1. **Automatic caption generation for news images**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2013 | Feng, Y., & Lapata, M. | IEEE | 77 | Statistic-based Model |

该篇论文是基于统计的模型，其提出两种模型（extractive model 和 abstractive model），一般基于统计生成caption的模型有两个重要的步骤：content selection 和 surface realization。关于前者作者给出了3种方法：常见的表示方法（通用）， LDA（针对文档）和 Image Annotation model（针对图片），当得到 topic（文档） 和 keyword（图片） 后，就能进行 surface realization，在后面这步，作何提出全文重心的2种生成模型，但extractive model 因为是通过抽取文档内相似的词，因此没有第二种 abstractive model 好。

1. **Deep fragment embeddings for bidirectional image sentence mapping**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2014 | Karpathy, A., Joulin, A., etc. | NIPS | 417 | SL, MS, EDA, WS |

该论文首先用 Region CNN (RCNN) 检测图片多个目标，然后针对不同的图片块使用CNN去生成 CNN representations （即 image fragment）, 同时使用 Dependency tree Relations (DTR) 针对相应的图片描述生成 sentence fragment。之后 inner product 两个fragments 得到 image-sentence similarity, 这个similarity 矩阵之后会经过Fragment Alignment Objective 及 Multiple Instance Learning extention对每个similarity 进行处理，处理后，使用 Global Ranking Objective 总和之前的结果使得提出的 image object 能和拆分的句子成分相对应。

该模型主要用在Image Annotation和Image Search上，无论怎样都要求输入得有图片及相应描述。其并不属于那种只要输入图片即可生成描述的模型。缺点及限制有：在Align视觉成分和部分句子成分时，有些单个视觉成分会被拆分成多个句子成分，例如：一只黑白狗图片成分和（CONJ，黑，白）及（AMOD, 白， 狗）。此外，有些词成分（如：each other）无法被理解，因此图片搜索的时候使用each other的字眼并不能带来任何帮助。不能理解数量概念和空间概念，因为模型在视觉上只是识别了object而已。

1. **Deep visual-semantic alignments for generating image descriptions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2015 | Karpathy, A., & Fei-Fei, L | IEEE | 2240 | SL, VS, EDA, DC & WS |

该论文跟一作上一年的论文十分相似，不同点在于在原因基础上进行了改动，加入一个多模态RNN模型使得能够实现描述生成功能。总的来说，分为两个部分，第一个部分是负责生成visual-sentence pairs，第二部分是用上一部分的输出训练新的RNN。具体来说，首先，用RCNN将图像检测出不同的object regions，再用CNN相应给每个oject region生成representations (V\_i)，于此同时用BRNN（双向RNN）将图像的描述生成h-dimensional words vectors (s\_t)， 使用V\_i^T\*S\_t将双方嵌入到公共空间里，最大化objective function，学习好模型后，使用Markov Random Field (MRF) 控制生成caption长度的力度，此时学习完毕后的BRNN模型可用于标注任何输入的图像并输出visual-sentence pairs的alignment结果，截止到这里为止，所得到的模型只是个图片标注模型（要求输入图片和相应描述），而为了实现图片描述生成，之后会利用标注好的pairs训练一个新的多模态RNN模型，最后训练好的这个模型只需要输入图像即可得到完整的caption。

缺点在于三点：该方法不同实现滑动小窗口不断给出新的caption，给定一张图片就只能生成特定region内的描述。第二点，在第二部分将图像特征放入多模态RNN中是，只是通过简单的加性方法加入到第一个隐藏层当中，没有考虑其它更复杂的方法。最后一点，这不是end-to-end。

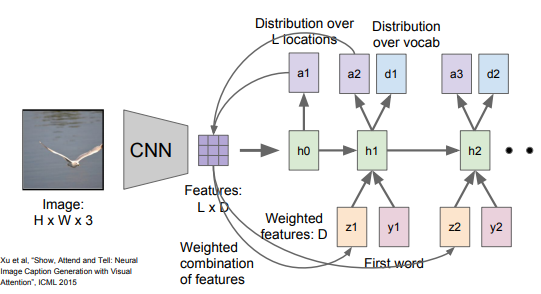
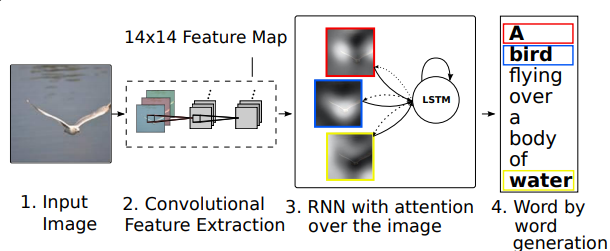
1. **Show and tell: A neural image caption generator**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2015 | Vinyals, O., Toshev, A., etc. | IEEE | 2324 | SL, VS, EDA, WS |

|  |  |
| --- | --- |
|  | 该论文启发了Show, attend and tell: Neural image caption generation with visual attention这篇论文，原理都差不多，只是没有加入关注机制进模型。另外该模型称为NIC模型，有兴趣的两点是：使用了word embedding vectors而不是简单的独热编码，另外，讨论了人工评估模型结果与state-of-art的一些matrics之间的不同，并采用了累计图可视化人工测量结果，这个可借鉴到我的毕业论文当中去。  该篇论文还有一篇专门将其具体应用在2015年 MSCOCO比赛的经验分享，大致内容不变，不过有更多关于实验的细节描述：**Show and tell: Lessons learned from the 2015 mscoco image captioning challenge** |

1. **Show, attend and tell: Neural image caption generation with visual attention**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2015 | Xu, K., Ba, J., etc. | ICML | 2865 | SL, VS, EDA, WS, A |



该模型提出了soft和hard两种关注机制，首先使用CNN提出输入图像的特征，然后作为context vector z (不同关注机制具有不同的计算方法，同样地损失函数和梯度计算也因此而不同)输入到LSTM模型当中，运行了第一个step后，就对应生成下一个关注位置的分布，将关注分布与图片特征权重结合后得到一个权重特征，该权重特征随后会被跟文本描述中的分词一起输入到LSTM第二步，第二步的输出会生成位置分布和词分布，重复上述过程直至结束。

关注机制与使用object detection的区别在于，后者不关注non-object region，而前者会关注，这因此基于关注机制的模型提高了图像描述生成的表现。在论文中有提到，hard关注比soft关注好，可能是因为soft加入了正则化，使得关注点object与其它object界定模糊，容易混淆识别的内容。另外，关注机制在运行中把特定的图片特征考虑到LSTM里的每一步里，这一点相比于那些把图像特征只放入第一个隐藏层就完了的模型来说，Attention模型对图像特征考虑更充足。

1. **Image captioning with deep bidirectional LSTMs**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2016 | Wang, C., Yang, H., etc. | ACM | 74 | SL, VS, EDA, WS |

|  |  |
| --- | --- |
| image.png | 该篇论文提出了多模态双向LSTM模型，该模型主要由四部分组成：第一部分是用于编码图像输入的CNN，用于编码文本输入的文本LSTM（Text-LSTM, T-LSTM）和用于把视觉和文本向量嵌入到一个共同的语义空间并解码成句子的多模态LSTM（Multimodal LSTM， M-LSTM）。最后一部分是softmax层。这个双向LSTM是由两个分开的LSTM层共同施行的，他们分别用于计算前向隐藏序列\*h\*（附前向箭头）和后向隐藏序列\*h\*（附后向箭头）。 |

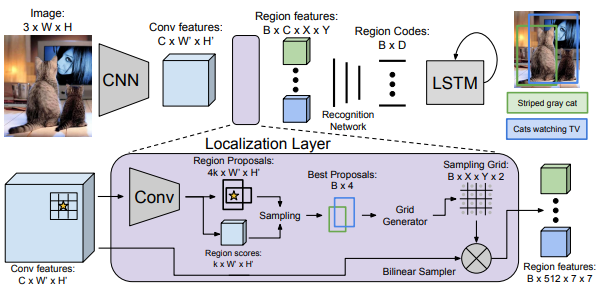
**7. Image captioning with semantic attention**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2016 | You, Q., Jin, H., etc. | IEEE | 413 | SL, VS, CA, WS, A |

|  |  |
| --- | --- |
|  | 作者把image captioning方法分为 top-down 和bottom-up两种，由于两种方法各有优势，前者可以实现端对端，但缺乏细节关注，可后者虽然不能实现端对端但能将细节融入训练过程中。本篇论文的模型结合了这两种方法，但都是作用在图像上，top-down方法就是用CNN把图像的全局特征抽取出来，bottom-up方法则是使用视觉属性预测（Visual attribute prediction）获取图像的属性。之后将全局输入LSTM先，然后属性输入到LSTM的输入节点和输出节点上，同时文本被Glove处理后也依次序列输入到模型中。  论文讨论了3种将属性融入到模型的方法，他们分别是attention model (ATT)，Element-wise max (MAX) 和 concatenation (CON)，最后表现好的是ATT。之后作者又讨论了3种属性预测方法，其分别是非参数方法k-NN和参数方法Multi-label Ranking和FCN，实验发现FCN最好。综合最后，作者把模型称为ATT-FCN。  该模型的优点在于提出了属性，这样在模型生成的描述会更加具体，坏处是受属性影响，一旦出现错误的属性，生成文本质量就会变差。 |

**8. Densecap: Fully convolutional localization networks for dense captioning**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2016 | Johnson, J., Karpathy, A., etc. | IEEE | 444 | SL, VS, EDA, DC |



该论文提出Fully Convolutional Localization Network (FCLN)实现端对端的dense captioning。模型先用CNN抽取图像特征，再使用一个自定义的Localization Layer抽取出不同的Region features，之后使用Recognition Network（即FCNN）得到Region Codes并输入到LSTM语言模型得到dense captions。本文的重点是Localization layer，在该层中，用Fast R-CNN在每个像素点提出k种不同scales的regions同时也给出相应的confidence scores。最后使用某种采样方法，采样出合适的Regions，之后用Grid Generator和Bilinear Sampler得到Region features。

该论文的Dense Caption生成的不错，同时信息检索的结果也很好。有源码公开。

**9. Knowing when to look: Adaptive attention via a visual sentinel for image captioning**

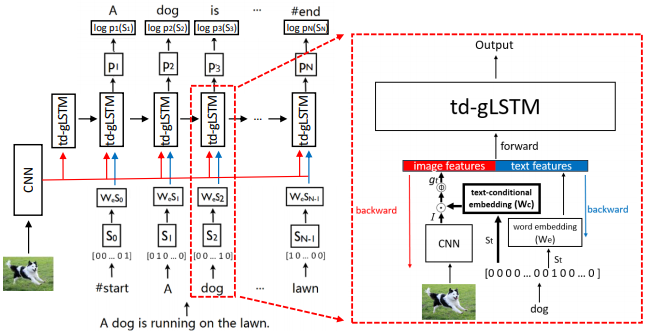
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2017 | Lu, J., Xiong, C., Parikh, D., etc. | IEEE | 225 | SL, VS, EDA, WS, A |

该论文同样使用了关注机制，但因为有时候像句子中的“the”和“of”就没必要进行视觉关注，所以模型除了关注图片哪里该关注，还应该实现自适应控制什么时候该关注，即该模型能够自适应调节什么时候依赖visual signals和什么时候依赖于语言模型。模型的重心在于有个”visual sentinel”其考虑了memory cell里的信息，独立于从图片提取的spatial image features。关于text部分，模型用LSTM。关于image部分，模型从三个方面考虑：（1）输入cell上的图片全局特征；（2）visual sentinel vectors (s\_t) - 控制where to look； （3）Spatial image features - 控制when to look。另外，在模型结构上，作者受残差网络的启发，改了原有的soft attention model。这应该是目前state-of-art的最好的基于attention机制模型。

|  |  |
| --- | --- |
|  |  |

**10. Watch what you just said: Image captioning with text-conditional attention.**

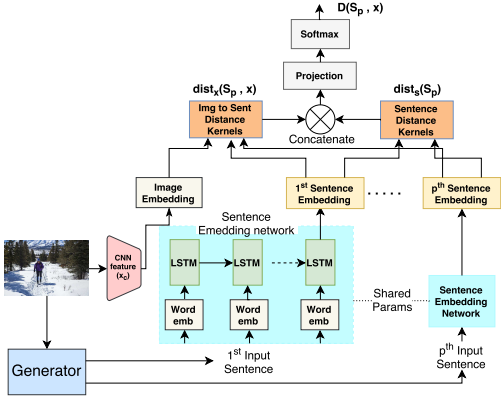
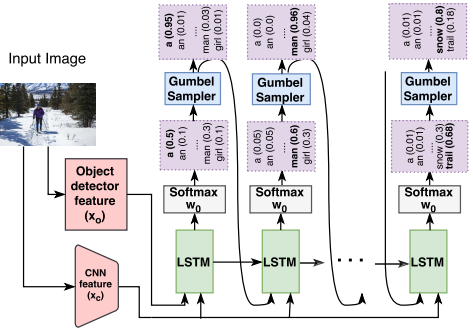
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2017 | Zhou, L., Xu, C., etc. | ACM | 18 | SL, MS, EDA, WS, A |



|  |  |
| --- | --- |
|  | 该模型叫time-dependent guiding LSTM（td-gLSTM）。 而之前的gLSTM是time-independent。有时候图像并没有足够的证据反映真实描述，比如图片上，一个男人坐着面对电视，而没怎么露出沙发，可是baseline的描述是该男人坐在沙发上看电视。因此如果光把重心放在图像特征上有些不妥，相反应该基于描述文本去学习推断图像中的不充足信息（即模型应学习当描述生成到哪就看图片语义特征的哪块）。另外，该模型是end-to-end。  简单来说，该论文的特别之处在于text attention，而image attention是由text condition的。最后结果表现还行但不够惊艳，只是进步了1-2的百分点。作者后续考虑引入attributes和regions in feature map进去，同时由于模型易过拟合，即使是MSCOCO也是，所以他们也考虑用一些弱标记的数据集。 |

**11. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2017 | Shetty, R., Rohrbach, M., etc. | IEEE | 42 | ODL, VS, EDA, WS |

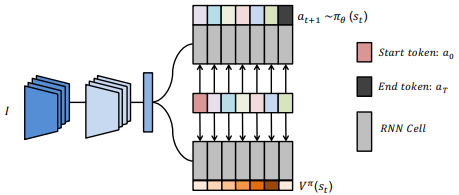


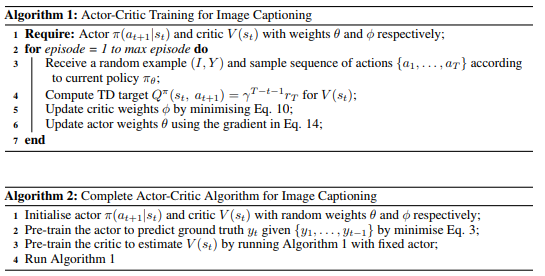
该论文提到传统模型生成的描述通用性不错，但多样性差，为此加入了GAN进入模型当中。上左图为生成器，上右图为判别器。关于生成器，输入图片抽取两种特征类型，一种是经过Faster-RCNN目标检测的特征，另一个是ResNet的CNN特征，输入到模型后，输出时需要使用Gumbel Sampler，该样本器的作用是使得判别器的输入是连续型的。如果使用softmax的连续型概率结果输入到判别器中，判别器很清楚地就能识别出真假来，无法达到训练结果。如果使用独热编码结果，由于是离散型，判别器无法实现反向传播机制。因此Gumbel Sampler在这是很必要的，也有其它使数值连续化的方法，例如RL的policy gradient，但是它耗计算机资源。关于判别器，它会衡量图片与句子的距离（即衡量句子是否能够正确地描述图片），也同时衡量多个生成描述之间的距离（即衡量生成描述的多样性）。

在结果方面，虽然在常见的测量方法上没有提高，但是如果使用人工测量，会得到GAN生成的描述更偏向于人工生成描述的结论。同时，作者在研究结果多样性的方法值得参考和借鉴。

**12. Actor-critic sequence training for image captioning**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2017 | Zhang, L., Sung, F., etc. | NIPS | 19 | ODL, MS, WS |





作者提出一种基于actor-critic reinforcement learning（RL）增强学习的图片描述生成方法。即使没集成模型，在MSCOCO也排到了第三的位置，准确率很高，该算法主要由两部分组成：Actor（Policy Network）和Critic（Value Network）。前者生成行动，后者评价行动。算法首先需要预训练好这两个网络，再一起用采样的样本去实施正常的模型训练。两种网络都使用了LSTM。

该模型有三个比较重要的地方：第一是采样（克服了train和test的不匹配问题），第二是Advantage Function（帮助policy gradient方向能朝着增加优于平均的行为的发生概率的方向移动），第三个是预训练actor和critic（否则梯度易消失，同时双方也不能提供较强的训练信号给对方）。值得一提的是：RL的准确率很高，这里的原因估计是它的reward机制结合了CIDEr测量方法的缘故，而非RL方法用的都是cross entropy loss，多少还是有点差距的。另外，虽然作者说该模型效率不错，但个人感觉还是挺耗计算机资源的，毕竟可选actions（即tokens）很多。

**13. Self-critical sequence training for image captioning**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2017 | Rennie, S. J., etc. | IEEE | 281 | ODL |

Rennie et al. utilized a REINFORCE algorithm called self-critical sequence training (SCST) to optimize the image captioning system. SCST is to use inference algorithm at test time to normalize the reward instead of estimating the reward signal and normalization during training. According to their empirical finding, this approach with the test-time decoding technique is effective in optimizing non-differentiable metrics such as CIDEr.

该篇论文还有一部分讨论的RL各部分（例如actor，reward）在序列生成中扮演哪些不同的角色，如有需要，可以补充加入进来。

**14. A Comprehensive Survey of Deep Learning for Image Captioning**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2019 | Hossain, M. D., Sohel, F., etc. | CSUR | 2 | - |

基于深度学习的图像描述的文献综述。

**15. Long-term recurrent convolutional networks for visual recognition and description**

Donahue er al. \cite{donahue2015long} presented an end-to-end model called Long-term Recurrent Convolutional Networks (LRCNs). This model stacks two-layer LSTMs for sequence learning.

**16. Deep captioning with multimodal recurrent neural networks (m-rnn)**

Mao et al. \cite{mao2014deep} introduced a multimodal Recurrent Neural Networks (m-RNN) that includes the global image features into each time step. There are four types of layers in m-RNN and they are the word embedding layer, the recurrent layer, the multimodal layer and the softmax layer. The multimodel layer maps the visual component (i.e., the image representation) and the language components (i.e., the outputs of the word embedding layer and the recurrent layer) into a common space. Then, the softmax layer is responsible to predict the next word based on the output of the multimodal layer.

**RL-based Models:**

**2015 Sequence level training with recurrent neural network**

The action space is determined by the vocabulary size so the search space of the whole generation will be very large. Hence, it is challenging for the model to start with the initial random policy. To deal with the problem of the large search space, Ranzato et al. \cite{ranzato2015sequence} proposed an algorithm called Mixed Incremental Cross-Entropy Reinforce (MIXER) that uses curriculum learning and a loss function which mixes cross-entropy loss (XENT) and REINFORCE. Firstly, this model trains RNN with XENT to gain the optimal policy. Next, the model replaces the initial random policy, which benefits the model to avoid selecting the poor initial policy from the search space. Then, MIXER continues the training with the XENT-REINFORCE loss and incremental learning.

@article{ranzato2015sequence,

title={Sequence level training with recurrent neural networks},

author={Ranzato, Marc'Aurelio and Chopra, Sumit and Auli, Michael and Zaremba, Wojciech},

journal={arXiv preprint arXiv:1511.06732},

year={2015}

}

**2017 Improved image captioning via policy gradient optimization of SPIDEr**

Liu et al. \cite{liu2017improved} found that although MIXER can optimize BLEU-4, it is hard to reproduce this model to target other metrics. Therefore, they use Monte Carlo rollouts to estimate the value at each intermediate action rather than mixing maximum likelihood estimation into the training. This method benefits the convergence to be more efficient and the model becomes robust to the hyperparameter tuning but MIXER does not. Further, this paper linearly combines SPICE and CIDEr as a new metric called SPIDEr. Optimizing SPIDEr leads to the generated captions are fluent and semantically reliable.

@inproceedings{liu2017improved,

title={Improved image captioning via policy gradient optimization of spider},

author={Liu, Siqi and Zhu, Zhenhai and Ye, Ning and Guadarrama, Sergio and Murphy, Kevin},

booktitle={Proceedings of the IEEE international conference on computer vision},

pages={873--881},

year={2017}

}

**2017 Deep Reinforcement Learning-based Image Captioning with Embedding Reward**

For generating captions that satisfy different metrics at the same time, Ren et al. \cite{ren2017deep} introduced an actor-critic RL model to train ‘policy network’ and ‘value network’. The policy network locally guides the model to sample the next word on the basis of the current state. On the other hand, the value network computes the reward of the action by involving the consideration of all possible future predictions, which can avoid of exposure bias problem and can guide the model globally. Specifically, the policy network consists of CNN and RNN while the value network contains CNN, RNN and Multilayer Perceptron (MLP). These two networks are firstly trained by supervised learning and then use RL with curriculum learning to train them. Additionally, in order to generalize different evaluation metrics instead of only optimizing a certain metric, \cite{ren2017deep} defines the reward as the visual-semantic embedding similarities. As a result, this model has well scores across many metrics.

@inproceedings{ren2017deep,

title={Deep reinforcement learning-based image captioning with embedding reward},

author={Ren, Zhou and Wang, Xiaoyu and Zhang, Ning and Lv, Xutao and Li, Li-Jia},

booktitle={Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition},

pages={290--298},

year={2017}

}

**2017 Actor-Critic Sequence Training For image captioning**

In 2017, Zhang et al. \cite{zhang2017actor} also proposed an image captioning model by using actor-critic training but the framework is encoder-decoder rather than the decision-making framework in \cite{ren2017deep}. About the architecture, this model utilizes CNN as the encoder and LSTM as the decoder. Particularly, the policy network and value network are both comprised of LSTM. Also, a finite Markov decision process (MDP) is applied to generate the caption. This simple model is not just computational cheap but also can gain competitive performance on the image captioning competition.

@article{zhang2017actor,

title={Actor-critic sequence training for image captioning},

author={Zhang, Li and Sung, Flood and Liu, Feng and Xiang, Tao and Gong, Shaogang and Yang, Yongxin and Hospedales, Timothy M},

journal={arXiv preprint arXiv:1706.09601},

year={2017}

}

**2018 Stack-captioning: coarse-to-fine learning for image captioning**

Many previous image caption generator follows the one-stage generation framework so it is challenging to generate detailed captions. Therefore, Gu et al. \cite{gu2018stack} introduced an RL-based multi-stage framework. The model encodes image with CNN and decodes sentence with a coarse-to-fine decoder that includes a coarse LSTM and a series of attention-based fine LSTM. Although the multi-stage framework refines the generated caption from coarse to rich, it faces the risk of vanishing gradient. To overcome the problem, each stage has to include intermediate supervisions with a cross-entropy-based loss function. Meanwhile, the training is RL-based and those LSTM decoders will be the agents. Consequently, this multi-stage model is able to output more descriptive image captions.

@inproceedings{gu2018stack,

title={Stack-captioning: Coarse-to-fine learning for image captioning},

author={Gu, Jiuxiang and Cai, Jianfei and Wang, Gang and Chen, Tsuhan},

booktitle={Thirty-Second AAAI Conference on Artificial Intelligence},

year={2018}

}

**2018 Context-aware visual policy network for sequence-level image captioning**

In the aforementioned RL-based models, only \cite{gu2018stack} utilizes visual attention but not just focus on the language decoding. Liu et al. \cite{liu2018context} also use visual attention in their model and propose a Context-Aware Visual Policy network (CAVP) for generating the visual representation. Unlike the tradition attention that only concentrates on a certain part of the image at each time step, CAVP is able to make the model consider multiple visual compositions. This model is comprised of a CAVP and a language policy network. The language policy network takes the output of CAVP for generating the image description. During the training, the actor-critic policy gradient is implemented to optimize the above two policy networks. The result shows the RL-based image captioning approach achieve state-of-art performance on MSCOCO.

@article{liu2018context,

title={Context-aware visual policy network for sequence-level image captioning},

author={Liu, Daqing and Zha, Zheng-Jun and Zhang, Hanwang and Zhang, Yongdong and Wu, Feng},

journal={arXiv preprint arXiv:1808.05864},

year={2018}

}

**GAN-based Models:**

**2017 Towards diverse and natural image descriptions via a conditional GAN**

The first GAN-based image captioning framework is proposed by Dai et al. \cite{dai2017towards}. It is called Conditional Generative Adversarial Networks (CGAN) where the generator conditions on images to generate captions and the evaluator measure the quality of the generated sentences. However, training GAN-based image captioning models have two main difficulties. First of all, the discrete output from RNN generator is non-differentiable so it hinders the back-propagation. For solving these two problems, \cite{dai2017towards} suggested Policy Gradient into the training to make gradients continuous. Also, based on Monte Carlo rollouts, the framework estimates the future reward during the caption generation, which returns early feedback back to the generator so it alleviates the vanishing gradient problem. In general, this framework not just involve GANs but also the RL technique.

@inproceedings{dai2017towards,

title={Towards diverse and natural image descriptions via a conditional gan},

author={Dai, Bo and Fidler, Sanja and Urtasun, Raquel and Lin, Dahua},

booktitle={Proceedings of the IEEE International Conference on Computer Vision},

pages={2970--2979},

year={2017}

}

**2017 Language generation with recurrent generative adversarial networks without pre-training**

So far, there is little research that applies GANs to image captioning but GAN-based text generation is popular. However, the image captioning is basically a text generation task but with an extra consideration (i.e., image), Therefore, the non-differentiable problem of GAN-based image captioning also is the difficulty for GAN-based language generation. Except for Gumbel Sampler in cite\{shetty2017speaking}, there are other methods in the area of GAN-based language generation can help to tackle the non-differentiable problem.These methods are more likely to be referred by GAN-based image captioning. For instance, In order to gain a differentiable generator in the character-level generation, Press et al. \cite{press2017language} sum up the product of the probability of the characters and the corresponding embedding of these characters as the input to the next time step.

@article{press2017language,

title={Language generation with recurrent generative adversarial networks without pre-training},

author={Press, Ofir and Bar, Amir and Bogin, Ben and Berant, Jonathan and Wolf, Lior},

journal={arXiv preprint arXiv:1706.01399},

year={2017}

}

|  |
| --- |
| **1978 Digital Image Processing by Use of Local Statistics** |
| Lee investigated two types of image noise that are additive noise and multiplicative noise. Firstly, the addictive-noise image is generated by adding a random value from a uniform distribution or a Gaussian distribution on each original image pixel. Secondly, the image with multiplicative noise is made by multiply each pixel by a random value in the range (0.7, 1.0). Furthermore, these two noises can also be combined to corrupt the image. |
| **2010 Image de-noising by various filters for different noise** |
| Patidar and Srivastava presented four types of noise that often occur in digital image processing. They are Amplifier noise (i.e., Gaussian noise), Salt-and-pepper noise, Shot noise (i.e., Poisson noise) and Speckle noise. Respectively, the physical generators of these types of noise are the colour camera, the analog-to-digital converter, electronic circuit/optical device and radar. About the Amplifier noise, basically, it is the additive Gaussian noise that each noisy pixel value is the sum the truel pixel and a random value from a Gaussian distribution cite\{引用Lee 和A Comparative Study of Various Types of Image Noise and Efficient Noise Removal Techniques}. Next, the ‘salt-and-pepper’ noisy image is produced by randomly overlaying ‘salt’ pixel (with value 255) or ‘pepper’ pixel (with value 0) on each true pixel with a certain probability cite\{引用Salt-and-Pepper Noise Removal by Median-Type Noise Detectors and Detail-Preserving Regularization}. Third, Shot noise is related to each pixel's intensity. Last, Speckle noise is the same as the multiplicative noise in \cite{Lee}.  提及：meidian filter can successfully de-noise for the image with salt-and-pepper noise. |
| **2011 Efficient Technique for Color Image Noise Reduction** |
| Despite prior the finding \cite{Patidar and Srivastava}, Mythili and Kavitha investigate four more types of noise that are Quantization noise, Film Grain noise, Non-isotropic noise and periodic noise. Quantization noise is also called as Uniform noise which quantizes the image pixels to different discrete values from a uniform distribution. Film Grain appears randomly on the image with a binomial distribution or a Poisson distribution, depending on the probability of the occurrence of dark grain. Non-isotropic noise is like the combination of horizontal scratches and vertical scratches while the result of periodic noise is like the bars. |

|  |
| --- |
| **2013 Intriguing properties of neural networks** |
| In 2013, Szegedy et al. presented the neural network highly misclassify ‘adversarial examples’ that are the images with a subtle, imperceptible perturbation. Especially, a fixed set of adversarial examples can negatively affect different networks no matter what hyperparameters, such as the number of layers, activation function, weight initialization and regularization, they use. Even, the model trained by a subset of the training dataset cannot be avoided. |
| **2016 Adversarial examples in the physical world** |
| Further, Kurakin et al. showed that not only inputting adversarial examples directly into the model can pose a threat to the result, but in the physical world, these perturbed examples can also attack the machine learning model through a camera. |
| **2017 Adversarial Patch** |
| Brown et al. developed an adversarial image patch and it can cause a significant attack to the classifier by placing the adversarial salient sticker on the normal image input. Meanwhile, the attack can insist on working in the real world, regardless of the scale, location and rotation of the patch. |
| **2016 Accessorize to a crime Real and stealthy attacks on state-of-the-art face recognition** |
| Sharif et al. generated an eyeglass-frame accessory to fool the face detection systems and face recognition systems. It is a form of adversarial perturbations and the human face who wears perturbated eyeglasses can be easily impersonated to another subject.  解释本项目噪声与adversarial干扰不同的之处：Most of adversaries have to expose to the feature space, which means that the model architecture and parameters are known to the adversarial examples generator. However, in real practice, the above condition cannot usually be satisfied so the thesis prefers to focus on the more common image noise such as salt-and-pepper noise.  可用于最后讨论用，对block noise的应付方法：About developing human-like algorithms, Sharif et al. suggested that the image classification based on image attribute may perform more benefit rather than relying on the pixel values. |
| **2018 Robust physical-world attacks on deep learning visual classification** |
| Eykholt et al. proposed a physical adversarial perturbation shaped like a black and white sticker which can attack the classifier of the stop sign and the attack success rate can reach 100%.  提及一点：but the attack is black-box, which means the adversarial sticker cannot be generated without knowing the model architecture and weights. |
| **2014 Explaining and harnessing adversarial examples** |
| Goodfellow et al. stated that the adversarial training can have better regularization effect rather than using dropout technique. |