# Using Attention-Based Encoder-Decoder Framework to Generate Hashtags and Stories for Instagram Images

SHIVAM GAUR

# Outline

- Introduction
- Motivation
- Methodology
- Experiments
- Results & Analysis
- Conclusion & Future work
- References

# Introduction

## Image Captioning : Computer vision + Natural Language Processing.


Source : Adapted from [1]

# Background

- **Earliest methods [2]:**
  1. Retrieval based
  2. Template based

- **Frameworks based on deep neural networks [2]:**
  1. Combined with early approaches
  2. Multimodal neural networks
  3. Encoder-decoder framework
  4. Attention-based framework
  5. Compositional architectures
  6. Frameworks to deal with novel images

# Objectives

Generate hashtags for Instagram images using attention mechanism

Explore the possibility of using the hashtags to generate narrative caption for the image.

# Why Hashtags?

1. #LOVE
2. #INSTAGOOD
3. #PHOTOOFTHEDAY
4. #FASHION
5. #BEAUTIFUL
6. #HAPPY
7. #LIKE4LIKE
8. #PICOFTHEDAY
9. #ART
10. #PHOTOGRAPHY

# Existing Work : Sentence level captions

Focus on generating image captions in the form of single sentence.

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

Source : Adapted from [3]

# Existing Work : Personalized hashtags

**(GT)** #fashionkids #stylish-cubs #kidzfashion …
**(Ours)** #pink #babygirl **#fashionkids** #cutekidsclub …


**(GT)** #connecticut #books #bookbarn
**(Ours)** **#books** #reading

Source : Adapted from [4]

# Existing Work : Visual storytelling

(a) (b) (c) (d) (e)

**Captions:**

(a) A small boy and a girl are sitting together.
(b) Two kids sitting on a porch with their backpacks on.
(c) Two young kids with backpacks sitting on the porch.
(d) Two young children that are very close to one another.
(e) A boy and a girl smiling at the camera together.

**Story #1:** The brother and sister were ready for the first day of school. They were excited to go to their first day and meet new friends. They told their mom how happy they were. They said they were going to make a lot of new friends . Then they got up and got ready to get in the car .

**Story #2:** The brother did not want to talk to his sister. The siblings made up. They started to talk and smile. Their parents showed up. They were happy to see them.

Source : Adapted from [4]

# Novel approach for visual storytelling

- Can a short story be generated for a single image by using a single keyword instead of a sentence long caption or a sequence of captions?

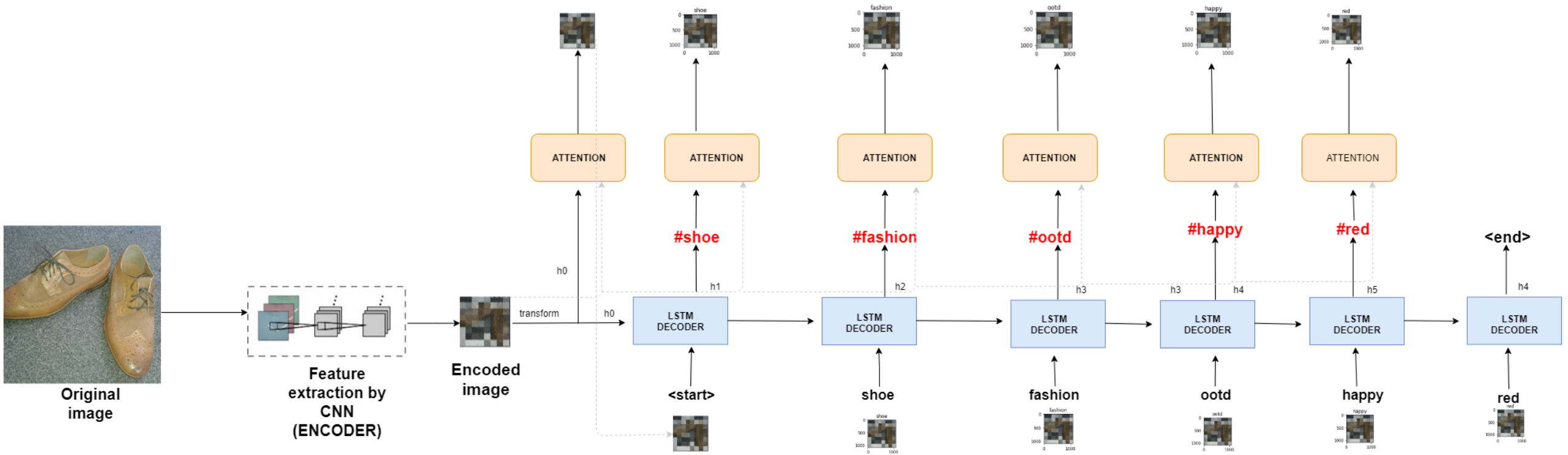
- This work aims to explore this possibility.

# Methodology

Two-phase process

First, generate a hashtag for an input image by using attention model

Second, leverage the hashtag from previous stage to produce a short story by using a character-level language model

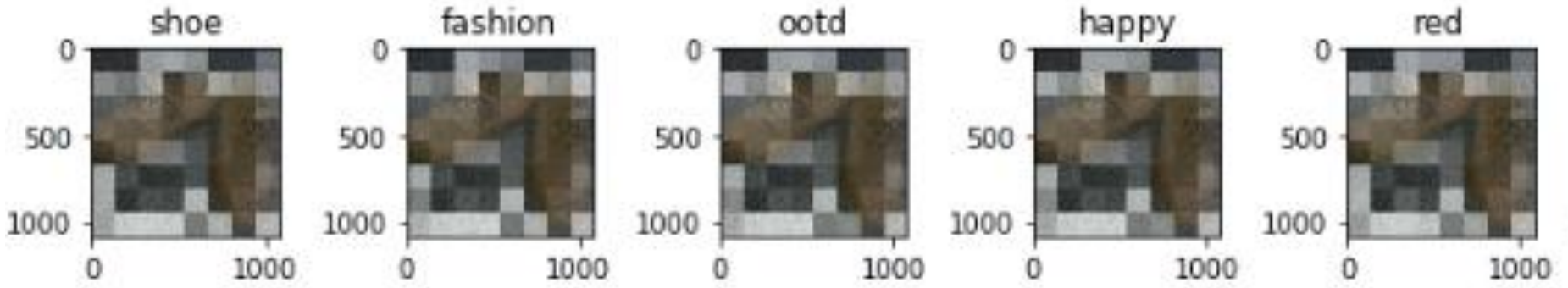



Source : Adapted from [3]

# Phase 1 : Hashtag generation using attention mechanism

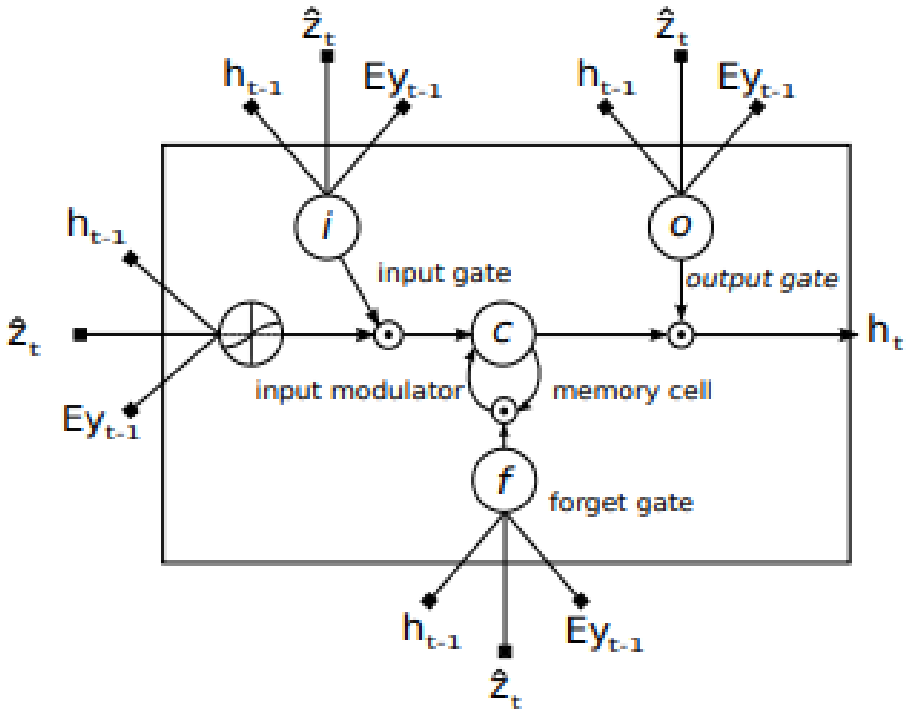
# Hashtag generation using attention mechanism

- Attention mechanism focusses on important features of the image
- The model takes an image I as input and produces a one-hot encoded list of hashtags denoted by X where $|X| >= 1$ and $X = \{x_1, x_2, x_3, x_4\ldots\ldots, x_N\}$, such that $x_i \in R^K$ [3]. K is the size of the vocabulary and N is the number of hashtags generated for the image.



# Encoder-Decoder

- Image features are extracted from lower CNN layers (ENCODER).
- The decoder uses a LSTM that is responsible for producing a hashtag (one word) at each time step t, which is conditioned on a context vector $z_t$, the previous hidden state $h_t$ and the previously generated hashtag [3].




Source : Adapted from [3]

# Context vector - $z_t$

- The set of L feature vectors, also called annotation vectors, can be represented as: A = $\{a_1, a_2, a_3, a_4\ldots\ldots, a_L\}$, such that $a_i \in R^D$ and represents a specific image location [3].
- Soft attention mechanism is used to generate a hashtag.
- Weighted image features are used as input to the LSTM to account for attention.
- In soft attention method, context vector $z_t = \phi(\{a_i\}, \{\alpha_i\})$ where $\phi(\{a_i\}, \{\alpha_i\}) = \sum_{i=1}^{L} \alpha_{t,i}\, a_i$ [3].

and $\alpha_{t,i} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}$

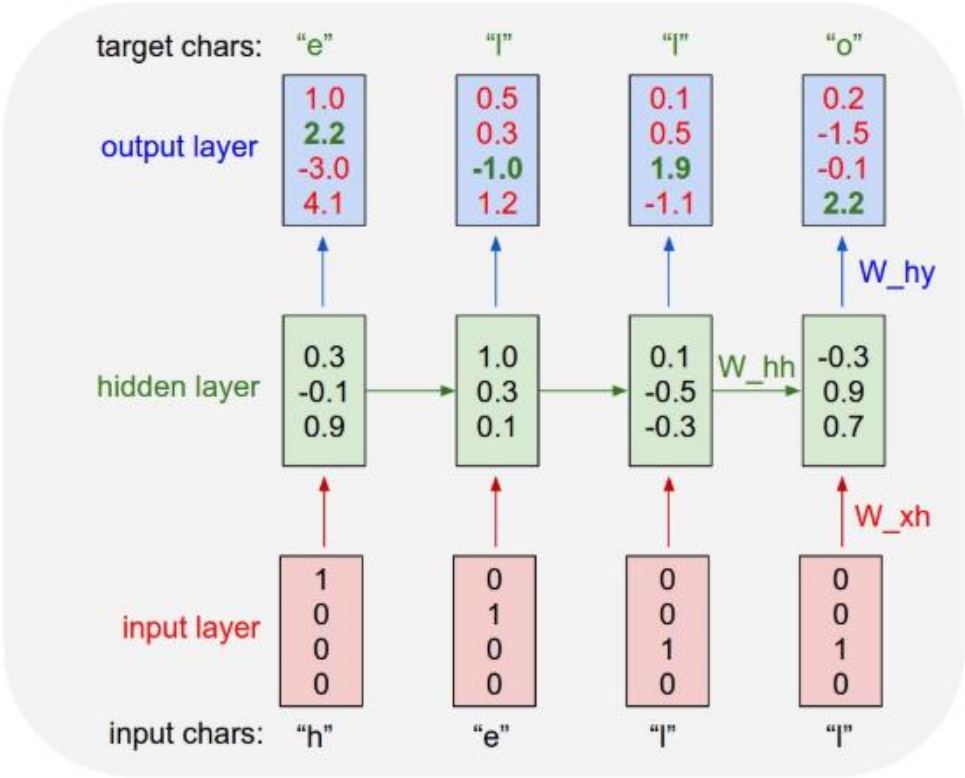
$e_{ti} = f_{att}(a_i, h_{t-1})$

# Soft attention model

- The model is trained end to end with the aim of minimizing following negative log-likelihood function [1].
- $L_d = -\log\left(P(y|x)\right) + \lambda \sum_i^L \left(1 - \sum_t^C \alpha_{ti}\right)^2$

# Phase 2: Story generation

Generating narrative caption for the image using character-level language model

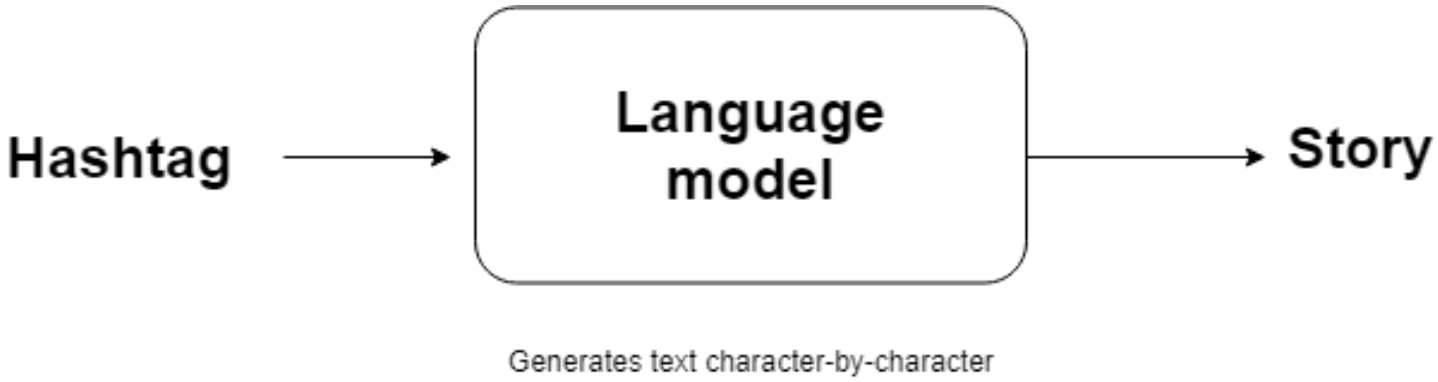



An example RNN with 4-dimensional input and output layers, and a hidden layer of 3 units (neurons). This diagram shows the activations in the forward pass when the RNN is fed the characters "hell" as input. The output layer contains confidences the RNN assigns for the next character (vocabulary is "h,e,l,o"); We want the green numbers to be high and red numbers to be low.

Source : Adapted from [6]

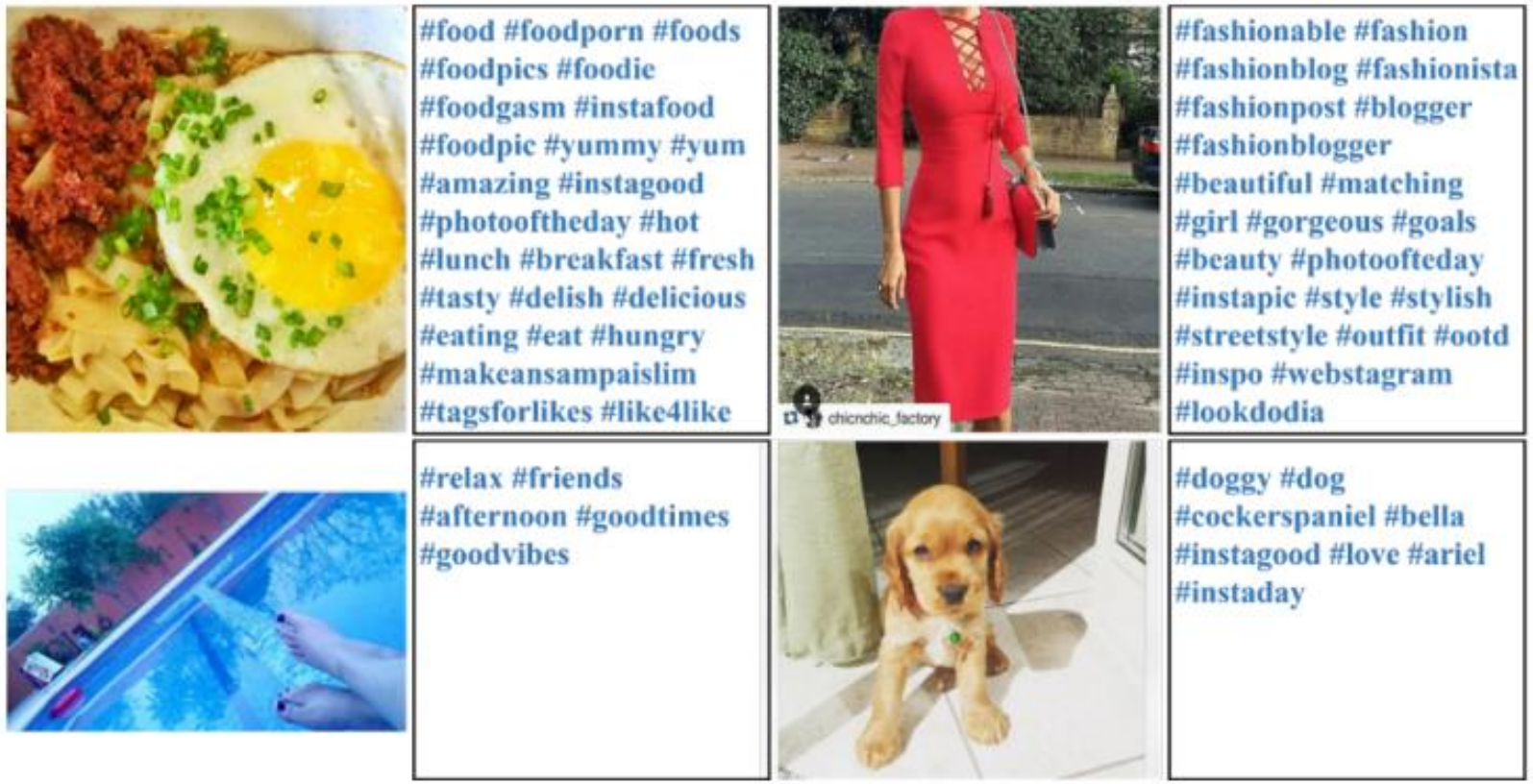
# Story generation

- The RNN models the probability distribution of the characters in sequence given a sequence of previous characters [7].
- The hashtag generated in phase 1 is chosen as seed text and using the character sequences of this seed text, new characters are generated in sequence.
- The model is trained to generate narratives by adopting the writing style in the corpus using the hashtag.



# Experiment setup: Hashtag generation

- The attention-based model is trained on HARRISON dataset.
- HARRISON is an acronym for ‘Hashtag Recommendation for Real-World Images in Social Networks’.
- The dataset consists of 57,383 images and around 260,000 hashtags.

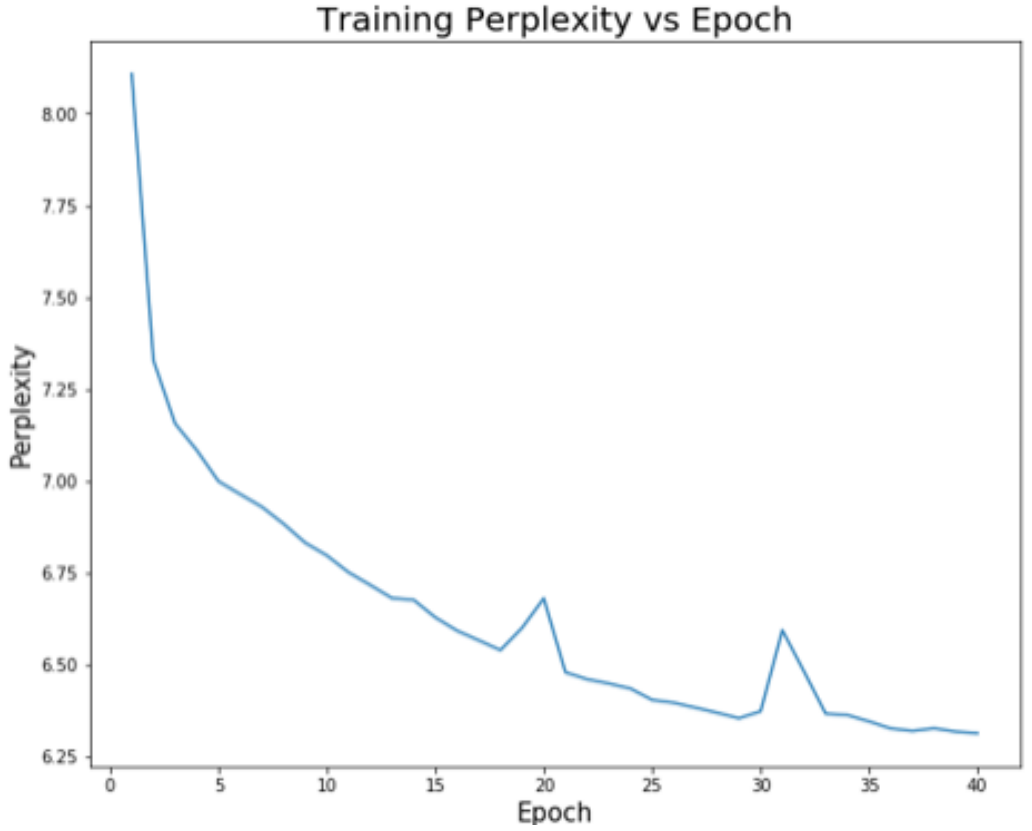



Source : Adapted from [8]

# Training the soft-attention model

- The entire network was trained from end-to-end. InceptionV3 (pretrained on Imagenet) was used to classify images in the HARRISON dataset and features were extracted from the last convolutional layer.
- To generate hashtags, the CNN-LSTM model with embedding dimension size of 256, 512 GRU(LSTM) units and Adam optimizer was trained for 40 epochs on a GEForce GTX Titan GPU with each epoch taking about 2.5 hours.
- The model was trained on 80 percent of data (around 43K images) while the remaining was used for testing.

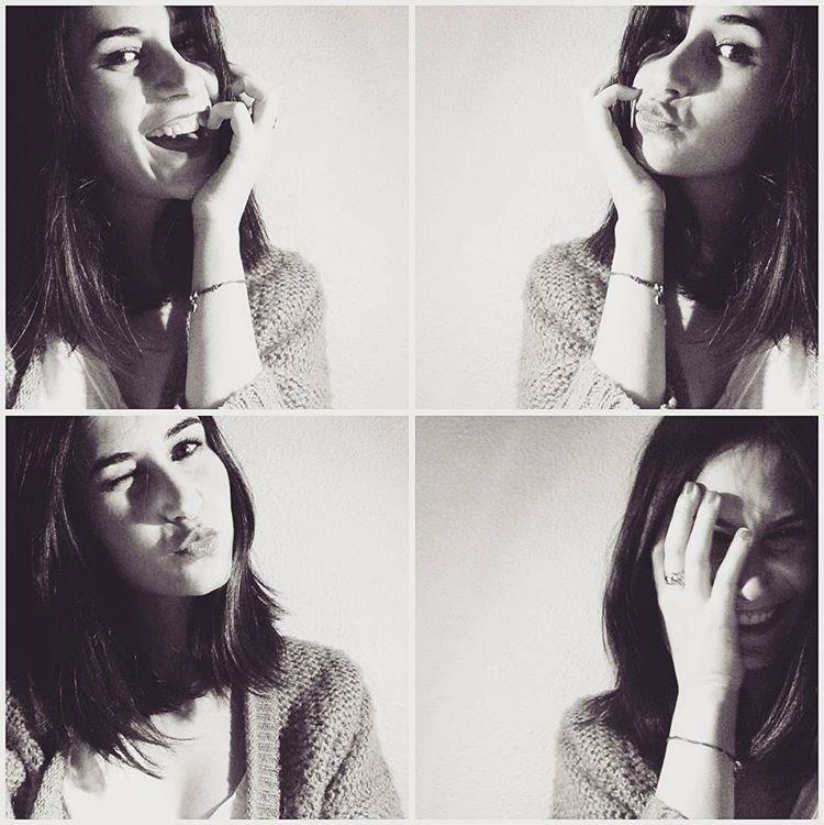
# Training the soft-attention model

# BLEU-1 score evaluation

| Soft-attention model [3] | Multi-label image classification using Alexnet [7] | Show and Tell model for image captioning [1] |
|---|---|---|
| **0.0825** | **0.0047** | **0.1321** |

# Predicted hashtags



Ground truth hashtags:

**#selfie**

Predicted hashtags:

**1. Soft-attention model [3]:**

#selfie (BLEU-1 : 1.0)

**2. Multi-label image classification model [7]:**

#show, #laugh, #cherryblossom (BLEU-1 : 0)

**3. Show and Tell model [1]:**

#girl, #polishgirl, #selfie, #black, #white, #eye

(BLEU-1: 0.0067)

# Predicted hashtags


Ground truth hashtags:

**#nba #lakers #shoot #nike #mambaday #kobe #history #family**

Predicted hashtags:

**1. Soft-attention model [3]:**

#kobebryant, #lakers, #nike (BLEU-1 : 0.25)

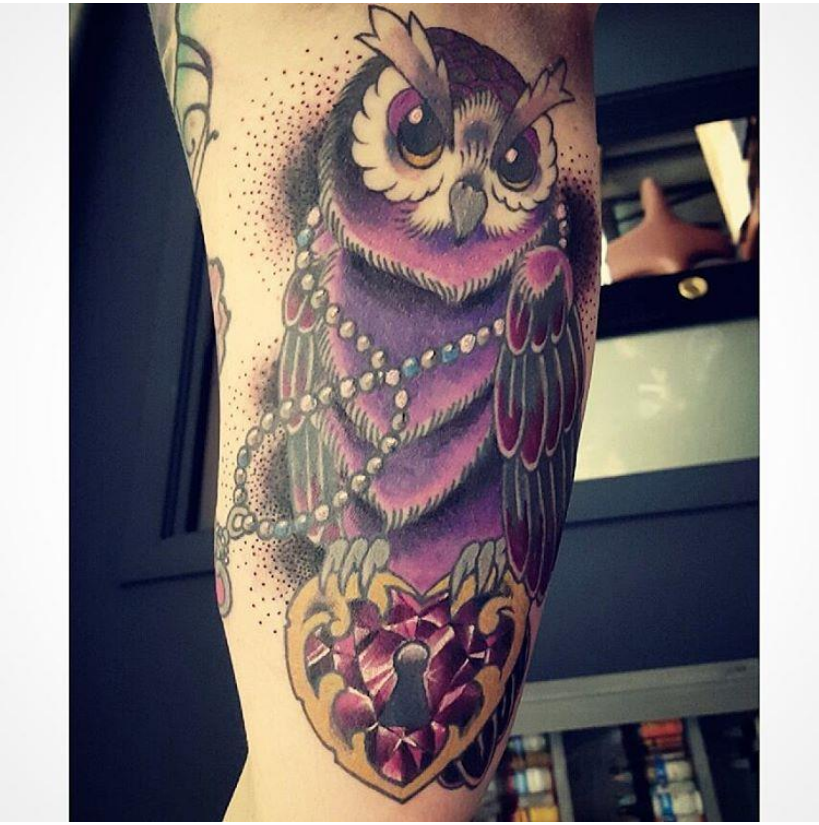
**2. Multi-label image classification model [7]:**

#vintage, #wakeup, #latepost (BLEU-1 : 0)

**3. Show and Tell model [1]:**

#mambaday, #nike (BLEU-1: 0.25)

# Predicted hashtags



Ground truth hashtags:

**#tattoo**

Predicted hashtags:

**1. Soft-attention model [3]:**

#tattoo, #ink, #inked (BLEU-1 : 0.1353)

**2. Multi-label image classification model [7]:**

#genre, #london, #mylove (BLEU-1 : 0)

**3. Show and Tell model [1]:**

#tattoo, #ink, #inked (BLEU-1: 0.1353)

# Predicted hashtags

Ground truth hashtags:

**#river #city #beautiful #old**

Predicted hashtags:

**1. Soft-attention model [3]:**

#sea, #sky (BLEU-1 : 0)

**2. Multi-label image classification model [7]:**

#shadow, #flowerstagram, #cherryblossom (BLEU-1 : 0)

**3. Show and Tell model [1]:**

#Italy, #italia, #travel, #sea (BLEU-1: 0)

# Experiment setup: Story generation

- The character - level RNN model is trained on ‘PersonaBank’ corpus which is a collection of 108 personal narratives from various weblogs [10].

- These stories cover a wide range of topics from romance and wildlife to travel and sports.

- Out of 108 stories, 55 are positive stories while the remaining are negative. Average length of story in the corpus is 269 words.

# Training the character-level language model

- The language model is trained using a standard categorical cross-entropy loss. Let ‘t’ be the vector representing the one hot encoding of the output character and ‘v’ be the vector of probability distribution over ‘n’ characters in the 'PersonaBank' vocabulary, then the loss for single character is equal to $-\sum_{i=1}^{n} t_i \log(v_i)$. The objective is to minimize this loss.
- The language model was trained for 100 epochs with word embedding dimension size of 1024, 2 LSTM layers, softmax activation function, RMSProp optimizer and a learning rate of 0.01on a GEForce GTX Titan GPU to generate stories with 500 characters in length.

# Predicted narrative caption

**Ground truth hashtag: #friend #happy**

**Predicted hashtag: #friend (BLEU-1: 0.5)**

**Seed text for story : #friend**

**Story generated by model:**

*and thanksgiving a bit when I get to the bottle is haircuts span the payment spectrum. Which leads me back 50 bucks. So bad haircuts span the phone with her and we hung out all day and contracting at the corner other day. We both remember that day well now. Jake to talk behind the lid. 'Ccapsual, it was a bit gear to a smile that I could have done a better job. Seriously. Now, I just want my brother and I love him don't all this year. He was that because I look at him everytime he breath.*

# Predicted narrative caption

**Ground truth hashtag: #tired**

**Predicted hashtag: #tired, #selfie (BLEU-1: 0.3678)**

**Seed text for story : #tired**

**Story generated by model:**

*we could when we went to over how the last time I saw his one of our Saturdays for make-up camping trip.  This is the first time I have been camping yesters down at the eyes I gether Jane an  deviant I realized her and his mom. :) It was so nice. But driving Like we were exhausted. John## sat down after dinner, told--limfy when I had dinner this all we'd have to remember to eell him everytiming of! We are always feri*

# Qualitative analysis of stories

- A group of seven people was organized wherein each person was presented with a set of 21 images with associated hashtag and story caption.
- Out of 21 images, 15 belonged to the HARRISON dataset while the remaining were chosen randomly from internet without revealing this to the group.
- Every person in the group had to answer the following three questions for every image and its associated hashtag and narrative caption.

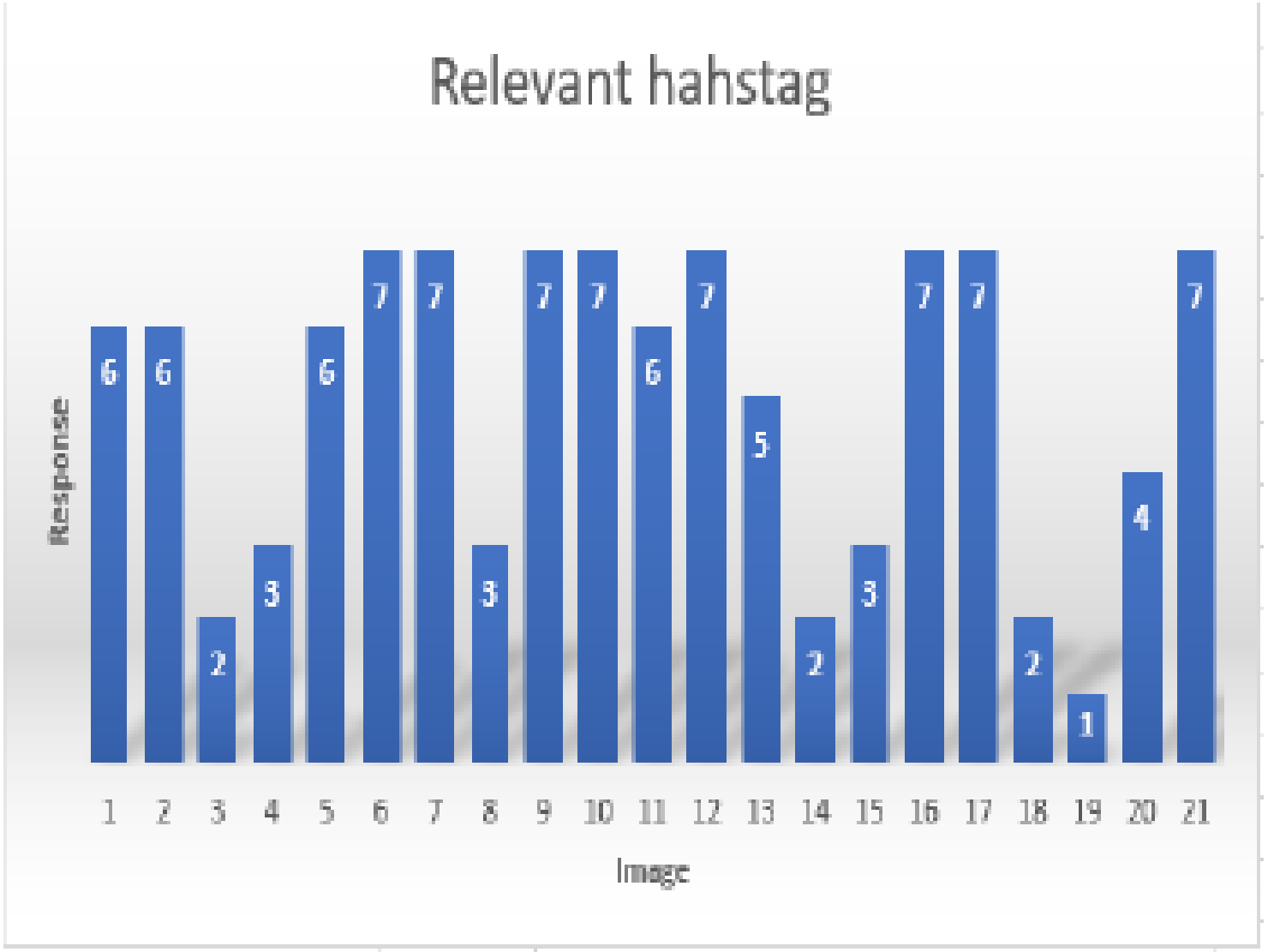
**Q.1. Is the hashtag relevant to the context of the image?**

**Q.2. Does the caption generated have a meaningful context?**

**Q.3. How close is the context of the story to that of the image?**

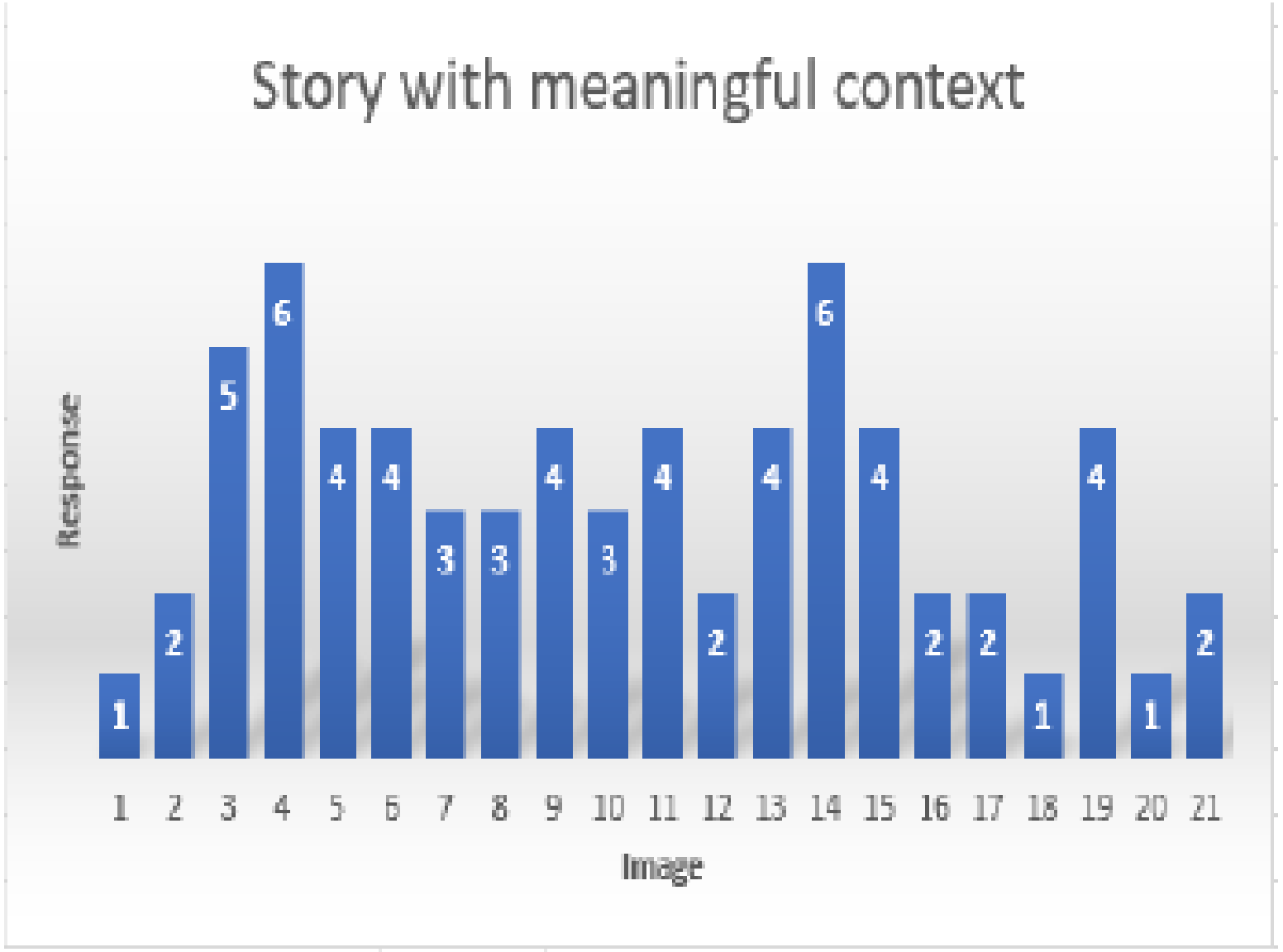
# Qualitative analysis of stories

- For more than 65% of the images, the hashtag generated was meaningful and clearly related to the context of the image.
- A hashtag is considered relevant if more than half of the people in the study group considered it as relevant.



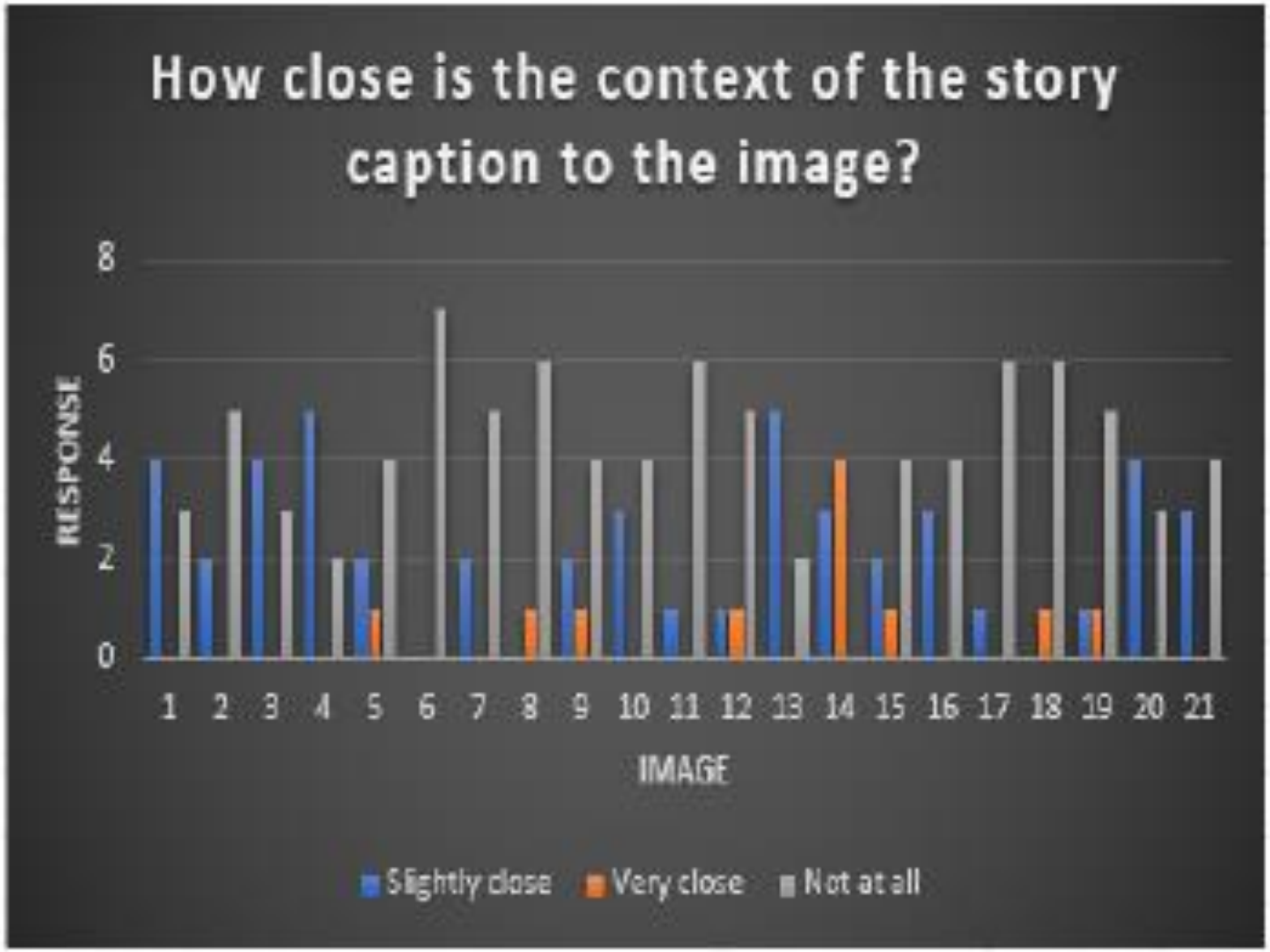
# Qualitative analysis of stories

- For nearly half of the images (47%), story generated as caption had some context.



# Qualitative analysis of stories

- Very few images had captions that could strongly be associated with the image context.
- Nearly 70% of images lacked captions with good context .
- For remaining images, the caption only slightly matched the image context.



# Conclusion & Future Work

- The results obtained from the experiment show that on one hand where an attention model is able to generate meaningful hashtags for the input image and a character-level language model can be exploited to generate short stories with relevant context, it is still a challenging task to match the context of the captions with that of the image.
- A bigger corpus of narratives required to train language model for story generation
- Need to train attention model for more epochs for better performance

# Demo

1. https://www.loom.com/share/a9a07793b2ea40dc84b3e5d282b86ab4
2. https://www.loom.com/share/82562399a48c4bdeb0d258b8b9ad4275

# References


| | |
|---|---|
| **[1]** | **O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in IEEE Conference on Computer Vision and Pattern Recognition, 2015.** |
| **[2]** | S. Bai and S. An, "A survey on automatic image caption generation," Neurocomputing, vol. 311, pp. 291-304, 2018. |
| **[3]** | K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," arXiv.org, 2016. |
| **[4]** | C. Park, B. Kim and G. Kim, "Attend to You: Personalized Image Captioning with Context Sequence Memory Networks.," arXiv.org, 2017. |
| **[5]** | X. Wang, W. Chen and W. Yuan-Fang, "No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling," arXiv.org, 2018. |
| **[6]** | A. Karpathy, "The Unreasonable Effectiveness of Recurrent Neural Networks," Andrej Karpathy blog, 21 May 2015. [Online]. Available: http://karpathy.github.io/2015/05/21/rnn-effectiveness/. [Accessed 09 October 2018]. |
| **[7]** | M. M. Krishna, M. Neelima, H. M and R. M. V. G, "Image classification using Deep learning," International Journal of Engineering & Technology, pp. 614-617, 2018. |
| **[8]** | I. Sutskever, J. Martens and J. Hinton, "Generating Text with Recurrent Neural Networks," in Proceedings of the 28 th International Conference, Bellevue, WA, USA, 2011. |
| **[9]** | K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 2002. |
| **[10]** | M. Park, H. Li and J. Kim, "HARRISON: A Benchmark on Hashtag Recommendation for Real-world Images in Social Networks," arXiv.org, 2016. |
| **[11]** | M. S. Lukin, K. Bowden, C. Barackman and A. M. Walker, "PersonaBank: A Corpus of Personal Narratives and Their Story Intention Graphs," arXiv.org, 2017. |