Collaborative Distillation for Ultra-Resolution Universal Style Transfer 논문 리뷰

- Collaborative Distillation for Ultra-Resolution Universal Style Transfer.
- CVPR 2020.
- https://github.com/MingSun-Tse/Collaborative-Distillation

Abstract

- Universal style transfer 가 무엇인가?
  - Universal style transfer 방식은 style transfer에서 방대한 데이터로부터 사전에 학습된 깊은 합성곱 신경망 구조가(deep Convolution Neural Network, eg,. VGG-19) 가진 풍부한 표현 능력을 활용하는 방식을 말함.
- 문제점은 무엇인가?
  - 기존 universal style transfer 방식을 적용하기에는 너무 거대한 모델의 사이즈로 인하여 ultra-resolution 영상을 처리에 메모리가 많이 필요하다는 단점 존재.
- 논문의 저자들은 어떻게 문제를 해결했는가?
  - 새로운 knowledge distillation 방식(저자들은 이를 Collaborative Distillation 이라고 지음, 여기서 왜 collaborative 라는 이름이 붙었는지 확인 할 것)을 제안.
    - Convolution filter를 줄여주는 encoder-decoder 구조 기반 neural style transfer임.

Abstract

- 제시된 방식의 주된 아이디어는 무엇인가?
  1. 논문에서 제안한 Collaborative Distillation 방식은 encoder-decoder 쌍들이 독특한 협력 관계(exclusive collaborative relation)을 구성하다는 특징을 발견한 것에 기반하여 시작됨.
  2. Collaborative Distillation 방식을 적용할 때 feature의 사이즈가 맞지 않는다는 점을 해결하기 위하여, linear embedding loss를 제안하였음.
     - 여기서, linear embedding loss의 역할은 학생 네트워크가(student network) 선생님 네트워크에게(teacher network) linear embedding을 학습하는 방식을 말함.
- 논문의 실험 결과와 그 의미는 무엇인가?
  - 기존 universal style transfer의 접근 방식(WCT와 AdaIN)과 달리 모델의 사이즈가 매우 줄어들었음.
  - 다른 stylization paradigm에 논문이 제안한 알고리즘을 일반화가 가능한 최적화 기반 stylization scheme이 가능함.

논문에서 중점적으로 봐야할 사항

- Universal style transfer 방식이 정확히 무엇인가?
- 기존 universal style transfer 방식으로 처리하면 메모리가 많이 드는 이유는 무엇인가?
- 논문에서 제안한 Collaborative Distillation 방식이 정확히 무엇이며 어떠한 배경에서 이러한 방식을 제시하게 되었는가?
- 저자들이 주장한 encoder-decoder 쌍들이 독특한 협력 관계(exclusive collaborative relation)가 정확히 무엇이며 이것을 어떻게 발견하게 되었는가?
- 제시한 방식의 핵심 방식인 linear embedding loss가 정확히 무엇이며 이것이 어떤 기반에서 제시되게 되었는가?
- 논문에서 WCT의 기존 기법을 많이 착용하고 비교한 것 같은데 WCT 방식이 정확히 무엇인가?

Introduction

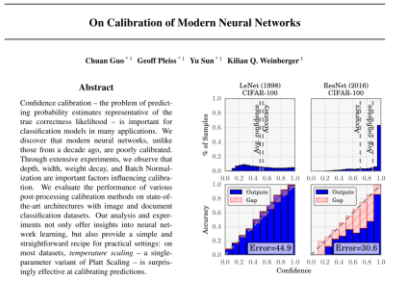
- Universal style transfer 란 무엇인가?
  - 사전에 학습되지 않은 어떠한 새로운 style 영상이 들어와도 동작이 가능한 style transfer 방식을 말함.
  - Universal style transfer 는 무작위의 style 의 통계적인 정보를 잘 얻어내어 효과적인 representation 을 추출 하는 것이 핵심임.
- 최근 universal style transfer 의 연구 방향은 어떠하였는가?
  - 최근 universal style transfer 논문에서 제안된 방식들은 VGG-19 와 같이 사전에 잘 훈련된 깊은 신경망에서 추출된 representation 을 활용하는 방식 이었음.
  - 하지만, VGG-19 와 같은 거대한 사이즈의 모델을 활용하면 해상도가 큰 영상이 입력으로 들어갔을 때, 메모리가 너무 많이 든다는 단점이 있음.

Introduction

- 논문에서 주목한 점은 무엇인가?
  - 논문에서는 거대한 CNN 모델이 주어졌을 때 모델의 파라미터를 줄이면서 모델을 압축하는 기법인 모델 압축(model compression) 기법에 주목하였음.
  - 논문의 저자들은 이러한 모델 압축 기법에 집중 하되 기존에 논문들은 classification 과 같은 high-level task 에서 주로 제안되었지만 저자들은 low-level task 에 특화된 새로운 모델 압축 기법을 제안하였음.
  - 논문의 저자들은 모델 압축 기법 중 knowledge distillation(KD) 방식에 특히 더 집중하였음.
    - Knowledge distillation(KD) 방식이란 teacher 라고 불리는 방대한 모델의 지식을 student 라고 불리는 작은 네트워크에게 전달해 주는 방식을 말함.
      - 논문의 저자들은 지식(knowledge)를 크게 다음 두가지로 나눌 수 있다고 언급함.
        1. 자신에게 내제된 타고난 클래스 유사도 구조(inherent class similarity structure)를 반영 하기 위하여 지식(knowledge) 을 smooth/soft probability 가 knowledge 가 될 수도 있고 혹은
        2. 서로 다른 샘플(different samples) 사이에 유사도 구조(similarity structure)을 반영 하기 위하여 샘플 간의 관계(sample relations) 가 knowledge 가 될 수도 있음.

Smooth/soft probability

- 그렇다면 , smooth/soft probability 가 무엇일까?
  - [1] 은 probability smoothing 은 학습 데이터에서 관측하지 않은 사건에 대하여 non-zero 확률을 부여하는 것으로 더 많은 사건으로 질량 함수를 나눴을 때, 확률 분포가 더 smooth 해 지는 효과가 있다고 언급.
    - Label smoothing 과 비슷한 맥락이라고 이해.


On Calibration of Modern Neural Networks

Chuan Guo [*1] Geoff Pleiss [*1] Yu Sun [*1] Kilian Q. Weinberger [1]

Abstract

Confidence calibration – the problem of predicting probability estimates representative of the true correctness likelihood – is important for classification models in many applications. We discover that modern neural networks, unlike those from a decade ago, are poorly calibrated. Through extensive experiments, we observe that depth, width, weight decay, and Batch Normalization are important factors influencing calibration. We evaluate the performance of various post-processing calibration methods on state-of-the-art architectures with image and document classification datasets. Our analysis and experiments not only offer insights into neural network learning, but also provide a simple and straightforward recipe for practical settings: on most datasets, *temperature scaling* – a single-parameter variant of Platt Scaling – is surprisingly effective at calibrating predictions.

LeNet (1998) CIFAR-100

ResNet (2016) CIFAR-100

Error=44.9

Error=30.6

Outputs

Gap

Accuracy

% of Samples

Confidence

[1] https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_936
https://datascience.stackexchange.com/questions/25639/what-does-smooth-soft-probablity-mean

Introduction

- 논문에서 주목한 knowledge distillation 을 논문의 목표에 맞게 어떻게 변형 하였는가?
  - 비록, 앞에서 언급한 지식(knowledge) 은 top of one-hot label 에 추가적인 정보를 제공하는 효과가 있으며 학생(student)의 성능을 끌어 올리는데 효과가 있다고 말하지만 여기서 말하는 추가적인 정보는 label-dependent(저자들이 사용한 표현) 하기 때문에 low-level task 에서는 적용이 힘들다고 주장함.
  - 저자들은 대다수의 neural style transfer 네트워크가 주로 encoder decoder 구조로 이루어져 있다는 점과 decoder 가 encoder 의 지식을 통해서 학습되고 있다는 사실에 집중하였음.
    - 여기서 주목해야 할 점은 stylization 과정에서 exclusive collaborative relationship(저자들이 사용한 표현) 즉, 독특한 협력 관계를 구성하고 있다는 점임.

Introduction

- 논문에서 주목한 knowledge distillation 을 논문의 목표에 맞게 어떻게 변형 하였는가?
  - 저자들은 대다수의 neural style transfer 네트워크가 주로 encoder decoder 구조로 이루어져 있다는 점과 decoder 가 encoder 의 지식을 통해서 학습되고 있다는 사실에 집중하였음.
    - 여기서 주목해야 할 점은 stylization 과정에서 exclusive collaborative relationship(저자들이 사용한 표현) 즉, 독특한 협력 관계를 구성하고 있다는 점임.
    - 과연 논문에서 주목한 독특한 협력 관계 란 무엇일까?
      - 디코더 D 가 특정 엔코더 E 와 독점적으로(exclusively) 학습되기 때문에 만약 다른 엔코더 $\hat{E}$ 가 디코더 D 와 작용(work)할 수 있다면 사실은 다른 엔코더 $\hat{E}$ 는 엔코더 E 의 기능을 할 수 있다고 주장.
      - 저자들은 이러한 생각에 기반을 두고, neural style transfer 에 특화하여 지식을 증류하는(distill) 기법을 제안함.

독특한 협력관계를 나타내 주는 그림

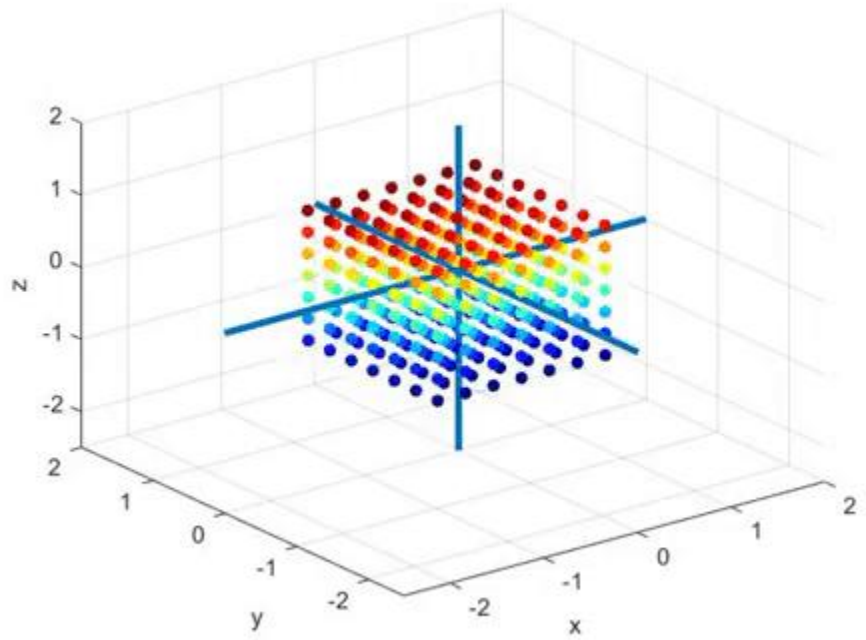
Figure 2: Examples of the exclusive collaboration phenomenon on two different encoder-decoder collaborative relationships: image reconstruction of WCT [39] (Row 1) and style transfer of AdaIN [24] (Row 2). Column 1 is the input, the other four columns show the outputs using different encoder-decoder combinations. If the two encoder-decoder pairs (E1-D1, E2-D2) are trained independently, the encoder can only work with its *matching* decoder.

Introduction

- 논문에서 제안한 기법은 무엇인가?
  - VGG-19 네트워크와 같은 거대한 엔코더가 주어졌다고 가정하면, 두단계의 압축 과정이 있음.
    1. 엔코더를 위한 협력 네트워크(collaborator network) 를 학습.
    2. 거대한 엔코더를 작은 사이즈를 가진 엔코더로 대체하고 난 후 작은 엔코더를 협력 네트워크(collaborator network)를 고정시킨 채로 학습.
       - 이때, 작아진 엔코더는 거대한 사이즈를 가진 channel 의 수가 더 줄어들기 때문에 그들의 출력 feature 의 dimension 도 마찬가지로 축소함.
       - 즉, 축소된 네트워크는 직접 협력 네트워크와 작업 할 수 없음.
       - 이를 해결하기 위하여, 선생님의 출력물을 미리 linear embedding 으로 학습을 시켜 학생의 차원에 맞추어 줌.
       - 결국 선생님과 학생이 협력하기 전에 이전에 선생님을 한번 linear embedding 시켜 학생의 차원에 맞게 차원을 낮추어 줌.

What is dimension reduction?

- 차원축소란 무엇인가?



Contribution

- 논문에서는 무엇을 새로 제시하였는가?

Related work

- 이전에는 어떤 연구가 이루어 졌으며 저자들이 제안한 방식과 관련점과 차이점은 무엇이 있을까?

Proposed Method – collaborative distillation

- 저자들이 제안한 collaborative distillation의 이론적인 근거는 무엇인가?
  - 많은 stylization 방식들이 encoder decoder 를 구조로 이루고 있는 이유는 encoder 에서 style rendering 을 위한 deep representation을 학습하고 decoder 에서 학습된 representation을 stylized 된 영상으로 되돌리기 때문임.

  1) 인코딩 단계
     - 인코딩 단계에서 무작위의 style 정보도 효율적으로 인코딩 하기 위하여 style 정보가 직접적으로 인코딩 되지 않기 때문에, encoder는 무작위로 style 영상이 입력되어도 style 영상을 나타내는 representation 을 충분히 추출하기 위하여 충분한 표현 능력을(expressive) 가지고 있어야함.
       - 기존 기법들을 이러한 방대한 수용능력(massive capacity) 과 계층적인(hierarchical) 구조를 가지고 있어야 하기 때문에 사전에 충분히 학습된 VGG-19 네트워크를 encoder 로 채택하고 있음.

Proposed Method – collaborative distillation

- 저자들이 제안한 collaborative distillation의 이론적인 근거는 무엇인가?
  2) 디코딩 단계
     - 디코딩 단계는 서로 다른 stylization schemes 에 크게 의존하기 때문에 encoder와의 협력 관계도 다를 수 있음.
       - 논문에서는 기존 방식인 WCT와 AdaIN을 예시로 들고 있음.
         - WCTs 기법은 stylization에 직접적으로 참여하지 않기 때문에 encoder와 decoder의 협력적인 관계(collaborative relationship) 가 필요하다고 주장.
           - Why? Style transfer 단계에서는 decoder 를 학습하지 않고 hand-crafted 방식으로 feed-forward 로 결과물을 출력하는 방식이기 때문이라고 추측. 뿐만 아니라 autoencoder 구조의 복원 단계에서도 인코딩 단계는 고정된 채(학습하지 않고) 디코딩 단계만 학습하기 때문에 협력적인 관계를 완전히 활용하지 못한다는 지적이 합리적임.
         - AdaIN 기법은 stylization 과정에서 디코딩을 직접 학습하고 있는데 여기서는 content와 style 영상이 직접 인코딩 과정에 입력으로 들어가 content feature 가 style feature 의 통계적인 특성(평균, μ과 분산, σ)에 의하여 변환 과정을 거침. 최종적으로 디코딩 단계에서는 style feature 에 통계적인 특성에 맞게 변환된 content feature 을 stylized 영상으로 출력해줌. 여기서, stylized 영상은 content (혹은 style) 의 거리적인 관점에서 content (혹은 style) 과 가까움(close) 하다고 가정하고 있음. 즉, AdaIN 에게 있어서 협력적인 관계는 style transfer 임.

Proposed Method – collaborative distillation

- 저자들이 제안한 collaborative distillation의 이론적인 근거는 무엇인가?
  - 저자들은 대표적인 style transfer 기법인 두 기법의 가장 큰 문제를 무엇으로 보았을까?
    - 저자는 두 방식 모두 decoder 가 encoder 의 지식(knowledge) 를 활용해서 학습하고 있다고 언급하였음. 즉, 이것은 decoder를 학습하는 단계에서 encoder의 지식이 decoder 에 새어 들어가고(is leaked into) 있다고 말함. 아마도(저자들도 단정짓는 표현을 하지는 않음) decoder D 가 오직 대응하는 관계에 있는 encoder E 와 작용 할 수 있을 것이라고 추측 하였고 이러한 관계를 볼트와 너트로 비유하였음. 이는 다른 encoder E'이 encoder E와 같은 구조를 가지고 있어도 encoder E' 는 decoder D 와 서로 작용하며 협력 할 수 없어 비효율적임.
      - 이러한 독점 관계는 decoder 가 대응하는 encoder에 특화되어 있기 때문에 decoder가 inherent 정보를 가지고 있음.
    - 저자들은 만약 새로운 encoder E' 이 decoder D 과 양립 가능하다면 encoder E' 이 기존 encoder E의 역할을 수행 할 수 있기 때문에 대체 가능하다고 주장함.
      - 이점을 활용하여, 기존 encoder E 보다 더 작은 사이즈의 encoder E' 을 사용한다면 모델을 압축 할 수 있음을 의미한다고 주장.

Proposed Method – collaborative distillation

- 저자들이 주장한 이론적인 근거를 바탕으로 collaborative distillation 과정이 어떻게 이루어졌을까?
  - 저자들은 collaborative distillation 과정을 두단계로 구분하였음.
    1. 기존에 사전에 학습된 큰 encoder 에 대응하여 collaborator 네트워크를 학습함.
    2. 기존 encoder E 를 더 작아진 encoder E' 로 대체함.

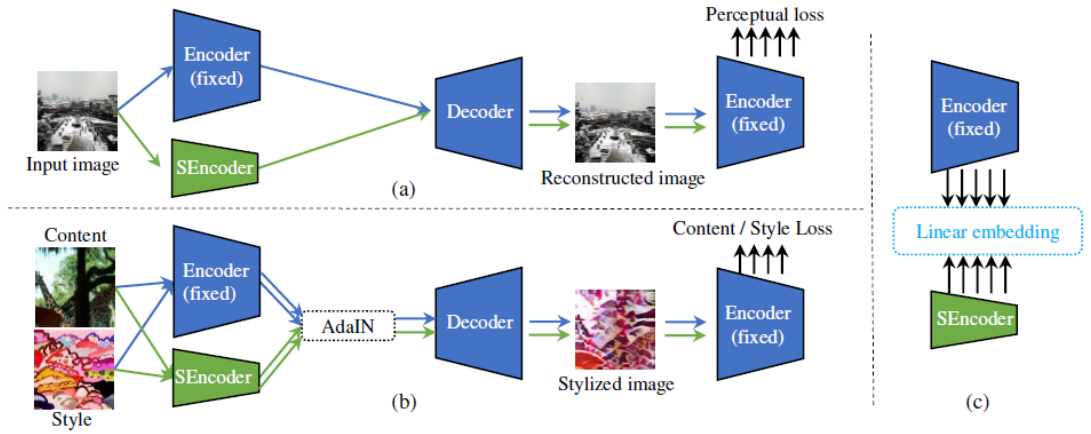


Figure 3: Illustration of the proposed Collaborative Distillation framework (best viewed in color). (a) and (b) depict two kinds of the encoder-decoder collaborative relationship for universal neural style transfer: image reconstruction for WCT [39] and style transfer for AdaIN [24], respectively. Blue arrows show the forward path when training the collaborator network (namely, the decoder). Green arrows show the forward path when the small encoder ("SEncoder") is trained to functionally replace the original encoder ("Encoder"). (c) shows the proposed linear embedding scheme to resolve the feature size mismatch problem and infuse more supervision into the middle layers of the small encoder.

Proposed Method – collaborative distillation

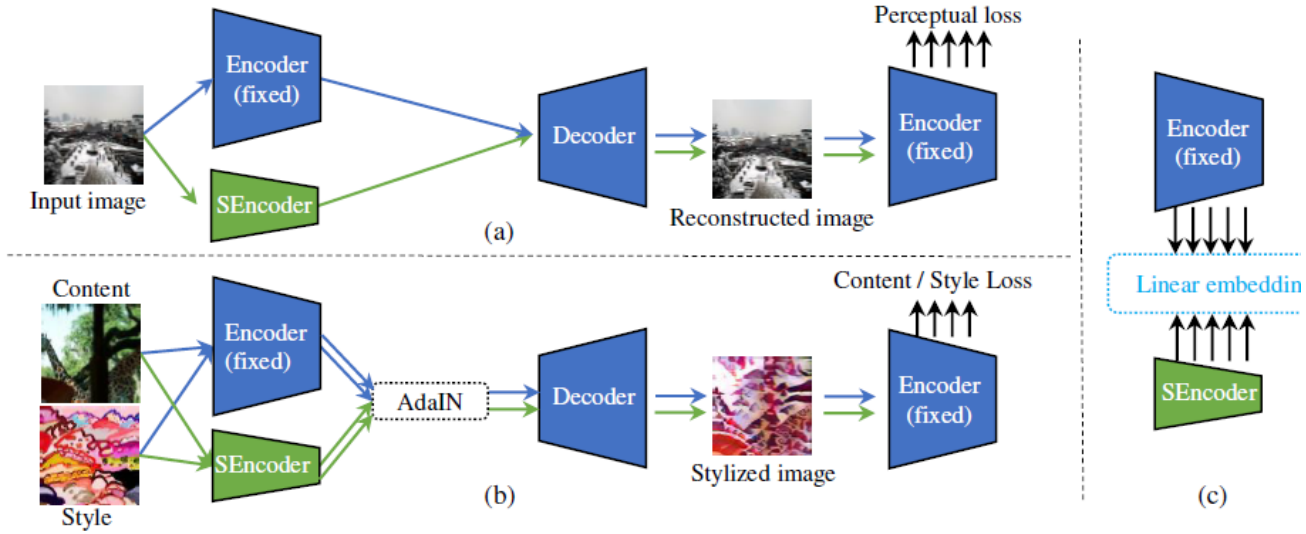


Figure 3: Illustration of the proposed Collaborative Distillation framework (best viewed in color). (a) and (b) depict two ki of the encoder-decoder collaborative relationship for universal neural style transfer: image reconstruction for WCT [39] style transfer for AdaIN [24], respectively. Blue arrows show the forward path when training the collaborator netw (namely, the decoder). Green arrows show the forward path when the small encoder ("SEncoder") is trained to function replace the original encoder ("Encoder"). (c) shows the proposed linear embedding scheme to resolve the feature mismatch problem and infuse more supervision into the middle layers of the small encoder.

- 저자들이 주장한 이론적인 근거를 바탕으로 collaborative distillation 과정이 어떻게 이루어졌을까?
  - 저자들은 collaborative distillation 과정을 두단계로 구분하였음.
    1. 기존에 사전에 학습된 큰 encoder 에 대응하여 collaborator 네트워크를 학습함.
    2. 기존 encoder E 를 encoder E' 에 대체함.

Proposed Method – collaborative distillation

1. 기존에 사전에 학습된 큰 encoder 에 대응하여 collaborator 네트워크를 학습함.
   1) WCTs 방식에 적용 하였을 때
      - Decoder를 사전에 학습 할 때 autoencoder 구조를 기반으로 학습을 하는 방식으로 decoder가 가능한 입력영상으로 충실하게(faithful) 복원하는 것이 핵심임.
   2) AdaIN 방식을 적용 하였을 때
      - Style transfer 과정에서 decoder 를 직접적을 학습시킴.

Universal Style Transfer via Feature Transforms Proposed Algorithm for decoder – reconstruction

- encoder
  - 이를 위하여 사전에 학습된 VGG-19 구조를 encoder 로 사용하였고 이에 대한 학습은 하지 않고 파라미터를 고정.
- decoder
  - 원본 영상으로 VGG features을 원본 영상으로 다시 복원을 해 주기 위하여 decoder 를 학습.
  - 구조는 VGG-19 네트워크와 대칭적인 구조로 구성.
    - 단, feature maps을 업 샘플링 할 때 nearest-neighborhood 방식 사용.
- 손실 함수
  - Euclidean distance.
    - $\|I_0 - I_i\|^2$.
  - Perceptual distance.
    - $\|\varphi(I_0) - \varphi(I_i)\|^2$.
    - φ는 사전에 학습된 VGG encoder 에서 추출된 feature map을 의미.
  - L_{collab}
- 이렇게 학습을 진행 한 후에 style transfer 과정에서는 decoder 는 고정된 채 feature를 invert시키는 용도로 사용.

Arbitrary style transfer in real-time with adaptive instance normalization Proposed Algorithm for decoder – style transfer

- 손실 함수
  - Content distance.
  - Style distance.
    - Gram-matrix 를 활용하여 style feature 의 통계적인 분포를 표현하였고 이를 줄이는 방향으로 학습을 진행.

Proposed Method – collaborative distillation

2. 기존 encoder E 를 더 작아진 encoder E' 로 대체함.
   - 이때, encoder E' 는 encoder E 와 구조상으로는 동일 하지만 layer 에 filter 즉, channel 수가 더 작아진 구조를 채택.
   - 목표는 더 작은 구조의 encoder E' 가 기존 encoder E 와 동일한 기능으로 동작 하는 것이 목표임.
   - How?
     - 이때, 어떻게 작아진 네트워크가 기존 네트워크의 역할을 수행 할 수 있을 것인가?
       - 저자들은 이를 해결하기 위한 방법으로 linear embedding 을 제안하였음.

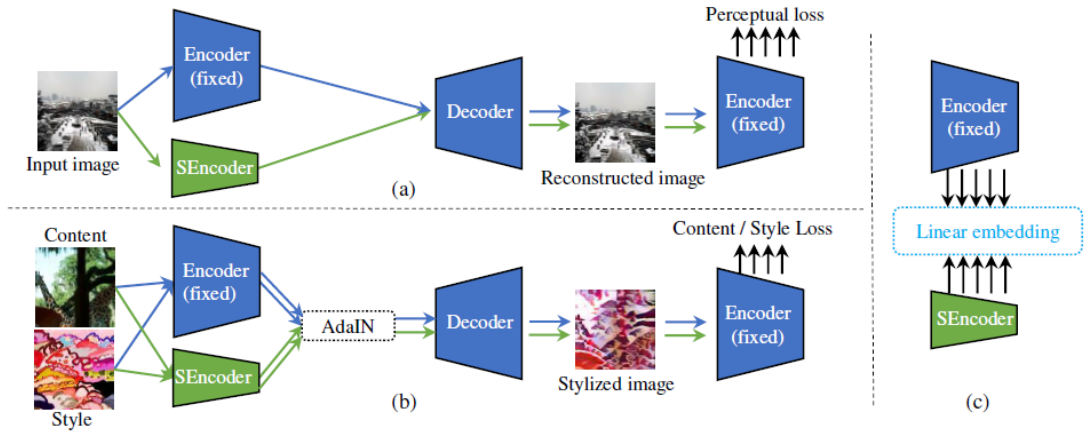
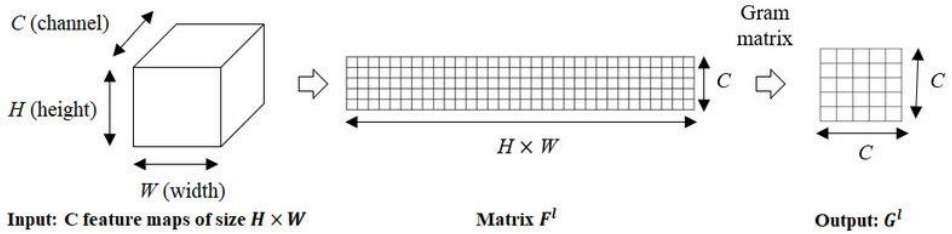


Figure 3: Illustration of the proposed Collaborative Distillation framework (best viewed in color). (a) and (b) depict two kinds of the encoder-decoder collaborative relationship for universal neural style transfer: image reconstruction for WCT [39] and style transfer for AdaIN [24], respectively. Blue arrows show the forward path when training the collaborator network (namely, the decoder). Green arrows show the forward path when the small encoder ("SEncoder") is trained to functionally replace the original encoder ("Encoder"). (c) shows the proposed linear embedding scheme to resolve the feature size mismatch problem and infuse more supervision into the middle layers of the small encoder.

Proposed Method – collaborative distillation – linear embedding

- 기존 encoder E 를 더 작아진 encoder E' 로 대체하기 위하여 작아진 encoder E' 을 decoder 와 연결해 주는 방식을 제안하였지만 이 방식은 feature size mismatch 라는 같은 문제점이 발생함.
  - 어떻게 feature size mismatch 문제를 어떻게 해결 할 것인가?
    - 구체적으로, 기존 encoder E 의 출력 feature 사이즈가 $C \times H \times W$ 일 때, 이에 대응되는 decoder 의 입력 feature 사이즈도 $C \times H \times W$ 로 이루어짐. 이때, 더 작아진 encoder E' 는 더 적은 개수의 filter 를 가지고 있기 때문에 feature 사이즈가 $C' \times H \times W$ $(C' < C)$ 로 decoder 와 사이즈가 맞지 않게 되는 문제가 발생함.
    - 이를 해결하기 위하여, 저자들은 어떻게 stylization 과정에서 channel 의 개수가 어떠한 역할을 하는지 집중함.
      - 딥러닝 기반 style transfer 초기 연구였던 Gatys et al. 은 영상의 스타일 정보를 VGG-19 에서 추출된 deep features 의 gram matrix (즉, covariance matrix 로 channel 간 변동이 얼만큼 닮았는지 기술함) 로 다음과 같이 표현 할 수 있다고 언급.
        - $G = F \cdot F^T$



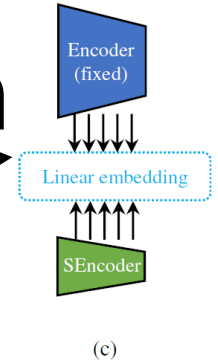
Proposed Method – collaborative distillation – linear embedding

- 어떻게 feature size mismatch 문제를 어떻게 해결 할 것인가?
  - 즉, feature 를 어떻게 압축 할 것인가의 문제를 lower dimension 에서 basis vectors 의 선형 조합(linear combination)으로 풀 수 있음.
    - $F = Q \times F'$,
      - 여기서, Q 는 $C \times C'$ 의 변환 행렬(transform matrix) , F' 은 $C' \times HW$ 의 feature basis matrix 로 , 변환 행렬(transform matrix) 을 통하여 기존 deep features F 의 linear embedding 으로 묘사 가능.
      - Q 는 $C \times C'$ 에서 $C' < C$ 으로 일종의 고차원으로 가는 변환 행렬이라고 이해.
    - 새롭게 정의된 작아진 encoder E' 의 출력 feature $F' \in \mathbb{R}^{C' \times H \times W}$ 을 F 와 Q 로 표현 가능 함.
      - 이때, 새로운 feature F' 의 gram matrix, G' 은 원래 encoder E 가 가지고 있는 gram matrix, G와 같은 개수의 고유 값(eigenvalue) 을 지님.
      - 즉, 기존 F 을 대신하여 사이즈가 줄어든 F'의 새로운 encoder E' 를 사용한다고 하더라도 style 나타낼 수 있는 힘(description power) 은 변함이 없게 됨.
        - Why? 새로운 feature F' 는 기존 feature F 와 linear 한 관계이기 때문이라고 이해.

Proposed Method – collaborative distillation – linear embedding

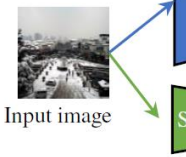
- 어떻게 feature size mismatch 문제를 어떻게 해결 할 것인가?
  - 그렇다면, Q 는 어떻게 $C \times C'$ 의 변환 행렬(transform matrix) 을 학습하여 작아진 encoder E' 를 만들 수 있을까?
    - Linear embedding loss 를 다음과 같이 표현 할 수 있음.
      - $L_{embed} = \|F - Q \cdot F'\|_2^2$,
        - 여기서, F 는 기존 encoder E 의 출력이고 F'는 작아진 encoder E' 의 출력이며, linear embedding이 선형 변환이라는 가정을 만족하기 위하여 기존 features에 선형적인 변환을 거쳐 축소된 features 를 만들어 내는 변환 행렬인 Q가 활성 함수(activation) 가 없는 (without non-linear activation function, 즉 비 선형 변환을 하지 않았다고 이해) fully-connected layer를 통해서 학습됨.

Proposed Method – collaborative distillation – linear embedding

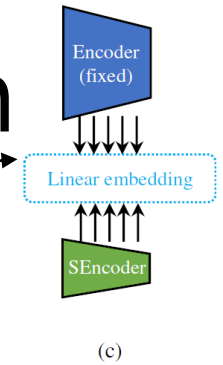


(c)

- 이 개념을 더 확장 하여, feature 의 사이즈를 조정해주는 linear embedding 이 encoder 의 출력 layer 에만 적용되는 것이 아니라 중간 layer 에도 적용 시켜 볼 수 있지 않을까?
  - 논문의 저자들은 그림과 같이 encoder 의 중간 layer 에 linear embedding 방식을 도입시켜 보았음.
    - 이를 통해서 크게 2개의 효과를 누릴 수 있었음.
      1. 작아진 encoder E' 을 학습시키는 입장에서 학습의 gradient source는 오직 decoder 이며, 이 학습의 gradient source 는 fully-connected layer, Q 를 통과함.
         - 하지만, fully-connected layer, Q 는 많은 파라미터를 가지고 있지 않기 때문에 실제로는 정보의 bottleneck 현상이 일으키면서 일종의 학생 네트워크 E'의 학습을 지연시킴.
           - Why? 커다란 decoder 가 작은 fully-connected layer를 통과하기 때문에 bottleneck 현상을 일으킨다고 이해.
         - 이러한 branches(fully-connected layers)가 학생 네트워크인 encoder E' 의 중간 layer 에 plugged into 되면서, 학생 네트워크 E' 에 더 많은 gradient 를 주입 할 수 있게 됨. 결국, 학습을 더 가속화(boost) 하면서 gradient vanishing 현상도 억제 할 수 있음.



Proposed Method – collaborative distillation – linear embedding



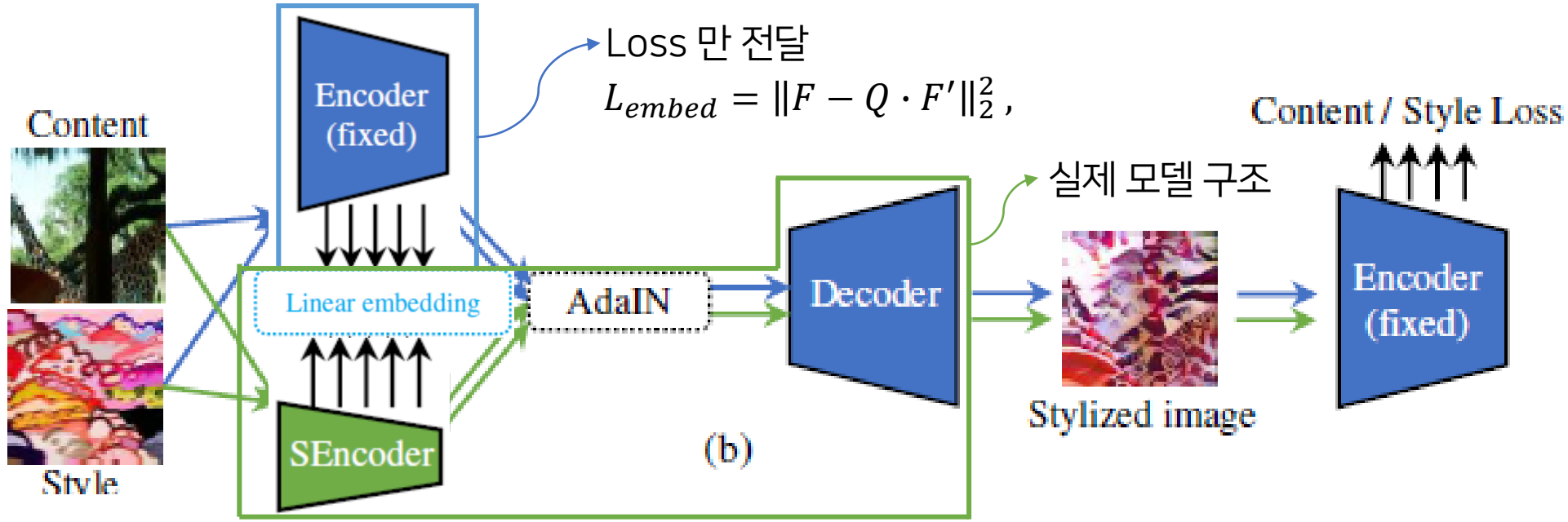
- 이 개념을 더 확장 하여, feature 의 사이즈를 조정해주는 linear embedding 이 encoder 의 출력 layer 에만 적용되는 것이 아니라 중간 layer 에도 적용 시켜 볼 수 있지 않을까?
  - 논문의 저자들은 그림과 같이 encoder 의 중간 layer 에 linear embedding 방식을 도입시켜 보았음.
    - 이를 통해서 크게 2개의 효과를 누릴 수 있었음.
      2. 영상의 스타일 정보는 많은 중간 layer 의 features 을 이용하여 표현이 됨.
         - 다시 말하면, 하나의 layer 에서 추출된 feature 만으로는 style 정보를 표현 할 수 없음.
         - 이로 인하여, 하나의 layer 보다는 다양한 layer에 대한 supervision 이 강력한 style description power 를 얻는데 필수적임.
           - Encoder 에서 강력한 style description power 을 가지는 것이 중요한 이유는 무작위의 style 영상에서도 강인하게(robust) 대응 할 수 있기 때문임.

Proposed Method – collaborative distillation – linear embedding

- 최종 손실 함수
  - $L_{total} = \beta \sum_{i=1}^{k} L_{embed} + L_{collab}$,
    - 여기서, β는 2개의 손실함수 term 의 균형을 맞추기 위한 weight factor 임.

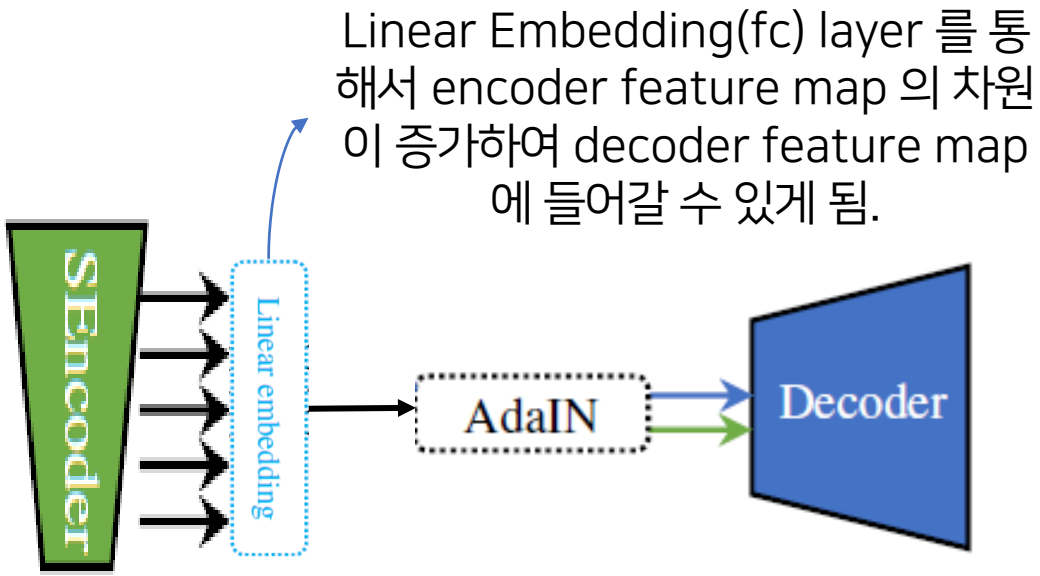
의문점

- Decoder 와 encoder size mismatch 문제를 정확히 어떻게 해결했다는 건가?
  - 정확히는 이렇게 학습되는 구조임.



의문점

- Decoder 와 encoder size mismatch 문제를 정확히 어떻게 해결했다는 건가?
  - Inference 시,



고찰

- 굳이 style transfer 에 국한되지 않고 다른 문제에도 적용해 볼 수 있을 것으로 예상.