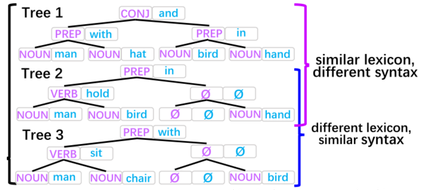
Variational Structured Semantic Inference for Diverse Image Captioning

Fuhai Chen[1], Rongrong Ji*[12], Jiayi Ji[1], Xiaoshuai Sun[1], Baochang Zhang[3], Xuri Ge[1],
Yongjian Wu[4], Feiyue Huang[4], Yan Wang[5]
[1]Department of Artificial Intelligence, School of Informatics, Xiamen University
[2]Peng Cheng Lab, [3]Beihang University, [4]Tencent Youtu Lab, [5]Pinterest

Introduction

"A picture is worth a thousand words"

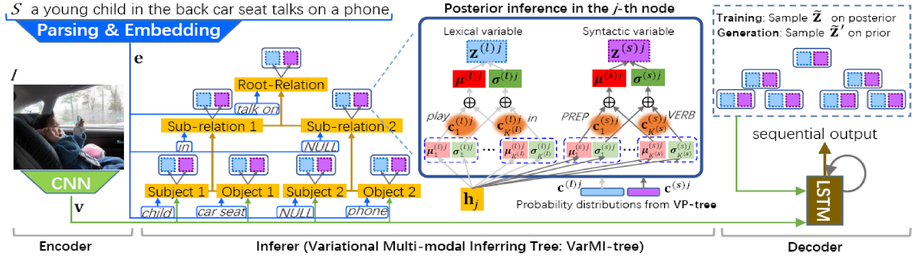
- **Diverse image captioning** aims at generating diverse captions from a given image, which is quite intuitive to derive heterogeneous understanding from human being.
- Traditional models typically tend to generate homogeneous sentences due to the limited variation in the maximum likelihood objective.
- GAN-based and VAE-based methods treated diverse image captioning as a blackbox without explicitly modeling the **key factors to diversify the expression**, i.e., the lexical and syntactic diversities.



- This paper aims at explicitly modeling the lexical and syntactic diversities from the visual content towards diversified image caption generation.

Method

- Variational Structured Semantic Inferring for diverse Image Captioning



- Encoder-inferer-decoder architecture

  Encoder: visual feature **v** and textual feature **e** are extracted from CNN and word embedding model respectively.
  Inferer: VarMI-tree is designed to infer the latent lexical/syntactic variable $\mathbf{z}^{(\ell)}/\mathbf{z}^{(s)}$ upon an additive Gaussian distribution in each node.
  Decoder: $\tilde{\mathbf{z}}$ is sampled from the posterior inference and is used for training, while $\tilde{\mathbf{z}}'$ is sampled from the prior inference and is fed to generate captions.

- Model is in principle optimized by maximizing the lower bound on the log-likelihood:

$$\mathcal{L}(\theta,\phi^{(\ell)},\phi^{(s)},\psi;S,\mathbf{v},\mathbf{c}^{(\ell)},\mathbf{c}^{(s)}) = \mathbb{E}_{\mathbf{z}^{(\ell)}\sim q_{\phi^{(\ell)},\psi},\mathbf{z}^{(s)}\sim q_{\phi^{(s)},\psi}}\left[\log p_\theta(S|\mathbf{z}^{(\ell)},\mathbf{z}^{(s)},\mathbf{v},\mathbf{c}^{(\ell)},\mathbf{c}^{(s)})\right]$$
$$- D_{\mathrm{KL}}\left(q_{\phi^{(\ell)},\psi}(\mathbf{z}^{(\ell)}|S,\mathbf{v},\mathbf{c}^{(\ell)})\|p(\mathbf{z}^{(\ell)}|\mathbf{c}^{(\ell)})\right) - D_{\mathrm{KL}}\left(q_{\phi^{(s)},\psi}(\mathbf{z}^{(s)}|S,\mathbf{v},\mathbf{c}^{(s)})\|p(\mathbf{z}^{(s)}|\mathbf{c}^{(s)})\right)$$

The former is maximized to reduce the reconstruction loss of the caption generation in decoder, while the later two measure the difference between the distributions of the posterior and the prior with the prior guidance.

Experiments

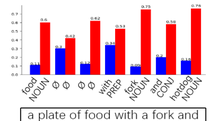
Accuracy Evaluation on MSCOCO

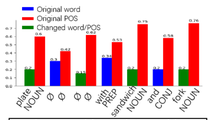
Metric	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rouge-L	CIDEr	Spice
ErDr-cap [29]	69.9	51.8	36.6	25.6	23.1	50.3	84.3	16.4
Up-Down [3]	**79.8**	-	-	**36.3**	**27.7**	**56.9**	**120.1**	**21.4**
G-GAN [6]	-	-	30.5	20.7	22.4	47.5	79.5	18.2
Adv [5]	-	-	-	-	23.9	-	-	16.7
CAL [9]	66.5	48.4	33.2	21.8	22.6	47.8	75.3	16.4
GMM-CVAE [8]	70.0	52.0	37.1	26.0	23.2	50.6	85.4	16.3
AG-CVAE [8]	70.2	52.2	37.1	26.0	23.4	50.6	85.7	16.5
VSSI-cap-L	69.9	51.9	37.3	26.1	23.5	50.7	87.3	16.8
VSSI-cap-S	70.4	52.7	37.9	27.1	23.8	51.1	88.8	17.0
VSSI-cap	70.4	52.7	38.1	27.3	23.9	51.3	89.4	17.1

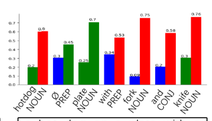
Metric	Num.	mB.⇓	div1⇑	div2⇑	Uni.⇑	Nov.⇑
human	n=5	**51.0**	**34.0**	**48.0**	**99.8**	-
ErDr-cap [29]	n=5	78.0	28.0	38.0	-	34.18
Up-Down [3]	n=5	80.9	27.1	35.8	-	63.60
G-GAN [6]	n=5	-	-	-	-	81.52
Adv [5]	n=5	70.0	34.0	44.0	-	73.92
CAL [9]	n=5	-	32.5	40.7	-	-
AG-CVAE [8]	n=5	70.2	33.1	42.9	66.9	79.67
AG-CVAE [8]	n=10	77.3	22.7	31.3	70.8	79.68
VSSI-cap-L	n=5	68.7	34.3	45.6	80.2	79.30
VSSI-cap-S	n=5	63.0	33.8	46.3	82.4	80.26
VSSI-cap	n=5	62.4	33.9	47.2	83.0	85.20
VSSI-cap	n=10	74.2	22.3	33.2	80.7	80.34

Diversity Evaluation on MSCOCO with n output captions

Changing the inputs of VP-tree to evaluate VarMI-tree







ErDr-cap (blue), AG-CVAE (green), VSSI-cap (red)