"black and white dog jumps over bar."

"man in black shirt is playing guitar."

Team: Furious Four

Topic: Automatic Image Captioning

"girl in pink dress is jumping in air."

"construction worker in orange safety vest is working on road."

Image(s) Credits: https://daniel.lasiman.com/post/image-captioning/

Introduction & Motivation

Image Credits: http://www.haveyougotthatright.com/about-the-project#project

What is Automatic Image Captioning?

*"The process by which a computer system **automatically** assigns metadata in the form of captioning or keywords to a digital image"*

Image Credits: https://www.inovex.de/blog/end-to-end-image-captioning/

Why is Automatic Image Captioning Important?

- ★ Automatic generation of image captions is a task close to the heart of **Scene understanding** - one of the primary goals of computer vision
- ★ **E-Commerce**: Labelling items
- ★ **Web Search/SEO:** Assigning relevant keywords to images can improve the indexing and retrieval results by the search engines
- ★ **Aid for Visually Impaired:** Scene -> Text; Text -> Audio (**Seeing AI -** https://www.youtube.com/watch?v=DybczED-GKE)

Image Credits: https://www.youtube.com/watch?v=33krnU_juFE

It's Not an Easy Problem to Solve !

- ★ Caption generation involves challenges from both **Computer Vision** and **Natural Language Processing (NLP)**
- ★ 1 - Need to determine the objects present in the scene
- ★ 2 - Need to come up with a logically coherent description for the given scene
- ★ Individually difficult problems. Further, we also need to express the **relation between the object attributes and activities**. Was viewed as a difficult/unsolvable problem for a long time.
- ★ However, things are changing

Image Credits: https://www.inovex.de/blog/end-to-end-image-captioning/

Solving the Problem

As part of our project we demonstrate the following:

- ★ An image captioning system built by using the ideas given in Show, Attend and Tell (Xu et al. 2015)
- ★ The main idea being implementation of an attention model on top of the regular NN framework
- ★ The Attention model improves the results by making the dependencies between source and target sequences independent of the in-between distance.
- ★ We obtain good results on the Flick8K dataset

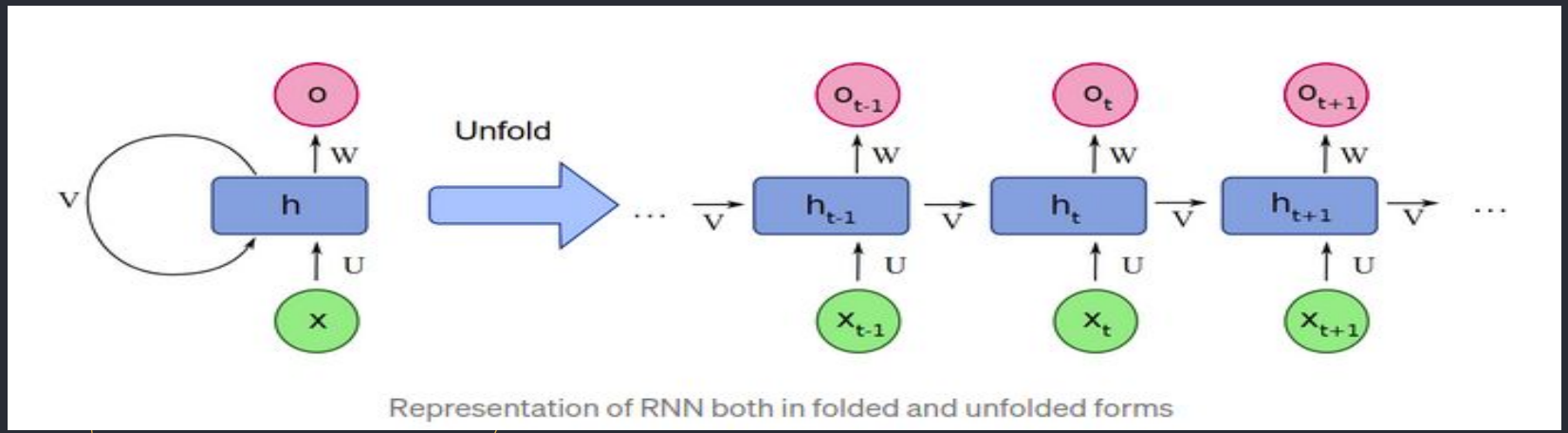
Image Credits: http://www.tsnetsystems.com/solutions/overview.html

Method

Image Credits:
https://www.businessanalystlearnings.com/blog/2015/5/9/4-process-improvement-methods-that-work-when-to-apply-them

Introduction to the Ideas Used

★ Advances in NLP and Machine Translation showed the power of RNNs to solve problems involving sequential data



Representation of RNN both in folded and unfolded forms

Image Credits: https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010

Introduction to the Ideas Used (contd.)

- ★ A typical RNN consists of an encoder RNN and decoder RNN.
- ★ Show and Tell (Vinyals et al. 2014) proposed to replace the **encoder RNN** with a CNN and showed a massive improvement in the results compared to the existing methods

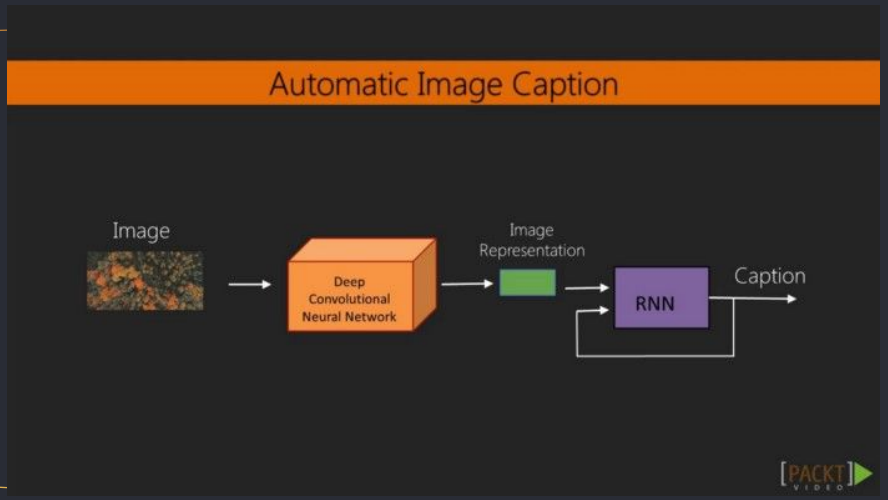


Image Credits: https://medium.com/swlh/image-captioning-in-python-with-keras-870f976e0f18

Introduction to the Ideas Used (contd.)

- ★ Show, Attend and Tell (Xu et al. 2015) further extended upon the ideas given in Show and Tell (Vinyals et al. 2014) and added an attention model. (Another idea borrowed from the field of NLP)
- ★ <What is attention>

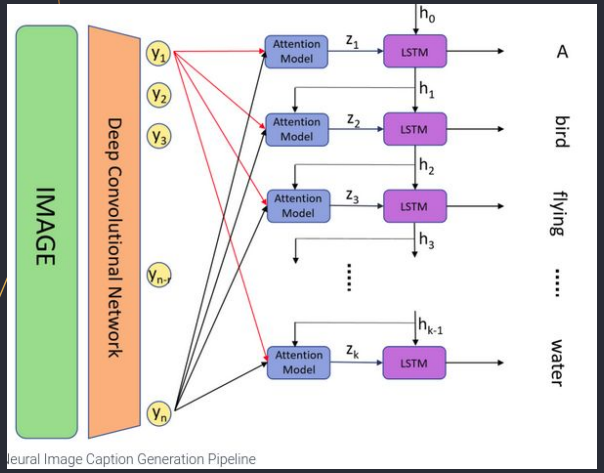


Image Credits: https://iml.dfki.de/active-learning-in-image-captioning/

Introduction to the Ideas Used (contd.)

- The paper proposes two methods:
  - **Soft / Global Attention**
  - **Hard / Local Attention**

- **Soft Attention:**
  - Attending to the entire input state space at once
  - **Pro:** The model is smooth, differentiable and deterministic
  - **Con:** It is expensive when the source input is large

- **Hard Attention:**
  - Attending to a part of input image
  - **Pro:** Less calculation at inference time
  - **Con:** Stochastic; Non-differentiable and requires complicated techniques to process

- For the purpose of the project, we have implemented the **soft attention** part

Overview of our Method (Training Step)

Training:

1. **Pre-processing of data:**
   a. Vocabulary, frequency, caption length, (talk about vector padding)
   b. Add <start> and <end> tags to each caption

2. **Feature extraction:**
   a. Using VGG-16
   b. Reshape Images; Image augmentation

3. Attention + LSTM **(Decoding Part)**

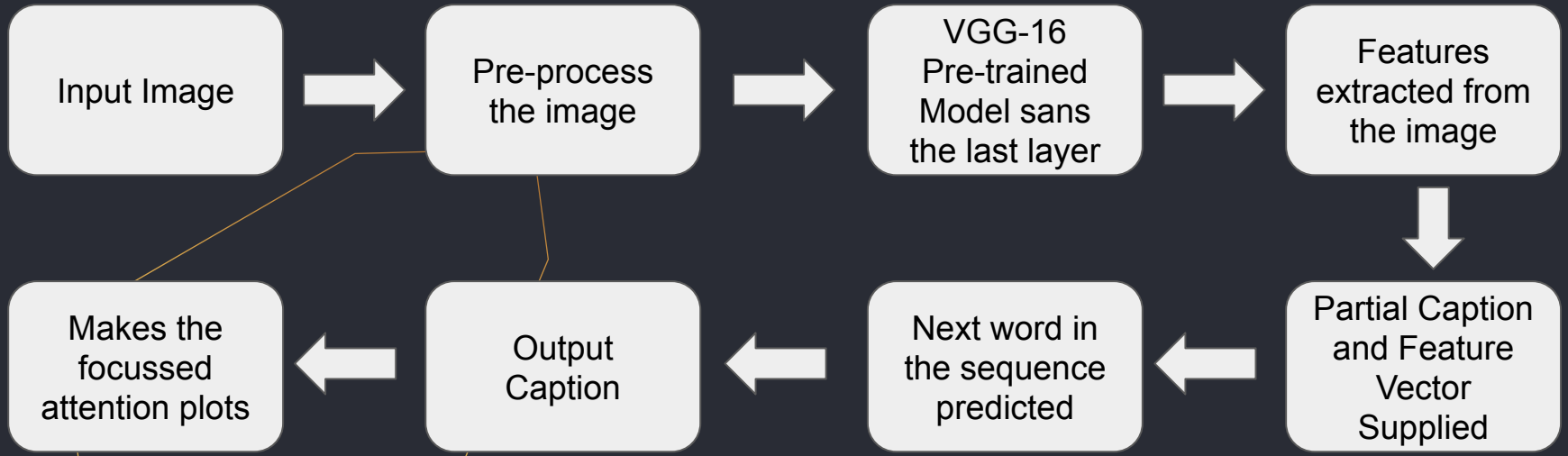
4. **Training** is happening at step-3

Overview of our Method (Prediction Step)

Prediction:

1. Get features of the test image using VGG-16
2. Feed this into our decoder (Attention Model + LSTM model)
3. Make predictions

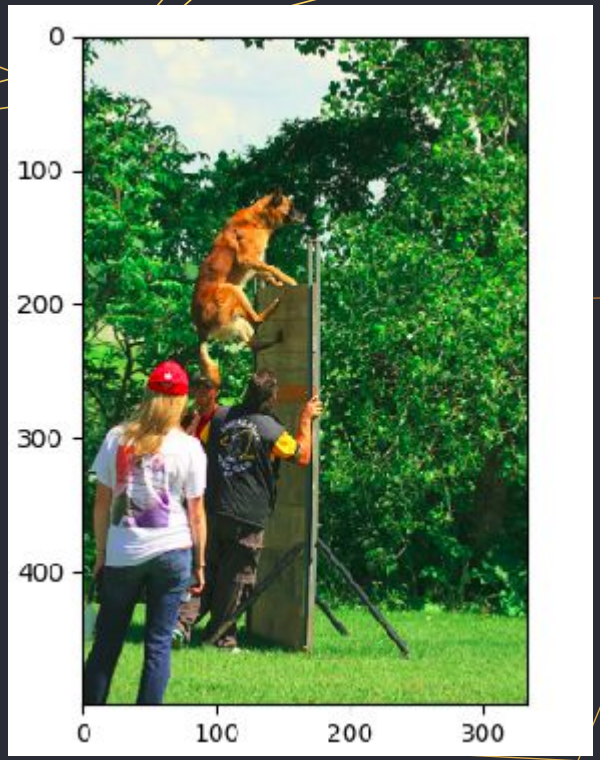
Visual Representation of Prediction Process



RESULTS

Image Credits: https://www.yourtrainingedge.com/powerful-written-goals-in-7-easy-steps/

Earlier Results



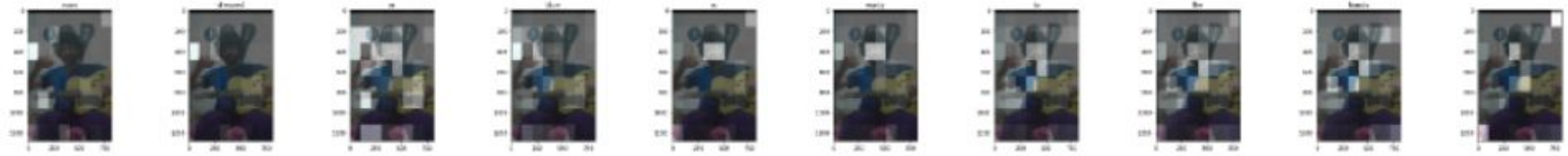
startseq roadwork burning cab nails join automobile roadwork wheeler treadmill roadwork bottoms bottoms ladders roadwork woman sunglasses treadmill hula stripes portrait roadwork woman sunglasses treadmill hula stripes ladders roadwork woman sunglasses treadmill hula stripes portrait roadwork silhouette treadmill roadwork dreadlocks poses nails catcher pickup roadwork workman treadmill paper suitcase roadwork wheeler wheeler ladders roadwork woman sunglasses treadmill hula stripes portrait roadwork silhouette treadmill roadwork dreadlocks poses nails catcher pickup roadwork it ladders roadwork harvesting

Results

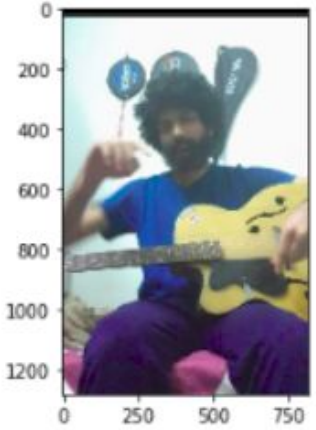
Real Caption: the small golden dog is attempting to take furry object from larger
Prediction Caption: big dog and smaller dog fight over piece of fabric

Results (contd.)

```
Prediction Caption: man dressed in blue is ready to the hands
```

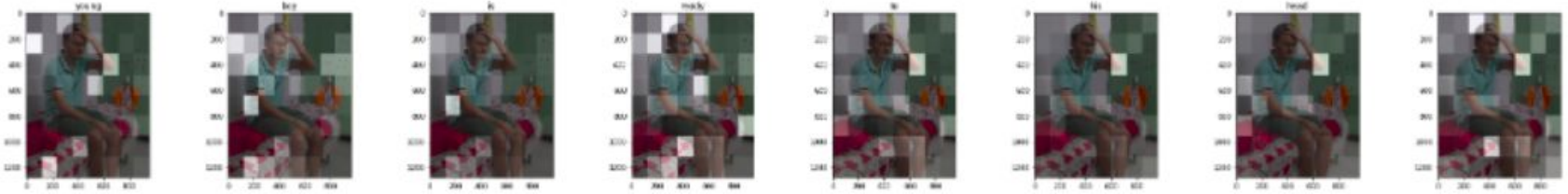


```
: <matplotlib.image.AxesImage at 0x7fd7c18f9c18>
```



Results (contd.)

```
Prediction Caption: young boy is ready to his head
```



```
<matplotlib.image.AxesImage at 0x7fd7c19b0748>
```

TASKS COMPLETED

Image Credits:
https://medium.com/live-your-life-on-purpose/do-our-achievements-mean-anything-59d06fc9e74e

Tasks Completed

- ★ Successfully developed an image captioning system with respectable accuracy. Given an input image, our system can provide a logical caption to it.
- ★ Increased understanding of CV, NLP and Attention-based ML models for all the team members
- ★ Specifically post mid-evals:
  - Corrected regular (non-attention) pipeline
  - Added attention model

Difficulties Faced

Image Credits:
https://www.monsterinsights.com/lesser-known-google-experiments-for-growing-your-business/

Challenges Faced **& Solved** by Us

- **Training Platform:**
  - Options considered - Google Colab, Ada, Kaggle
  - Google Colab Issue - 12 Hrs reset; and Laptop hanging
  - Ada
  - Large datasets
  - **Solution:** Eventually fixed ada
- **Feature Encoding Dictionary:**
  - Some images didn't get saved properly after creating the feature encoding dictionary and led to an error when we later loaded the encodings file
  - **Solution:** Encoding at runtime
- **Training Time:**
  - Huge training times due to redundant calculation of feature vectors.
  - **Solution:** Compute feature vectors just once and store them. Led to 25% reduction in time

Challenges Faced **& Solved** by Us (contd.)

- **Executing Code:**
  - Initial started coding in .py file.
  - Time wastage due to redundant code execution
  - **Solution:** Shifted to .ipynb
- **Dataset:**
  - First used Flickr 30K dataset.
  - Took long time in training.
  - **Solution:** Changed the dataset to Flickr 8K

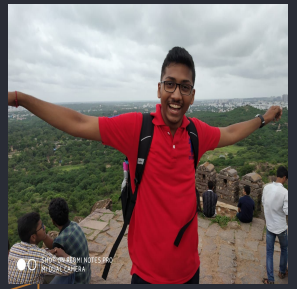
THE TEAM

Image Credits: https://iacsp.org/teamwork-within-the-surveillance-room/

TEAM

ABHISHEK SHAH
2018101052

TIRTH UPADHYAYA
2018101069



SAI TANMAY REDDY CHAKKERA
2018101054

AMOGH TIWARI
2018111003

#Team-FuriousFour

Questions?

Image Credits: https://code.likeagirl.io/lets-play-20-questions-cf91e193025

Thank You!

Image Credits: https://www.123rf.com/photo_28453044_stock-vector-thank-you-note-with-smiley.html

Presentation Credits: This presentation template was borrowed from SlidesGo, including icons by Flaticon, images and infographics by Freepik

References

Image Credits: https://resumegenius.com/blog/resume-help/references-on-resume

References

- https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html
- https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010
- https://www.coursera.org/lecture/language-processing/sequence-to-sequence-learning-one-size-fits-all-4z2ox
- https://www.coursera.org/lecture/nlp-sequence-models/recurrent-neural-network-model-ftkzt
- https://www.coursera.org/lecture/nlp-sequence-models/long-short-term-memory-lstm-KXoay