Learning to Sketch with Shortcut Cycle Consistency

2022.05.02 202132032 김형범

목차

1. Intro

2. Difference from other models

3. Method

4. Experiments & Evaluation

5. Conclusion

Intro

Sketch generation

Trained_model

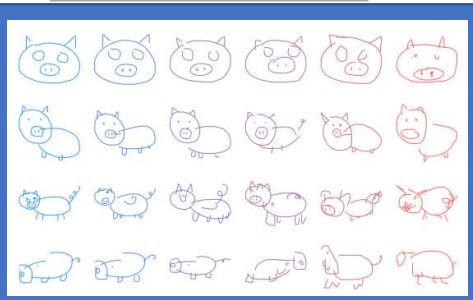
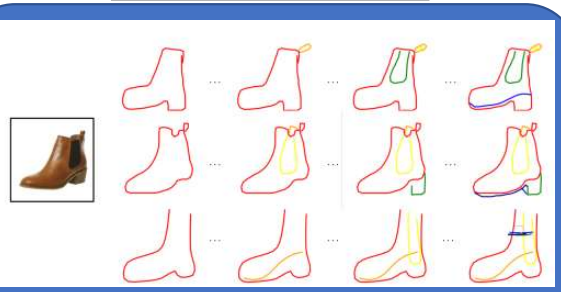


Photo2sketch



Intro

Sketch generation

Trained_model

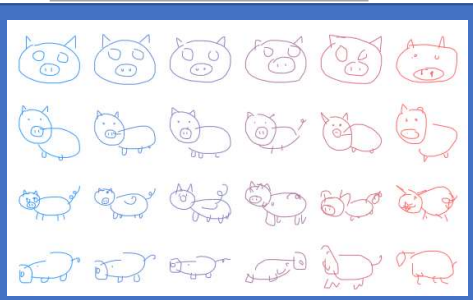
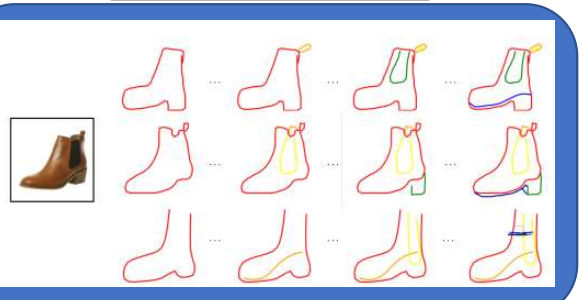
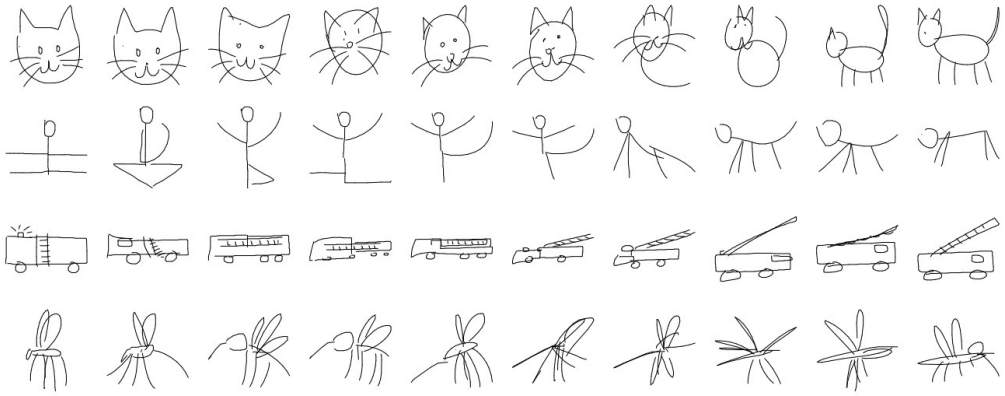


Photo2sketch



Intro

- 컴퓨터(모델)가 스케치를 생성하도록 하는 연구는 꾸준히 진행되어 왔다.



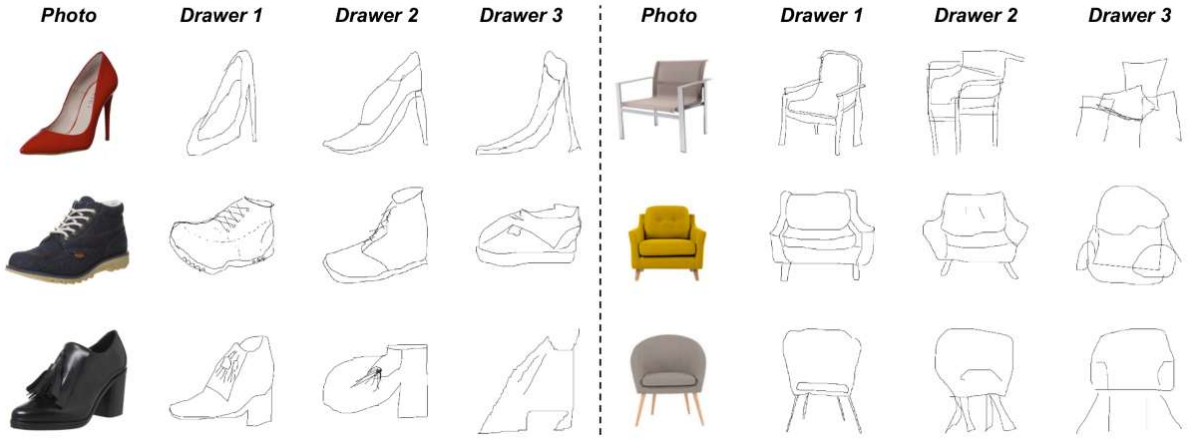
대표적인 모델인 Sketch-RNN의 결과 이미지

- **하지만 생성된 sketch의 품질이 매우 낮다.**
- **생성된 스케치를 평가할 척도도 불분명하다.**

Intro

ㅇ 모델을 통해 생성된 스케치는 왜 품질이 낮을까?

- 사람의 스케치는 동일한 instance 를 묘사할 때에도 다양한 수준의 정교함과 추상도를 보여준다. → 문제 자체가 쉽게 해결할 수 없는 문제



Intro

ㅇ 모델을 통해 생성된 스케치는 왜 품질이 낮을까?

- 사람의 스케치는 동일한 instance 를 묘사할 때에도 다양한 수준의 정교함과 추상도를 보여준다. → 문제 자체가 쉽게 해결할 수 없는 문제

- 모델에 사용할 sketch 데이터의 품질도 나쁘고 개수도 적다. (photo-sketch 데이터는 특히 현저히 적다 → supervised learning이 힘들다.)

Intro

ㅇ 모델을 통해 생성된 스케치는 왜 품질이 낮을까?

- 사람의 스케치는 동일한 instance 를 묘사할 때에도 다양한 수준의 정교함과 추상도를 보여준다. → 문제 자체가 쉽게 해결할 수 없는 문제

- 모델에 사용할 sketch 데이터의 품질도 나쁘고 개수도 적다. (photo-sketch 데이터는 특히 현저히 적다 → supervised learning이 힘들다.)

- 스케치-사진 도메인 사이의 도메인 격차가 매우 크다. → 도메인 격차가 큰 도메인 끼리의 변환에는 noise가 발생할 확률이 높다.

Intro

모델이 스케치를 잘 그리게 하려면 어떻게 해야할까?

Intro

모델이 스케치를 잘 그리게 하려면 어떻게 해야할까?

사람이 하는 것처럼 모델이 스케치를 하려면 어떻게 해야할까??

Intro

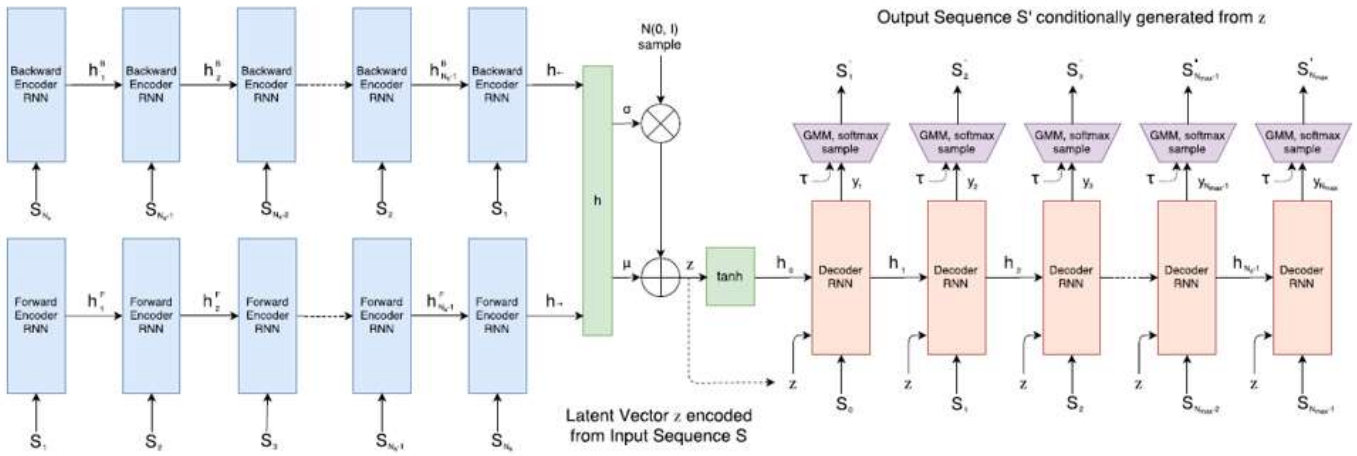
모델이 스케치를 잘 그리게 하려면 어떻게 해야할까?

사람이 하는 것처럼 모델이 스케치를 하려면 어떻게 해야할까??

펜을 종이에 닿게 한다.
→ 한 획(stroke)를 그린다
→ 펜을 들어올린다
→ 다음 획(stroke) 위치로 이동한다.
…

Difference from other models

ㅇ 기존의 방법



- Sketch generation 모델 중 가장 유명한 모델인 Sketch-RNN
  - ㄴ 각 스케치를 펜을 제어하는 일련의 동작(이동방향, 펜을 들어올리는 시기, 그리기를 중지하는 시기)로 표현
  - ㄴ 스케치를 스트로크 하나씩 시간순으로 그려야 하므로 RNN 사용
  - ㄴ 그러나 합성 스케치는 특정 물체 사진을 조건으로 하는 것은 아니다.

Difference from other models

ㅇ 해결하려고 하는 문제는 물체 사진으로부터 스케치를 생성하는 것

- 기존의 기술 덕분에 neural sketcher 역할을 하는 생성 시퀀스 모델은 구성할 수 있으나 합성된 스케치는 특정 물체 사진을 조건으로 하는 것이 아니다. → **CNN을 통해 사진을 인코딩 하여 neural sketcher에 넘겨주자**

- 도메인 차이가 큰 것 때문에 결과 스케치의 품질이 안 좋은 것을 극복하자. → **4개의 sub-model로 나누어 도메인 차이를 극복하자.**

- supervised learning에 사용되는 sketch-photo 데이터들은 도메인 차이가 커 noisy한 supervised signal만을 제공. → **multi-task supervised and unsupervised hybrid learning을 통해 더 나은 인코더와 디코더를 학습하자**

Difference from other models

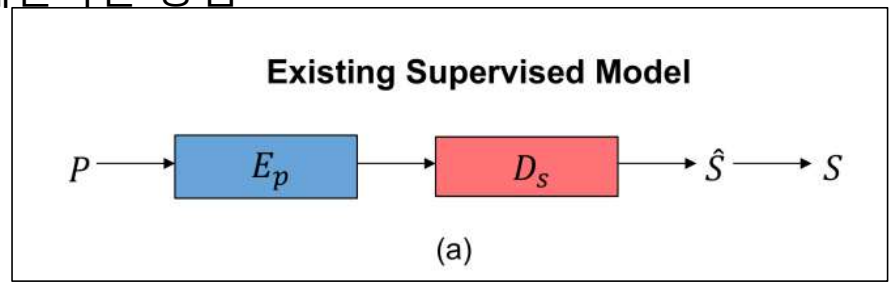
ㅇ 제안하는 hybrid learning framework는 다음과 같은 점에서 기존 접근 방식과 크게 다르다.

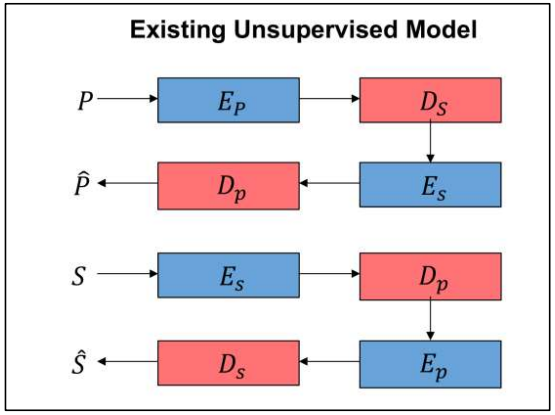
1. Noise가 많은 supervised signal을 최대한 활용하기 위해 다중 학습 프레임워크에서 supervised 방식과 unsupervised 방식을 결합한다. 특히 다양한 작업에서 인코더와 디코더를 공유함으로써 기존의 photo2sketch 합성 작업에 보다 **강력하고 효과적인 인코더와 디코더를 얻을 수 있다.**

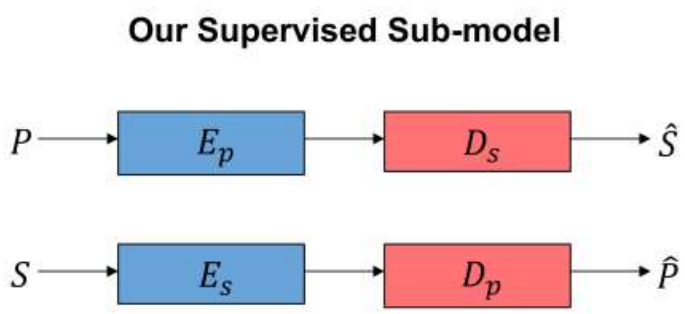
2. Cycle consistency를 기반으로 하는 기존의 unsupervised 모델과 달리 우리의 unsupervised 모델을 Shortcut Cycle consistency를 활용한다. Reconstruct를 위해 입력 도메인으로 돌아가기 위해 **다른 도메인을 통과하는 대신 우리의 모델은 바로 가기를 택하고 각 도메인 내에서 재구성을 완료하여 도메인 차이를 극복한다.**

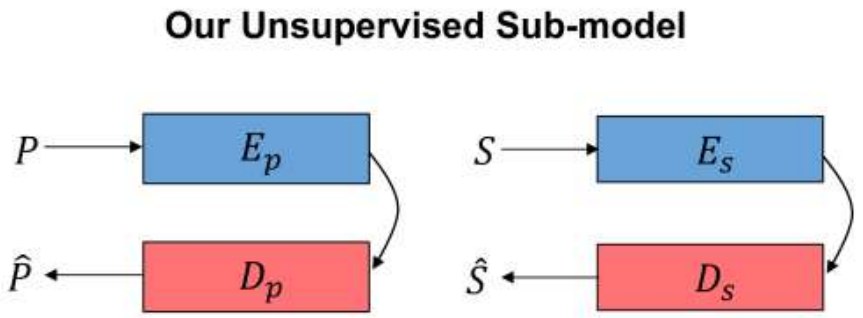
Difference from other models

ㅇ 제안하는 방법

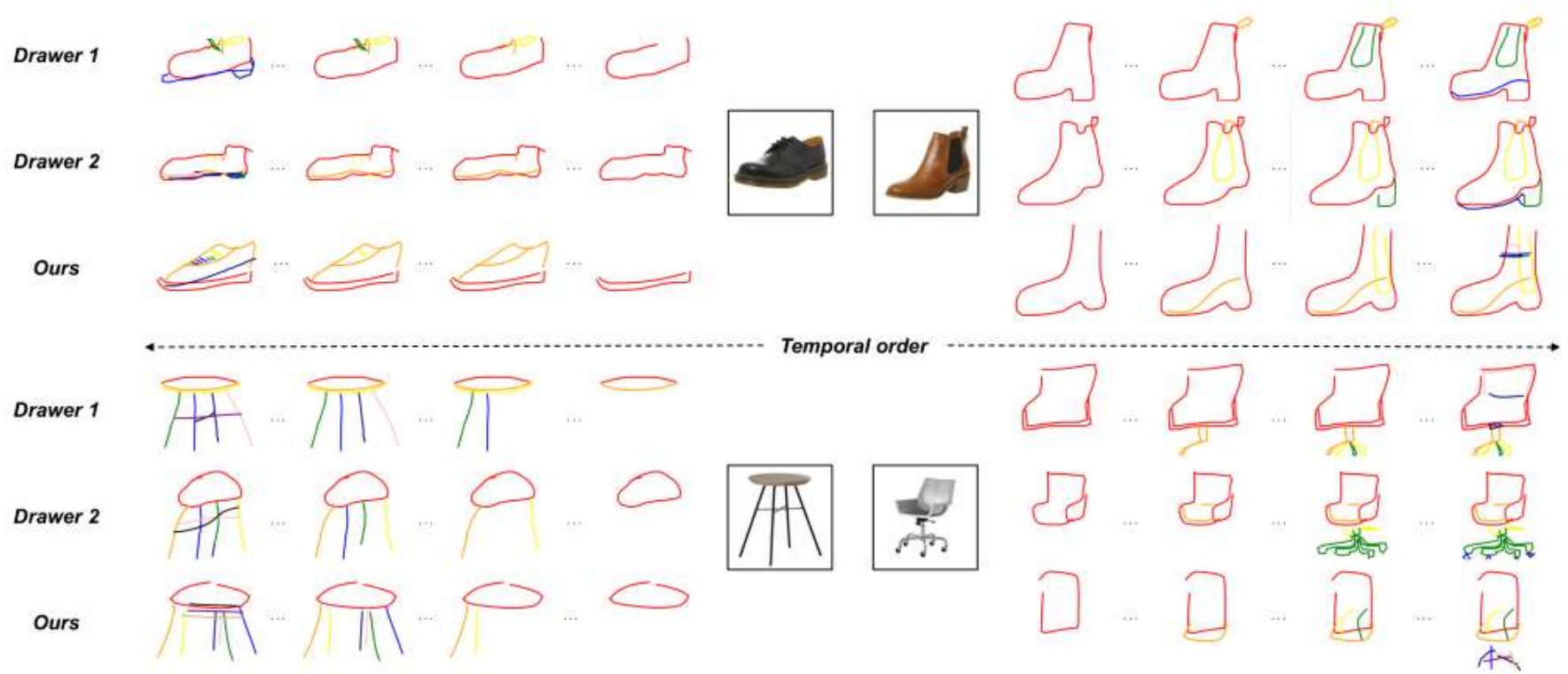








Difference from other models

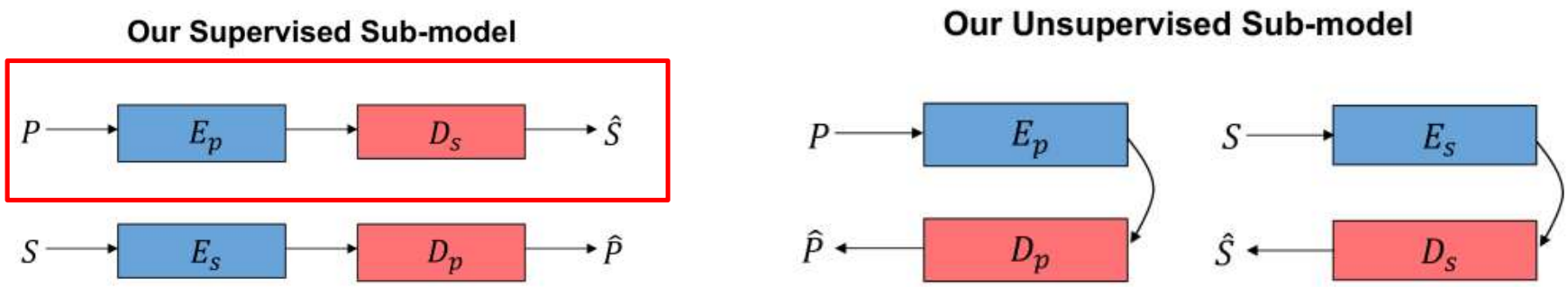


Difference from other models

ㅇ Contribution

1. 최초로 stroke-level의 cross-domain sketch generate 하는 모델
2. 주관적이고 다양한 사람의 그림 스타일에 의해 야기되는 noisy한 문제들을 식별하고 **multi-task supervised and unsupervised hybrid learning 으로 솔루션을 제안. Unsupervised learning은 shortcut cycle consistency를 통해 보다 효과적으로 달성**
3. **Photo-sketch paired 데이터 제작할 수 있는 모델**

Method



- 인코더와 디코더로 구성된 4개의 submodel로 구성된다.

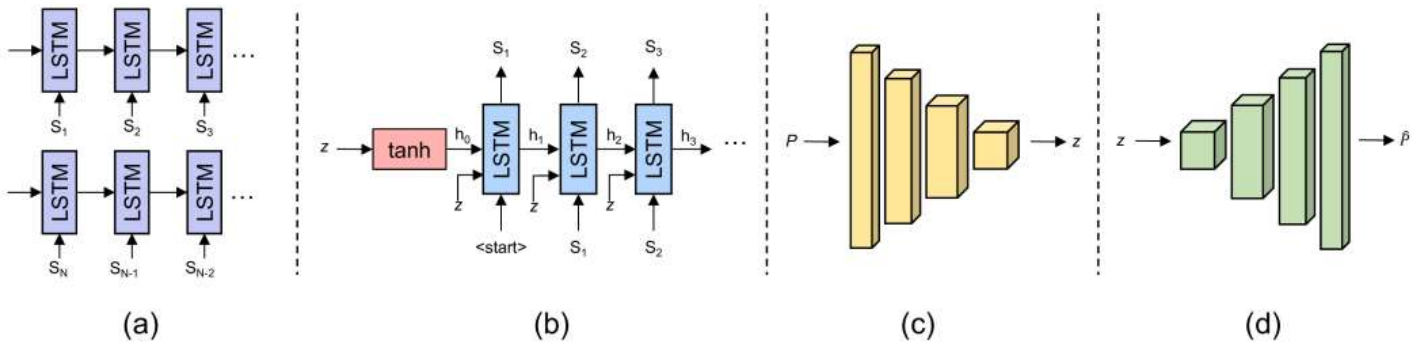
(1) 사진을 스케치로 변환하는 supervised submodel

(2) 스케치를 다시 사진 도메인으로 매핑하는 supervised submodel

(3) 사진을 reconstruct하는 unsupervised submodel

(4) 스케치를 reconstruct하는 unsupervised submodel

Method



○ Encoders

- 두 개의 인코더 $E_p(c)$와 $E_s(a)$는 각각 CNN, RNN으로 사진, 스케치를 입력으로 받아 latent vector를 출력한다.

- 특히 E_s는 양방향 LSTM이다. FC 레이어에서 IID 가우시안 분포에서 샘플링된 랜덤 벡터 N(0,I)와 융합하여 최종 임베딩 레이어를 구성한다.

Method

ㅇ 스케치 연구에서 사용되는 스케치의 포맷

- 스케치는 스트로크의 집합

- 스트로크는 벡터형식으로 표현 (△x, △y,p1,p2,p3)

- **△x, △y**는 각각 스트로크의 변량을 의미

- **p1,p2,p3**는 펜의 상태를 의미하며 one-hot 이다.
  ㄴ **p1** : 펜이 현재 용지와 닿고 있고 다음 포인트와 현재 포인트를 연결하는 선이 그려지고 있음을 나타냄
  p2 : 현재 지점 이후에 펜이 용지에서 들어 올려지고 다음에는 선이 그려지지 않음
  p3 : 스케치가 끝남

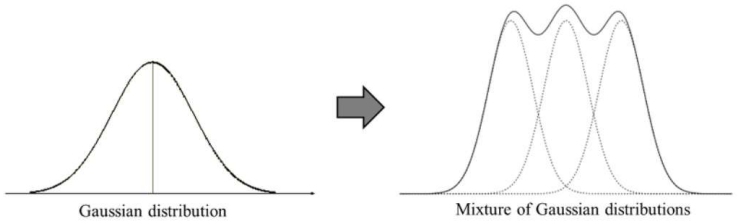
Method

ㅇ △x, △y는 GMM(Gaussian mixture model)을 사용해 다음 수식으로 나타낸다. P_{xy} 는 (x,y의 상관관계를 뜻함)

$$p(\Delta x, \Delta y) = \sum_{j=1}^{M} \Pi_j \mathcal{N}(\Delta x, \Delta y \mid \mu_{x,j}, \mu_{y,j}, \sigma_{x,j}, \sigma_{y,j}, \rho_{xy,j}), \text{ where } \sum_{j=1}^{M} \Pi_j = 1 \quad (3)$$

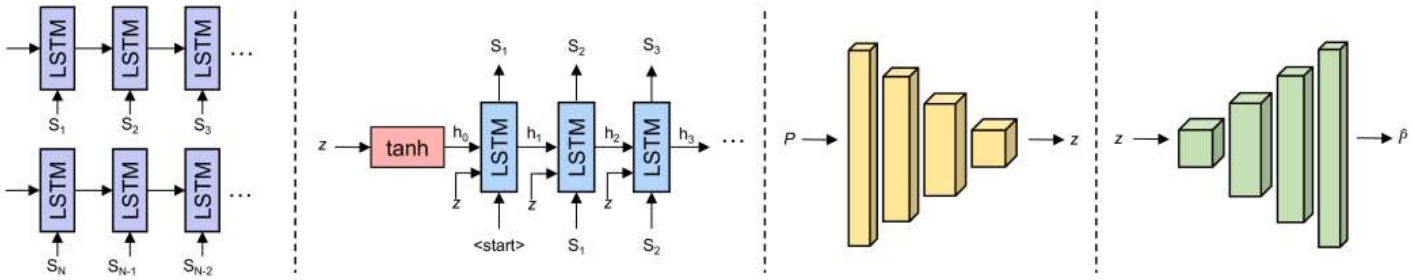
- △x, △y의 분포는 각각 독립적인 정규분포를 따른다.(IID Gaussian)

- 다음 스트로크 후보들은 독립적인 가우시안 분포들의 집합(Gaussian mixture model).



→ 모델이 스케치를 그리는 과정은 현재 위치, time stamp에 따라 다음 스트로크를 예측하는 것

Method



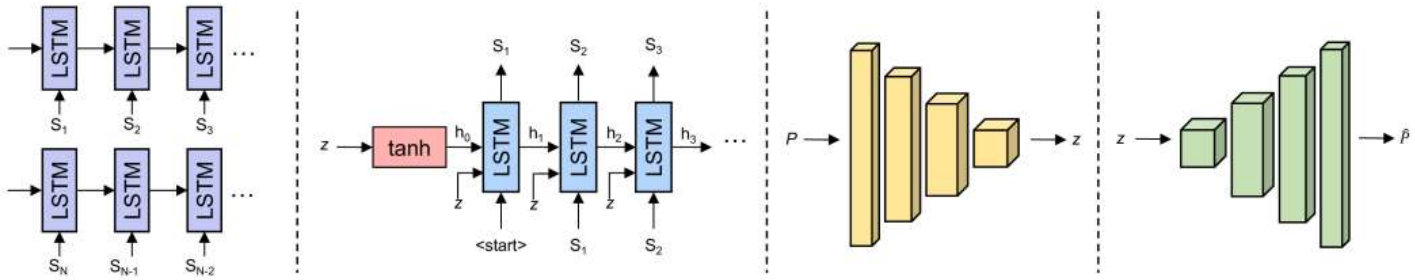
ㅇ Encoders

- 두 개의 인코더 $E_p(c)$와 $E_s(a)$는 각각 CNN, RNN으로 사진, 스케치를 입력으로 받아 latent vector를 출력한다.

- 특히 E_s는 양방향 LSTM이다. FC 레이어에서 IID 가우시안 분포에서 샘플링된 랜덤 벡터 N(0,I)와 융합하여 최종 임베딩 레이어를 구성한다. → 다음 스트로크를 예측하는 것

$$z = \mu + \sigma \odot \mathcal{N}(0, I)$$

Method



ㅇ Sketch Decoder

- Sketch decoder $D_s(b)$는 latent vector z에 조건화된 sketch stroke를 샘플링 하기 위해 LSTM 기반 시퀀스 모델을 구축한다.

- 이는 gaussian mixture model을 사용하여 각 sketch stroke offset(△x, △y)를 예측하고 각 timestamp에 대한 펜 상태 q_i를 categorical distribution로 모델링 함으로써 수행된다.

Method

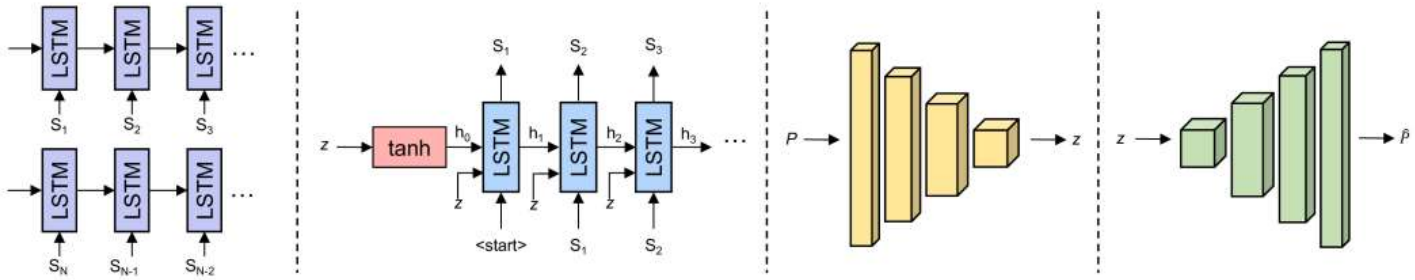
- Sketch Decoder loss

$$\mathcal{L}_{rnn}(S,\hat{S}) = \mathbb{E}_{x\sim S, y\sim \hat{S}}\left[-\frac{1}{N_{max}}\left(\sum_{i=1}^{N_s}\log(p(\Delta s_{x_i},\Delta s_{y_i}|x,y)\right) - \sum_{i=1}^{N_{max}}\sum_{k=1}^{3}p_{k,i}\log(q_{k,i}|x,y)\right)\right]$$

- N_{max}는 training set의 하나의 스케치의 최대 stroke 수이다.
- N_s는 특정 스케치의 stroke 수로써 N_{max}보다 클 수 없다.
- Index i는 time stamp를 의미하고 k는 pen-state를 의미한다.

위 loss를 최소화 한다는 것은 현재 스트로크를 기반으로 다음 스트로크를 예측하는 능력을 기르는 것이다.

Method



ㅇ Photo Decoder

- Photo Decoder $D_p(d)$는 기존에 자주 사용되던 CNN-based deconvolutional-upsampling block을 사용한다.

- 원본 이미지와 생성된 이미지의 차이를 측정하는 L2 loss를 사용한다.

$$\mathcal{L}_{\rightarrow p}(P, \hat{P}) = \mathbb{E}_{x \sim P, y \sim \hat{P}}[||x - y||_2]$$

Method

- **Short cycle consistency**

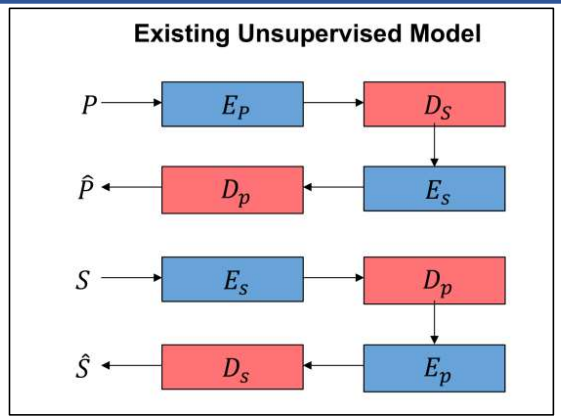
- Supervised signal을 제공하기 위한 paired example이 존재하기 때문에 사진부터 스케치까지 단방향 매핑을 학습하는 것으로 충분할 것으로 예상할 수 있다.

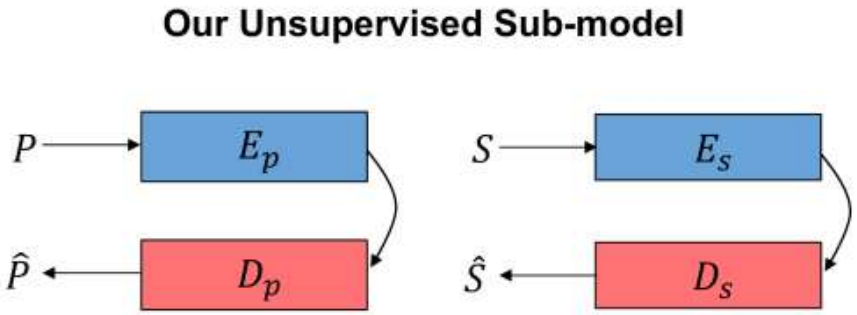
- 하지만 앞써 말한 것과 같이 photo-sketch pair는 약하고 noise가 많은 supervised signal을 제공하므로 단방향 매핑 기능을 효과적으로 학습할 수 없다.

→ 우리의 솔루션은 지도 학습 및 비지도 재구성 작업을 사용하여 양방향 매핑을 도입하는 것이다.

→ 네 개의 인코더와 디코더는 이러한 supervised 및 unsupervised 작업에 의해 공유되기 때문에, 그들은 다중 작업 학습의 혜택을 받는다.

Method



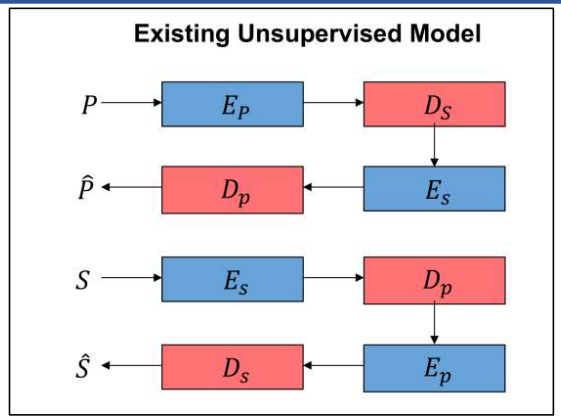


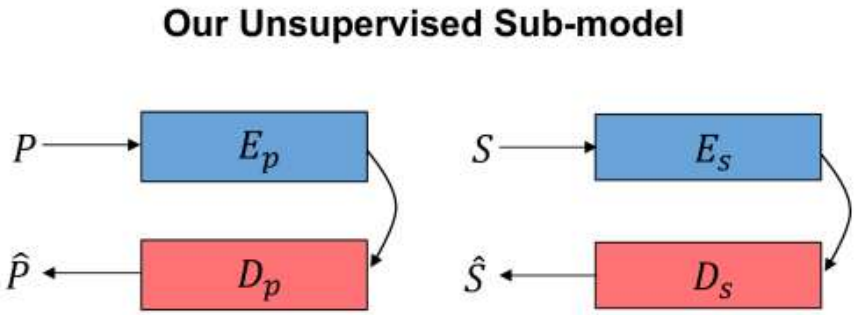
- unsupervised self reconstruction task에서 결과를 식별할 수 없다는 점에서 cycle consistency가 도입된다.

- 예를 들어 스케치 변환을 위한 연산은 $x \rightarrow E_p(x) \rightarrow D_s(E_p(x)) \rightarrow E_s(D_s(E_p(x)) \rightarrow D_p(E_s(D_s(E_p(x)))$이다. → 하지만 위 연산은 도메인 영역을 3번이나 넘어가기 때문에 결과의 품질이 낮다.

- 본 논문에서는 multi-task, 즉 supervised learning때 사용한 인코더와 디코더를 공유하고, paired data가 있기 때문에 위 연산을 $x \rightarrow E_p(x) \rightarrow D_s(E_p(x))$로 단축할 수 있다.

Method





- 이 결과 더 빠른 속도를 초래하는 것 외에도 기존 방식보다 훨씬 더 나은 photo2sketch 합성을 할 수 있다는 것을 발견했다.(도메인 간 이동이 적기 때문)

$$\mathcal{L}_{shortcut}(X,Y) = \mathcal{L}_{\to s}(Y, D_s(E_s(Y))) + \mathcal{L}_{\to p}(X, D_p(E_p(X)))$$

Method

- **Full Learning Objective**

- 네 개의 하위 모델들은 공동으로 학습된다. 따라서 다음과 같은 loss 역시 사용된다.

$$\mathcal{L}_{supervised}(X,Y) = \mathcal{L}_{\to s}(Y, D_s(E_p(X))) + \mathcal{L}_{\to p}(X, D_p(E_s(Y)))$$

- 또한 효율적인 posterior sampling을 가능하게 하도록 KL loss를 추가하여 4개의 하위 모델이 디코더에 공급하기 위해 유사한 분포를 사용하도록 한다.

$$\mathcal{L}_{KL} = \mathbb{E}_{x\sim X, y\sim Y, \hat{x}\sim \hat{X}, \hat{y}\sim \hat{Y}} [-\frac{1}{2}(1+\sigma^2 - \exp(\sigma))|x, y, \hat{x}, \hat{y}]$$

Full objective :

$$\mathcal{L}_{full}(X,Y) = L_{supervised}(X,Y) + \lambda_{shortcut}\mathcal{L}_{shortcut}(X,Y) + \lambda_{KL}\mathcal{L}_{KL}$$

Experiments & Evaluation

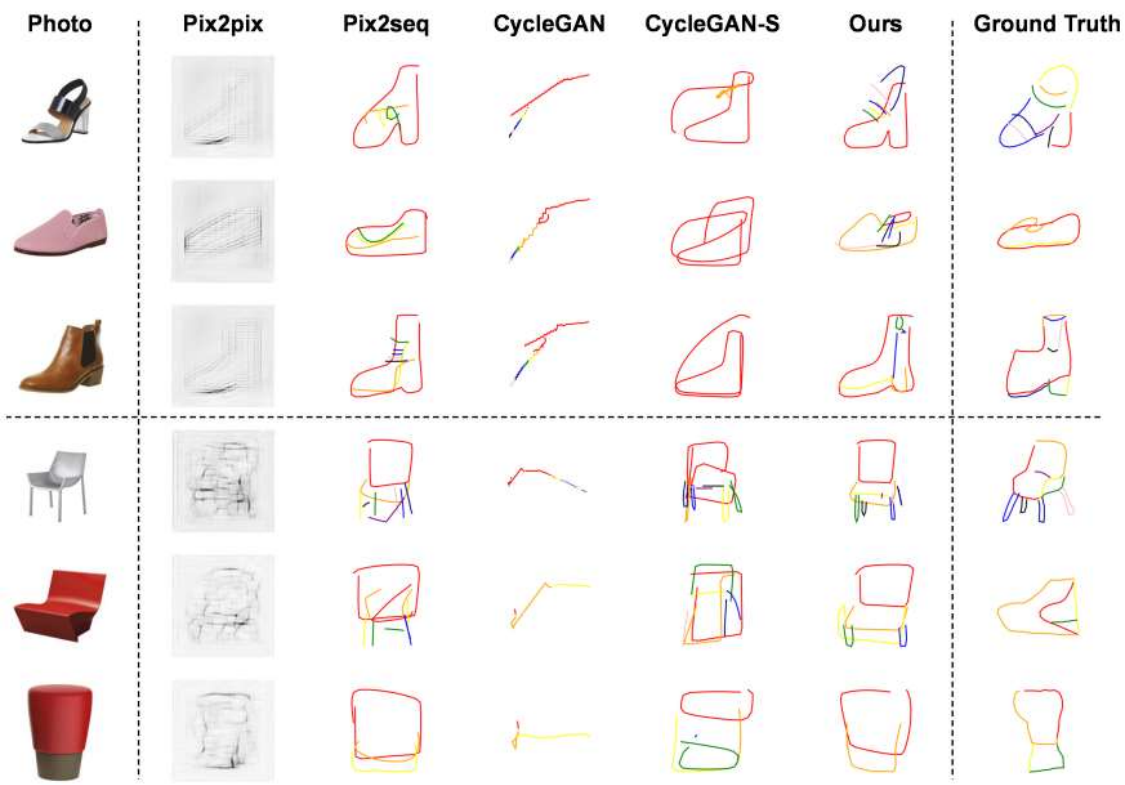
- 가장 큰 sketch photo dataset 인 QMUL-shoe-chair-V2 데이터 셋을 사용

- 위 데이터 셋의 sketch-pair 수가 제한되어 있기 때문에 QuickDraw 데이터셋을 사용해 모델을 pretrain 한다.

QMUL-shoe-chair-V2

QuickDraw

Experiments & Evaluation



Experiments & Evaluation

	Recognition		Retrieval	
ShoeV2	acc.@1	acc.@10	acc.@1	acc.@10
Human sketch [46]	36.50%	70.00%	30.33%	76.28%
Pix2pix [13]	0.00%	0.00%	0.50%	7.50%
Pix2seq [2]	51.50%	86.00%	4.50%	26.00%
CycleGAN [51]	0.00%	0.00%	0.50%	4.00%
CycleGAN-S	18.00%	51.50%	2.00%	18.00%
Our full model	53.50%	90.00%	6.00%	28.50%
ChairV2	acc.@1	acc.@10	acc.@1	acc.@10
Human sketch [46]	10.00%	35.00%	47.68%	89.47%
Pix2pix [13]	0.00%	0.00%	2.00%	16.00%
Pix2seq [2]	5.00%	51.00%	3.00%	31.00%
CycleGAN [51]	0.00%	8.00%	1.00%	7.00%
CycleGAN-S	12.00%	55.00%	6.00%	33.00%
Our full model	13.00%	55.00%	8.00%	36.00%

- Recognition 모델을 통해서 생성된 그림의 category를 맞출 수 있는지(domain level)

- Retrieval 모델을 통해서 생성된 그림의 원본 사진이 어떤 것인지 맞출 수 있는지(instance level)

Conclusion

- 사진 속 물체에 대한 사진으로부터 시각적으로 품질이 뛰어난 스케치를 합성할 수 있는 최초의 stroke-level 모델을 제안했다.

- Photo-sketch 사이의 약한 supervised signal에 대처하기 위해 multi-task supervised and unsupervised hybrid learning을 제안했다.

- 논문 속 모델이 많은 대안모델보다 정성적인 평가와 정량적인 평가 모두 우수한 성능을 달성했다는 것을 보여주었다.

- FG-SBIR 작업에 대한 데이터로 사용할 수 있는 품질이 좋은 sketch-photo pair 데이터를 생성할 수 있다.

감사합니다