Video DeCaptioning Using U-Net with Stacked Dilated Convolutional Layers

Shivansh Mundra, Arnav Kumar Jain, and Sayan Sinha

1 Introduction

Videos often have captions embedded into them such that one is unable to turn them off when not required. Although it is more comfortable viewing the video without the captions, there is hardly any way out. Video decaptioning aims to solve the task of replacing the text overlays in frames with semantic coherent regions. In this work, we explore the application of state-of-the-art computer vision algorithms to address this challenge in an automated fashion. The task requires first finding the region with captions and then predicting the high-level context, hence making it significantly more difficult when compared to classical image or video inpainting methods. However, decaptioning becomes increasingly more difficult, when the subtitles cover most of the parts of the frame and are of different size, font and colours.

The authors Sayan Sinha and Arnav Kumar Jain contributed equally.

S. Mundra (✉)
Department of Mechanical Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India
e-mail: coolshivansh8@iitkgp.ac.in

A. K. Jain
Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur, India

S. Sinha
Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India
e-mail: sayan.sinha@iitkgp.ac.in

S. Escalera et al. (eds.), *Inpainting and Denoising Challenges*,
The Springer Series on Challenges in Machine Learning,
https://doi.org/10.1007/978-3-030-25614-2_6

Video decaptioning can be seen analogous to the task of image inpainting. Image inpainting is referred to filling of missing holes with pixel values harmonious with the context. The major difference between the two tasks is that for decaptioning there is an additional work to first find the textual region. For text detection, methods are typically either connected component-based (CC-based) [1] or texture based [2, 3]. The CC-based methods can extract text efficiently, but have difficulties when text touches itself or other graphical objects, which may happen in digital video since text is often embedded in complex backgrounds. Jain and Bhattacharjee [3] presents a text extraction system that treats text as a distinctive texture and uses unsupervised clustering to classify each pixel as text or non-text. However, in video frames, natural scenes like the leaves of a tree or grass in a field have textures similar to text, and in the feature space, text and non text often overlap. Thus, traditional inpainting methods may not work well as finding the text regions are difficult. Moreover, the traditional methods often fail in capturing the high level semantics of the scene. This is because they tend to find matching patches from unmasked regions.

Recently, Convolutional Neural Networks [4] have advanced the performance of various tasks in Computer Vision [5–7]. Initial deep learning methods, used $\mathcal{L}_2$ loss on the reconstructed pixel values. The models trained on Mean Squared Error (MSE) loss are unable to capture high texture details of the scene, as they are trained on pixel-wise differences. Also, the reconstructions are found to be blurry. To solve this, adversarial loss [5, 6] have been widely used, where a discriminator is trained simultaneously to distinguish between real and inpainted images, aiding in sharper outputs. Also, Johnson et al.[8] took euclidean distance between the features extracted from a VGG19 [9] network to yield perceptually better results.

In this work, we propose an end-to-end training method for the purpose of video decaptioning. Our purpose of video decaptioning has been broken down such that we try to focus on the regeneration of the entire image from the input, sans the captions. An encoder-decoder network suits our case the most. The main idea lies in the fact that an encoder-decoder model supplements the common contracting network by successive layers, where pooling operators are replaced by up-sampling operators. Hence, these layers help in increasing the resolution. We use an U-Net based architecture, which have the following advantages: (1) U-net is symmetric, (2) different image sizes can be used as input because there is no dense layer, and (3) the skip connections help to combine general information with localization and context.

Notably, Convolutional network based methods are found to create boundary artifacts, distorted structures and blurry structures inconsistent with surrounding areas. It is likely due to the ineffectiveness of CNNs in modelling long-term correlations between distant contextual information and the hole regions. For example, to allow a pixel being influenced by the content of 64 pixels away, it requires at least 6 layers of 3 Œ 3 convolutions with dilation factor of 2 or equivalent factor [10, 11]. To tackle the issue of recovering complex image semantics and structures, we propose to use

dilated convolutions in the encoder. Dilated Convolutions expand the receptive fields of the convolutions by dilating the convolution kernels. The architecture is based on the fact that dilated convolutions support the exponential expansion of the receptive field without loss of resolution or coverage. Moreover, for perceptually pleasing outputs, we further trained our model to minimize distance between the features from a pre-trained VGG19 [9] network.

2 Related Work

Initially, major works of inpainting could be categorized into three verticals. In the works of Hirani and Totsuka [12], frequency and spatial domain information are blended to fill in a given region with a selected texture. Dis occlusion was another popular method introduced by Nitzberg et al. [13]. It can be seen that non-learning approaches to image inpainting rely on propagating appearance information from neighbouring pixels to the target region. They are specific to image sets and can be used to fill in only small gaps.

Computer vision has made tremendous progress on semantic image understanding tasks such as classification [14], video summarization [7, 15] and segmentation [16] in the past decade. Conventional Sparse coding methods [17] were sensitive to image orientation and environment and couldn't be generalized into cross domain works. Recently, Convolutional Neural Networks (CNNs), have greatly advanced the performance in these tasks. The success of such models on image classification paved the way to tackle harder problems, including unsupervised understanding and generation of natural images. More recent methods typically initialize the gaps with values such as a constant or mean pixel value after which the resultant is passed through a deep CNN network. In our paper, such an effort is not required as the captions are atop the image and are to be fed directly. Pathak et al. [5] first introduced the concept of image inpainting using an encoder-decoder network with adversarial losses.

Recent works based on Generative Adversarial Networks (GANs) [18], like [19, 20] have shown convincing results in patch based inpainting. They used GANs in two contexts, one global discriminator and the other one is a local discriminator. But GAN based methods often fail when it comes to inpainting on dataset with diverse classes. Hence these methods couldn't be directly applied in the task of decaptioning. However, Ledig et al. [6] showed that GAN can produce more visually sharper and pleasing images where they used the loss of discriminator to produce sharper results.

Xie et al. [21] has shown that Auto Encoder-Decoder based methods have produced good results in image denoising and image inpainting tasks. Also, this work has shown that the shape of the mask (a region that needs to inpainted) is not required to be given as input to the model. They directly take the corrupted frame and output the reconstructed images. We follow this paradigm as the region with subtitles can occupy different areas in frames.

Feature Learning methods such as [5] have shown good results on high-resolution images where a large section of an image was needed to be inpainted. They have trained on reconstruction and adversarial losses which resulted in real looking images and closer to manually inpainted image. One good advantage of learning features is to understand it's semantics which is essential for unsupervised inpainting.

3 Proposed Method

3.1 Architecture

3.1.1 UNet

U-net [16] is a popular encoder-decoder network, which were first used for the task of bio medical image segmentation. Since then, they have produced state-of-the-art in a wide variety of computer vision tasks like image super-resolution, semantic segmentation and image inpainting. In our case, the encoder takes the captioned frame as input and converts it into a feature representation, which is feed-forwarded through the decoder to get the decaptioned frame.

The input frame is passed through blocks of convolutions followed by maxpool downsampling layers to encode the input image into a latent representation. The purpose of contraction is to capture the context in the input frame. Our encoder is inspired from pix2pix [10] which were used for the task of image-to-image translation. The network consists of five convolutional blocks and each max pooling layer reduces the spatial dimension by two with an increase in the channel length by the same factor. The part of the network between the encoder and decoder is called bottleneck layer. This layer consists of two convolutional layers with batch normalization [22] and dropout.

The latent representation is then passed through the decoder to get the output frame. The decoder network consists of regular convolution operations clubbed with up sampling and concatenation. Up sampling, also called as convolutions with fractional strides, results in higher resolution at each step. There are skip-connections between the symmetrical layers of encoder and decoder i.e. the high resolution features from encoder are concatenated with the up sampled features in decoder. This encourages precise localization combined with contextual information aiding in better reconstructions.

3.1.2 Dilated Convolutions

Dilated convolutions, also known as atrous convolutions, have been explored widely in the computer vision tasks like semantic segmentation, object detection and machine translation. The main idea is to improve the receptive field of the

convolutions. This is achieved by exponentially expanding the receptive field without losing resolution as well as coverage i.e. kernel weights are expanded by a dilation factor. An increase in dilation factor makes the kernel more sparse with an increase in the kernel size. Dilated convolutions can be defined as,

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)k(t), \tag{1}$$

where $*_l$ is referred as l-dilated convolution. Moreover, we define dilated convolutions to have exponentially expanding receptive field, as discussed in the original paper. In our case, we replaced the convolution layers in encoder with dilated convolutions.

3.1.3 Residual Skip Connections

Deep networks are often difficult to train. In fact, a deeper network might not perform better than its shallower counterpart. Gradients get stalled, and the error is larger. In order to make it easy to train such networks and to get over the issue of vanishing gradient, we incorporated residual skip connections [23]. In our case, we apply residual skip connections in the bottleneck layer.

3.2 Loss Functions

3.2.1 Mean Squared Error Loss

Pixel-wise Mean Squared Error (MSE) loss, also called as $\mathcal{L}_2$ is the most widely used optimization target in various similar tasks like image inpainting [5] and image super resolution [6]. The $\mathcal{L}_2$ helps to capture the structure and coherence of the context frame. It is calculated as:

$$l_{MSE} = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} (I_{i,j}^{d} - I_{i,j})^2, \tag{2}$$

where I^d denotes the decaptioned output from the model, W and H denote the dimensions of the image.

3.2.2 VGG Based Perceptual Loss

Networks trained with $\mathcal{L}_2$ result in overly smooth reconstructions and have visually unpleasing high frequency content. Perceptual loss is a feature reconstruction loss

defined by deep neural networks [8]. It guides neural models to generate images visually similar to their corresponding targets (e.g., ground truth) and has been widely utilized in style transfer [24]. We used VGG19 [9] network to extract middle-level features of both generated frame and ground truth frame and then took pixel wise reconstruction loss on both of them. The perceptual loss is defined as

$$l_{i,j}^{VGG/i,j} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I_{x,y}^{d}) - \phi_{i,j}(I_{x,y}))^2, \tag{3}$$

where $\phi_{i,j}$ indicates the feature map obtained by the jth convolution (after activation) before the ith max pooling layer within the VGG19 network. $W_{i,j}$ and $H_{i,j}$ denote the width and height of the feature maps outputs (Figs. 1, 2 and 3).

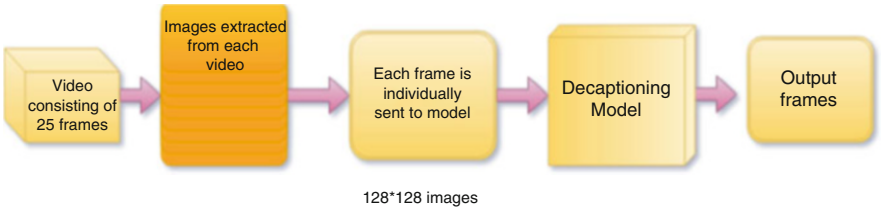

Fig. 1 Training work flow

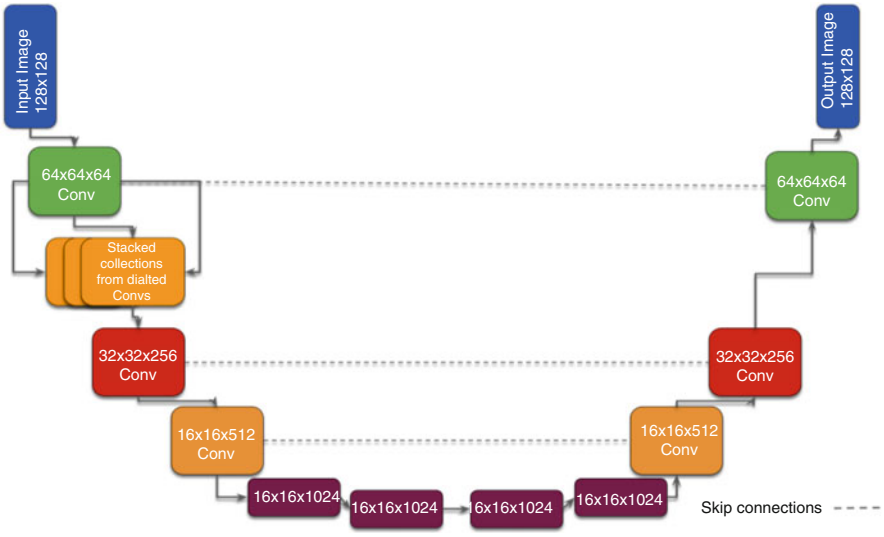

Fig. 2 Model architecture

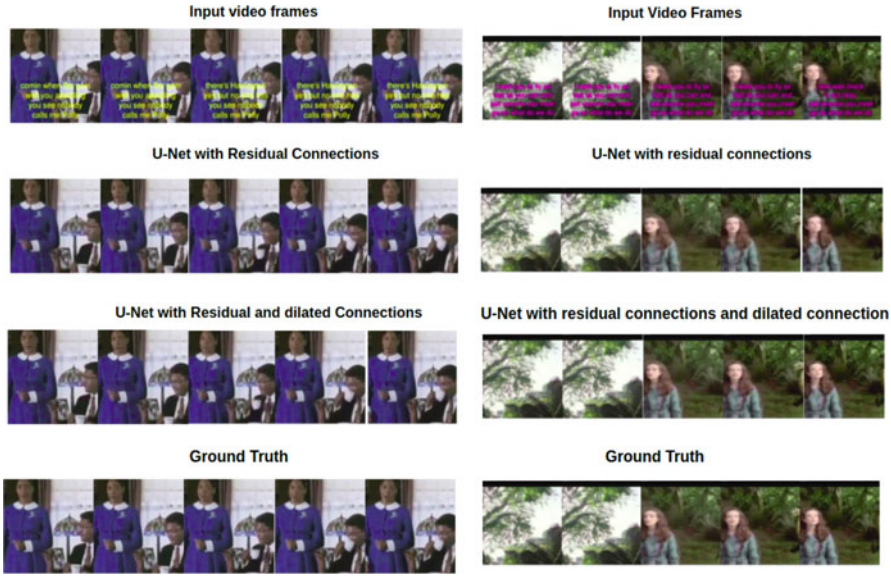

Fig. 3 Left: test case 1 and right: test case 2

4 Results

As explained above, our model was trained on MSE (pixel wise reconstruction) Loss and fine tuned our model on VGG feature loss. In this training procedure, we used Adam Optimizer [25] while decreasing learning rate by a factor of 10 in fine tuning. As you can see in above images from the test case, there is very little difference visible between images generated by adding dilated convolution layers because the resolution of the data set provided in the challenge, but we can see difference in the losses computed. We have kept residual connections in both the part of our experimentation. In our testing pipeline, we used a pre-trained model provided by organizers as a part of the baseline. First, the image was divided into 16 equal parts, and each part was fed to a pretrained model to check if there was text overlay in the corresponding image. If there was text overlay in the part, it was replaced by a similar part from the predicted image from the model. If the score of text classification was below a threshold score, it was replaced by the corresponding region from the input image. This process was similar to Poisson Blending [26]. Our method took lesser time to reach optimal minimal compared to GAN based methods as there were no generator and discriminator trying to optimize simultaneously by min-max strategy. Also, our solution doesn't require a binary mask for inpainting hence decreasing inference time. Our method took approximately 5 s to generate a decaptioned video.

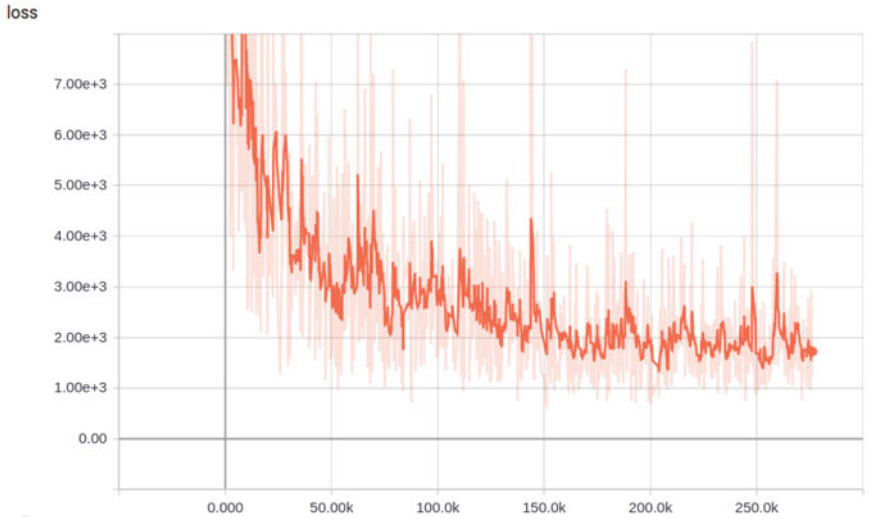

Fig. 4 MSE loss vs iterations

Table 1 Results

Method	MSE loss	PSNR loss	DSSIM loss
Baseline	0.0022	30.1856	0.0613
U-Net without dilation	0.014	32.850	0.0511
U-Net with dilations	**0.0012**	**32.1713**	**0.0482**

As this was the first attempt in the field of video decaptioning, there weren't many baselines we could refer to. Hence we had shown a comparison with Baseline and our model without dilated convolution. With this approach we came **2nd** in training phase and **4th** in test phase of Chalearn Video Decaptioning Challenge (Fig. 4 and Table 1).

5 Conclusion

From our experience in this competition, we came to following conclusion in the task of Video Decaptioning and related problem statements:

- Simple Auto Encoder-Decoder based solution is not good when it comes to noise removal from a large section as the model is generating the image from just encoded latent representation.
- Hence we need a model which have incorporated image semantics in the part of encoding and can be used while generating a decaptioned image. U-Net

based model was proven a good choice in the related field as it included skip connections between symmetric layers in the encoder-decoder part.
- As we needed to capture end to end semantics in the image to get the global feature, we used stacked dilated convolution layer to incorporate global semantics in the encoding part. Here noise removal was to be done considering generated image was supposed to look real and dilated convolution layers were useful to that.
- Simple Encoder Decoder architecture generally decrease the sharpness and resolution in the image generated; residual connections were added to improve sharpness. Although the advantage of adding residual connection was not adding significant difference, but it could increase resolution and the visual appearance by a significant margin when it comes to high-resolution data set.
- We did not extract explicit mask for the region of text removal as the encoder-decoder model implicitly learns it.
- We didn't explore the effect of temporal dimension in the process of video denoising but incorporating temporal dimension should help.

References

1. Anil K Jain and Bin Yu. Automatic text location in images and video frames. *Pattern recognition*, 31(12):2055–2076, 1998.
2. Victor Wu, R Manmatha, and Edward M Riseman. Finding text in images. In *ACM DL*, pages 3–12, 1997.
3. Anil K Jain and Sushil Bhattacharjee. Text segmentation using gabor filters for automatic document processing. *Machine vision and applications*, 5(3):169–184, 1992.
4. Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
5. Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.
6. Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.
7. Arnav Kumar Jain, Abhinav Agarwalla, Kumar Krishna Agrawal, and Pabitra Mitra. Recurrent memory addressing for describing videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2200–2207. IEEE, 2017.
8. Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
9. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
10. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.
11. Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1538–1546, 2015.

12. Anil N. Hirani and Takashi Totsuka. Combining frequency and spatial domain information for fast interactive image noise removal. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 269–276, New York, NY, USA, 1996. ACM.
13. Mumford David Shiota Takahiro Nitzberg, Mark. Filtering, segmentation and depth. In *Springer-Verlag*.
14. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
15. Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.
16. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
17. Muhammad Hanif, Anna Tonazzini, Pasquale Savino, and Emanuele Salerno. Sparse representation based inpainting for the restoration of document images affected by bleed-through. In *Multidisciplinary Digital Publishing Institute Proceedings*, volume 2, page 93, 2018.
18. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
19. Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.
20. Xinshan Zhu, Yongjun Qian, Xianfeng Zhao, Biao Sun, and Ya Sun. A deep learning approach to patch-based image inpainting forensics. *Signal Processing: Image Communication*, 2018.
21. Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems*, pages 341–349, 2012.
22. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
23. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
24. Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.
25. Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. iclr (2015), 2015.
26. Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on graphics (TOG)*, 22(3):313–318, 2003.