Generating Diverse Questions using VAE

Unnat Jain*[1] Ziyu Zhang*[2] Alexander Schwing[1]
[1]UIUC [2]Snapchat Research

Motivation

Task: Given an image as input, generate diverse questions

Applications:

- AI assistants using visual cues
- VQA and Robotics
- Engaging machine-human interaction
- Active teaching of content

Is there a candle burning?

How many people are in the room?

What kind of wine is the woman drinking?

What are they celebrating?

Is this a formal occasion?

Contributions:

- Generative capability of VAE + Representative from LSTM nets
- Low dimensional latent space embedding of questions
- Sample from 'latent space' to generate new questions

Introduction & Related Work

Previous methods:

- Mostafazadeh et al. (ACL '16): Pretrained CNN + GRU Appealing model, one question per image
- Vijaykumar et al. (ICLR '17): Diverse beam search Samples from complex energy landscape, a hard task in general

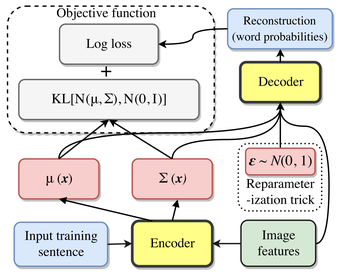
Our Approach: (Inference)

- Sample from an embedding space z
- Decode sample to form a question x

How to construct the embedding space?

Variational Auto-encoder:

- Construct image embedding
- Encode image embedding and sentence
- Sample from constructed embedding
- Reconstruction sentence while taking into account image



Maximizing likelihood: parametric distribution $p_\theta(x)$ Introduce posterior $q_\phi(z|x)$ (manifold assumption)

$$\ln p_\theta(x) = \sum_z q_\phi(z|x) \ln p_\theta(x) = \mathcal{L}(q_\phi, p_\theta(x,z)) + \mathrm{KL}(q_\phi, p_\theta(z|x))$$

Lower bound: $$\mathcal{L}(q_\phi, p_\theta(x,z)) = \sum_z q_\phi(z|x) \ln \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}$$
$$= -\mathrm{KL}(q_\phi, p(z)) + E_{q_\phi}[\ln p_\theta(x|z)]$$

Our Approach

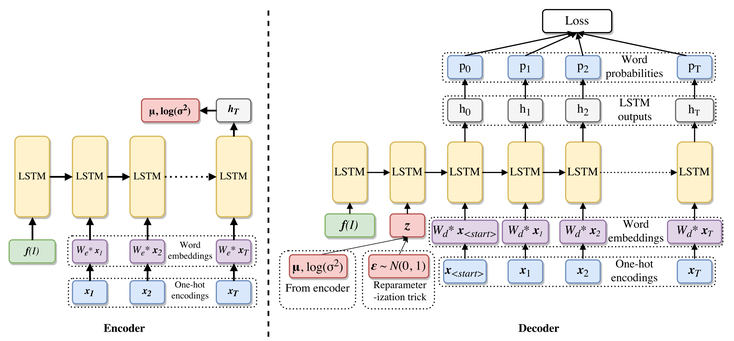
VAE + LSTM model:

- Encoder: Given image & question $\rightarrow$ generate z by $q_\phi(z|x)$
- Decoder: Given image & sample $z \rightarrow$ reconstruct x by $p_\theta(x|z)$
- Re-parametrization trick: Encoder gives $\mu(x;I)$ and $\sigma(x;I)$ and decoder changes to $p_\theta(x|z) \sim \mathcal{N}(\mu(x;I), \sigma(x;I))$
- Minimizing objective function adds structure to latent space & encourages reconstruction

Objective function:

$$\min_{\phi,\theta} \mathrm{KL}(q_\phi(z|x), p(z)) - \frac{1}{N}\sum_{i=1}^{N} \ln p_\theta(x|z^i), \text{s.t. } z^i \sim q_\phi$$

Model details:
Image feature = 4096; Vocab size = 10, 849; Word embed. = 512; LSTM hidden = 512, Question space = 20; LR = 0.01 (1/2 after 5 epochs)



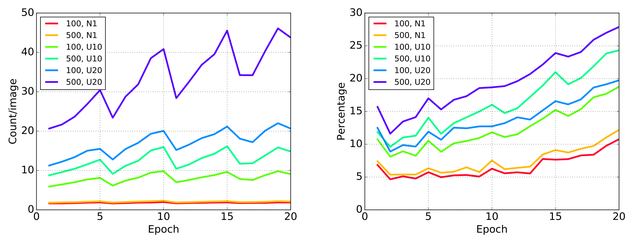
Results

Training Data: VQA + VQG dataset

- VQA - literal questions - aids training VAE+LSTM model
  Size: 82, 783 + 40, 504 (train+val) images $\times$ 3 questions
- VQG - engaging and natural questions - helps add diversity
  Size: 9000 train images $\times$ 5 questions
- Total: 125, 697 images and 399, 418 questions (238, 699 unique)

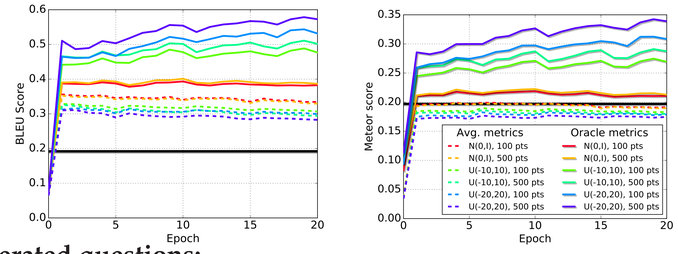
Testing Data: All three of VQG datasets - COCO, Flickr & Bing

Diversity metrics:
Generative strength: Avg. number of unique questions per image.
Inventiveness: $\frac{\text{Unique questions which were never seen in training set}}{\text{Total unique questions for that image}}$



Results (continued)

Accuracy metrics: Bleu, Meteor



Generated questions:

Error Analysis

Recognition failures: Pre-learned visual features are incapable of capturing fine information
Co-occurrence based failures: Objects not in the image appear in questions.

Sunburst plots