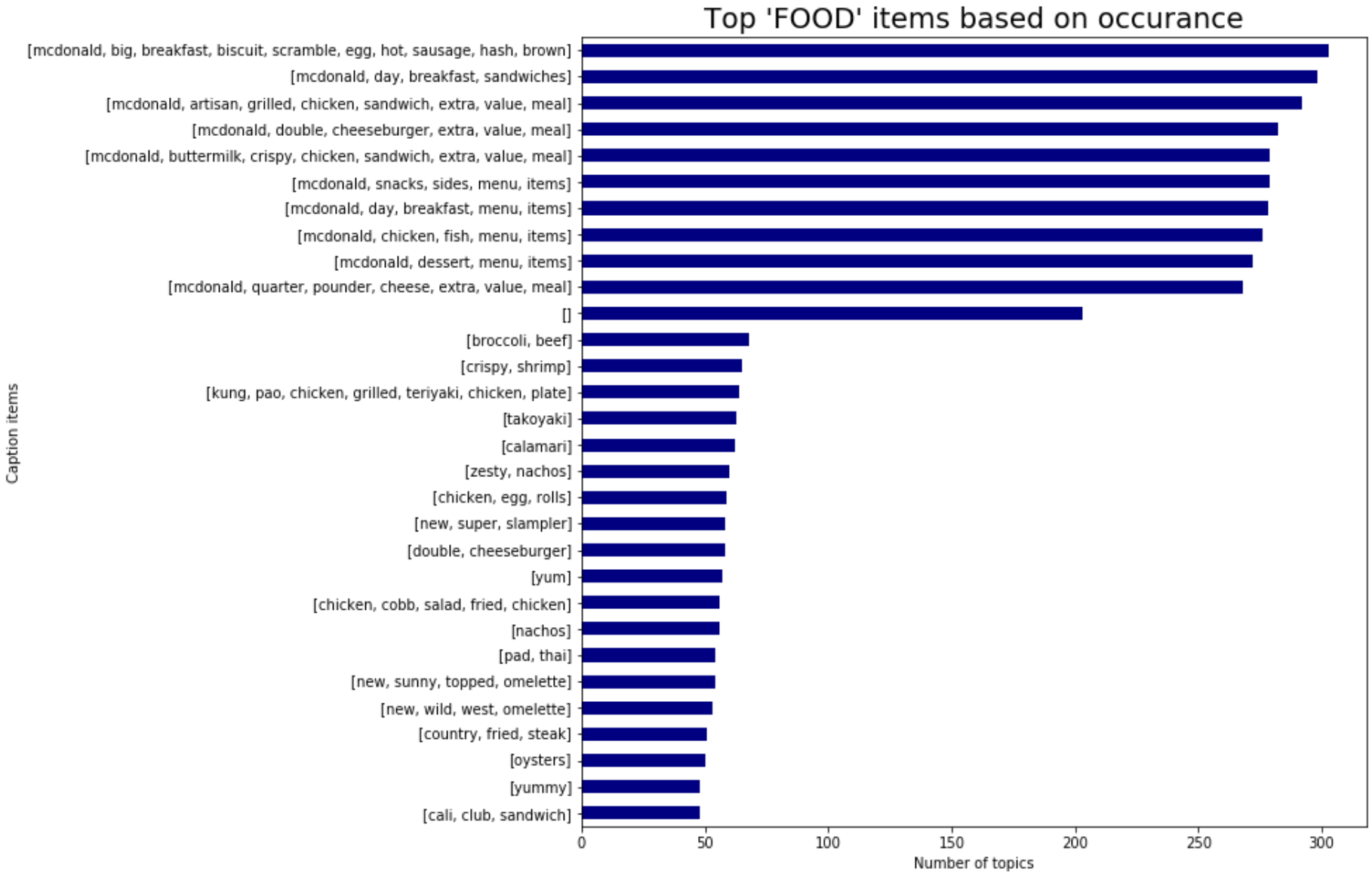
- To clean the caption.
- Retain the caption which are alphanumeric and more than 5 characters using regular expression for pattern matching.
- Lemmatization, Stop words removal etc.
- The count of images with caption after pre-processing retained is 92k i.e., no image is left.
- Avg. number of words per image

  Before pre-processing: 5.5

  After pre-processing: 4.1

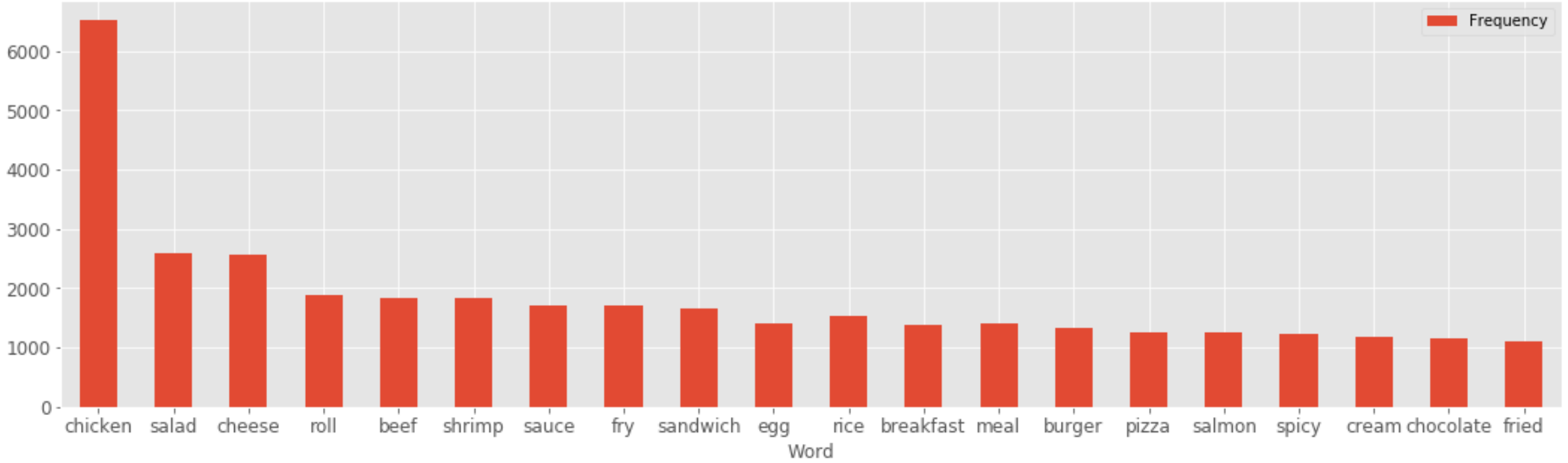
Pre-processing(On caption)

Finding top Food item sets based on occurrence



Pre-processing (On caption)

Finding top occurring Food related words:



- Top 20 words which occur frequently in entire caption set.
- Purpose: Required in the proposing approach.

Inferences from Preprocessing

- The preprocessing on caption brings it into digestible form so that building model can perform better.
- Top occurring words can be considered as labels for the particular image.
- The top occurring word extraction help in multilabel classification for pre-processing module in pipeline.

Proposed Methodology

Input Food Image

CNN Encoder

LSTM Decoder

Intermediate Caption

chicken biryani with curry

Multi Label Classification

Intermediate Label

basmati rice with masala chicken

Natural Language Generator

Generated Caption

Basmati rice and masala chicken biryani with curry

Reference:

- Baig, Muhammad & Shah, Mian & Wajahat, Muhammad & Zafar, Nauman & Arif, Omar. (2018). "Image Caption Generator with Novel Object Injection". 1-8. 10.1109/DICTA.2018.8615810.

Image Captioning

Caption Generation Model: CNN- LSTM Framework

- CNN Encoder – InceptionV3 for Feature extraction

  CNN Feature extraction followed by Fully Connected Layer and Relu activation.

- LSTM Decoder – Attention , GRU and Fully Connected Layer

- Gradient descent to minimize the loss.

- Trained on the Yummly28k dataset.

- Loss is observed to be decreasing.

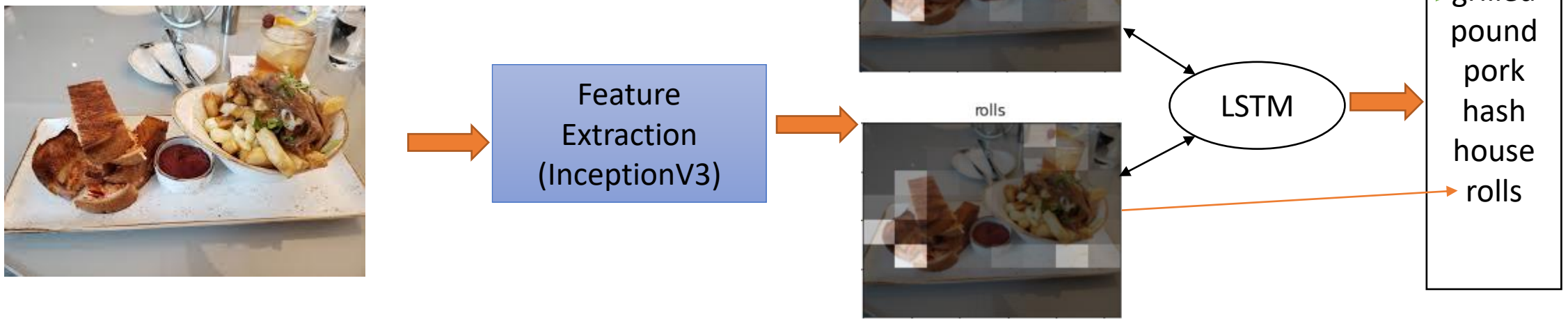
Image Captioning

Input Image

Feature Map using InceptionV3

RNN with Attention

Caption Generation



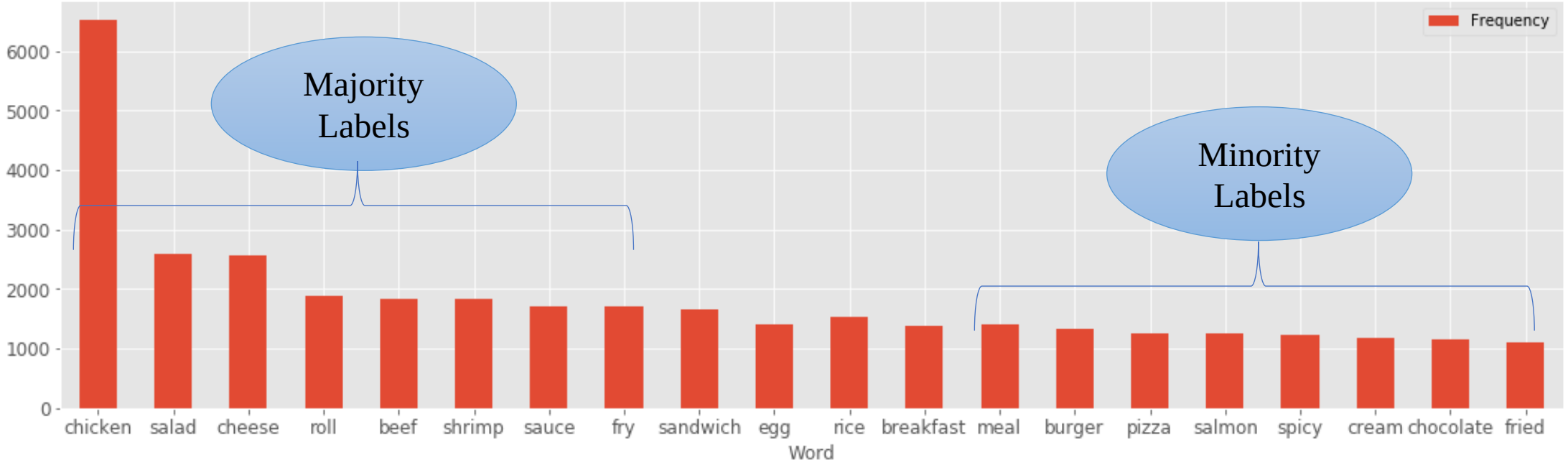
CNN – LSTM Framework for Caption Generation

Role of Multi Label Classification

- To generate a part of a caption as a label considering only the image features. Adding more meaning to the caption.
- To handle lack of information in the caption provided.
- To handle long tail problem.

Long Tail Problem

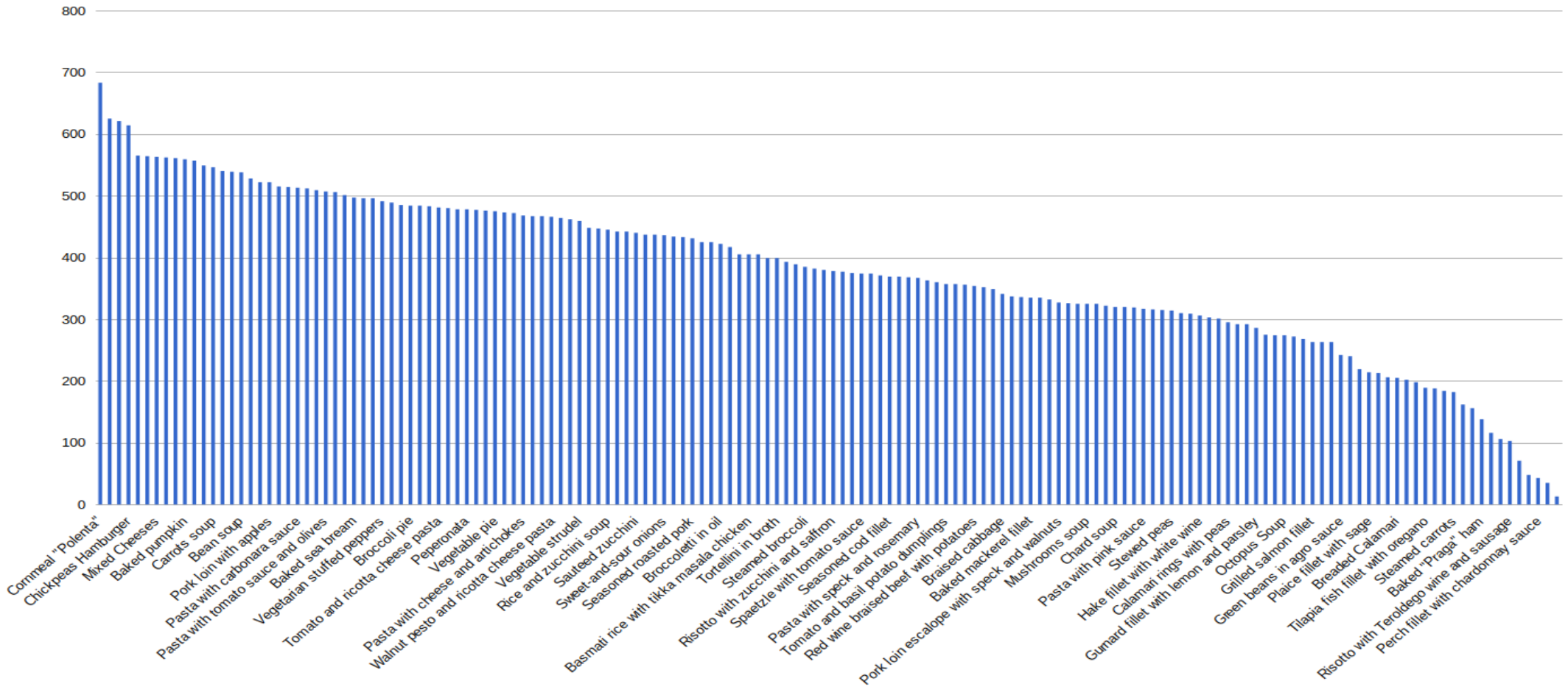
- Hard case in earlier approach Multi Label Classification with one hot encoding. As well in YELP.
- One hot encoding classifier failed on test data.
- Reason : Long Tail Problem or Imbalanced labels.



- Gradient dominance by majority labels.
- Under representation of minority labels.
- One hot encoding → Multiple Categories.

Multi Label Classification

Well balanced food data for multi label classification – FFoCat [1]

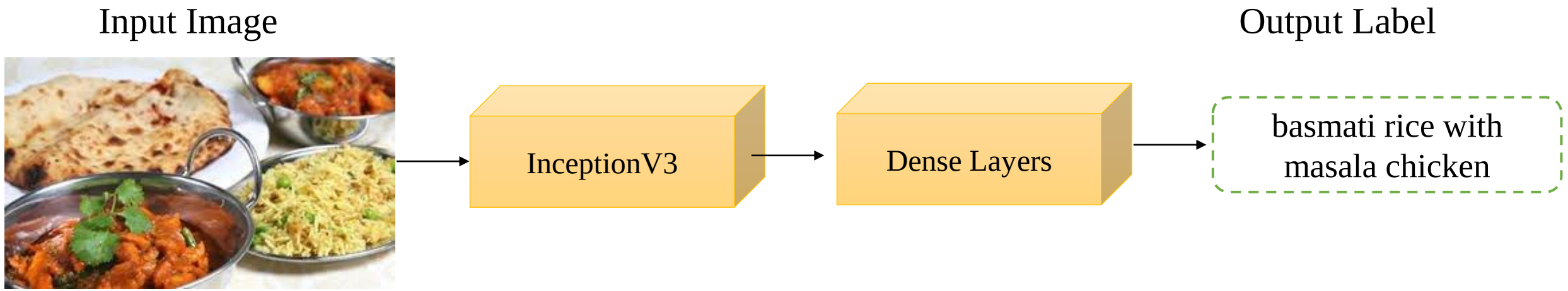


Each label(Eg. Rice and Zucchini Soup) can have multiple categories(Eg.‘Rice’, ‘Zucchini Soup’) of food.

Multi Label Classification

Multi Label Classification – Method

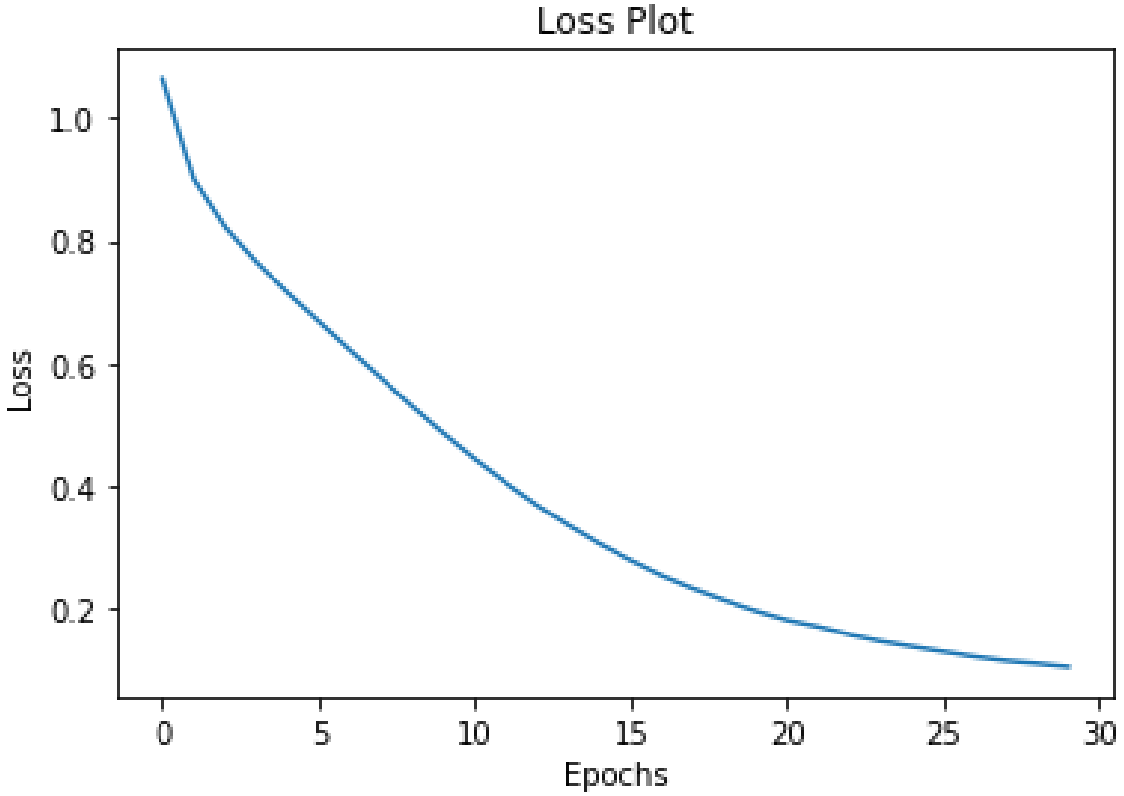
- Transfer learning approach for multi label classification.
- Trained on FFoCat dataset using InceptionV3 as feature extractor followed by dense layer against the 156 labels consisting of multiple categories of food.



- The label predicted is considered as an intermediate label in generating caption.

Performance

Image Captioning- Loss plot

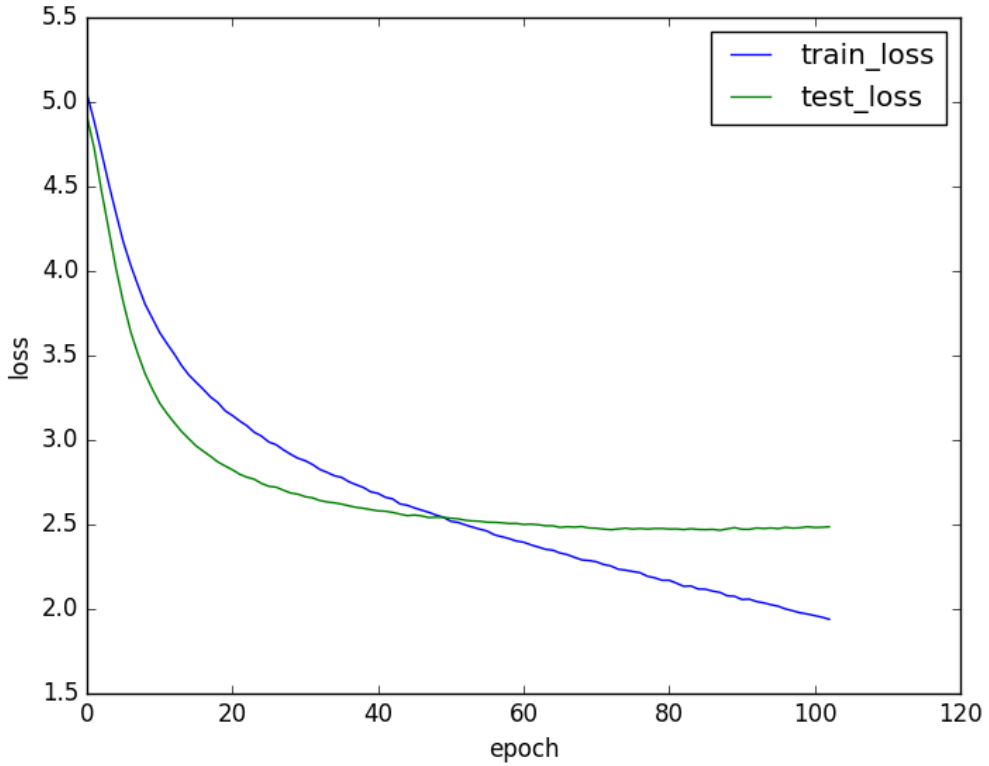


- Gradually decreasing loss indicating performance of the model is good.
- Able to achieve loss 0.1.

Performance

Multi Label Classification- Performance

- Gradual decrease in test loss at initial epochs followed by slow decrement.



- Test loss below the train.
- Evaluation of model on yelp data set.

Multi Label Classification

Evaluation On FFoCat Data

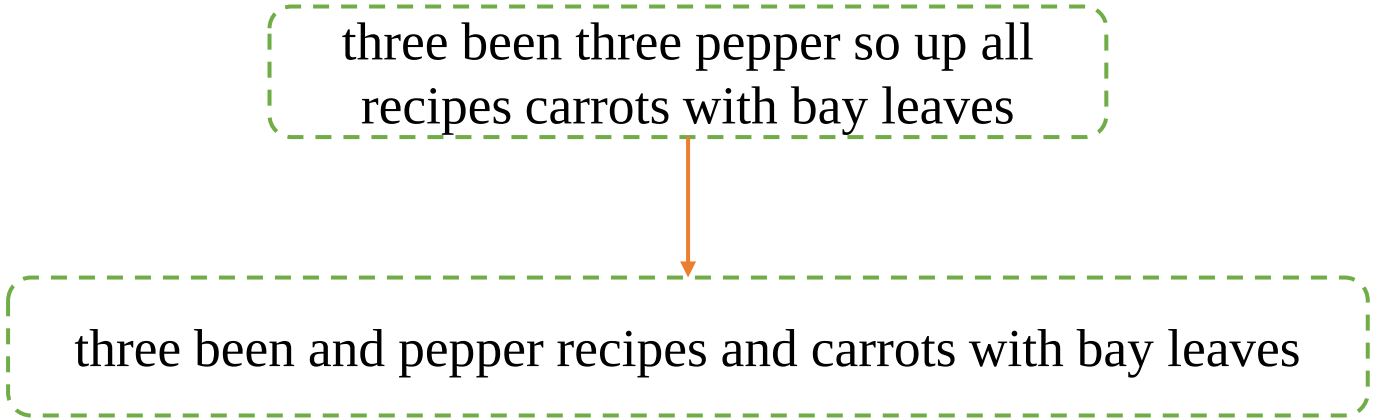
Generally it’s a classification dataset. For caption generation we need BLEU as a performance measure

Image	Label Prediction
	Given Label: vegetable loaf Predicted Label: green bean loaf

- BLEU of 0.52 is achieved.

Natural Language Generator

- Google’s T5 is a Text-To-Text Transfer Transformer.
- Pretrained on the data Colossal Clean Crawled Corpus(C4).
- Finetuning on Yummly28k food data.



Evaluation

On Yummly28k data

Image	Caption Prediction
	Given Caption: beet and carrot slaw martha stewart Intermediate Caption: lentil fried rice martha stewart Intermediate Label: Carrots with bay leaves **Generated Caption:** beet and carrot slaw martha stewart

Evaluation

On Yummly28k data

Image	Caption Prediction
	Given Caption: broccoli salad tasteofhome Intermediate Caption: mexican salad tasteofhome Intermediate Label: Gratineed broccoli **Generated Caption:** broccoli rabe tart tasteofhome

Evaluation

On Yummly28k data

Image	Caption Prediction
	Given Caption: fried chicken and cole slaw sandwiches Intermediate Caption: fried chicken and pen slaw sandwiches Intermediate Label: Chickpeas Hamburger **Generated Caption:** fried chicken and cole slaw sandwiches

Evaluation

On Yummly28k data

Image	Caption Prediction
	Given Caption: chocolate almond pastries martha stewart Intermediate Caption: chocolate almond pies down cake martha stewart Intermediate Label: Radicchio chicory pie **Generated Caption:** chocolate almond upside down cake martha stewart

Performance Evaluation Measures

- Achieved good score in BLEU, METEOR and ROUGE performance evaluation on Yummly28k data.

Metric	Score
BLEU_1	0.68
BLEU_2	0.65
BLEU_3	0.62
BLEU_4	0.60
METEOR	0.40
ROUGE	0.69

Results on Yelp Data

On Uncaptioned YELP data

Image	Caption Prediction
	Intermediate Caption: peanut butter cookies Intermediate Label: Chickpeas Hamburger **Generated Caption:** cowboy cookies i adore food

Results on Yelp Data

On Uncaptioned YELP data

Image	Caption Prediction
	Intermediate Caption: pizza glazed pecans my recipes Intermediate Label: Grilled scamorza cheese **Generated Caption:** pizza glazed pecans my recipes

Results on Yelp Data

On Captioned YELP data

Image	Caption Prediction
	Given Caption: mcdonald's quarter pounder with cheese extra value meal Intermediate Caption: are tandoori mushroom Intermediate Label: Chickpeas Hamburger **Generated Caption:** vegan mushroom burgers

Results on Yelp Data

On Uncaptioned YELP data

Image	Caption Prediction
	Intermediate Caption: soy mexican rice bowl with peanut dressing Intermediate Label: Chili with meat and beans **Generated Caption:** healthy mexican rice bowl with cilantro lime vinaigr

Results on Yelp Data

On Captioned YELP data

Image	Caption Prediction
	Given Caption: wings were Intermediate Caption: blt and butter arugula Intermediate Label: Chicken wings **Generated Caption:** cumin spiced roasted chicken with almonds and raisins

Results on Yelp Data

On Captioned YELP data

Image	Caption Prediction
	Given Caption: my wife had the benedict Intermediate Caption: baked lunch with ricotta side Intermediate Label: Cauliflower with cream **Generated Caption:** baked new potatoes with ricotta sour cream

Results on Yelp Data

On Captioned YELP data

Image	Caption Prediction
	Given Caption: my friends to try lao chuan cuisine in vegas we love all the dishes we got that evening i went back again Intermediate Caption: quinoa and cream with ribs wings rib Intermediate Label: Pasta with mussels **Generated Caption:** linguine with clams and mussels epicurious

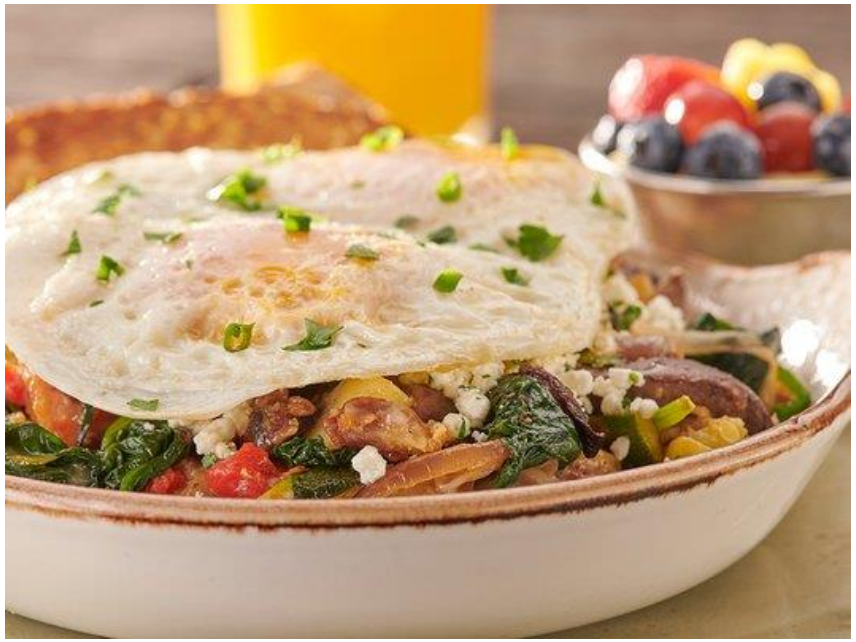
Results on Yelp Data

On Captioned YELP data

Image	Caption Prediction
	Given Caption: mcdonald's quarter pounder with cheese extra value meal Intermediate Caption: are tandoori mushroom Intermediate Label: Chickpeas Hamburger **Generated Caption:** vegan mushroom burgers

Limitations

On YELP data

Image	Caption Prediction
	Intermediate Caption: simple black bean spring Intermediate Label: Chili with meat and beans **Generated Caption:** indian black bean nachos

Limitations

On YELP data

Image	Caption Prediction
	Intermediate Caption: indian spiced tacos with grilled shrimp Intermediate Label: Vegetable strudel **Generated Caption:** indian spiced cauliflower omelet with grilled shrimp

Limitations

On YELP data

Image	Caption Prediction
	Intermediate Caption: Broccoletti in oil Intermediate Label: Spicy pork with ginger cream **Generated Caption:** thai chicken coconut broccoli and coriander

Limitations

On YELP data

Image	Caption Prediction
	Intermediate Caption: Mexican pork loin rolls Intermediate Label: marshmallow cabbage **Generated Caption:** marshmallow cabbage slaw martha stewart

Conclusions

- We have proposed a novel approach for caption prediction on food images.
- It is composed of CNN-LSTM, a Multi-Label classifier and an NLG model for predicting captions from an input food image.
- The proposed model was able to achieve a BLEU score of 0.68 on the Yummly28k dataset and is performing reasonably good on uncaptioned food images.
- In the future semantic features of the image can be considered to generate better caption.

References

[1] Baig, Muhammad & Shah, Mian & Wajahat, Muhammad & Zafar, Nauman & Arif, Omar. (2018). “Image Caption Generator with Novel Object Injection”.

[2] Kelvin Xu, Jimmy Lei, Ba Ryan Kiros , Kyunghyun Cho ,Aaron Courville , Ruslan Salakhutdinov ,Richard S. Zemel ,Yoshua Bengio “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention ”(ICML 2015)

[3] Yao, Ting & Pan, Yingwei & Li, Yehao & Mei, Tao. (2018). “Exploring Visual Relationship for Image Captioning.”

[4] Gan, Zhe & Chen, Yen-Chun & Li, Linjie & Zhu, Chen & Cheng, Yu & Liu, Jingjing. (2020). Large-Scale Adversarial Training for Vision-and-Language Representation Learning.

[5] Salvador, Amaia & Hynes, Nicholas & Aytar, Yusuf & Marín, Javier & Ofli, Ferda & Weber, Ingmar & Torralba, Antonio. (2017). Learning Cross-Modal Embeddings for Cooking Recipes and Food Images.

[6] Lei Meng, Long Chen, Xun Yang, Dacheng Tao, Hanwang Zhang, Chunyan Miao, and Tat-Seng Chua. 2019. “Learning Using Privileged Information for Food Recognition.” In Proceedings of the 27th ACM International Conference on Multimedia (MM ’19), October 21–25, 2019, Nice, France.

References

[7] Yuan, Xinpan & Liu, Qunfeng & Long, Jun & Hu, Lei . (2019). Deep Image Similarity Measurement Based on the Improved Triplet Network with Spatial Pyramid Pooling. Information.

[8] Donadello, Ivan & Dragoni, Mauro. (2019). “Ontology-Driven Food Category Classification in Images.”

[9] Peng, Yuxin & Qi, Jinwei. (2018). Show and Tell in the Loop: Cross-Modal Circular Correlation Learning. IEEE Transactions on Multimedia.

[10] Y. Cui, G. Yang, A. Veit, X. Huang and S. Belongie, "Learning to Evaluate Image Captioning," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, CVPR.2018

[11] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[12] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text Summarization Branches Out (2004).

[13] Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005.

[14] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A Survey on Food Computing. ACM Comput. Surv. 1, 1(July 2019).

Thank You