Captioned Image Sequence Generation Using Deep Neural Networks

*Vincent M. Casser, Camilo L. Fosco, Justin S. Lee, Karan R. Motwani**
Harvard Computational Science and Engineering/Electrical Engineering* Masters Programs

CS109b/STAT121b/AC209b - Spring 2018

Introduction

While photorealistic image generation is, in many domains, a solved problem, this is not currently the case in the generation of sequential visual information. This is due to the difficulty of modeling "temporal-realism," i.e. depicting a sequence of motions in images that seems realistic to a person when played in sequence. In this work, we use deep learning to attempt to generate plausible sequences of images from various datasets comprised of textual descriptions and image sequences.

Figure 1. An example of a "temporal-realistic" image sequence from the KTH Human Pose dataset.

Data Exploration

Four different datasets were used in this project:

- KTH Human Pose dataset
- T-GIF: 100k GIFs from Tumblr with crowd-sourced captions
- Synthetic "MNIST-in-motion" dataset – built from scratch
- Synthetic "Icons-in-motion" dataset – built from scratch

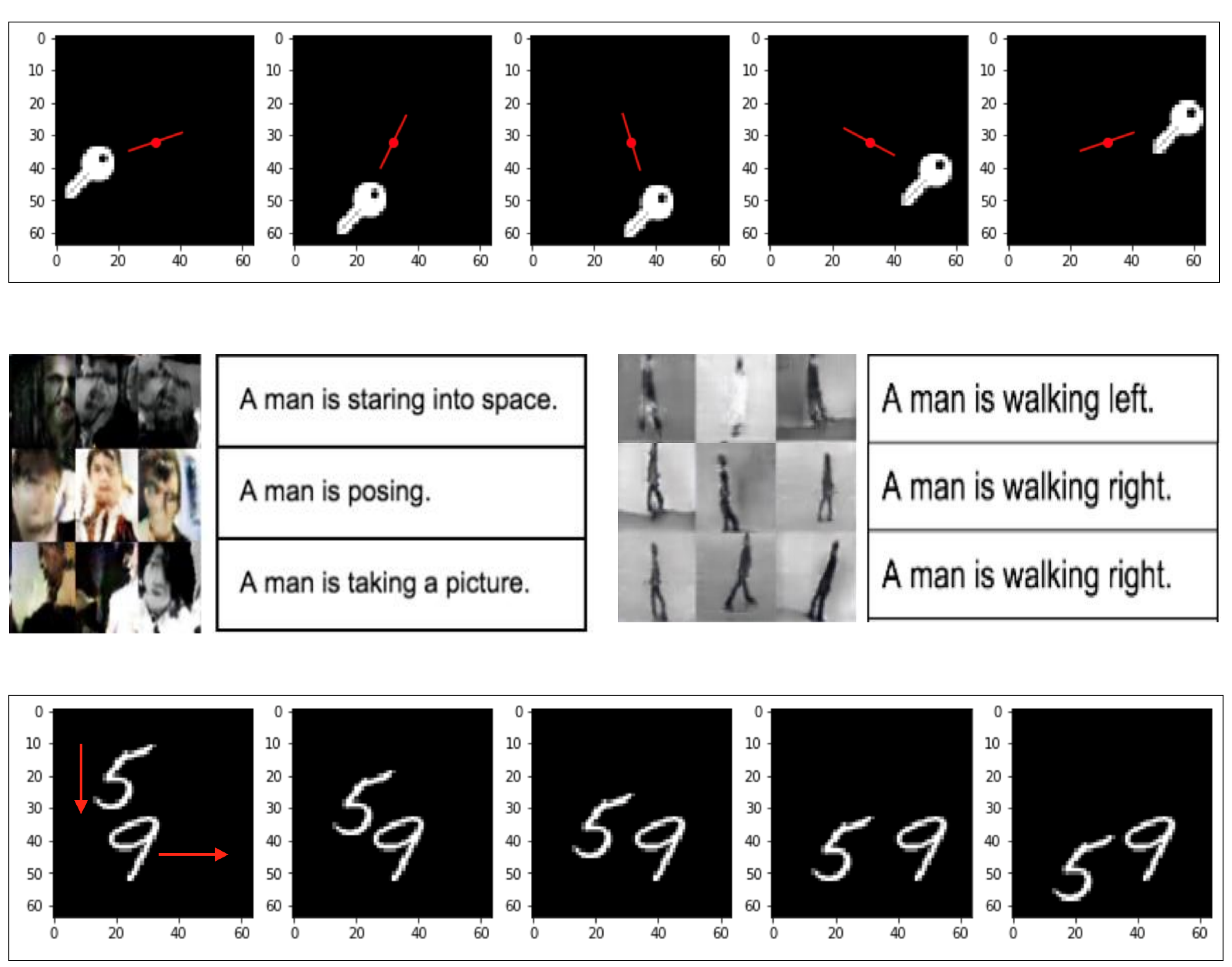


Figure 2. Top: Icons-in-motion dataset. *Left:* Text to Image GAN output trained on T-GIF subset with captions. *Right:* GAN output trained on KTH subset with captions. *Bottom:* MNIST-in-motion dataset.

Methodology

We focused on two modified architectures:

- An extension of U-Net, modified to operate on 3D-volumes instead of single images as input. We generate the next frame based on N previous ones and a caption.
- The same architecture as in (1), but in an adversarial setting – i.e. with the presence of a discriminator that operates at a frame level.

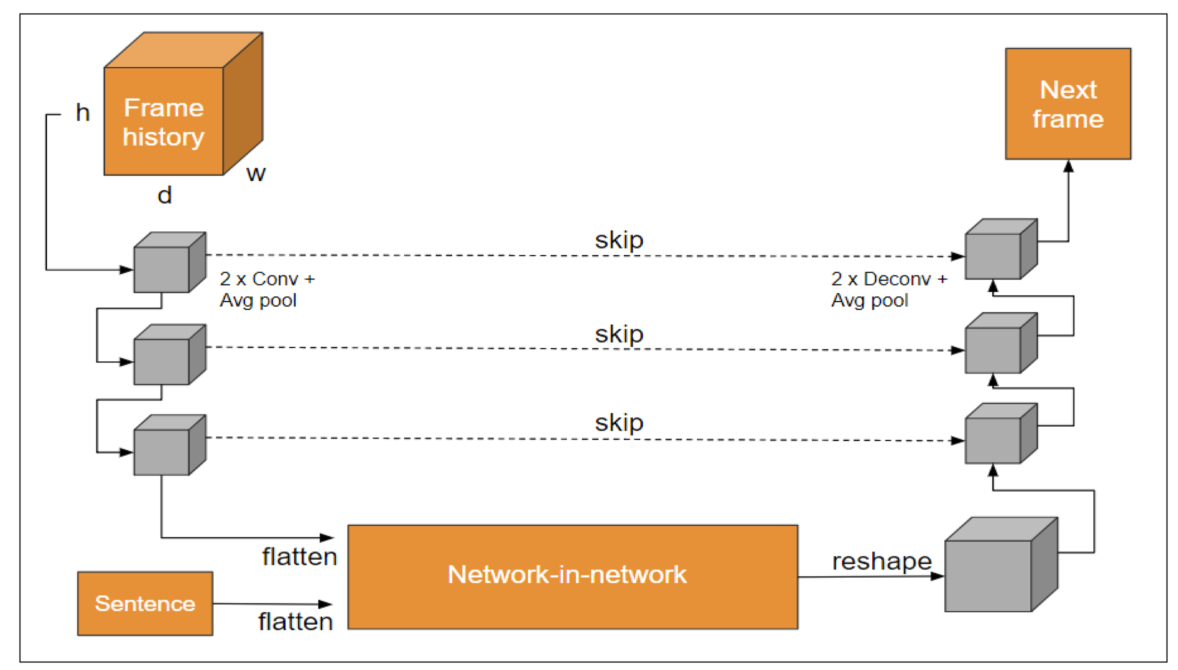


Figure 3. A diagram of the modified U-Net architecture; captions were incorporated mid-network. The network uses MSE as loss.

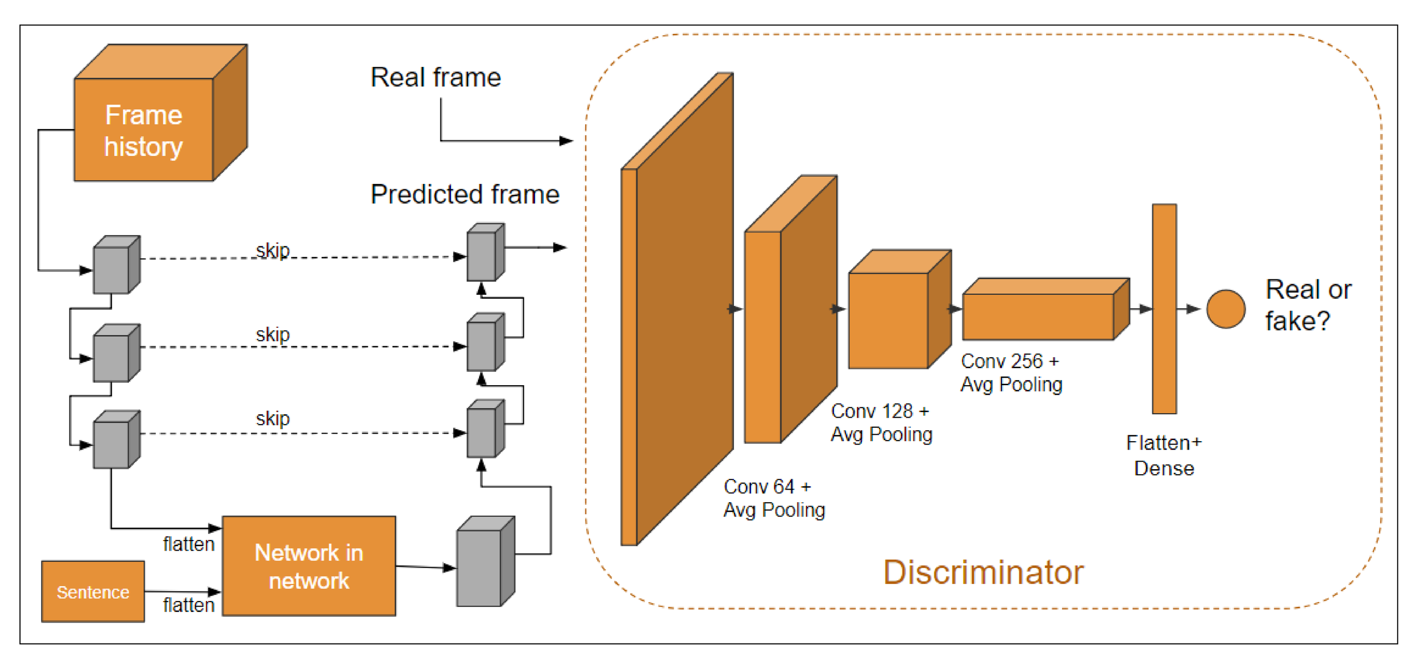


Figure 4. A diagram of the modified U-Net architecture in the adversarial setting. The discriminator tries to discern real frames from synthetic ones.

Results

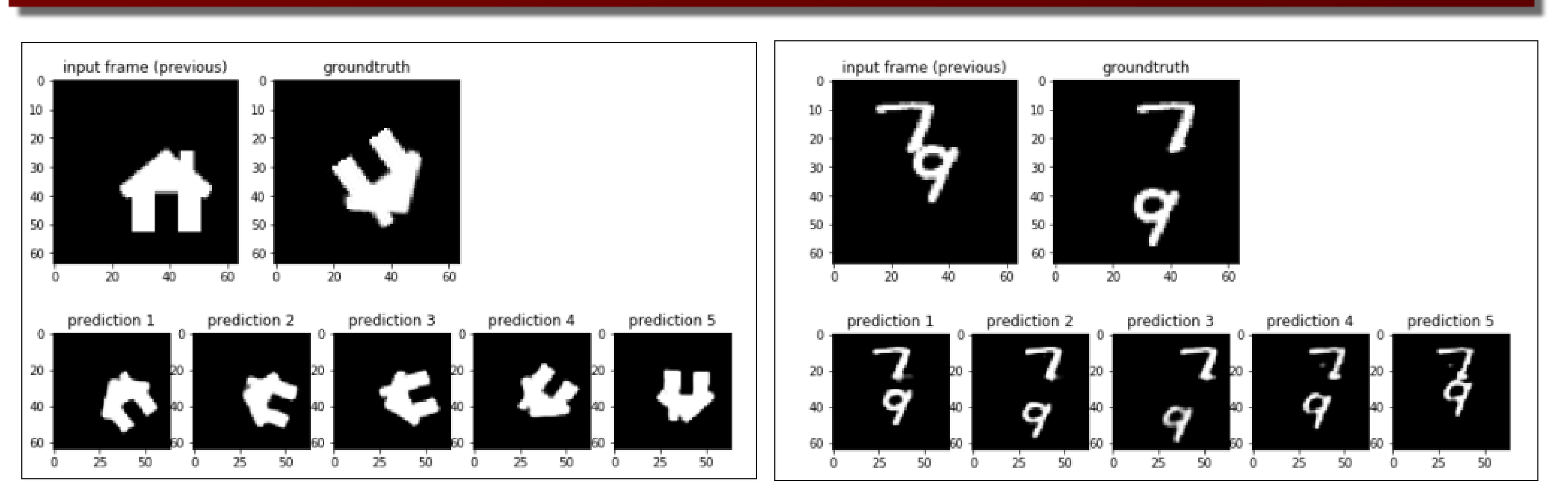
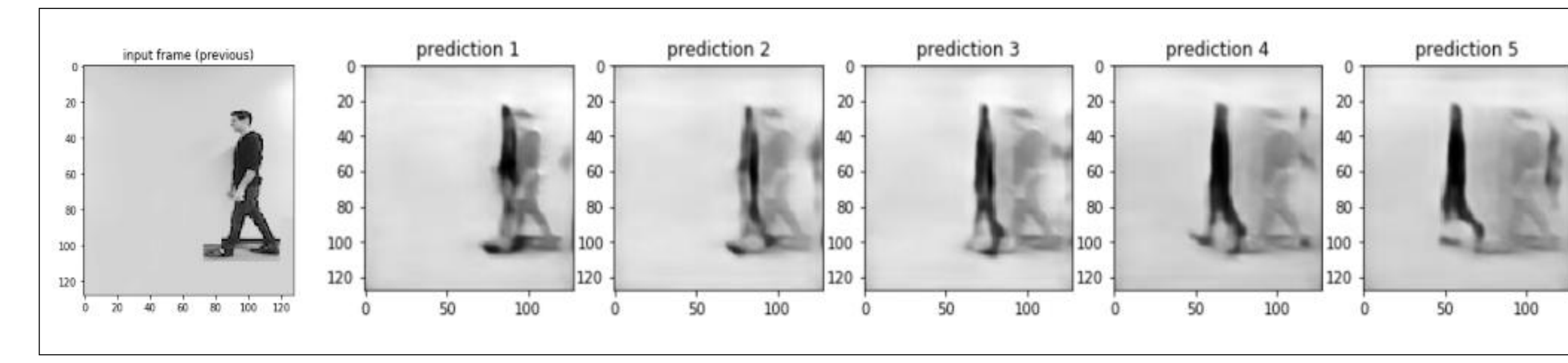
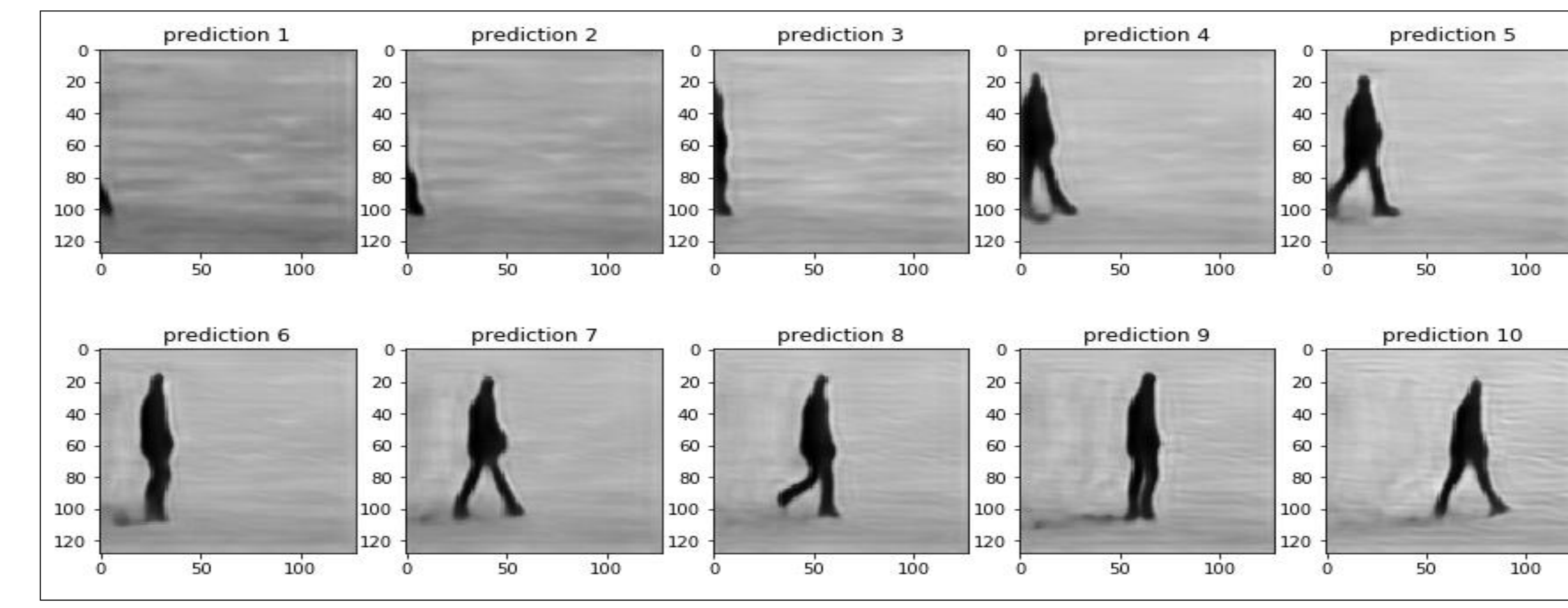


Figure 5. *Left:* Test example from Icons-in-motion dataset with caption "*The house is rotating counter-clockwise.*". *Right:* Test example on a house icon with caption "*The house is rotating counter-clockwise.*" Both outputs come from the U-Net based model, which achieved strong results on these datasets.





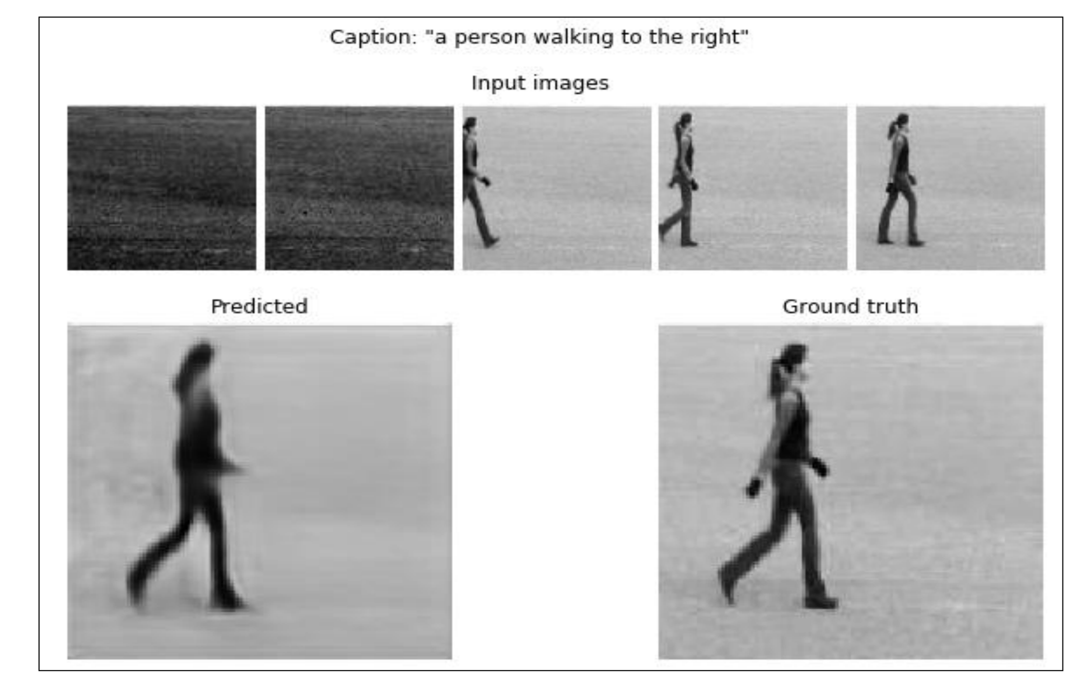


Figure 6. Top: Gif generated from a picture of our own Camilo Fosco with caption "A person walks left.". *Middle:* full prediction outputted by the adversarial model. *Bottom:* An example of a prediction from the adversarial network given 5 input frames and the caption "A person walks right."

Conclusions

In this work, we created novel architectures based on U-Net to generate temporally cohesive sets of images. To accomplish this, we (1) created synthetic data depicting basic events that served our needs (2) built new architectures for frame prediction, and (3) tested our systems on the synthetic datasets and the KTH dataset. We achieved temporal cohesion in image sequences, and compared a GAN approach with an MSE based U-Net for frame generation. We conclude that in this setting, GANs require significant fine-tuning to outperform U-Net-based architectures.

Citations

- Pan, Y., Qiu, Z., Yao, T., Li, H., Mei, T., 2017. To create what you tell: generating videos from captions.
- Marwah, Tanya; Mittal, Gaurav; Balasubramanian, Vineeth N. Attentive semantic video generation using captions
- Vondrick, C., Pirsiavash, H., Torralba, A., 2016. Generating videos with scene dynamics.
- Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H., 2016. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.