RefineCap: Concept-Aware Refinement for Image Captioning

Yekun Chai[†], **Shuo Jin**[‡], **Junliang Xing**[†]
[†]Institute of Automation, Chinese Academy of Sciences
[‡]University of Pittsburgh
chaiyekun@gmail.com shj42@pitt.edu jlxing@nlpr.ia.ac.cn

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.

1 Introduction

Holding the promise of bridging the domain gap between computer vision and human language, image-to-text translation, *a.k.a* image captioning, has lately received great attention in both communities (Johnson et al., 2016; Chen et al., 2017; Anderson et al., 2018; Fan et al., 2019). Combining image scene understanding and language generation, it aims to translate descriptive texts given corresponding images.

Existing work maintained that frequently occurred n-grams of reference captions in the training set are preferred in the caption generation, regardless of image contents (Fan et al., 2019). Visual concept (*i.e., tag*) prediction is proposed to leverage the visual semantics from images for generating relevant words (Wu et al., 2016; You et al., 2016; Gan et al., 2017; Yao et al., 2017; Fan et al., 2019). It predicts the probability of each semantic concept that occurs in the corresponding image out of the selected image-grounded vocabulary from reference captions, for the use of following caption generation.

However, most of them adopted Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) as the language decoder, in which

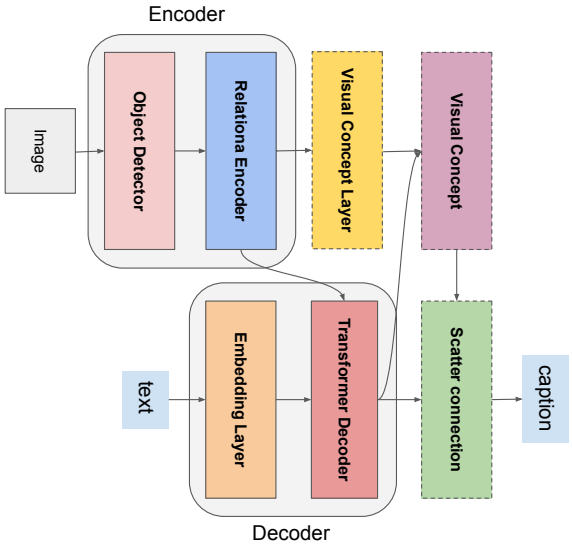


Figure 1: Schematic illustration of proposed models, where "text" indicates previous words of the captions.

sequence-align recurrence inherently precludes parallelization in practice. Initially proposed for neural machine translation (NMT), Transformer (Vaswani et al., 2017) architecture has made rapid progress in numerous applications, such as NMT (Ott et al., 2018; Gu et al., 2019), image generation (Parmar et al., 2018), automatic speech recognition (Dong et al., 2018), computer games (Vinyals et al., 2019), *etc*. We adopt Transformer blocks rather than RNN to support parallel training in our model.

Intuitively, when describing an image, people often take much time to "think" of the words to generate coherent functional words, and just "look" at the image regions occasionally for content words. Similarly, 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.

The adaptive attention model (Lu et al., 2017) found that not all generated words are actively related to visual contents—some words can be reliably predicted just from the language model, such as "phone" after "taking on a cell". Also, non-

visual words like "the" can be generated with the language model inference.

Grounding on this, we introduce a language-decoder-guided gate to regulate visual-concept vectors, followed by a scatter connection layer to map each element of the image-grounded vector to the corresponding position in the decoder's vocabulary. This jointly takes into account what the captioner "thinks" via the language model and what it "looks" at with the dynamic visual-semantic concept vector. In this paper, 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.

Our work illustrates that image-grounded concept detection advances the performance of Transformer-based encoder-decoder captioning architecture by incorporating the visual-semantic representation using reinforcement learning. Our key contributions are as follows:

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

2 Methodology

Fig. 1 shows an overview of the proposed model. We use an object detector to extract visual object features and a relational encoder to capture the pairwise relations among detected objects (Sec. 2.1). Then visual-concept components learn the implicit visual tags in the image. The Transformer language decoder receives the extracted object features (Sec. 2.2), followed by a scatter connection mechanism (Sec. 2.3) for decoder refinement.

2.1 Transformer as Relational Encoder

Following (Anderson et al., 2018), we use Faster RCNN (Ren et al., 2015) with ResNet-101 (He et al., 2016) as the object detector. We extract the object proposals using Region Proposal Network and mean-pooled convolution to generate the 2048-dimensional proposal feature.

Let $\mathbf{X} = [\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_M]^\top \in \mathbb{R}^{M\times 2048}$ denote the extracted M visual proposal features of each image. We use a fully connected layer to reduce the spatial feature of 2048 into $D = 512$. Then we apply the standard Transformer encoder with layer number N_{enc} to capture the object-wise relation as in (Zambaldi et al., 2018). The output of encoder is represented as $\mathbf{F} = [\mathbf{f}_1\mathbf{f}_2\cdots\mathbf{f}_M]^\top \in \mathbb{R}^{M\times D}$.

2.2 Transformer Decoder

Given the caption sequence with the length of T, we apply a standard Transformer decoder of N_{dec} layers as the caption generator. The input sequence firstly passes into word embedding and sinusoidal positional embedding layers, followed by a masked self-attention sublayer attending to previous histories. Then a cross attention sublayer as in Vaswani et al. (2017) is applied to capture the multi-modal attention between each word and extracted object features, followed by a position-wise feed-forward layer (FFN). All sublayers are encompassed by a residual connection (He et al., 2016) and layer normalization (Ba et al., 2016).

2.3 Visual-Concept Refinement

Visual Concept Layer The visual concept layer is designed to extract the probability of common concept words, such as noun and verbs. We employ the K most frequently occurring words whose POS tags are nouns, verbs, or adjectives as the image-grounded concept vocabulary set $\mathcal{V}_{\text{tag}}$. Following (Gan et al., 2017), K is set to 1,000 in our experiment. For each image, the visual concept layer projects its visual object features $\mathbf{f}_m$ $(m = 1, \cdots, M)$ into a (K/M)-dimensional vector, and then concatenates M different visual outputs to get the K-dimensional concept representation. Then we apply an activation function to get the probability of each concept word in $\mathcal{V}_{\text{tag}}$, which is a multi-label binary prediction.

$$\mathbf{v} = \mathbf{f}_1\mathbf{W_0}\|\mathbf{f}_2\mathbf{W_0}\|\cdots\|\mathbf{f}_M\mathbf{W_0} \quad (1)$$

$$\hat{\mathbf{v}} = \sigma(\mathbf{v}) \quad (2)$$

where $\mathbf{W_0} \in \mathbb{R}^{D\times K/M}$ denotes trainable weights, $\|$ denotes the concatenation along the last axis, $\sigma(x) = 1/(1+\exp(-x))$, $\mathbf{f}_m \in \mathbb{R}^D$ $(m = 1, 2, \cdots, M)$ is the m-th object feature out of M visual proposals. $\hat{\mathbf{v}} \in \mathbb{R}^K$ represents the visual concept vector, in which each element represents the confidence of image-grounded concept words.

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 non-linearity g (we use sigmoid function here) to produce the decoder-guided gate for concept-aware modulation.

$$\mathbf{c^t} = \mathbf{u}^\top \tanh(\mathbf{W_1 h^t} + \mathbf{W_2 F}) \quad (3)$$

$$\mathbf{g^t} = g(\mathbf{W_3 c^t}) \quad (4)$$

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' represents the intermediate 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 $\mathbf{o^t}$:

$$\mathbf{o^t} = \mathbf{g^t} \odot \hat{\mathbf{v}} \quad (5)$$

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:

$$\mathbf{h^t}[j] = \begin{cases} \mathbf{h^t}[j] + \mathbf{o^t}[k] & \text{if } \mathcal{V}_{\text{cap}}(j) = \mathcal{V}_{\text{tag}}(k) \\ \mathbf{h^t}[j] & \text{otherwise,} \end{cases} \quad (6)$$

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. Then a softmax function is applied afterward to get the probability over all caption vocabularies.

2.4 Training with Policy gradient

Captioning as Reinforcement Learning The image captioning task is cast to the RL problem: the policy network *RefineCap*, defined as π_θ parameterized by θ, takes an action a_t at t-th time step for each observation (*i.e.*, image) to predict the next word w_t until reaching the rollout end. The return G for each rollout is defined as CIDEr-D scores (Vedantam et al., 2015) between hypothesis and ground-truth captions.

We leverage the REINFORCE with baseline algorithm to reduce the gradient variation. The parameter θ_{t+1} is updated as follows:

$$\theta_{t+1} = \theta_t + \alpha(G_t - b)\nabla \ln \pi(\cdot) \quad (7)$$

where b takes the average of batch returns, α is the learning rate.

3 Experiments

3.1 Experimental Setup

Dataset and Evaluation We experiment on the MS-COCO dataset (Lin et al., 2014) and report the performance on *Karpathy* offline splits, consisting of 113,287/5,000/5,000 images for training/val/test sets, in which each image is paired with 5 human annotations. We empploy BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Denkowski and Lavie, 2014), CIDEr-D (Vedantam et al., 2015), and SPICE (Anderson et al., 2016) as evaluation metrics.

Implementation Details We set the embedding dimension D to 512, the layer numbers of both encoder and decoder as 3 (for fast training), batch size of 50, head number h as 8, the hidden dimension of FFN layer as 2,048, the maximum number of extracted features as 50. Word embeddings are randomly initialized. We employ Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9, \beta_2 = 0.98$ as in (Li et al., 2019). To avoid over-fitting, we set the dropout rate as 0.1, and early stopping patience as 5 during training. The beam width is set as 5 during beam search decoding. To initialize the model weights for RL training, we pretrain the model with supervised learning using cross-entropy loss with the same setting. All experiment are trained on a single NVIDIA Tesla V100 GPU.

3.2 Results

Table 1 exhibits the performance of visual-concept based image-to-text translation models on MS-COCO dataset, in which the proposed model outperforms baseline models by a clear margin.

Ablation Test Table 2 shows that the proposed method boosts the CIDEr scores on standard Transformer for image caption generation, indicating the proposed visual-semantic module and scattered-connection mechanism promote the CIDEr-D score in contrast with other baseline models.

Qualitative Analysis Fig. 2 illustrates the generated caption samples, where the detected visual concepts advance the quality of predicted captions. As shown in Fig. 2 bottom left, discovered visual tags enhance the adequacy of generated captions, such as "tie" with the confidence of 1, which is not mentioned in ground truth. Our model adds "near the ocean" as the adverbial modifier in comparison with baseline in the bottom right.

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**

Table 1: Overall performance of the proposed model and **visual-concept based** models.









Figure 2: 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.

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Ours	0.802	0.378	0.283	0.580	**1.272**	0.225
w/o refinement	0.786	0.366	0.277	0.570	1.202	0.210

Table 2: Comparison with standard Transformer.

Better Accuracy Transformer baseline without the proposed method sometimes generates mismatched words, which can be reliably rectified by the proposed method. For example, in upper left of Fig. 2, 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 Fig. 2 (bottom left), the proposed model predicts the occurrence of "tie" which is overlooked by both the baseline and ground truth.

Implicit Visual Concept Modeling We found that the detected concepts can exactly match semantic objects in the image. Notably, such visual-concept detection can be treated as the side effect of the scatter-connection mappings since the proposed method implicitly learns the visual mapping from the visual objects to visual-concept vocabulary (as aforementioned in Sec. 2.3), instead of using an explicit training objective for learning such mapping.

To investigate the necessity of explicit visual-concept learning, we design further experiments by pre-training the visual-concept detector for multi-label binary classification before training the caption generator. We empirically find it unnecessary to pretrain the visual-concept detector before the caption generation, as shown in Sec. A.3.

4 Conclusion

We proposed a new visual-concept based image captioning framework to generate meaningful descriptive sentences. Leveraging visual concepts and scatter mapping, *RefineCap* demenstrates its effectiveness on the MS-COCO dataset.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE.

Zhihao Fan, Zhongyu Wei, Siyuan Wang, and Xuan-Jing Huang. 2019. Bridging by word: Image grounded vocabulary construction for visual captioning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6514–6524.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems*, pages 11179–11189.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8928–8937.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. 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*, pages 375–383.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. *arXiv preprint arXiv:1802.05751*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, pages 1–5.

Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W Cottrell. 2017. Skeleton key: Image captioning by skeleton-attribute decomposition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7272–7281.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 203–212.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017. Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6580–6588.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, et al. 2018. Deep reinforcement learning with relational inductive biases. In *Proceedings of international conference on learning representations*.

A Appendices

A.1 Case Study of Generated Captions

Fig. 2, 3, 4 manifest the comparison between the baseline and proposed models, with the presence of visual concepts and their corresponding confidence (presented in brackets) for each image.

Visual-Concept Indication The proposed method implicitly learns the connection between visual concepts and image regions without explicit training on the visual concept detector. Further analysis can be found at Appendix 2.

Better Adequacy It can be seen that 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 the bottom left in Fig. 2, the proposed model predicts the occurrence of "tie" which is overlooked by both the baseline and ground truth. Other examples like "ocean" at the bottom right of Fig. 2, "store" at bottom left example of Fig. 3a, "ocean" at upper left in Fig. 4a, "a ground of people ... around" at the upper right Fig. 4a, "kitchen" at the bottom left of Fig. 4a, "clock" at upper left in Fig. 4b, "store" at upper right in Fig. 4b, "street sign" at bottom left in Fig. 4b, "basket" at bottom right in Fig. 4b, *etc.* We found that captions generated by the proposed model could achieve better scores in terms of the adequacy.

Better Accuracy Meanwhile, captions decoded by the proposed models equip better accuracy. For instance, in upper left of Fig. 2, our model predicts the presence of "two giraffes and another animal" but baseline identifies them as "three giraffes" by mistake. Other proofs, such as "suitcase" in the top right of Fig. 2, "living room" in the top left of Fig. 3a, " a body of water" in the top right of Fig. 3a, "flags" in the bottom right of Fig. 3a, "suitcase" and "floor" in the top left of Fig. 3b, "floor" in the bottom left of Fig. 3b, *etc.* Wrongly predicted words are marked in red in the examples. It can be concluded that the proposed model can produce image descriptions in a more accurate manner.

A.2 Additional Details for Experiments

We use SpaCy[‡] as the tokenizer and pos tagger to process the reference captions, with lowercase and punctuation removal. Tokens whose occurrence less than 5 are treated as unknown words in our vocabulary. The reference vocabulary size is 10,201, whereas the most frequent 1,000 words from reference sentences are used as the visual-concept vocabulary. We set the maximum length of sentences as 20, in which all input sequences are right padded with 0s or cut on the right hand side. For hyperparameter tuning, we test the decoder layer number $N_{\text{dec}} \in \{3, 6\}$ and found that model with 6 layers slows down the training speed but achieves the similar performance.

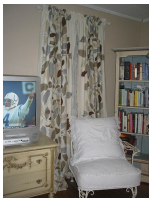
RL training takes the majority of time costs due to the reward computation. Specifically, it took around 4/2/1.5 hours to train one epoch for RL/MLE/tag training in our experiments.

[‡] https://spacy.io/

A.3 Ablation Study Curves

To investigate the necessity of explicit training on visual tags, we add additional pretraining stage on the visual concept detection as a multi-label classification task. Fig.5 compares the performance of proposed model with pretraining for 10 epochs using a binary cross-entropy loss (referred as *+tag pretraining*), and the counterpart without the pretraining, *i.e., RefineCap*, on the validation set. As shown in Fig.5, *RefineCap* conducts supervised training using cross-entropy loss as the weight initialization over the period of the beginning 25 epochs, followed by reinforcement learning process. Model with tag pretraining has three different stages: tag pretraining from epoch 0-10, supervised training from epoch 10-35, and finally reinforcement learning after 35-th epoch.

It can be seen that the overall trend of CIDEr-D scores, BLEU-1, BLEU-4 curves for models with tag pretraining witnesses the degeneration of performance, with no obvious change but a slight decrease, and the extra time and computing costs in the first 10 epochs. Besides, over the period of reinforcement learning, the rewards and reward baseline values of *+tag pretraining* models have decreased with 0.05 shift due to the different starting points but maintain a similar increase slope as *RefineCap*. Thus we extrapolate that explicit visual-concept pretraining could impede the weight initialization for the following training process in our model. We also observed a clear performance gain with REINFORCE algorithm in the final stage.



Visual concept:
blue(1), white(1),room(1), brown(1), old(0.997), television(0.980), living(0.970)

Baseline: a bedroom with a white bed and a tv.

Ours: a living room with a white bed and a television.

Reference:
A fluffy white chair that faces away from a television.
A pillow covered reading chair in the corner of the living room.
The living room is empty with the television on.
White ornate seat in nicely decorated room with television.
A white chair, books and shelves and a tv on in this room.



Visual concept:
black(1), white(1), bird(1), looking(1), picture(1), air(0.998), water(0.995), flying(0.979)

Baseline: a flock of birds flying over a beach.

Ours: a flock of birds flying over a body of water.

Reference:
A flock of small birds flying in the sky over the water.
Several birds that are flying together over a body of water.
A flock of birds flying over a field.
A black and white image showing birds flying over a body of water.
A group of birds flying over the water.



Visual concept:
blue(1), white(1), picture(1), looking(1), standing(1), red(1), black(0.999), street(0.998), walking(0.917), brown(0.824)

Baseline: a cow standing on the side of a city street

Ours: a cow walking down a sidewalk in front of a store .

Reference:
A cow standing near a curb in front of a store.
There is a cow on the sidewalk standing in front of a door.
A cow on the sidewalk on a corner in front of a store.
Cow standing on sidewalk in city area near shops.
A cow on a city sidewalk in front of a business.



Visual concept:
white(1), blue(1), red(1), grass(0.991), green(0.990), bird(0.972), small(0.964), water(0.950), wooden(0.843), field(0.733)

Baseline: a boat with an american flag on the back of it.

Ours: a boat with flags on a boat in the water.

Reference:
A swan is floating down the river by the boat.
A boat with flags and tents is docked next to a grassy bank.
A boat that is decorated with flags on the water.
A parked boat with some items inside of it.
A large red boat sitting next to a lush green shore.

(a)



Visual concept:
young(1), little(1), sitting(1), brown(1), small(1), black(1), green(0.998), red(0.991), playing(1), room(0.942), looking(0.902), luggage(0.881)

Baseline: a baby sitting on a chair with a laptop.

Ours: a baby sitting in a suitcase on the floor.

Reference:
Young boy sitting on top of a briefcase
A young baby sits on top of a briefcase.
A little boy sitting on a suitcase on the floor.
A small child sitting on top of a brief case.
A toddler boy is sitting on a brief case.

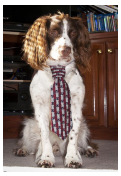


Visual concept:
looking(1), red(1), large(1), airplane(1), car(0.981), old(0.972), field(0.970)

Baseline: airplanes parked at an airport with a large window.

Ours: airplanes are parked at an airport on the tarmac.

Reference:
An airport filled with planes sitting on tarmacs.
The view of runway from behind the windows of airport.
A truck driving towards some planes parked on the runway
Planes on a wet tarmac unloading at arrival gates.
Window view from the inside of airplanes, baggage carrier and tarmac.



Visual concept:
large(1), blue(1), black(1), white(1), looking(1), brown(1), old(1), dog(0.999), wearing(0.998), sitting(0.983),room(0.966), big(0.944)

Baseline: a dog wearing a tie sitting on a table.

Ours: a dog wearing a tie sitting on the floor.

Reference:
A brown and white dog wearing a neck tie.
A brown and white dog wearing a tie on carpet.
A dog wearing a tie poses for the camera.
A dog is sitting with a neck tie on.
A close up of a dog wearing a tie.



Visual concept:
white(1), large(1), red(1), small(1), plate(1), pizza(1), little(1), green(1), picture(1), open(0.982), big(0.893),

Baseline: a pizza on a white plate on a table.

Ours: a pizza with tomatoes and cheese on a white plate.

Reference:
An Italian dish is presented on a white plate.
Baked pizza with red tomatoes and green olives.
A plate of food cut into slices has onions and olives on it.
A pizza sliced in four slices on a plate.
Onions, tomatoes and olives atop a plate on a table.

(b)

Figure 3: Supplemental examples of the generated captions.



Visual concept:
large(1), blue(1), white(1), looking(1), kite(0.998), flying(0.997), person(0.993), water (0.937)

Baseline: two kites flying in the sky on a beach.

Ours: two kites flying in the sky over the ocean.

Reference:
A kite flying through a blue sky with a long tail.
A colorful kite is flying over the ocean.
A kite flying through the sky on a clear day.
A very pretty kite is flying high in the sky.
There are different kites flying in the sky.



Visual concept:
large(1), white(1), looking(1),green(0.998), old(0.997), woman(0.996), standing(0.992), blue(0.978), brown(0.960), young(0.851), grass(0.791), people(0.735), group(0.585)

Baseline: a crowd of people looking at an airplane on a runway.

Ours: a group of people standing around a large plane on a runway.

Reference:
A group of people observing two planes at an air show.
A family walks on a runway near huge planes.
People walking towards aircrafts on display outside during the day.
People walking around two different Air Force airplanes.
Large, air force airplanes sit on a runway while tourists look at them.



Visual concept:
woman(1), black(1), person(1), picture(1), table(1), red(0.999), yellow(0.996), holding(0.988), hand(0.978), green(0.967), food(0.960), wearing(0.531), kitchen(0.516)

Baseline: a woman standing at a table with a tray of food.

Ours: woman holding a tray of food in a kitchen.

Reference:
A woman standing over a pan filled with food in a kitchen.
A woman smiling while she prepares a plate of food.
A smiling woman standing next to a plate of food she made
A woman in a bright pink summer shirt smiles and displays a party platter she has made.



Visual concept:
black(1), red(1), person(1), picture(1), white(1), man(1), street(1), people(0.998), motorcycle(0.995), wearing(0.994), parked(0.954), city(0.792), light(0.782), games(0.664), bus(0.627)

Baseline: two people riding on a motorcycle on a city street.

Ours: two people riding on a motorcycle in the rain.

Reference:
two people on motorcycles car and trees and water.
Motorcyclists waiting in traffic on a rainy day.
Two individuals sit on motorcycles on a busy street in the rain.
Two motorcyclists next to one another on damp street.
Two people in helmets sitting on a motorcycle behind a car.

(a)



Visual concept:
white(1), blue(1), large(0.998), standing(0.997), people(0.988), man(0.976), building(0.962), outside(0.882), woman(0.832), group(0.822)

Baseline: a group of people standing in front of a building.

Ours: a group of people standing outside of a building with a clock.

Reference:
A bunch of people sit in an open court yard.
A small group of people standing around a ball patio.
A group of people walking around a parking lot.
A group of people in front of a white building.
Many people on a courtyard under a clock.



Visual concept:
looking(1), red(1), large(1), airplane(1), car(0.981), old(0.972), field(0.970)

Baseline: a woman riding a skateboard on a sidewalk.

Ours: a woman riding a skateboard in front of a store.

Reference:
A girl is skateboarding down the Hollywood walk of fame.
A woman who is skateboarding down the street.
A woman with glasses and a scarf skateboards along Hollywood's Walk of Fame.
Lady in front of a store standing on a pink skateboard
A woman is riding her skate board down the sidewalk.



Visual concept:
red(1), white(1), big(1), snow(0.998), street(0.997), street(0.997), stop(0.980), large(0.965), sign(0.951), snowy(0.893), skis(0.903), ski(0.882), field(0.767)

Baseline: a stop sign is covered in snow in a

Ours: a stop sign and a street sign in the snow .

Reference:
A stop sign and street sign encased in snow and ice.
A couple of street signs sticking out the side of a ski slope.
A stop sign and street sign encased in a wall of snow.
A stop sign is buried in a large pile of snow.
A large pile of snow with a stop sign and a street sign poking out.



Visual concept:
yellow(1), young(1), food(1), brown(1), white(1), giraffe(1), standing(1), looking(1), large(0.997), walking(0.867), couple(0.856)

Baseline: two giraffes standing next to each other in a zoo.

Ours: two giraffes are eating from a basket in a zoo.

Reference:
Two giraffes stand and eat food out of a basket.
Two giraffes looking for food in an empty feeding basket
A couple of giraffes reach for a basket
Two giraffes that are eating from a basket.
Two giraffes eating from a basket on a pole.

(b)

Figure 4: Supplemental examples of the generated captions.

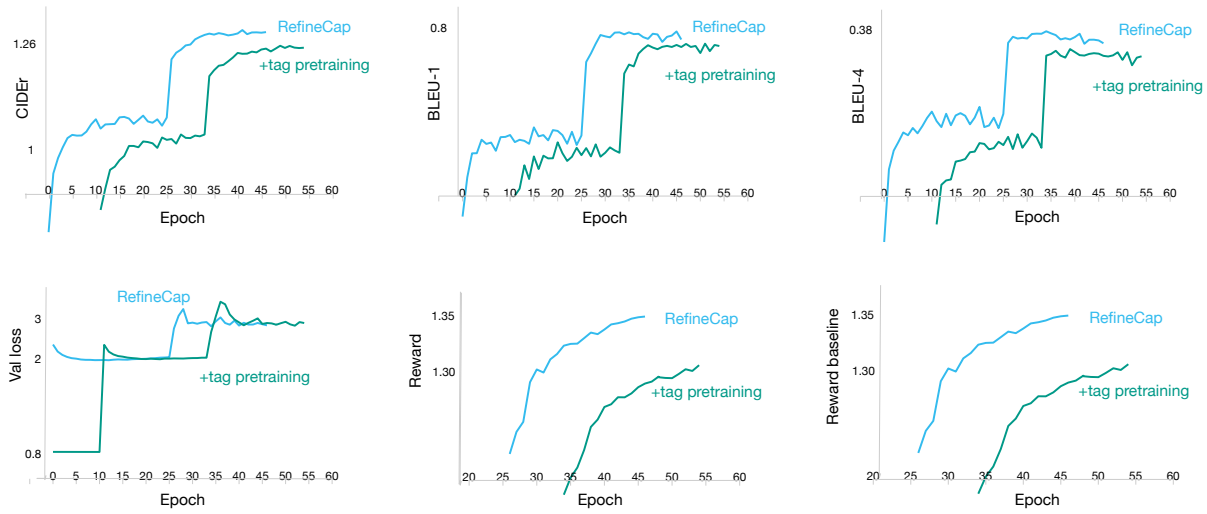


Figure 5: The performance curves of *RefineCap* w/ and w/o tag pretraining (in blue / green separately) on evaluation set, including CIDEr-D/BLEU-1/BLEU-4 metrics, validation loss, rewards and baseline reward values.