# *RefineCap*: Concept-Aware Refinement for Image Captioning

Yekun Chai†, Shuo Jin‡, Junliang Xing†

chaiyekun@gmail.com shj42@pitt.edu jlxing@nlpr.ia.ac.cn

†Institute of Automation, Chinese Academy of Sciences ‡University of Pittsburgh

## Abstract

Automatically translating images to texts involves image scene understanding and language modeling. In this paper, we propose a novel model, termed *RefineCap*, that refines the output vocabulary of the language decoder using decoder-guided visual semantics, and implicitly learns the mapping between visual tag words and images. The proposed Visual-Concept Refinement method can allow the generator to attend to semantic details in the image, thereby generating more semantically descriptive captions. Our model achieves superior performance on the MS-COCO dataset in comparison with previous visual-concept based models.

## Introduction

- For image captioning, only content-related words are actively matched with the image, whereas other functional words can often be automatically inferred using a language model.
- Not all generated words are actively related to visual contents—some words can be reliably predicted just from the language model [1], such as “phone” after “taking on a cell”. Also, non-visual words like “the” can be generated with the language model inference.
- We propose *RefineCap*, a visual-concept-aware refined encoder-decoder architecture to dynamically modulate the visual-semantic representations and thus enhance the output of language model. Within the proposed model, visual signals are decoder-regulated and the content-based gate removes the independent assumption of visual objects in the attention mechanism.

## Illustration

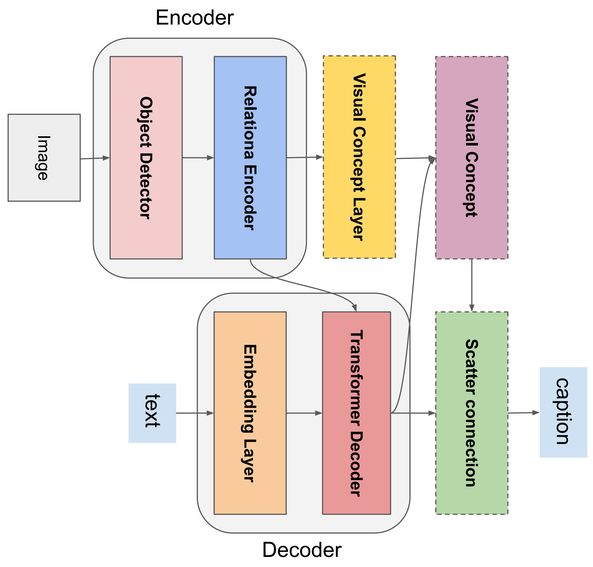


Fig. 1: Schematic illustration of proposed models, where “text” indicates previous words of the captions.

## Methodology

- Use the Transformer encoder [2] as the relational encoder followed by Faster RCNN object detector.
- Use standard Transformer decoder followed by the proposed scatter-connection mechanism.
- Denoting the decoder output at $t$-th time step ($t = 1, 2, \cdots, T$) by $\mathbf{h^t} \in \mathbb{R}^D$, we compute the context vector $\mathbf{c^t}$ by considering the interaction between the $t$-th decoder output and all encoded features $\mathbf{F}$ in one image, followed by a sigmoidal non-linearity $g$ to produce the decoder-guided gate for concept-aware modulation.

$$\mathbf{c^t} = \mathbf{u}^\top \tanh(\mathbf{W_1 h^t} + \mathbf{W_2 F}) \tag{1}$$

$$\mathbf{g^t} = g(\mathbf{W_3 c^t}) \tag{2}$$

where $\{\mathbf{W_1}, \mathbf{W_2}\} \in \mathbb{R}^{D' \times D}, \mathbf{u} \in \mathbb{R}^{D'}, \mathbf{W_3} \in \mathbb{R}^{K \times M}$ are parameters, $D'$ is the hidden dimension.

- At $t$-th time step, the visual-concept vector $\hat{\mathbf{v}}$ is modulated by the decoder-guided gate $\mathbf{g^t}$ to render the final refined representation $\mathbf{o^t}$:

$$\mathbf{o^t} = \mathbf{g^t} \odot \hat{\mathbf{v}} \tag{3}$$

where $\odot$ indicates the element-wise product.

**Scatter-Connected Mapping** Since the selected image-grounded vocabulary $\mathcal{V}_{\text{tag}}$ is the subset of caption vocabulary $\mathcal{V}_{\text{cap}}$, *i.e.*, $\mathcal{V}_{\text{tag}} \subset \mathcal{V}_{\text{cap}}$, we apply scatter-connected mapping by adding the corresponding element of $\mathcal{V}_{\text{tag}}$ onto $\mathcal{V}_{\text{cap}}$ to enhance the confidence of concept word prediction:

$$\mathbf{h^t}[j] = \begin{cases} \mathbf{h^t}[j] + \mathbf{o^t}[k] & \text{if } \mathcal{V}_{\text{cap}}(j) = \mathcal{V}_{\text{tag}}(k) \\ \mathbf{h^t}[j] & \text{otherwise,} \end{cases} \tag{4}$$

where $\mathcal{V}_{\text{cap}}(j)$ and $\mathcal{V}_{\text{tag}}(k)$ indicate the corresponding concept in the $j$-th position of caption vocabulary set and $k$-th word of concept vocabulary set. $[.]$ is tensor indexing operation.

**Training with Policy Gradient** We apply the REINFORCE with baseline algorithm to train the model with the CIDEr-D scores [3] as rewards.

## Results

| Models | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METER | ROUGE-L | CIDEr | SPICE |
|---|---|---|---|---|---|---|---|---|
| SemAttn (You et al., 2016) | 0.709 | 0.537 | 0.402 | 0.304 | 0.243 | - | - | - |
| Att-CNN+LSTM (Wu et al., 2016) | 0.74 | 0.56 | 0.42 | 0.31 | 0.26 | - | 0.94 | - |
| LSTM-C (Yao et al., 2017) | - | - | - | - | - | 0.230 | - | - |
| Skeleton Key (Wang et al., 2017) | 0.673 | 0.489 | 0.355 | 0.259 | 0.247 | 0.489 | 0.966 | 0.196 |
| SCN-LSTM (Gan et al., 2017) | 0.728 | 0.566 | 0.433 | 0.330 | 0.257 | - | 1.041 | - |
| Bridging (Fan et al., 2019) | - | - | - | 0.330 | 0.264 | **0.586** | 1.066 | - |
| Ours | **0.802** | **0.645** | **0.499** | **0.378** | **0.283** | 0.580 | **1.272** | **0.225** |

Fig. 2: Overall performance of the proposed model and **visual-concept based** models.

## Comparison with Transformers

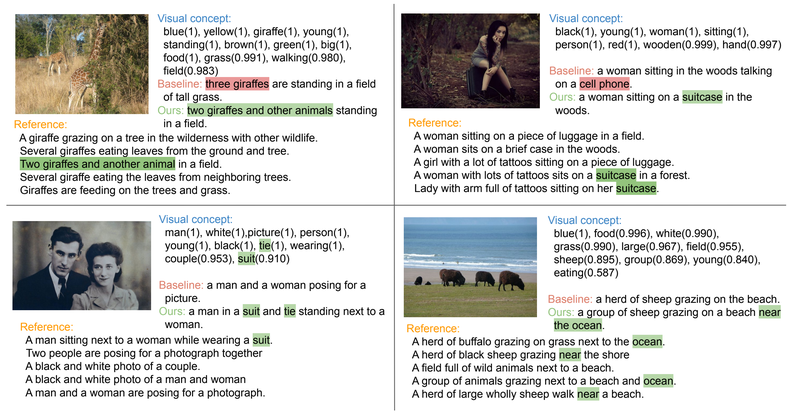


Fig. 3: Detected tags and generated captions using baseline (Transformer) and proposed models on MS-COCO, where red and green backgrounds indicate wrong and correct predictions respectively. The value in brackets means the confidence (*i.e.*, probability) of corresponding tags in the image.

**Better Accuracy** Transformer baseline without the proposed method sometimes generates mismatched words, which can be reliably rectified by the proposed method. For example, in the upper left figure, our model correctly predicts the presence of “two giraffes and another animal” but baseline identifies them as “three giraffes” by mistake.

**Better Adequacy** Our model can capture more specific details and meaningful contents in the image background that might be ignored by the baseline or even omitted in the ground truth. For example, as shown in bottom left, the proposed model predicts the occurrence of “tie” which is overlooked by both the baseline and ground truth.

## Conclusion

Our experiments show that image-grounded concept detection advances the performance of Transformer-based encoder-decoder architecture by incorporating the visual-semantic representation using reinforcement learning. Our contributions are:

- A scatter-connected mechanism to refine the language decoder using extracted visual semantics, which produces more specific descriptive words in captions.
- A competing model that outperforms the previous visual-concept based captioning models on the MS-COCO dataset.

## References

[1] Jiasen Lu et al. “Knowing when to look: Adaptive attention via a visual sentinel for image captioning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 375–383.

[2] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[3] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. “Cider: Consensus-based image description evaluation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575.