***GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation***

A deep learning network model, GLAC Net, that generates visual stories by combining global-local (glocal) attention and context cascading mechanisms.

In the image sequence encoders, the global context of the storyline is encoded using bi-directional LSTMs on features of five images, we give attention on the context (global attention). Additionally, we give local attention to image features directly. Then both of them are combined and sent to RNN-based sentence generators. While standard attention configuration needs a large number of parameters, we implement them in a very simple way via hard connections from the outputs of encoders or image features onto the sentence generators. To improve further the coherency of the generated stories, we design to convey the last hidden vector in the sentence generator to the next sentence generator as an initial hidden vector

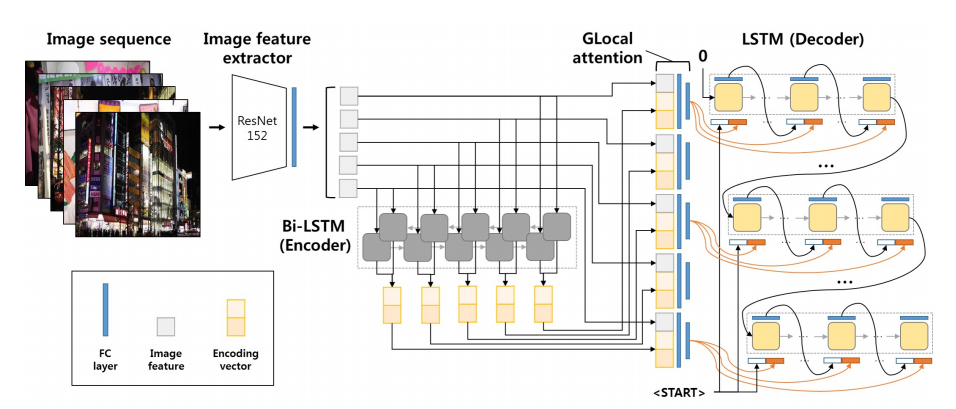
2 key ideas:

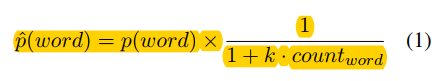
(1) utilizing two-level attention mechanism in the encoder part.

(2) conveying the hidden state to the next sentence generator.

Model:

(1)利用ResNet-152提取每张图像的特征(He et al.， 2015)。(2)将提取的特征按顺序输入到bi-LSTM中，使图像的上下文均匀地反映在整个故事中。由bi-LSTM输出和特定于图像的特征组成的glocal向量经过完全连接的层。然后，它被连接到单词标记，以便用作解码器的输入。





(where k is a constant for sensitivity.)

we apply batch normalization and dropout layers to prevent overfitting and improve the performance.

***Adversarial Reward Learning for Visual Storytelling:***

Specifically, we first incorporate a Boltzmann distribution to associate reward learning with distribution approximation, then design the adversarial process with two models – a policy model and a reward model.

Our main contributions are four-fold:

• We propose an adversarial reward learning framework and apply it to boost visual story generation.

•We evaluate our approach on the Visual Storytelling (VIST) dataset and achieve the state-of-the-art results on automatic metrics.

• We empirically demonstrate that automatic metrics are not perfect for either training or evaluation.

• We design and perform a comprehensive human evaluation via Amazon Mechanical Turk, which demonstrates the superiority of the generated stories of our method on relevance, expressiveness, and concreteness.

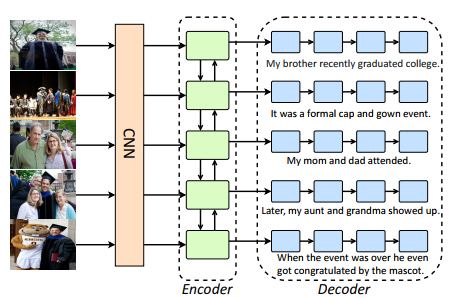


Figure 3: Overview of the policy model. The visual encoder is a bidirectional GRU, which encodes the high-level visual features extracted from the input images. Its outputs are then fed into the RNN decoders to generate sentences in parallel. Finally, we concatenate all the generated sentences as a full story. Note that the five decoders share the same weights.

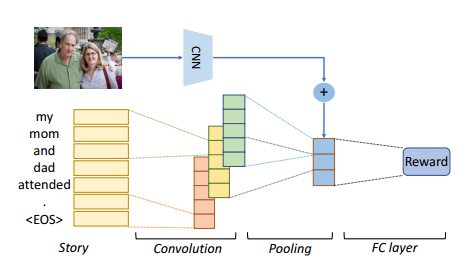


Figure 4: Overview of the reward model. Our reward model is a CNN-based architecture, which utilizes convolution kernels with size 2, 3 and 4 to extract bigram, trigram and 4-gram representations from the input sequence embeddings. Once the sentence representation is learned, it will be concatenated with the visual representation of the input image, and then be fed into the final FC layer to obtain the reward.

使用不同内核大小的多个卷积层来提取n-g特征，然后通过汇聚层将这些特征投射到句子级表示空间中，将句子表示与输入图像的视觉特征通过拼接结合起来，并将其送入最终的全连通决策层。最后,奖励模型输出估计奖励价值Rθ(W)。



where φ denotes the non-linear projection function, Wr, br denote the weight and bias in the output layer, and fconv denotes the operations in CNN. ICNN is the high-level visual feature extracted from the image, and Wi projects it into the sentence representation space. θ includes all the parameters above.

Reward Boltzmann Distribution/Adversarial Reward Learning见论文内第三部分