SSA-GAN: End-to-End Time-Lapse Generation with Spatial Self-Attention

PS1-13

Daichi Horita, Keiji Yanai
The University of Electro-Communications, Tokyo, Japan

Overview

- 単一画像からの動画生成
- Spatial self-attentionはEnd-to-Endでの学習を可能にした

Demo iPad

Motivation

- VGAN[1]: 固定された背景と前景を別々に生成

➡ **同時に両方のコンテンツを生成**

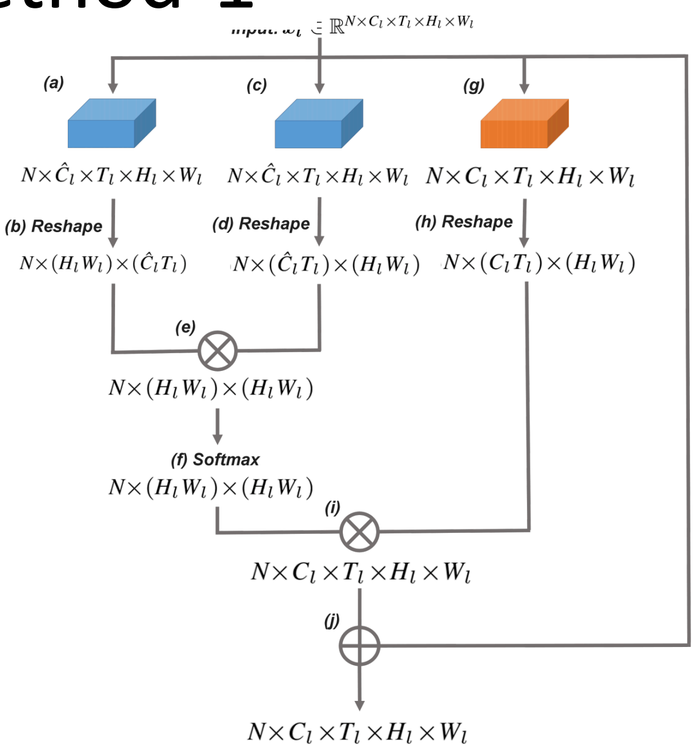
- TGAN[2]: 時間特性と空間特性の違いを考慮して特徴量を別々に扱う

➡ **3D畳み込みを用いて同時に扱えるようなネットワークの研究**

- MD-GAN[3]: 3D畳み込みを用いているが，モデルを二段階で学習

➡ **End-to-Endでの学習**

Method-1

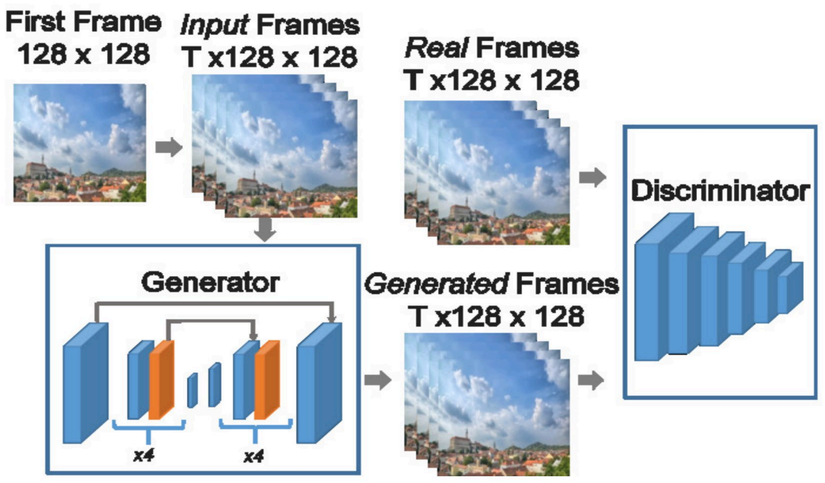


(f) $\hat{\boldsymbol{X}}_{\boldsymbol{l}} = \frac{\exp(\boldsymbol{X}_{\boldsymbol{l}})}{\sum \exp(\boldsymbol{X}_{\boldsymbol{l}})}$, where $\boldsymbol{X}_{\boldsymbol{l}} = \boldsymbol{x}_{\boldsymbol{l}_1} \otimes \boldsymbol{x}_{\boldsymbol{l}_2}$.

(i) $\boldsymbol{o}_{\boldsymbol{l}} = \hat{\boldsymbol{X}}_{\boldsymbol{l}} \otimes \boldsymbol{x}_{\boldsymbol{l}_3}$. (out) $\boldsymbol{y}_{\boldsymbol{l}} = \gamma \boldsymbol{o}_{\boldsymbol{l}} + \boldsymbol{x}_{\boldsymbol{l}}$,

- 各3D畳み込み(a,c,g)はカーネル1で構成
- 空間での特徴マップでSoftmaxを計算することで空間方向のAttentionを計算
  - 時間を含むTWH特徴マップで計算しようとするとGPUメモリに乗らない問題
- Spatial self-attention機構は最初は近隣のピクセル情報を頼りにするが徐々に全体的な場所により多くの重みを割り当てれることを可能とする
- ネットワークが最初は簡単なタスクを学習して除々にタスクの複雑さを増やすことでより良い特徴量を獲得できるようにする

Method-2



- 3D UNetの各3D畳み込みの後にspatial self-attention機構を追加
- ロス関数

$$\mathcal{L}_{adv} = \min_G \max_D \underset{Y \sim P_r}{E} [\log D(Y)] + \underset{\bar{X} \sim P_g}{E} [\log (1 - D(\bar{X}))],$$

$$\mathcal{L}_{con} = \underset{Y \sim P_r, \bar{X} \sim P_g}{E} [\|Y - \bar{X}\|],$$

$$\mathcal{L}_D = -\mathcal{L}_{adv},$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{con} \mathcal{L}_{con},$$

Experiment

- Cloud Dataset[3], Beach dataset[1]を使用
- 10人の被験者に100ペアの動画に対してPreference Opinion Score(POS)を評価，値は[0,1000]

"Which is more realistic?"	POS
Prefer Ours over MD-GAN Stage *I*	871
Prefer Ours over MD-GAN Stage *II*	526
Prefer MD-GAN Stage *I* over Real	286
Prefer MD-GAN Stage *II* over Real	322
Prefer Ours over Real	334

- Cloud

Method	MSE↓	PSNR↑	SSIM↑
MD-GAN Stage *I*	0.0970	16.9019	0.3583
MD-GAN Stage *II*	0.0307	22.7372	0.5920
SSA-GAN (Ours)	**0.0232**	**24.9100**	**0.6805**

- Beach

Method	MSE↓	PSNR↑	SSIM↑
RNN-GAN	0.1849	7.7988	0.5143
VGAN	0.0958	11.5586	0.6035
MD-GAN Stage *II*	0.0422	16.1951	**0.8019**
Ours (a)	0.0379	23.6601	0.7320
Ours (b)	**0.0374**	**25.6432**	0.7346

Future work

- 領域分割と組み合わせて予測タスクの単純化
- 人物にはポーズなどを組み合わせる
- 縛られた条件でやらず、使えるものを組み合わせる

[1]Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In Proc.of Neural Information Processing Systems, 2016.
[2]Masaki Saito and Eiichi Matsumoto.Temporal generative adversarial nets. In Proc.of IEEE International Conference on Computer Vision(ICCV), 2017.
[3]Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu,and Jiebo Luo. Learning to generate time-lapse videos usingmulti-stage dynamic generative adversarial etworks. In Proc.of IEEE Computer Vision and Pattern Recognition(CVPR), 2018.