Generative Adversarial Networks for Multimodal Representation Learning in Video Hyperlinking

Vedran Vukotić[1,2], Christian Raymond[1,2], Guillaume Gravier[1,3]

vedran.vukotic@irisa.fr christian.raymond@irisa.fr guillaume.gravier@irisa.fr

[1] UMR IRISA [2] INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES [3] Inria INVENTEURS DU MONDE NUMÉRIQUE, CNRS

ICMR 2017
Bucharest, Romania

Problem

- given a video segment (anchor), suggest a set of relevant video segments (targets)

Goal:

- use two modalities (automatic speech transcripts and video keyframes) to find relevant video segments
- ***visualize learned crossmodal relationships*** (e.g. what does the model expect to be visible in the video segment given a particular speech transcript segment)

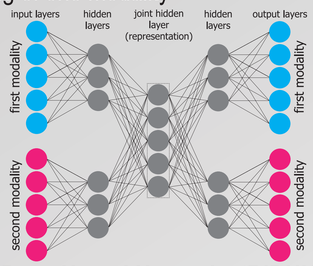
Means:

- perform multimodal fusion of the given input modalities
- synthesize images for a given speech transcript and find top words for a given image

Existing Approaches

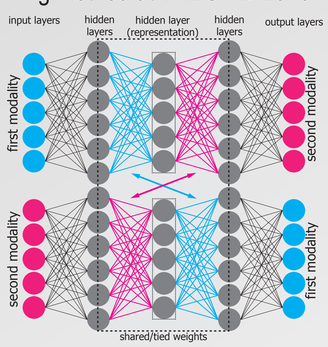
Classical Multimodal Autoencoders:

- separate branches for each modality
- reconstructing both modalities with added noise and sporadic zeroing of one modality



Bidirectional Deep Neural Networks (BiDNN):

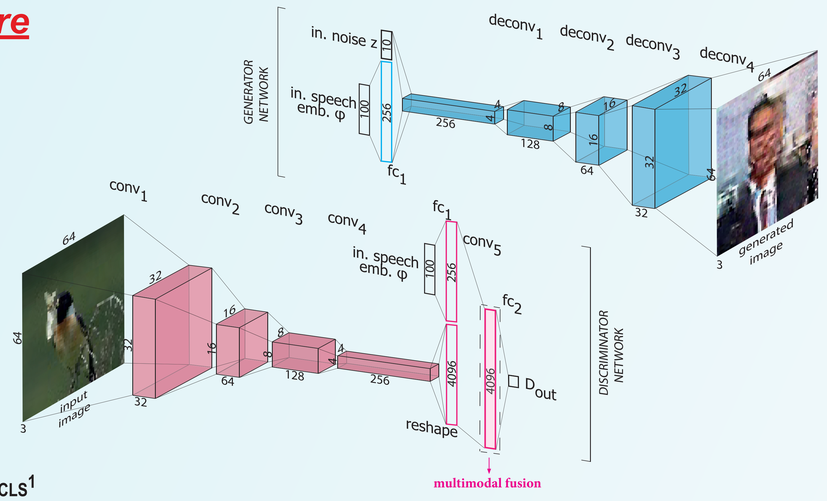
- ***crossmodal translations*** (with added restrictions) as a mean of performing ***multimodal fusion***
- best performing method at TRECVID 2016



References

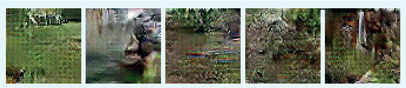
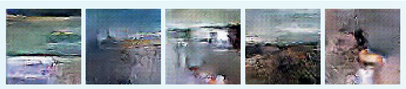
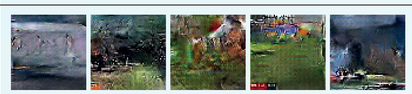
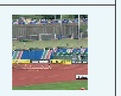
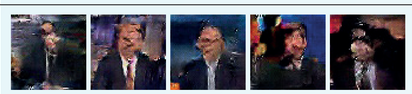
1. Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. **Generative Adversarial Text to Image Synthesis.** In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1060-1069. June 24, New York City, NY, USA.

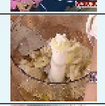
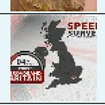
Architecture



- text to image GAN-CLS[1]

Crossmodal Visualizations

Input - Automatic Speech Transcript	Generated Images	Real Image
"…insects emerged to take advantage of the abundance . the warm weather sees the arrival of migrant birds stone chests have spent the winter in the south …"		
"…second navigation of the united kingdom . the north sea , it was at the north yorkshire moors between the 2 , starting point for the next leg of our journey along the coast …"		
"…this is a dangerous time for injuries for athletes . having said that , some of these upbeat again a game . there she is running strongly she looks more comfortable …"		
"…the role of my squadron afghanistan is to provide the the reconnaissance capability to use its or so forgave so using light armor of maneuvering around the area of …"		

Input Image	Top Words in the Speech Modality
	britain, protecting, shipyard, carriers, jobs, vessels, current, royal, aircraft, securing, critics, flagships, foreclosures, economic, national
	north, central, rain, northern, eastern, across, scotland, southwest, west, north-east, northeast, south, affecting, england, midlands
	pepper, garlic, sauce, cumin, chopped, ginger, tomatoes, peppers, onion, crispy, parsley, grated, coconuts, salt, crust
	mountains, central, foreclosures, ensuing, across, scotland, norwegian, england, country, armor, doubting, migration, britain, southern

Idea

- **generative network**
  - performs a text to image crossmodal translation
  - used to ***generate visualizations***
  - provides an infinite amount of artificial samples
- **discriminative network**
  - used to perform ***multimodal fusion***
  - obtained embedding used for multimodal retrieval
  - improves the generative network

Dataset

- **MediaEval 2014, formed post evaluation**
  - 10,321 video segments that contain both keyframes and automatic speech transcripts
- **automatic speech transcripts**
  - averaged Word2Vec
- **keyframe representing video segment**
  - image of 64 x 64 pixels (directly given to GAN)
  - VGG-19 features of such image for AEs / BiDNNs

Evaluation

Representation	P_{10} (%)	σ (%)
Speech Transcripts Only	56.55	-
Visual Only (VGG-19)	52.41	-
Multimodal AE	57.94	0.82
BiDNN	59.66	0.84
CGAN	62.84	1.36

Downsides

- **very limited image size**
  - 50 x 50 pixels
  - with images of size 224 x 224 pixels, BiDNN achieves 80% in P_{10}
- **slow training**
  - ~20 hours on a GPU compared to a few hours on a CPU for BiDNN

Conclusion

- good for visualizing crossmodal translations in the initial domain
- could potentially provide better multimodal fusion than multimodal autoencoding methods
- currently very limited and expensive to train