"black and white dog jumps over bar."

IMAGE CAPTIONING PROJECT

Team: Furious Four

"man in black shirt is playing guitar."

"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/

ABOUT THE PROJECT

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

ABOUT

- ★ Automatic generation of image captions is a task close to the heart of scene understanding - one of the primary goals of computer vision
- ★ Caption generation involves challenges from both computer vision (determining the objects present in the scene) and NLP (expressing the information in a natural language) and hence has been viewed as a difficult problem for long
- ★ Recently - surge in interest in solving this problem
- ★ We refer to: Show, Attend and Tell

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

OBJECTIVES OF OUR PROJECT

Image Credits: https://businessfirstfamily.com/write-business-objectives-results/

OBJECTIVES

- ★ Develop an image captioning system - Given an input image, our system should be able to provide a logical caption to it.
- ★ If possible, add some novelty to the given method and improve upon it further
- ★ Make the implementation publically available for use by research community (after the course is over)
- ★ Increased understanding of CV, NLP and Attention-based ML models for all the team members

METHOD OVERVIEW

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

METHOD-I: GENERAL OVERVIEW

- ★ The given paper uses a **attention model** (sequence to sequence model). It describes an analogy between it's encoder-decoder to machine translation systems, as image captioning can be considered an image to language translation problem
- ★ The paper implements **two** attention based **models**
- ★ One - A ***deterministic* soft** attention model which inputs a revised latent space vector representation of the image into the decoder.
- ★ The other - A **stochastic hard** attention model which inputs the latent space vector representation of a singular location in the image into the the decoder.

METHOD-II: Soft Attention

- ★ This attention mechanism takes the latent space vector representations for various locations all over the image and combines them based on the parameters α_i
- ★ The parameters are computed for each run of the LSTM, as an LSTM produces words one by one. These parameters are then fed to the attention mechanism which produces the input vector for the next LSTM step.
- ★ This method is **end to end differentiable**, hence can be trained as it is in the model.

METHOD-III: Hard Attention

- This attention mechanism takes the latent space vector representations of image locations and outputs a singular vector for a particular location into the decoder. $\Phi(\{\alpha_i\}, \{\alpha_i\})$ is a function that returns a sampled α_i at every point in time based upon a multinoulli distribution parameterized by α

- The parameters are computed for each LSTM cycle and the location in the image is changed accordingly. The method being stochastic in nature, it **can't be differentiated**. Hence a *multinoulli approximation* is made which makes this method differentiable

GOALS OF OUR PROJECT

Image Credits: https://www.workcompass.com/setting-goals/

GOALS

As described in the objectives section, our **OVERALL GOAL** is to make an image captioning system (by referencing Show, Attend and Tell)

Our **SHORT-TERM GOALS** are as described below:

1. **Finalizing** a **topic**, a **reference paper** and gaining a thorough **understanding** of the reference paper (**Done**)
2. **Download relevant datasets:** Flicker 8K, Flicker 30K, Microsoft Common Objects in Context (COCO)
3. **Cleaning and Pre-processing** of the dataset
4. Create a v**ocabulary dictionary** and create **word embeddings**.
5. **Build** the CNN, RNN based **model** based on our reference paper
6. **Training and Testing** of the above model
7. Try to **further improve** upon the model
8. **Training and Testing** of the new model

TIMELINE

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

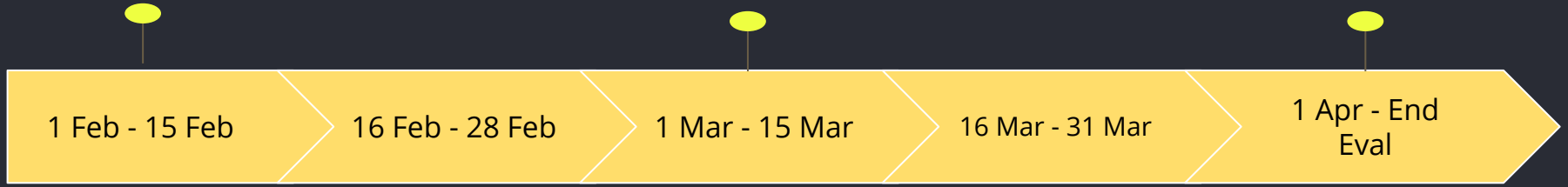
- Finalizing topic, reference paper, understanding the reference paper
- Creating project proposal

- Download and familiarize yourself with relevant datasets
- Perform any required cleaning and pre-processing of the dataset

- Create vocabulary dictionary and word embeddings
- Start working on the CNN, RNN Models

- Complete working on the CNN, RNN models
- Perform training and testing and see initial results

- Fine-tune the models
- Performing training and testing for the full dataset
- Add some novelty if possible



DELIVERABLES

MID-EVAL: Vocabulary dictionary, word embeddings created, along with some initial progress on the CNN, RNN models (Ideally, the CNN part should be completed)

END-EVAL: Demonstration of the complete end to end working of the system

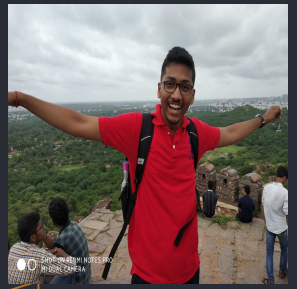
THE TEAM

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

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