Applied Deep Learning Project

Team 20

April 1st, 2022

Image Captioning

Fares Feki, Mohamed Issa, Juhyun Kim, Rihem Mansri

April 1st, 2022

Team 20

4 Layers of Stonks

M2 Data Science
IP Paris

Team Member 1

Fares FEKI

Team Member 2

Mohamed ISSA

Team Member 3

Juhyun KIM

Team Member 4

Rihem MANSRI

Table of contents

Introduction

- **Introducing the VizWiz dataset**

  VizWiz dataset includes images that were taken by visually impaired people.

  The VizWiz-Captions dataset includes:
  - 23,431 training images
  - 117,155 training captions
  - 7,750 validation images
  - 38,750 validation captions
  - 8,000 test images with no given captions.

Image Captioning?

- **What is Image Captioning?**

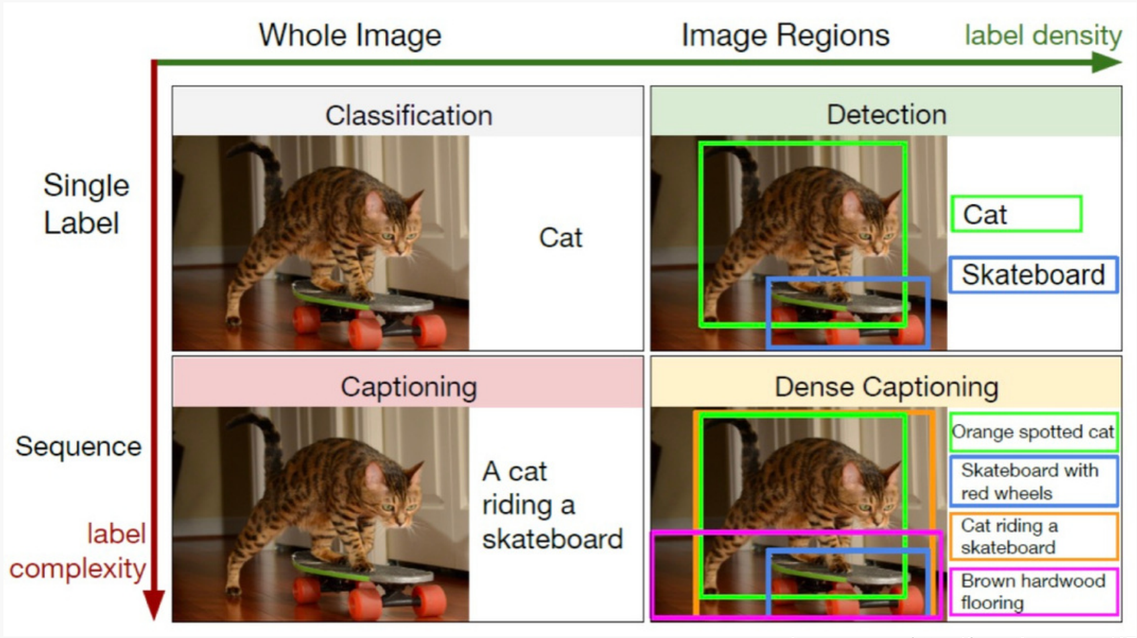


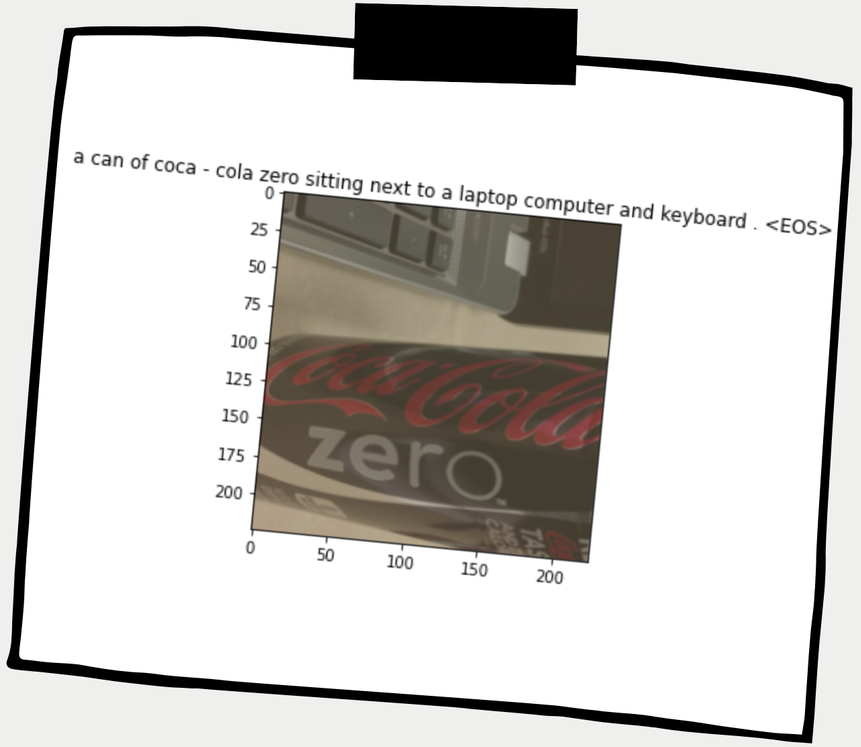
image source: https://bsnn.tistory.com/45

- Literally putting captions or short descriptions for images
- Input: Image
- Output: Text sentence

Image Captioning?

- **Why do we need Image Captioning?**

  - Visually impaired people have relied on (human-based) image captioning services.
  - Virtual assistance may be very useful to accomplish daily tasks.
  - Can be used also by police to get the registration number of vehicles prohibiting rules.



Before we dive in...

Vocabulary class

- **Constructed with captions corpus of our training dataset**
  - Dictionary to go from words to numerical values
  - Dictionary to go from numerical values to words
- **Text preprocessing**
  - Put words in lower case
- **Kept only the most frequent words in the corpus using a frequency threshold**

Dataset class

- **Ease loading of the dataset.**

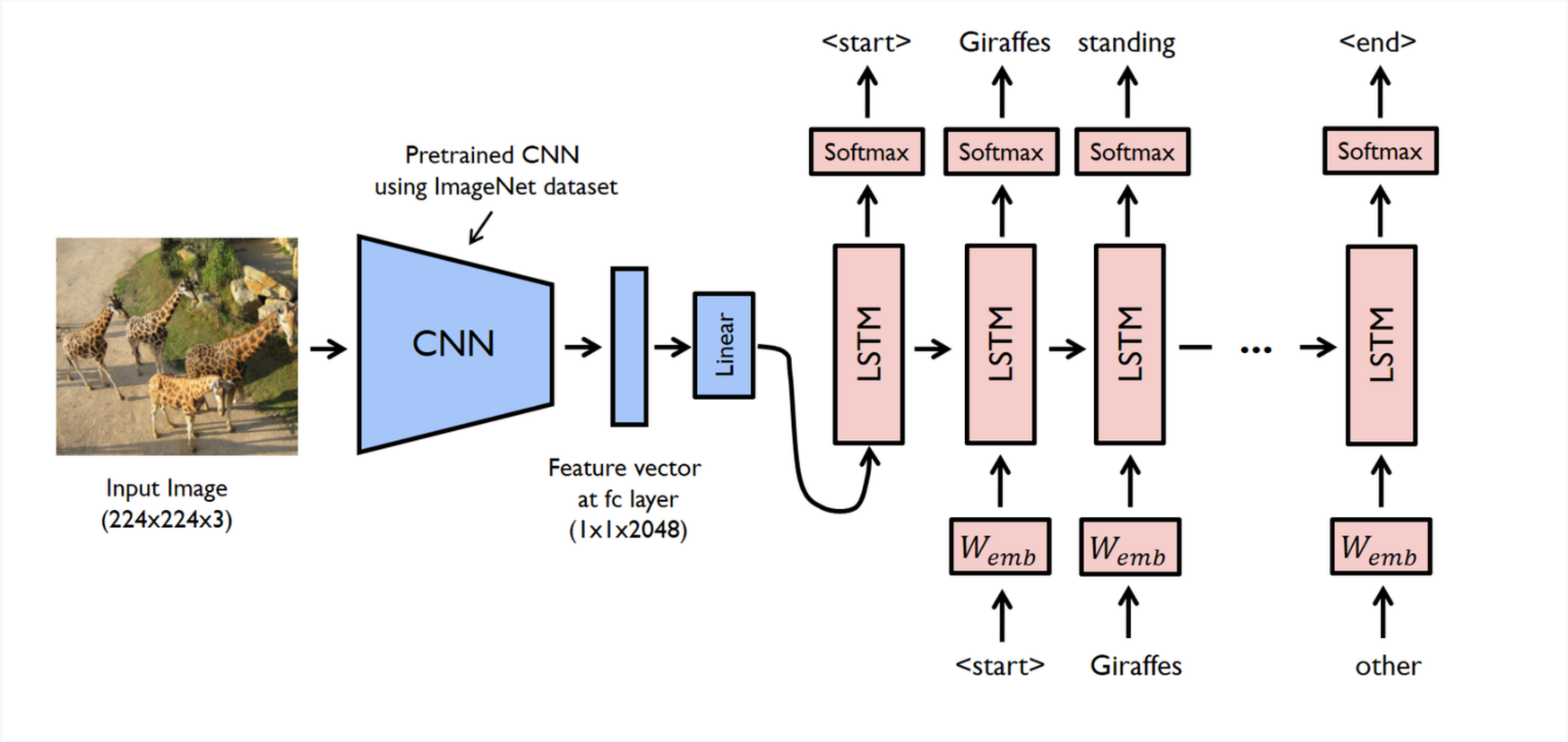
Pad batch class

- **Pad captions per batch**

+

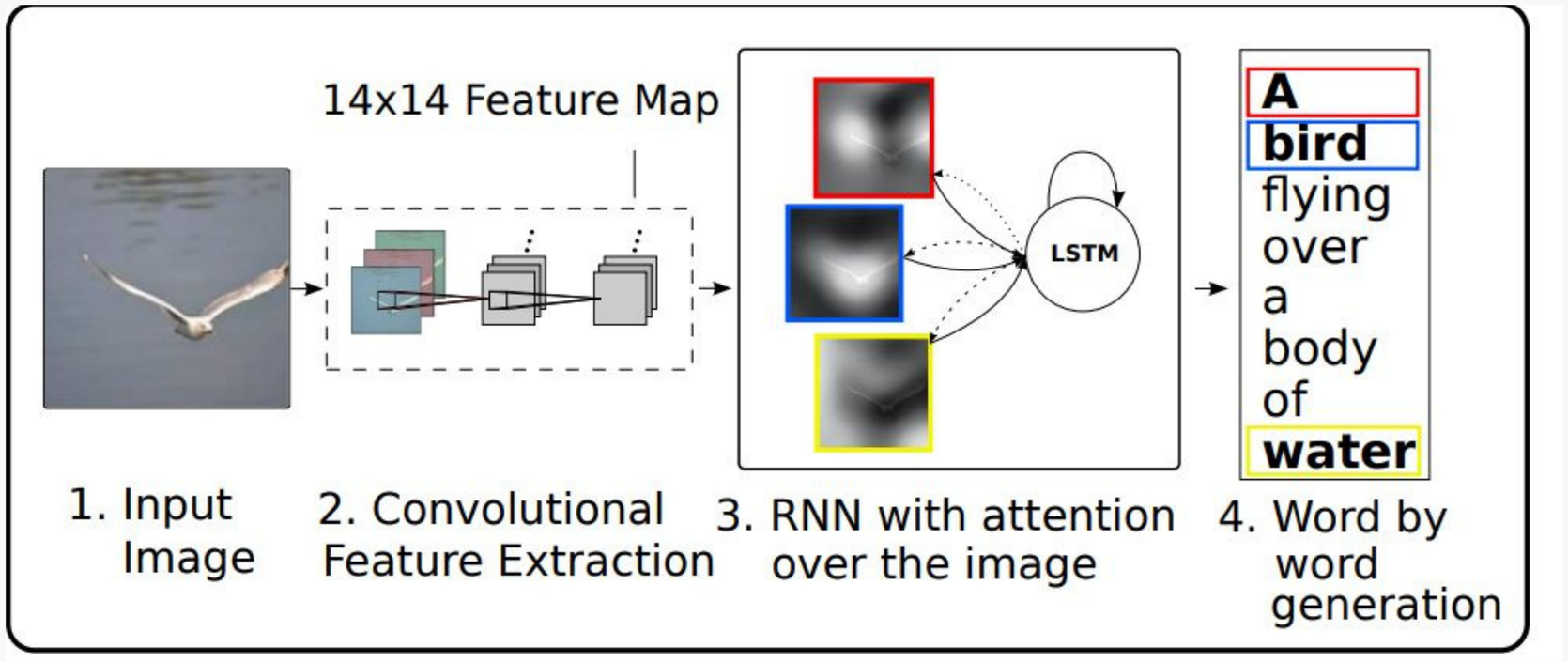
- **Apply transformations on the training images**
  - **Data augmentation technique:** Random cropping
- **Use Dataloader**
  - Enable creating batches for training
  - Shuffle parameter

Our Model



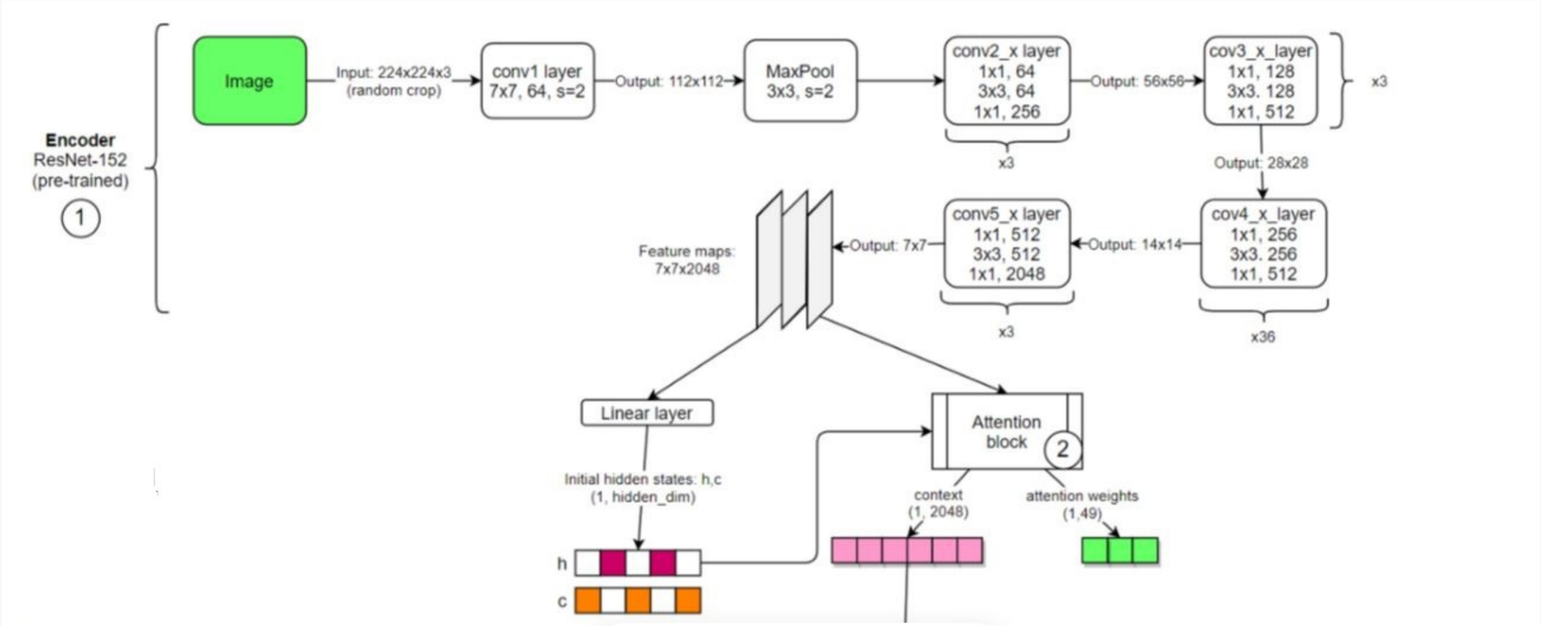
source: http://shikib.com/captioning.html

Our Model



source: Show, Attend and Tell: Neural Image CaptionGeneration with Visual Attention

Our Model

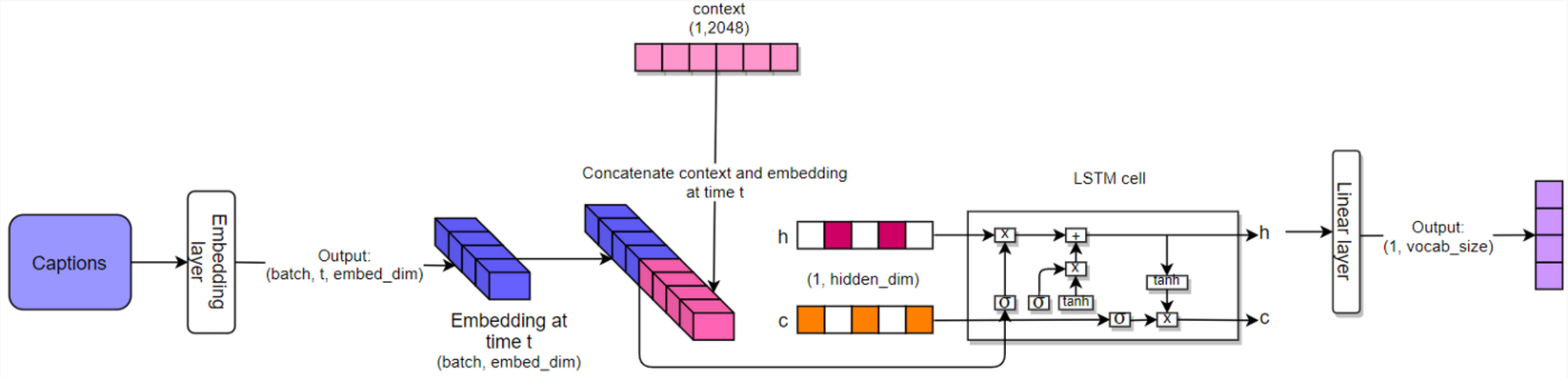


source: https://medium.com/analytics-vidhya/image-captioning-with-attention-part-1-e8a5f783f6d3

Our captioning Model is a **Seq2Seq model**

- The **encoder** uses a **pre-trained CNN** (ResNet model) to **extract the features**
- We used an implementation of the **Bahdanau Attention Decoder**
  - The **decoder** is composed of **LSTM**
  - Use **attention mechanism** between the feature maps produced by the encoder and the decoder's hidden states

Our Model



source: https://medium.com/analytics-vidhya/image-captioning-with-attention-part-1-e8a5f783f6d3

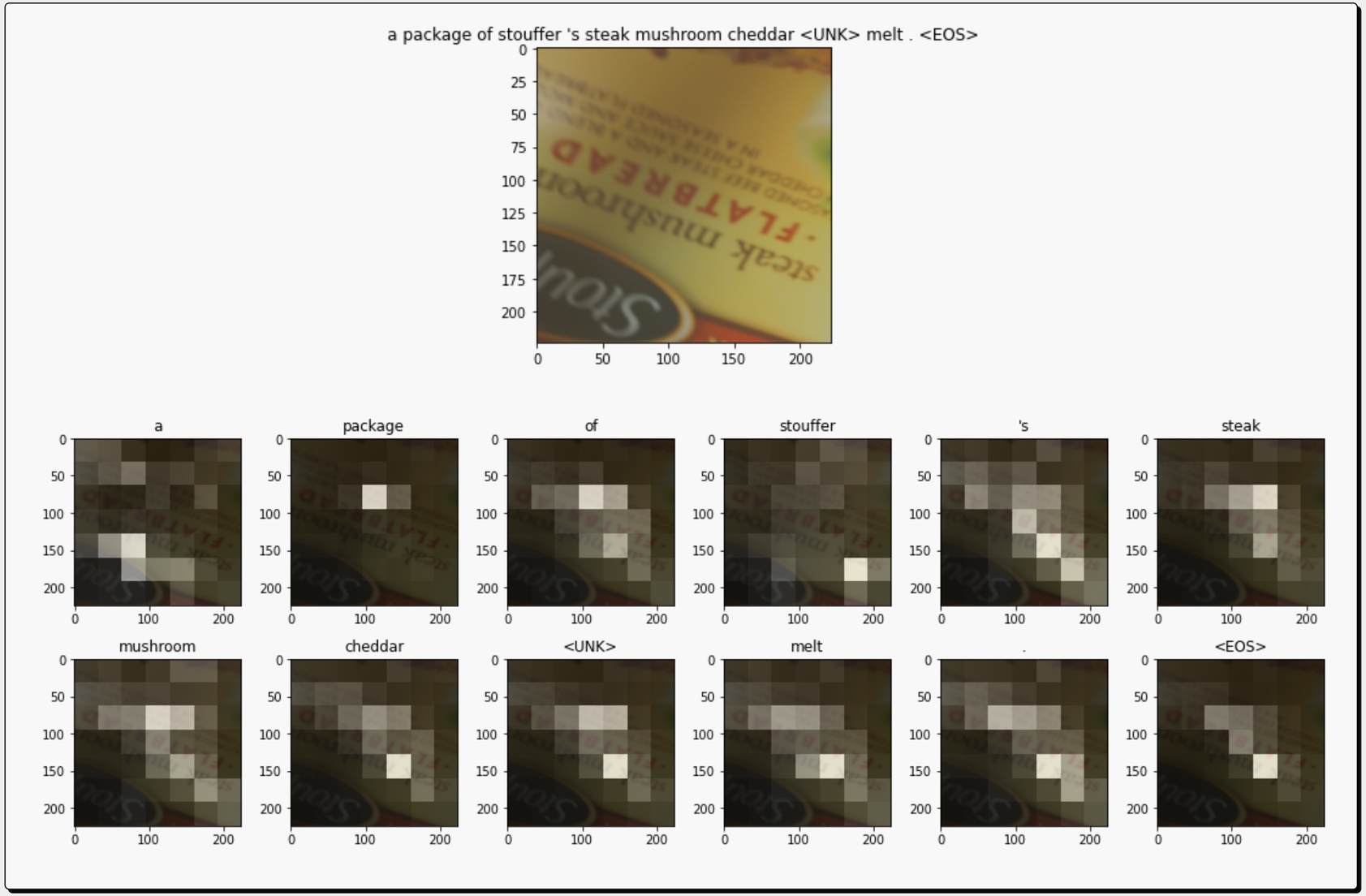
Our captioning Model is a **Seq2Seq model**

- The **encoder** uses a **pre-trained CNN** (ResNet model) to **extract the features**
- We used an implementation of the **Bahdanau Attention Decoder**
  - The **decoder** is composed of **LSTM**
  - Use **attention mechanism** between the feature maps produced by the encoder and the decoder's hidden states

Training

- **Use the VizWiz validation set to train the model**
- **Load trained model weights with several epochs: 20, 60, 80, and 100**
- **Use helper functions for visualization:**
  - Generate captions and attention scores of the given image
  - Plot the attention scores of the image

Example on the training set:



Testing

- **Take a subset of the VizWiz train set (3197 images)**

Example on the training set:

Model prediction: a bottle of lotion is on the table .

"Blue score: 0.642

Predicted result

Calculated BLEU score

Test image sample to predict the captions

Actual captions

-The back of a lotion bottle from the company Nivea.

-The back of a bottle of cream with information about the product.

-Back of bottle of Nivea moisturizing body wash

-A white plastic bottle with yellow and blue lettering is lying on it's side on a wood table.

-A bottle of Nivea is laying on its side on a brown surface.

Evaluation Metric: BLEU score

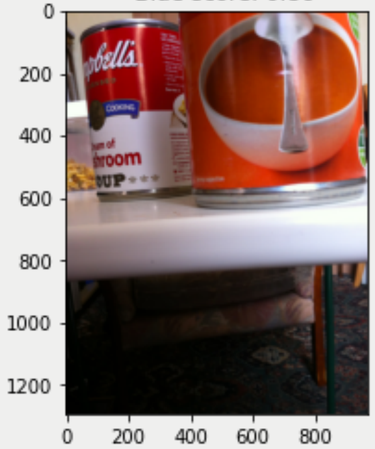
- BLEU-score: imported from nltk
- Mostly used for evaluating machine translation but can also be used for evaluating text generation tasks
- It can compare a few reference sentences and one candidate sentence
- It calculates the ratio of matching n-grams where 1-gram is a token and takes into consideration the occurence of the words.
- It ranges between 1.0 (perfect score) and 0.0 (complete mismatch)

Disadvantage : It doesn't account for the order of words in the sentence.

We take the **average** of the BLEU scores of the test samples

Results

Model prediction: a can of campbell 's soup and a can of soup on a counter .
"Blue score: 0.86



-two cans of tomato soup on a table top

-A can of mushroom soup next to a can of tomato soup.

-A can of tomato soup and a can of cream of mushroom soup sitting on top of a counter.

-Two cans in view, back one is Campbell's cream of mushroom, front one is tomato soup but cannot see brand.

Model prediction: a package of hazelnut milk is on a counter .
"Blue score: 0.627

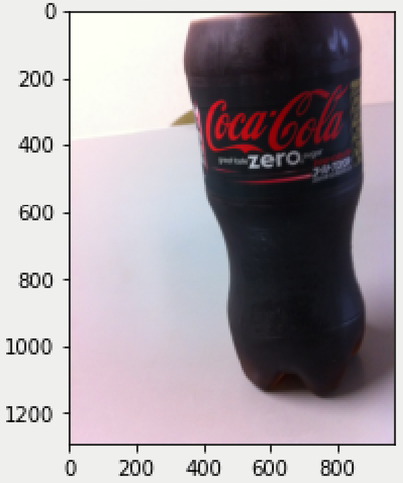
-A package of a Hazelnut and chocolate flavored coffee ground

-A container of chocolate hazelnut beverage with green and gold and brown colored packaging.

-A box of Hazelnut Chocolate sits in a dark room.

-a box of hazelnut drink standing upright in front of a blue wall

Model prediction: a black bottle of coca - cola zero with a red label .
"Blue score: 0.739



-A full bottle of coca cola zero with English and Japanese lettering on the label.

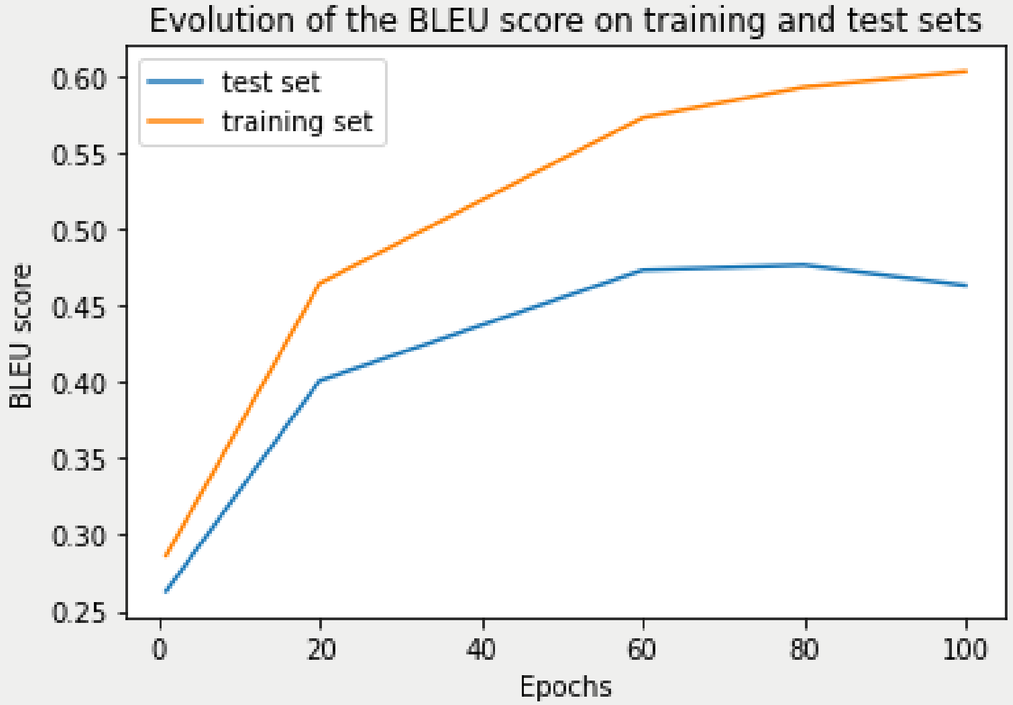
-a bottle of coke zero soda sitting on a grey countertop

-A clear bottle containing a dark colored coke

-A bottle of Coca Cola Zero sitting on a counter.

-I see a full bottle of Coca Cola with zero sugar on a white counter.

Results



- **Lower BLEU score when using the weights trained from more epochs**
  - Comparison between epochs 60, 80, and 100
- **The possibility of our test set is not very representative of our Training set**
  - The test set might contain new samples

Limitations

Model prediction: a gold colored tube with a <UNK> label
"Blue score: 0.348

-This looks like vanilla long lasting body spray.

-A bottle of Vanilla scented long lasting body spray.

-A bottle of vanilla Body Fantasies body spray.

-A long skinny container of vanilla perfume by body fantasies

-A bottle of Body Fantasies Signature, body spray in vanilla, is shown lying on mottled brown and cream carpet.

- **Did not have good results on every test set**
- **Takes too much training time**
- **Less efficient than other architectures**

Future work

- **Apply different text preprocessing and cleaning techniques:**
  - Lemmatization, stemming, removing the non-alphanumeric characters, etc
- **Apply other evaluation metrics:**
  - CIDEr (Consensus-based Image Description Evaluation)
  - ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
- **Use the whole VizWiz train set as the training set of the model**
- **Fine-tuning the pre-trained model could improve the model performance**
- **Try other architectures such as Transformers instead of LSTM**

Applications

Image to Text then to Speech

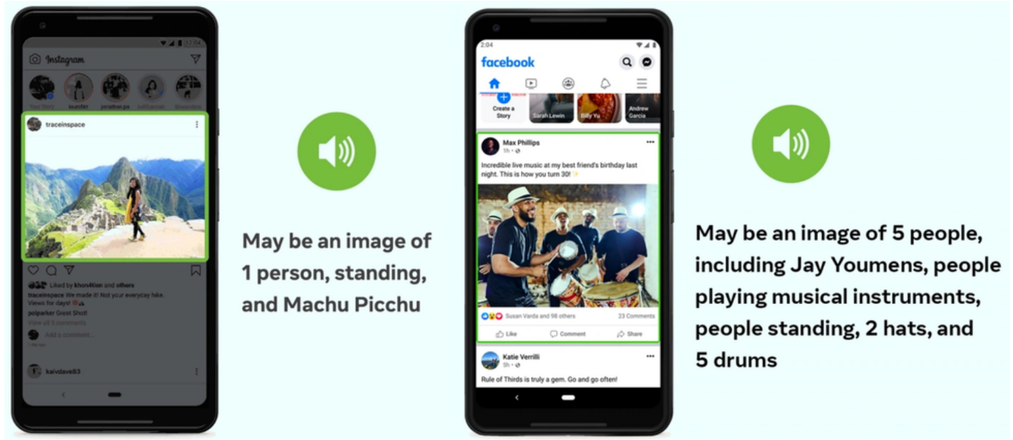


Image source: https://techcrunch.com/2021/01/19/facebook-and-instagrams-ai-generated-image-captions-now-offer-far-more-details/

New function implementation in social media:

1) After posting the post

- Upload image
  → Provide automatically the captions of the image
  → Transform the text to speech
- People who are blind can also "see" the social media posts

2) Before posting the post

- Preview of the image
  → Provide automatically the caption of the image
  → Transform the text to speech
- People who are blind can also "see" the preview of their social media posts before they post it

Applications

AI Art therapy helper

AI Art therapy helper using image captioning to help therapist with patient analysis:

Art therapy not only targets the people who are blind, but also people who need mental relaxation

- Ask the patients to draw or sketch certain objects
- Provide captions of the drawing/sketch
- Send it to the therapist who will be able to analyze the captions

AI Art therapy helper's role could also be extended by giving additional functions other than image captioning, such as providing the percentage of color usage and mood detection

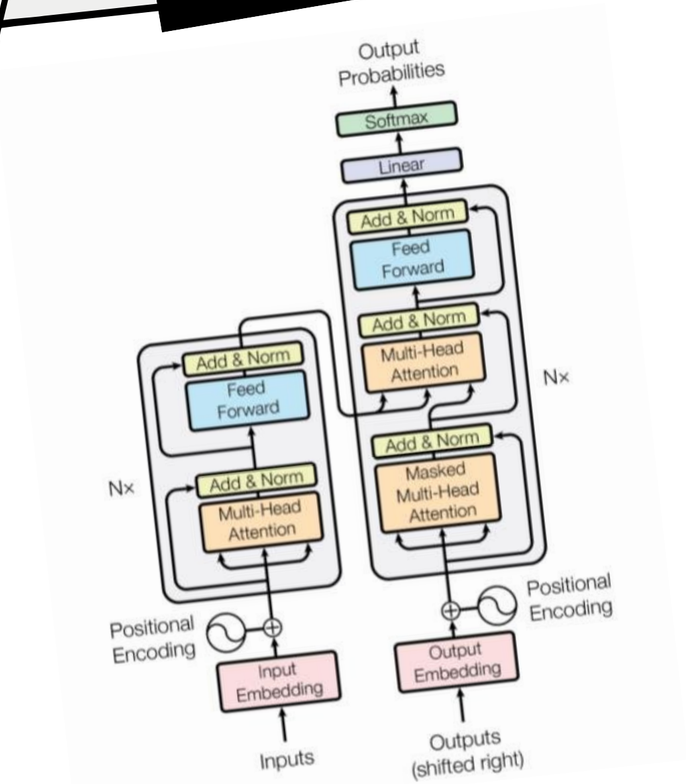
Applied Deep Learning Project

Team 20

April 1st, 2022

Thank you

Fares Feki, Mohamed Issa, Juhyun Kim, Rihem Mansri



for your attention!