TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

APPLICATION OF DEEP LEARNING IN CAPTIONING IMAGE

Instructors: MSc. Ngô Văn Linh

Group 18:
1. Nguyễn Trần Khang
2. Dương Minh Hoàng
3. Nguyễn Đình Trường
4. Đặng Huỳnh Đức

INTRODUCTION

Image Caption Demo

Image Caption

Browse image

Result: a man standing in front of a mountain <end>

INTRODUCTION

Image Caption Demo

Image Caption

Browse image

Result: a little boy with a bib on is watching the camera <end>

PRELIMINARIES

Convolutional Neural Networks (CNN) & Recurrent Neural Networks (RNN)

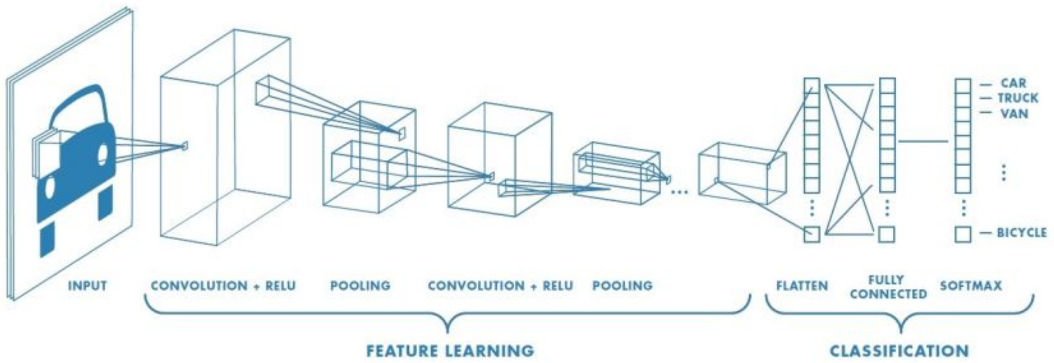


Figure 2.1: An example of CNN architecture

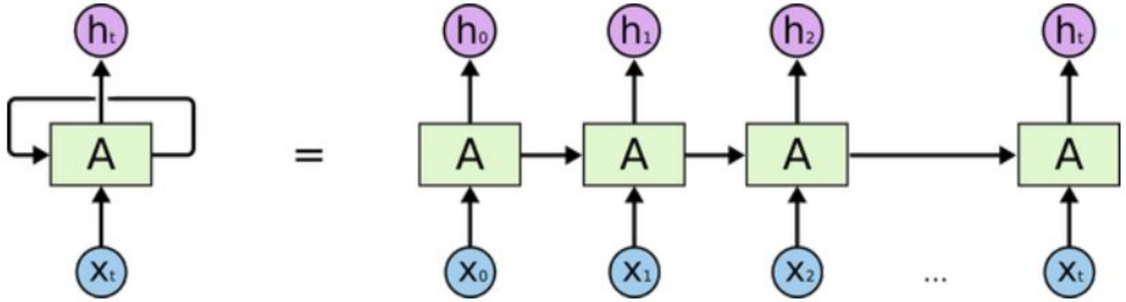
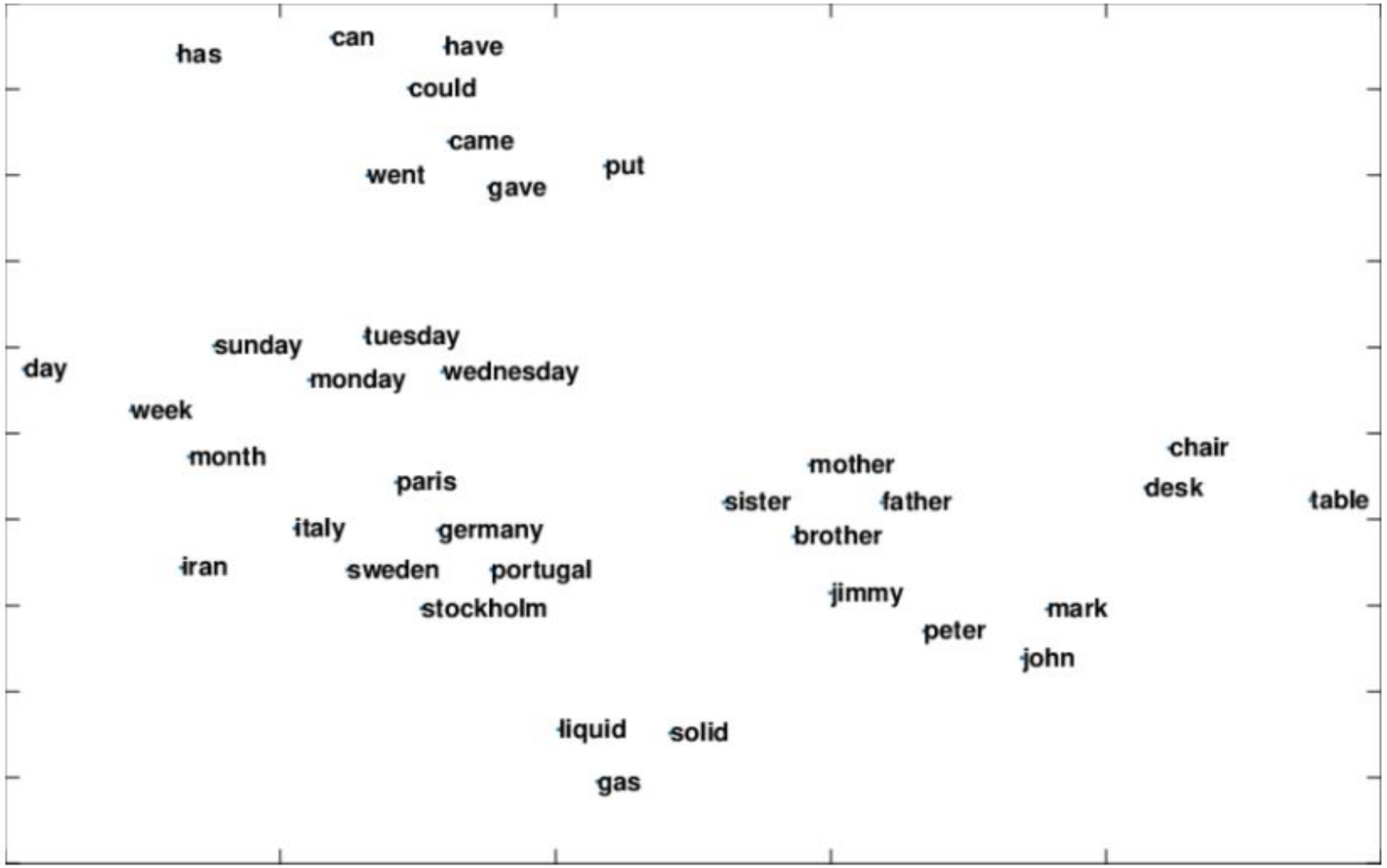


Figure 2.2: An example of RNN architecture

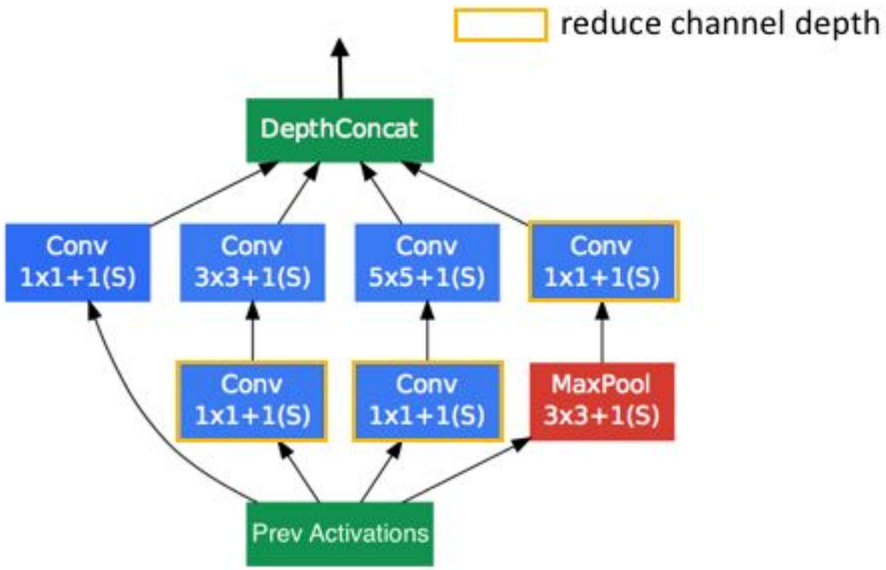
PRELIMINARIES

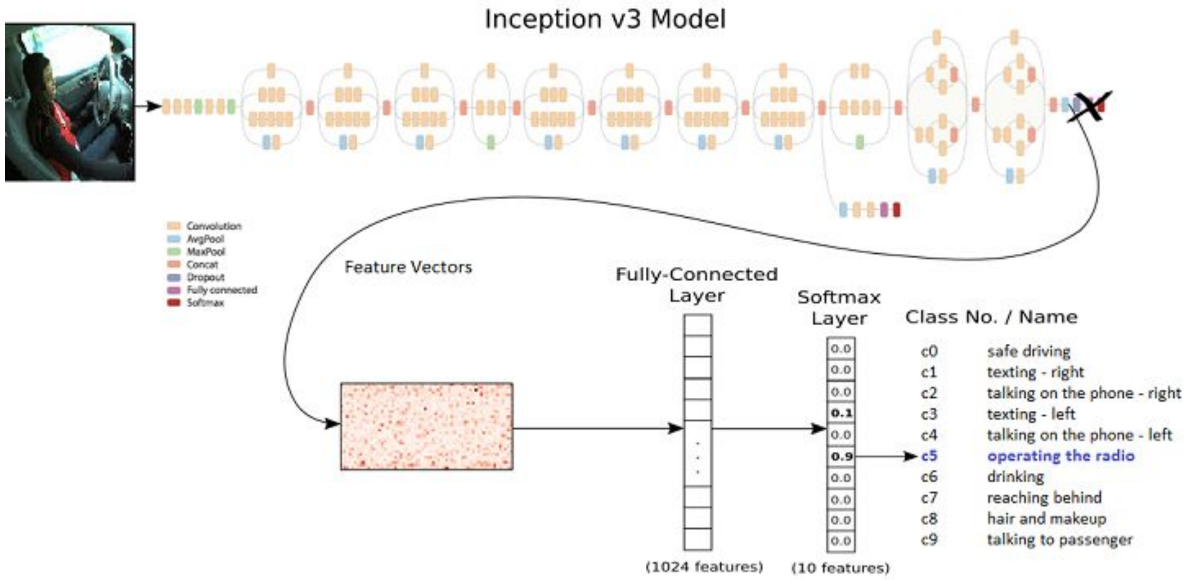
Word Embedding



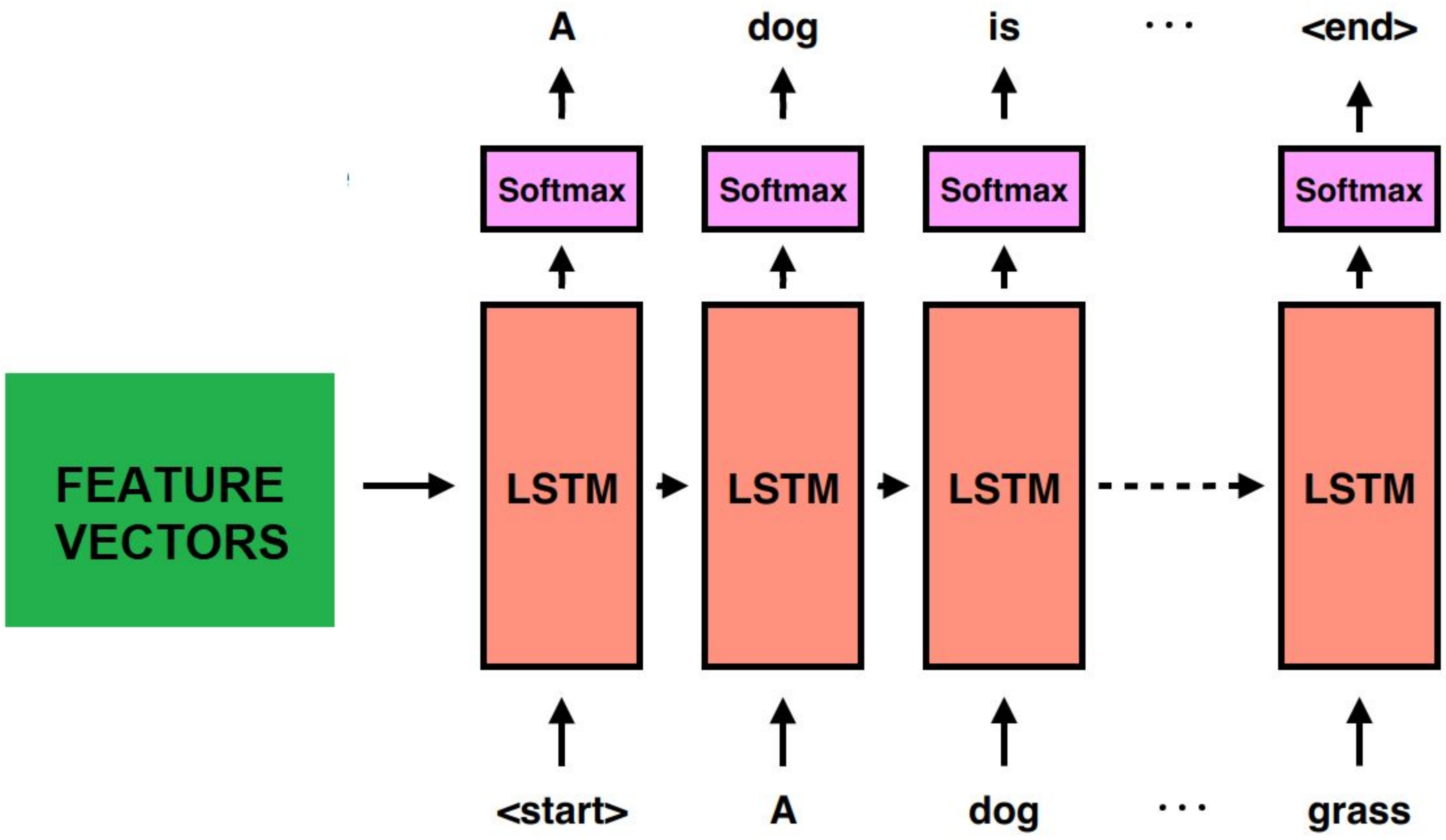
THEORY AND ACHITECTURE

Encoder: Take an image to extract feature vectors





DECODER: Predict word by word

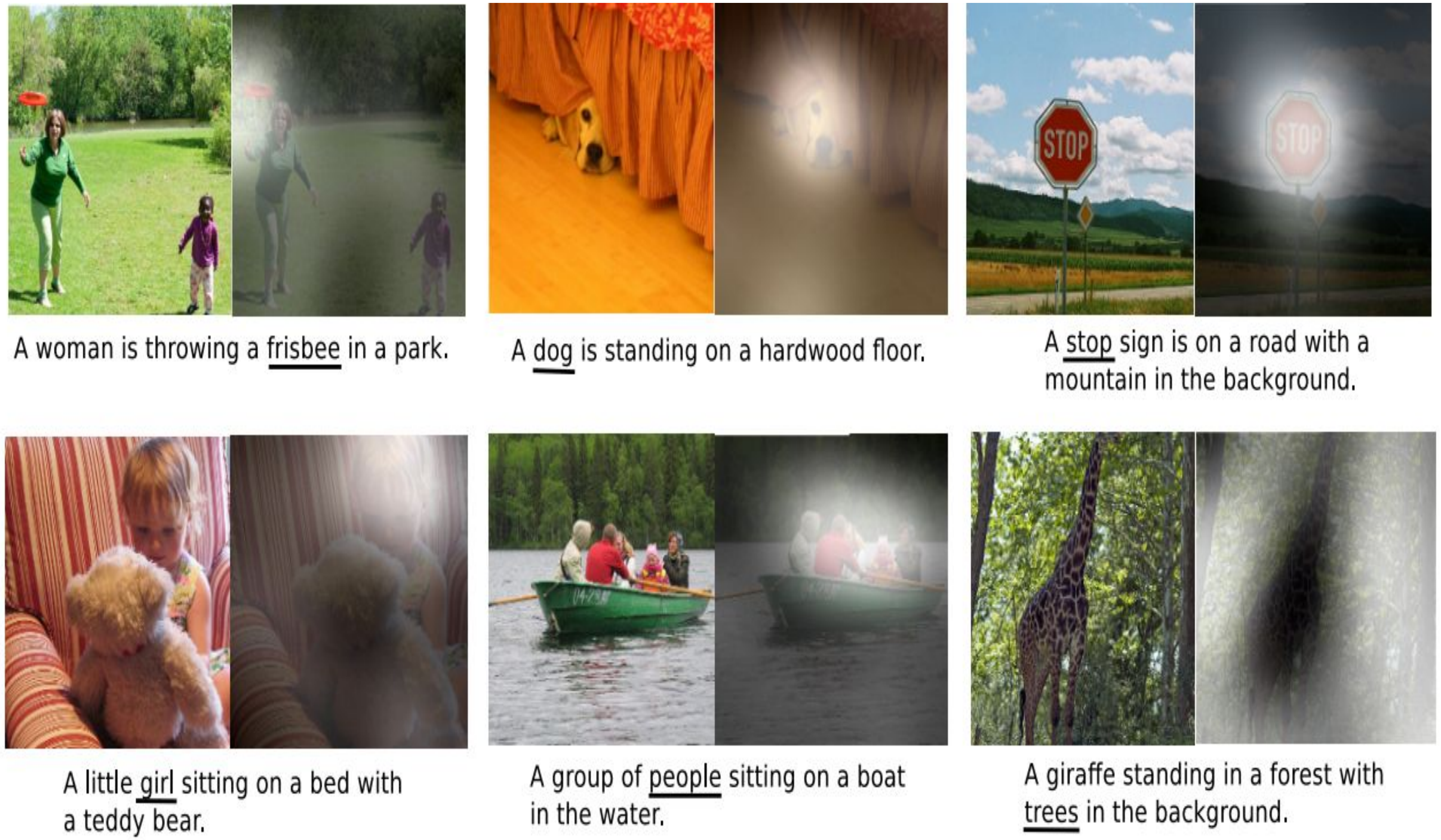


Attention Mechanism

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

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

$$\hat{z_t} = \theta(a_i, \alpha_i)$$



Stochastic ‘Hard’ Attention

St is a random variable in time step t that indicates where should be focused on.

$$p(s_{t,i} | s_{j<t}, a) = \alpha_{t,i}$$

$$\hat{z_t} = \sum_i s_{t,i} a_i$$

A LITTLE MATH

Maximum likelihood estimation

$$
\begin{aligned}
L_s &= \sum_s p(s \mid \mathbf{a}) \log p(\mathbf{y} \mid s, \mathbf{a}) \\
&\leq \log \sum_s p(s \mid \mathbf{a}) p(\mathbf{y} \mid s, \mathbf{a}) \\
&= \log p(\mathbf{y} \mid \mathbf{a}) \qquad (10)
\end{aligned}
$$

$$
\frac{\partial L_s}{\partial W} = \sum_s p(s \mid \mathbf{a}) \Big[\frac{\partial \log p(\mathbf{y} \mid s, \mathbf{a})}{\partial W} + \log p(\mathbf{y} \mid s, \mathbf{a}) \frac{\partial \log p(s \mid \mathbf{a})}{\partial W} \Big]. \qquad (11)
$$

Monte Carlo sampling

$$\tilde{s}_t \sim \text{Multinoulli}_L(\{\alpha_i\})$$

$$\frac{\partial L_s}{\partial W} \approx \frac{1}{N} \sum_{n=1}^{N} \left[\frac{\partial \log p(\mathbf{y} \mid \tilde{s}^n, \mathbf{a})}{\partial W} + \log p(\mathbf{y} \mid \tilde{s}^n, \mathbf{a}) \frac{\partial \log p(\tilde{s}^n \mid \mathbf{a})}{\partial W} \right]$$

A LITTLE MATH

Desterministic ‘Soft’ Attention

View alpha as weights of feature vectors

$$\mathbb{E}_{p(s_t|a)}[\hat{\mathbf{z}}_t] = \sum_{i=1}^{L} \alpha_{t,i}\mathbf{a}_i \qquad (13)$$

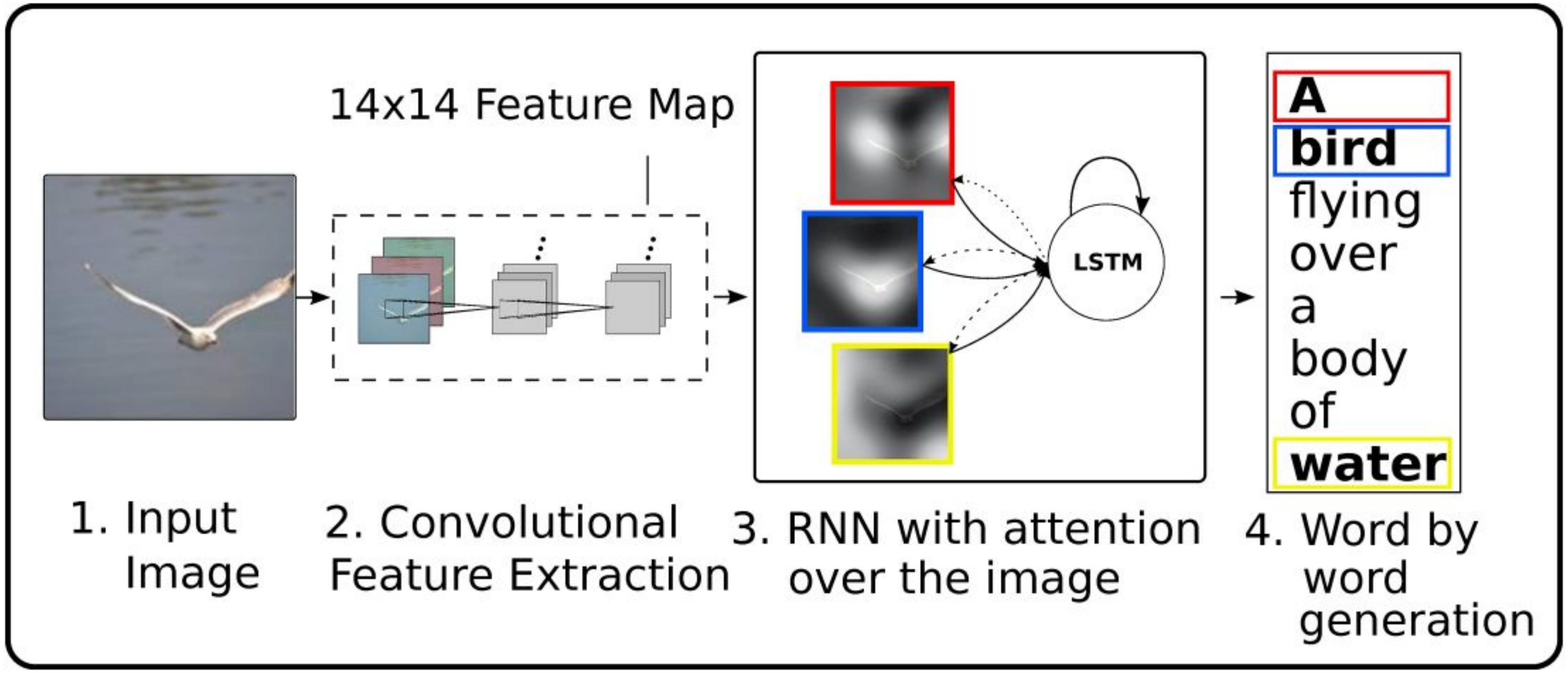
$$\phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right) = \sum_i^L \alpha_i \mathbf{a}_i$$

EXPERIMENT

Flickr dataset

Train: 6000 images
Test: 1000 images
Each image has 5 captions

Model



Overview of model

Evaluation metrics: BLEU score

Automatic Evaluation: Bleu Score

N-Gram precision

$$p_n = \frac{\sum_{n\text{-gram} \in hyp} count_{clip}(n\text{-gram})}{\sum_{n\text{-gram} \in hyp} count(n\text{-gram})}$$

Bounded above by highest count of n-gram in any reference sentence

brevity penalty

$$\mathbf{B} = \begin{cases} e^{(1-|ref|/|hyp|)} & \text{if } |ref| > |hyp| \\ 1 & \text{otherwise} \end{cases}$$

Bleu score: brevity penalty, geometric mean of N-Gram precisions

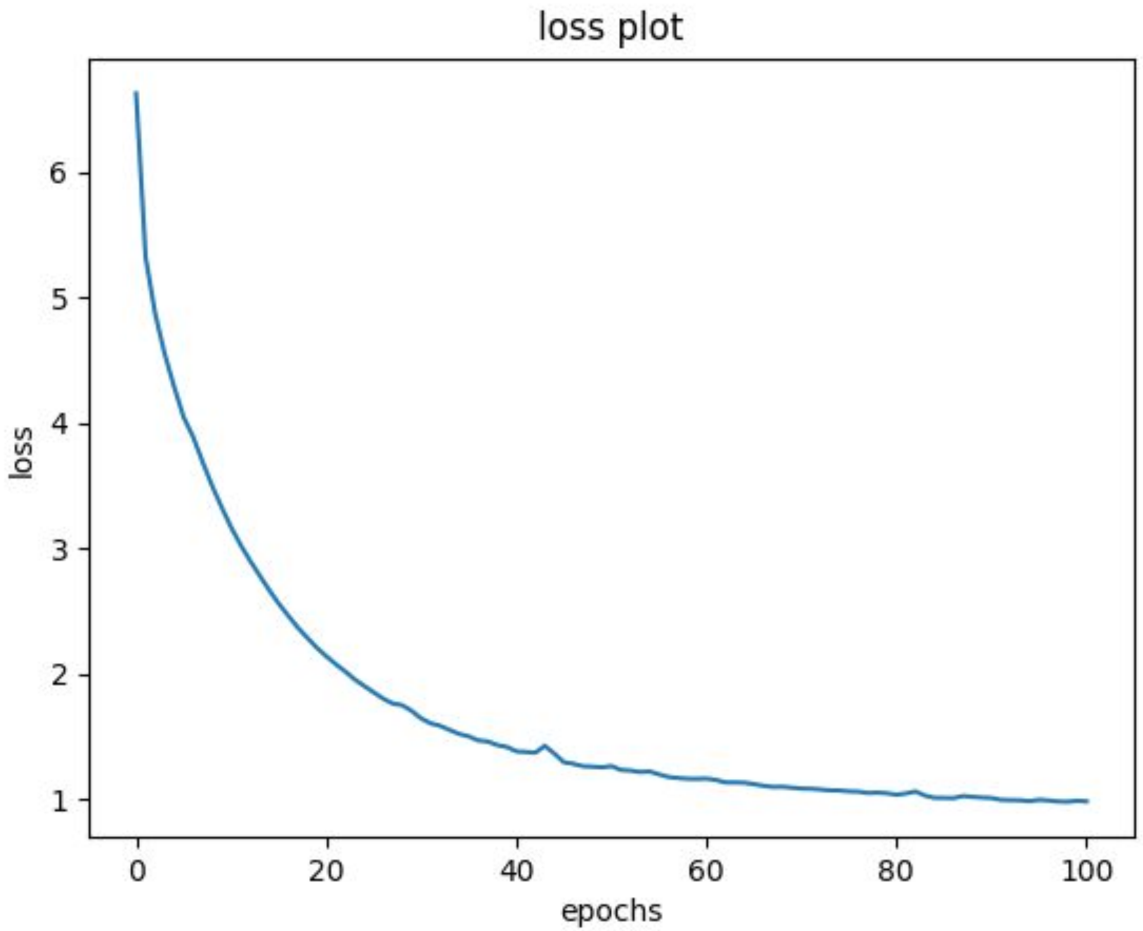
$$\mathbf{Bleu} = B \cdot \exp\left[\frac{1}{N}\sum_{n=1}^{N} p_n\right]$$

BLEU IN SUMMARY

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

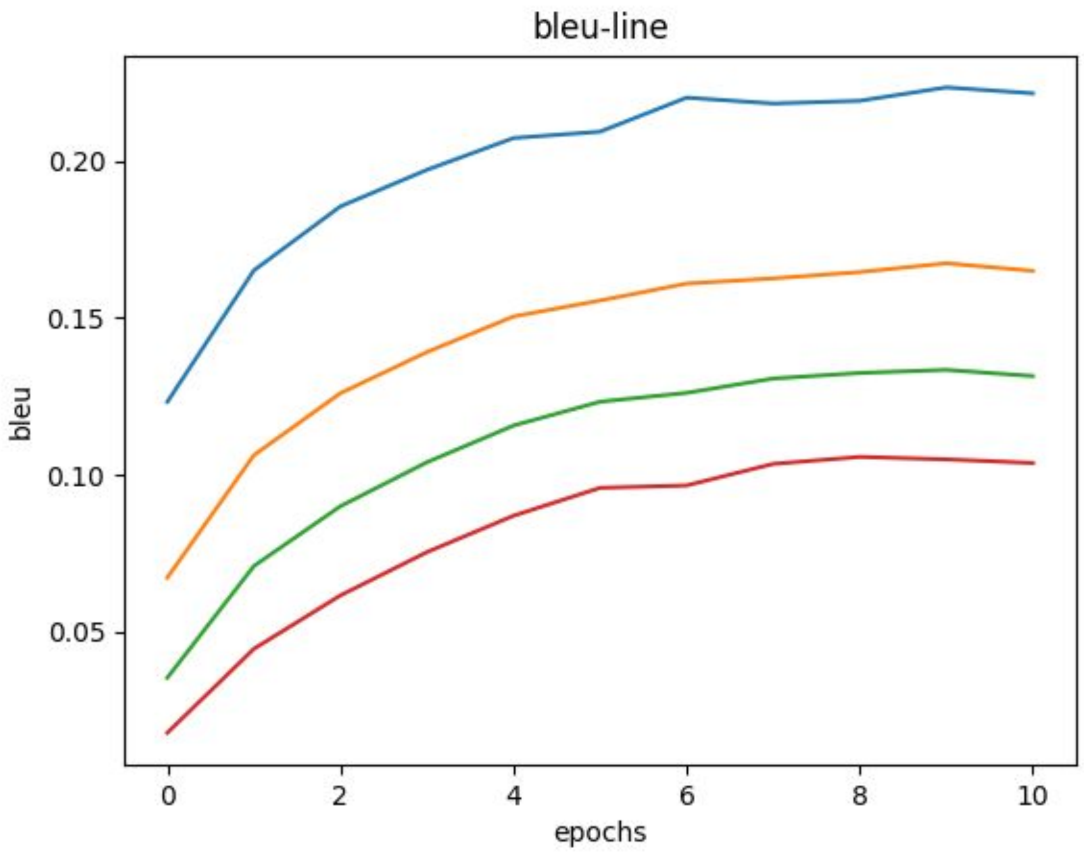
EXPERIMENT

Loss



EXPERIMENT

BLEU score per epochs on test set



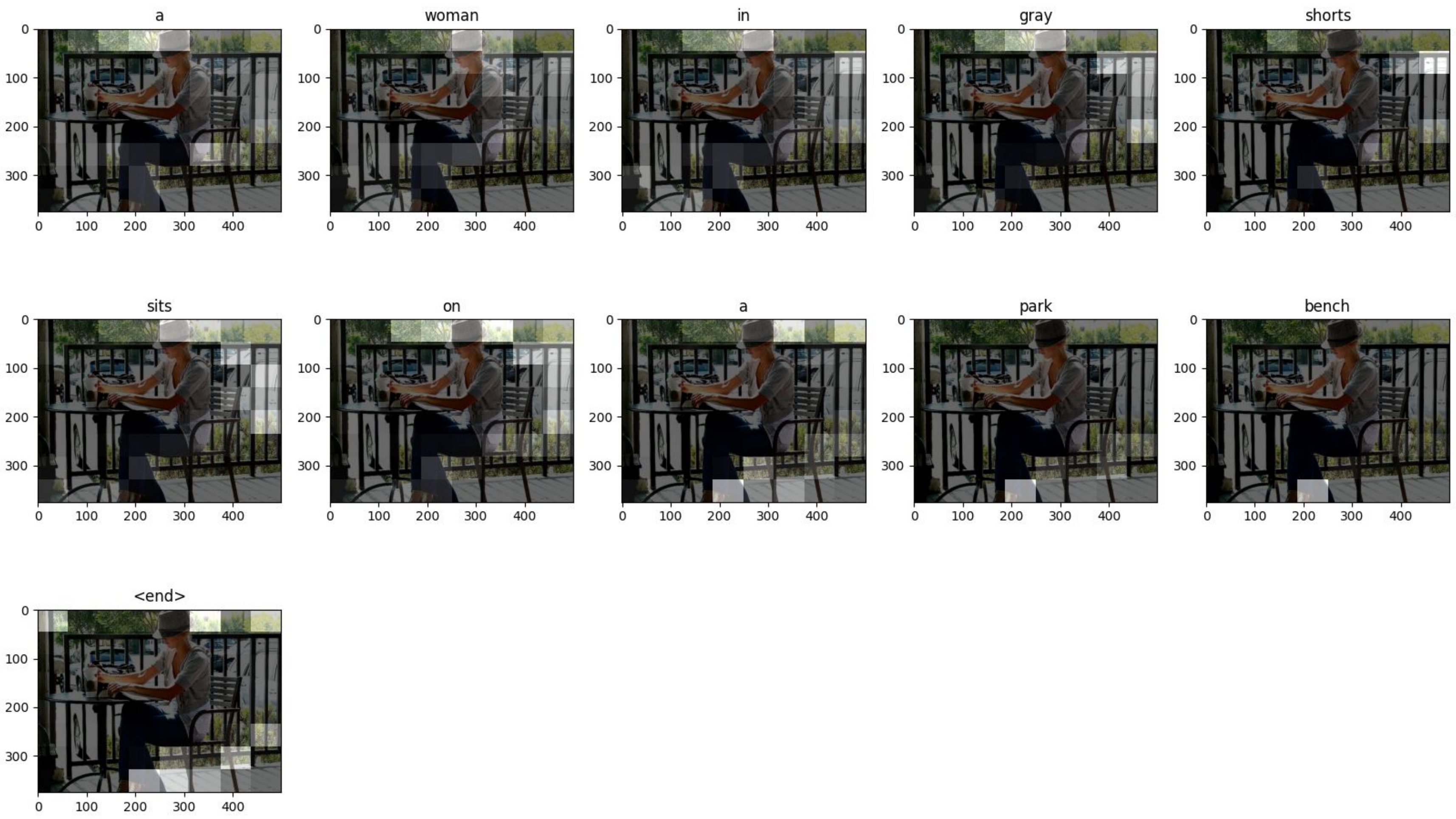
Compare different architectures

Evaluation model				
Model	Bleu1	Bleu2	Bleu3	Bleu4
GRU	0.3/0.22	0.23/0.16	0.2/0.12	0.15/0.11
GRU+embedding	0.28/0.2	0.21/0.15	0.17/0.1	0.14/0.09
LSTM	0.32/0.22	0.24/0.17	0.21/0.13	0.17/0.11
LSTM+dropout	0.29/0.22	0.23/0.18	0.18/0.12	0.15/0.1

Table 4.1: Model evaluation table

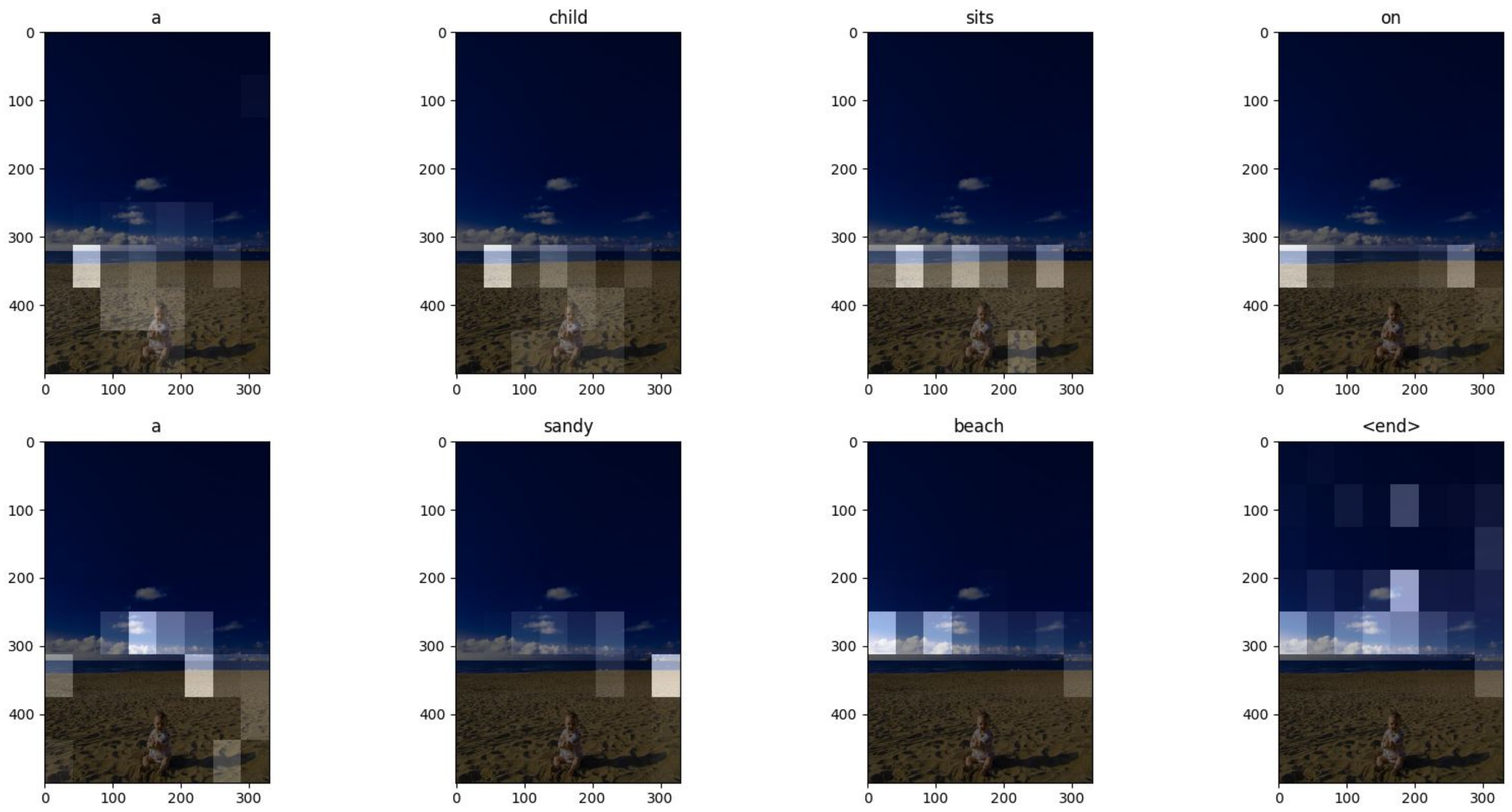
EXPERIMENT

ATTENTION VISUALIZATION



EXPERIMENT

ATTENTION VISUALIZATION




THANKS FOR WATCHING
AND LISTENING
makeameme.org