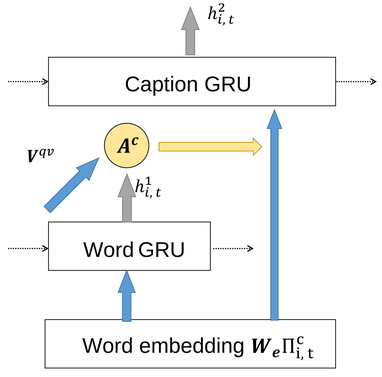
Joint Image Captioning and Question Answering

Jialin Wu, Zeyuan Hu, Raymond J. Mooney. The University of Texas at Austin

Introduction:

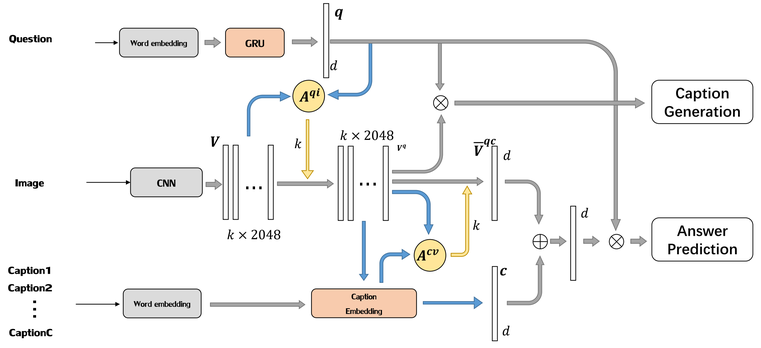
- Answering visual questions requires daily common knowledge, which is hard to be visually presented, and model the semantic connection among different parts in images.
- Image captioning systems tend to generate similar captions and fail to diversely describe images.
- Our system have these two tasks compensate each other, which is capable of jointly producing diverse image captions and answering visual questions.

Caption module:



- Word GRU sequentially encodes the words embedding.
- Caption attention module utilizes image features to generate attentions on words.
- Caption GRU produces the final caption representation.

Overview Structure:



Our system takes questions, images, and captions as inputs and uses questions and images' joint representation to generate question related captions. Blue arrows denote fc with learnable parameters and yellow arrows denote attention embedding.

Examples:

Q:What system are they playing | **Q:What system are they playing** | **Q:What system are they playing** | **Q:What system are they playing**

C1: Two people playing a video game on a table
0.1 0.2 0.0 0.0 0.7 0.8 / 0.1 0.0 0.1
C2: Two people are playing a video game on a table
0.1 0.2 0.0 0.0 0.1 0.7 / 0.7 0.1 0.1 0.1
C3: Two people are playing a video game in a living room
0.1 0.2 0.0 0.0 0.1 0.7 / 0.8 0.0 0.0 0.1 0.1

C1: Two people playing a video game on a table
0.3 0.2 0.1 0.0 0.8 0.8 / 0.1 0.0 0.1
C2: Two people are playing a video game on a table
0.3 0.2 0.0 0.2 0.1 0.8 / 0.8 0.1 0.1 0.1
C3: Two people are playing a video game in a living room
0.3 0.1 0.0 0.1 0.1 0.8 / 0.8 0.0 0.0 0.1 0.1

C1: A young man and woman playing a game with a Nintendo Wii controller
0.1 0.0 0.1 0.0 0.1 0.4 0.0 / 0.7 0.1 0.0 0.3 0.8 0.7
C2: A man and woman playing a Wii video game
0.1 0.1 0.0 0.1 0.4 0.0 0.9 / 0.7 0.7
C3: A man an a woman playing with an Xbox
0.1 0.1 0.0 0.1 0.1 0.4 0.0 / 0.0 0.5

A1: wii 0.18
A2: video game 0.34
A3: tv 0.71

A1: wii 0.29
A2: video game 0.64
A3: tv 0.21

A1: wii 0.31
A2: video game 0.67
A3: tv 0.20

A1: wii 0.78
A2: video game 0.64
A3: tv 0.19

BUTD | Ours w/o semantic connection | Ours w semantic connection | Ours with annotated captions

The answers' scores in the questions are that wii full score 1, video games score 0.3 and tv score 0. The attention weights on each caption word are shown below the word.

Online Caption Selection:

- We require the inner product of the current gradients from the VQA and captioning loss to be greater than a constant ξ and select a caption which maximizes that inner products.
- Our system is guaranteed to update with a shared descent direction from both the VQA parts and the image captioning parts, ensuring the consistency in the optimization process.

$$\arg\max_j \quad \sum_i \left(\frac{\partial \mathcal{L}^{vqa}}{\partial v_i^q}\right)^T \frac{\partial \mathcal{L}_j^c}{\partial v_i^q}$$

$$s.t. \quad \sum_i \left(\frac{\partial \mathcal{L}^{vqa}}{\partial v_i^q}\right)^T \frac{\partial \mathcal{L}_j^c}{\partial v_i^q} > \xi$$

Experiments:

	Y/N	Num	Other	All
Ours	84.7	46.8	59.3	68.4
Ours-10	86.2	47.4	60.4	69.7

Accuracy percentage on test-standard set.

	All
BUTD model	63.2
Ours with BUTD captions	64.6
Ours with our generated captions	65.8
Ours with annotated captions	69.1

Accuracy percentage on validation set.