S2IGAN: Speech-to-Image Generation via Adversarial Learning

Xinsheng Wang\textsuperscript{1,2}, Tingting Qiao\textsuperscript{2,3}, Jihua Zhu\textsuperscript{1}, Alan Hanjalic\textsuperscript{2}, Odette Scharenborg\textsuperscript{2}

\textsuperscript{1}School of Software Engineering, Xi’an Jiaotong University, China
\textsuperscript{2}Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands
\textsuperscript{3}College of Computer Science and Technology, Zhejiang University, China

https://github.com/xinshengwang/S2IGAN

https://xinshengwang.github.io/project/s2igan/
Background

Cross-modal learning between visual and speech

Identify matching pairs

“A golden brown dog is jumping over a barrier.”

“A black and white dog is playing in the water.”

Describe image using speech

“A black and white dog is playing in the water.”

Imagine the image according to the speech description

“A black and white dog is playing in the water. ”

Cross-modal retrieval

Image-to-speech generation

Interspeech 2019, Merkx et al.

ICNLSSP 2017, Hasegawa-Johnson et al.
New task

Speech-to-image generation (S2IG)

“A bird with a bright yellow belly and gray head with brownish wings”

S2IG: generate images on the basis of spoken descriptions bypassing any text information
Method

Overview

Two stages of speech-to-image generation:

Speech embedding stage

Speech Embedding Network -> Speech embedding

Image generation stage

Speech embedding -> Generative Model

Image of bird
The speech encoder is trained in a visually grounded way.

- **Image encoder (IED):**
  To obtain image embeddings

- **Speech encoder (SED):**
  To obtain speech embeddings

Framework of the speech embedding network (SEN).
Method

Generation model

- Densely-stacked generator
  Generates images from low-resolution to high-resolution step by step.

- Relation Supervisor
  To ensure that the generator produces high-quality images that are semantically aligned with the spoken description.
Experiments

Database

CUB-200 Bird
- “A bird with a blue head and a red breast with black wingbars.”
- “This bird has a bright blue crown and a brownish/red breast color.”
- ...

Oxford-102 Flower
- “This flower has small pink petals with a yellow center.”
- “this flower has small pink petals with a yellow center.”
- ...

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Training set</th>
<th>Test set</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>class</td>
<td>150</td>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>image</td>
<td>8855</td>
<td>2933</td>
<td>11788</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Training set</th>
<th>Test set</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>class</td>
<td>82</td>
<td>20</td>
<td>102</td>
</tr>
<tr>
<td>image</td>
<td>7034</td>
<td>1155</td>
<td>8189</td>
</tr>
</tbody>
</table>

Speech Caption:
Each image of CUB and Oxford has 10 text captions but no speech caption. Speech captions are synthesized with Tacotron2 with text captions as input.
Experiments

Evaluation

• Inception Score (IS, higher is better)
  ✓ higher IS means the model can generate more diverse and meaningful images

• Fréchet inception distance (FID, lower is better)
  ✓ lower FID means a smaller distance between the generated and real image distributions

• mean Average Precision (mAP, higher is better)
  ✓ higher mAP means higher semantic consistency between generated images and their corresponding speech descriptions.
Method

Results

<table>
<thead>
<tr>
<th>Evaluation Metric</th>
<th>Input</th>
<th>CUB (Bird)</th>
<th>Oxford-102 (Flower)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mAP</td>
<td>FID</td>
</tr>
<tr>
<td>StackGAN-v2</td>
<td>text</td>
<td>15.30</td>
<td>4.04±0.05</td>
</tr>
<tr>
<td>StackGAN-v2†</td>
<td>text</td>
<td>7.01</td>
<td>20.94±0.03</td>
</tr>
<tr>
<td>MirrorGAN</td>
<td>text</td>
<td>—</td>
<td>4.56±0.05</td>
</tr>
<tr>
<td>SEGAN</td>
<td>text</td>
<td>—</td>
<td>4.67±0.04</td>
</tr>
<tr>
<td>[12]</td>
<td>speech</td>
<td>—</td>
<td>18.37±0.04</td>
</tr>
<tr>
<td>StackGAN-v2</td>
<td>speech</td>
<td>8.09</td>
<td>18.94±0.04</td>
</tr>
<tr>
<td>S2IGAN</td>
<td>speech</td>
<td><strong>9.04</strong></td>
<td><strong>14.50</strong></td>
</tr>
</tbody>
</table>

On the speech-to-image task, our method achieves the state-of-the-art performance.
✓ establishes a solid new baseline for the S2IG task

StackGAN-v2 that takes our speech embeddings as input outperforms the original text-based image generation.
✓ the learned speech embeddings are competitive compared to text embeddings
✓ the proposed speech embedding model is effectiveness

StackGAN-v2: zhang et al, 2018 IEEE TPAMI.
MirrorGAN: qiao et al, 2019 CVPR.
SEGAN: tan et al, 2019 ICCV.
Method

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground Truth</td>
<td>A small blue bird with long tail feathers and short black head.</td>
</tr>
<tr>
<td>StackGAN-v2† (T2IG)</td>
<td>This bird has wings that are black and has an orange belly.</td>
</tr>
<tr>
<td>StackGAN-v2</td>
<td>This bird is brown and white in color, with a short black beak.</td>
</tr>
<tr>
<td>S2IGAN</td>
<td>A bird with a bright yellow belly and grey head with brownish wings.</td>
</tr>
<tr>
<td>The flower has petals that are pink with yellow and black lines</td>
<td>This flower has petals that are purple and very stringy.</td>
</tr>
<tr>
<td>The flower has yellow stamens and soft round layers of yellow petals</td>
<td>The flower has thin purple petals surround the red stamens in the middle.</td>
</tr>
</tbody>
</table>
Method

Results

S2IGAN’s ability to catch subtle semantic differences in the speech descriptions

- Generated images conditioned on different speech descriptions in which color keywords were changed, e.g.,
  - A small bird with a Yellow belly and Brown wings.
  - A small bird with a Yellow belly and Black wings.
  - …

![Generated images conditioned on different speech descriptions](image-url)
Conclusion and future work

Conclusion

• A novel speech-to-image generation (S2IG) task is proposed.
• The proposed speech-to-image generation model (S2IGAN) is able to generate high-quality images.
• S2IGAN achieves the state-of-the-art performance on the speech-to-image task.

Future work

• Use natural speech instead of synthesized speech database.
• Develop automatic speech word unit discovery method to make the speech-to-image generation system can benefit from the word-level attention mechanism.
Thanks for your attention!