Joint Caption Detection and Inpainting using Generative Network

Anubha Pandey, Vismay Patel

Indian Institute of Technology Madras

cs16s023@cse.iitm.ac.in

19th September 2018

Overview

Problem Statement

Chalearn LAP Inpainting Competition Track2 - Video Decaptioning

- **Objective** To develop algorithms that can inpaint video frames that contain text overlays in various size, background, color, location.

Problem Statement

Chalearn LAP Inpainting Competition Track2 - Video Decaptioning

- **Objective** To develop algorithms that can inpaint video frames that contain text overlays in various size, background, color, location.
- **Dataset**
  - Consists of a set of (X,Y) pairs where X is a 5 second video clip and Y is the corresponding target video clip.
  - 150 hours of diverse videos clips in 128x128 RGB pixels, containing both captioned and de-captioned versions, taken from YouTube.
  - 70000 training samples and 5000 samples in Validation and Test set

Introduction

- Video Decaptioning involves two tasks caption detection and the general video inpainting.

Introduction

- Video Decaptioning involves two tasks caption detection and the general video inpainting.
- Existing patch-based video inpainting methods search for complete spatio-temporal patches to copy into the missing area.

Introduction

- Video Decaptioning involves two tasks caption detection and the general video inpainting.
- Existing patch-based video inpainting methods search for complete spatio-temporal patches to copy into the missing area.
- Despite recent advances in machine learning, it is still challenging to aim at fast (real time) and accurate automatic text removal in video sequences.

Related Works

- More recently, Globally and locally consistent image completion [1] CVPR 2017 paper, has shown promising results for the task of Image Inpainting. It has improved the results by introducing local and global discriminators. In addition, it uses dilated convolutions to increase the receptive fields and replace the fully connected layers adopted in the contextual encoders.

Proposed Solution: Frame Level Inpainting and Caption Detection

- We propose a generative CNN to do joint caption detection and decaptioning task in an end-to-end fashion. Our network is inspired by the work of **Globally and locally consistent image completion** [1].

Proposed Solution: Frame Level Inpainting and Caption Detection

- We propose a generative CNN to do joint caption detection and decaptioning task in an end-to-end fashion. Our network is inspired by the work of **Globally and locally consistent image completion** [1].
- The network has two branches each for the image generation and the mask generation tasks

Proposed Solution: Frame Level Inpainting and Caption Detection

- We propose a generative CNN to do joint caption detection and decaptioning task in an end-to-end fashion. Our network is inspired by the work of **Globally and locally consistent image completion** [1].
- The network has two branches each for the image generation and the mask generation tasks
- Both the branches share the parameters up to first three convolution layers and the layers thereafter, are trained independently.

Proposed Solution: Frame Level Inpainting and Caption Detection

- We propose a generative CNN to do joint caption detection and decaptioning task in an end-to-end fashion. Our network is inspired by the work of **Globally and locally consistent image completion** [1].
- The network has two branches each for the image generation and the mask generation tasks
- Both the branches share the parameters up to first three convolution layers and the layers thereafter, are trained independently.
- **Inputs**
  - Frames from the captioned videos.
  - The caption masks, extracted by taking the difference between the corresponding frames of the ground truth decaptioned videos and the input captioned videos.

Network Architecture

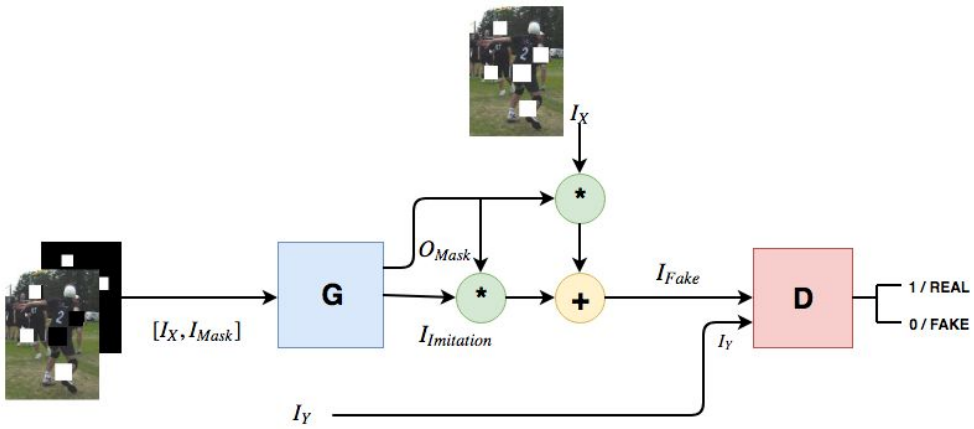

Figure: Architecture of the discriminator module of the inpainting network. Each building block is described in Figure 7.

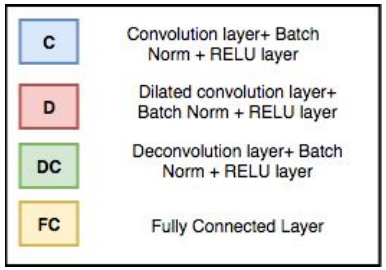

Figure: Building blocks of the network.

Network Architecture

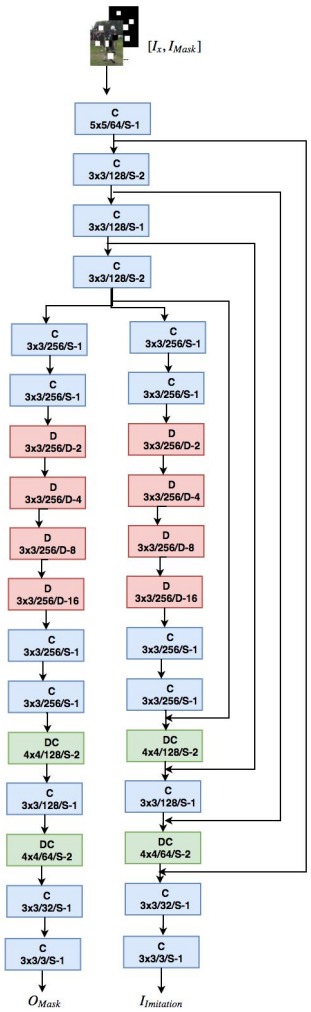

Figure: Architecture of the generator module of the inpainting network. Building block is shown in Fig 7.

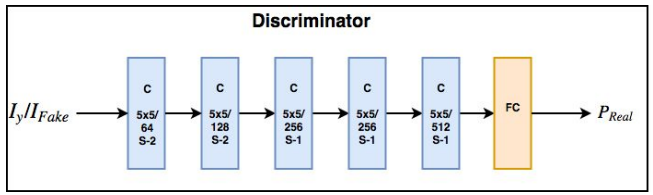

Figure: Architecture of the discriminator module of the inpainting network.Building block is shown in Fig 7.

Loss Functions

Following loss functions have been used to train the network-

- **Reconstruction Loss** [2]
  $L_r = \frac{1}{K}\sum_{i=1}^{K}|I_y^i - I_{imitation}^i| + \alpha * \frac{1}{K}\sum_{i=1}^{K}(I_{Mask}^i - O_{Mask}^i)^2$
  where, K is the batch size and alpha $= 0.000001$.

Loss Functions

Following loss functions have been used to train the network-

- **Reconstruction Loss** [2]
  $L_r = \frac{1}{K}\sum_{i=1}^{K}|I_y^i - I_{imitation}^i| + \alpha * \frac{1}{K}\sum_{i=1}^{K}(I_{Mask}^i - O_{Mask}^i)^2$
  where, K is the batch size and alpha $= 0.000001$.
- **Adversarial Loss** [2] $L_{real} = -log(p)$, $L_{fake} = -log(1-p)$
  $L_d = L_{real} + \beta * L_{fake}$
  where, p is the output probability of the discriminator module and $\beta = 0.01$ (hyper parameter)

Loss Functions

Following loss functions have been used to train the network-

- **Reconstruction Loss** [2]
  $L_r = \frac{1}{K}\sum_{i=1}^{K}|I_y^i - I_{imitation}^i| + \alpha * \frac{1}{K}\sum_{i=1}^{K}(I_{Mask}^i - O_{Mask}^i)^2$
  where, K is the batch size and alpha = 0.000001.
- **Adversarial Loss** [2] $L_{real} = -log(p)$, $L_{fake} = -log(1-p)$
  $L_d = L_{real} + \beta * L_{fake}$
  where, p is the output probability of the discriminator module and β = 0.01 (hyper parameter)
- **Perceptual Loss** [3]
  $L_p = \frac{1}{K}\sum_{i=1}^{K}(\phi(I_y) - \phi(I_{imitation}))^2$
  where, ϕ represents features from VGG16 network pretrained on Microsoft COCO dataset.

Training

- The network is trained using Adam Optimizer with learning rate 0.006 and batch size 20.

Training

- The network is trained using Adam Optimizer with learning rate 0.006 and batch size 20.
- For first 8 epochs only the generator module of the network is trained minimizing only the reconstruction loss and perceptual loss

Training

- The network is trained using Adam Optimizer with learning rate 0.006 and batch size 20.
- For first 8 epochs only the generator module of the network is trained minimizing only the reconstruction loss and perceptual loss
- For the next 12 epochs, the entire GAN network [2] is trained end-to-end minimizing all three losses- Reconstruction loss, Adversarial Loss and Perceptual loss.

Results

- With our proposed solution we secured **3rd position** in the competition.

Results

- With our proposed solution we secured **3rd position** in the competition.
- To evaluate the quality of the reconstruction, metrics as mentioned on the competitions website are used for pairwise frame comparison.

Evaluation Metrics	**Training Phase**	**Testing Phase**
PSNR	30.5311	32.0021
MSE	0.0016	0.0012
DSSIM	0.0610	0.0499

Conclusion

- We have proposed an end-to-end network for de-captioning which can simultaneously do frame level caption detection and inpainting.

Conclusion

- We have proposed an end-to-end network for de-captioning which can simultaneously do frame level caption detection and inpainting.
- However, this method requires individual frames from the clip to do its task which lacks the temporal context required to produce the desired result.

Future Work

- In future work, we aim to improve performance by exploiting the temporal information of the video clips.

Future Work

- In future work, we aim to improve performance by exploiting the temporal information of the video clips.
- We aim to explore models that use both temporal and semantic information.

Future Work

- In future work, we aim to improve performance by exploiting the temporal information of the video clips.
- We aim to explore models that use both temporal and semantic information.
- Techniques used in intermediate frame prediction can be employed to make the network temporally-aware.

Thank You

References


Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa.
Globally and locally consistent image completion.
ACM Transactions on Graphics (TOG), 36(4):107, 2017.


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.


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.