Unpaired 데이터셋을 사용한 Adversarial Networks 기반 영상 요약 모델 개발 및 평가

소프트웨어융합학과 유태원

6/11 소프트웨어융합캡스톤디자인

개요

1. 영상 요약

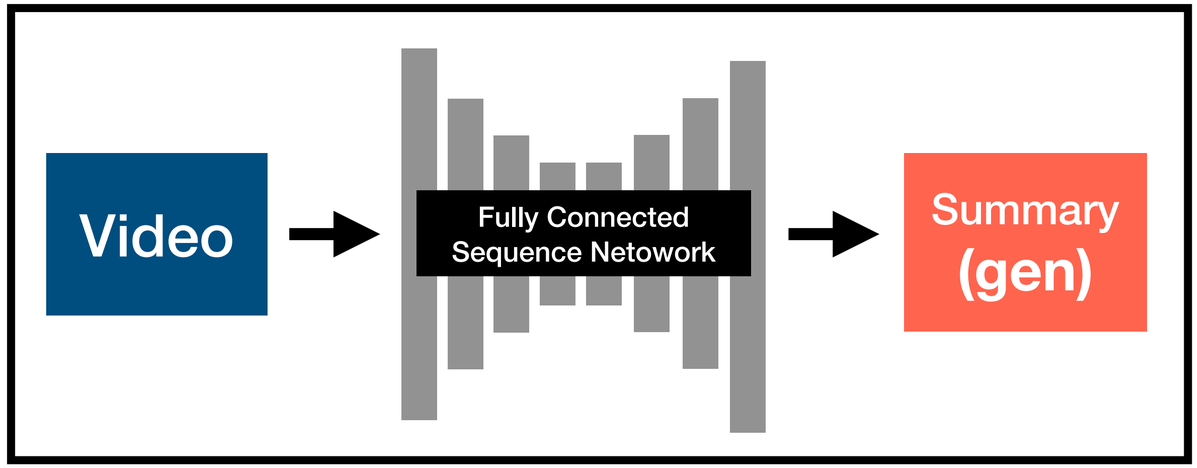
- 지난 수 년간, 온라인에 많은 양의 영상이 업로드 되면서, 사용자들이 영상을 효율적으로 찾을 수 있도록 하거나 분석의 용이성을 위해 영상을 요약하여 보여주는 시스템이 요구되고 있다.
- 영상 요약 방법에는 Key Frame Selection(프레임 선택)과 Key Shots Selection(구간 선택)이 있는데, 이번 연구에서는 Key Frame Selection방법을 사용하여 영상을 요약하는 모델을 개발하고 평가한다.

2. Unpaired 데이터셋 & Adversarial Networks

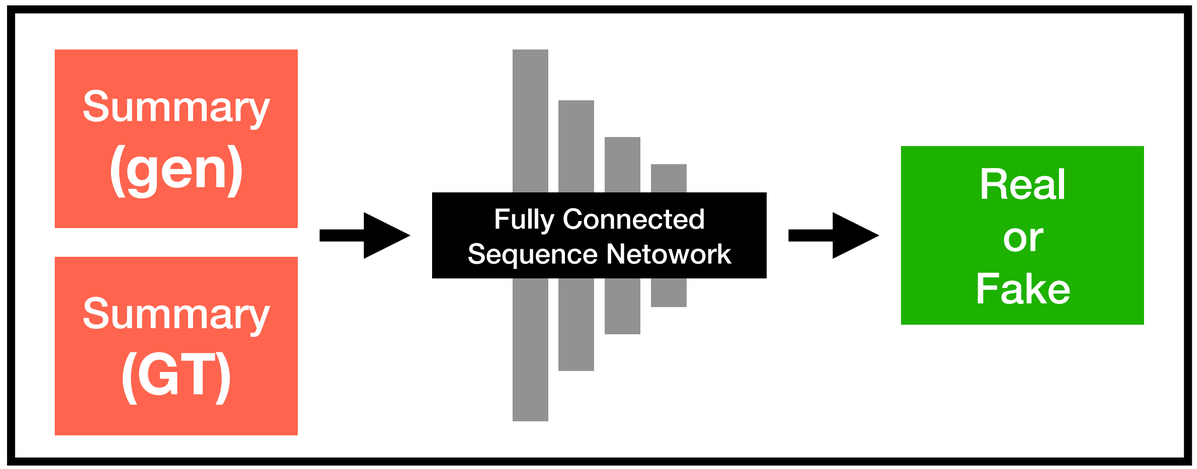
- 지도학습을 사용한 영상 요약 모델 학습은 데이터셋을 생성하는데 큰 비용과 시간이 든다는 문제가 있다.
  또한 제공되는 데이터셋 영상의 개수와 분야가 매우 한정적이어서 모델의 데이터 종속성이 커진다.
- 이 문제를 해결하기 위해, Unpaired 데이터셋과 Adversarial Networks 구조를 사용하여 모델이 영상의 분야에 종속되는 한계점을 해결하고자 했다.

모델 구조: Summary Generator & Discriminator

Summary Generator (SG)



Summary Discriminator (SD)



- **Summary Generator (SG)**

  영상 전체 프레임에서 중요하다고 판단되는 프레임을 선택한다. (Importance Score 상위 15%)

- **Summary Discriminator (SD)**

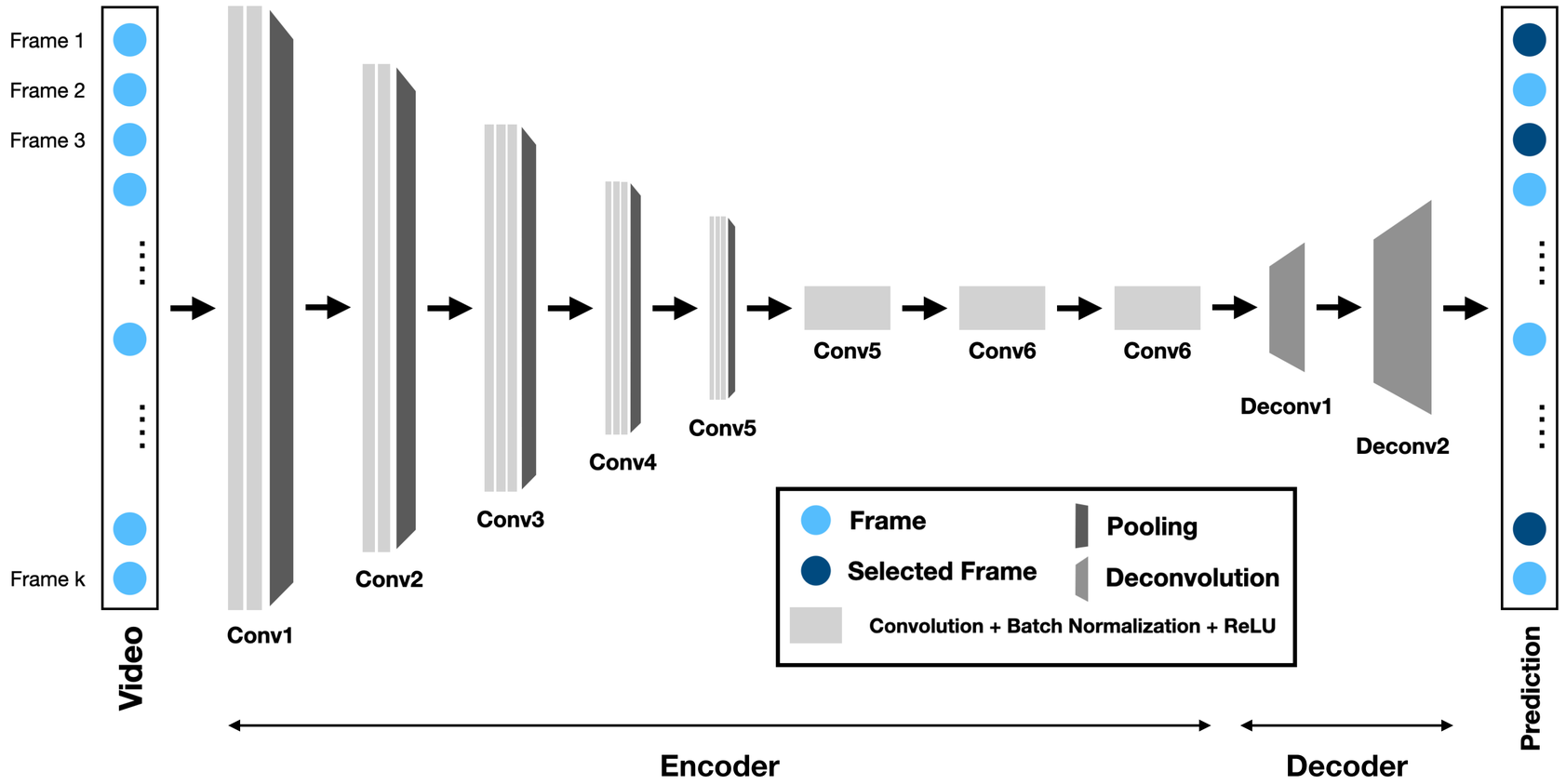
  실제 Ground Truth 요약 영상과 SG가 생성한 요약 영상을 받아 실제 요약 영상인지 SG가 생성한 요약인지 구분한다.

- **Fully Connected Sequence Network (FCSN)**

  SG와 SD 구조로 복잡하고 긴 영상 프레임간의 관계를 모델링하기 적합한 FCSN 모델을 사용했다.

참고문헌: Rochan, Mrigank, and Yang Wang. "Video summarization by learning from unpaired data." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

모델 구조: Fully Connected Sequence Networks

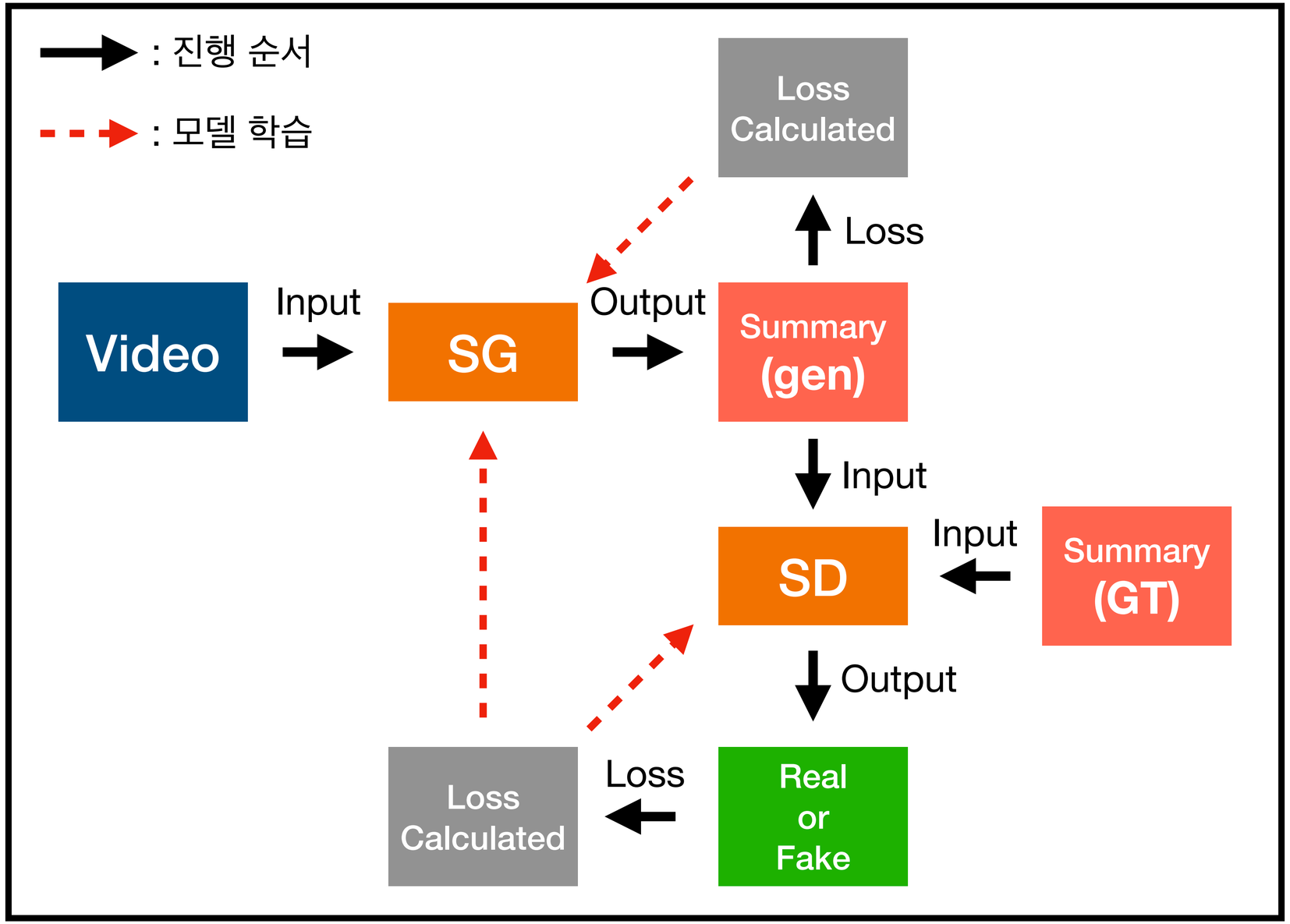


- 영상 프레임의 Feature를 추출하는 Encoder 부분(Conv1 ~ Conv6)과, 중요하다고 판단되는 프레임을 선택하기 위한 Decoder 부분(Deconv1 ~ Deconv2)으로 이루어져 있다.

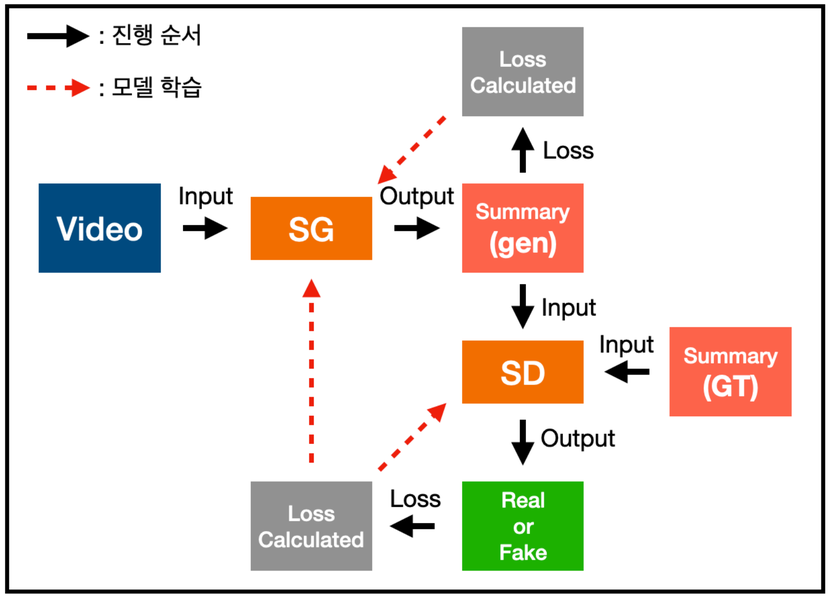
- SG는 Encoder와 Decoder를 거쳐 영상에서 중요한 부분을 선택하고, SD는 Encoder를 거쳐 영상이 SG가 생성한 것인지, Ground Truth인지 판단한다.

참고문헌: Rochan, Mrigank, Linwei Ye, and Yang Wang. "Video summarization using fully convolutional sequence networks." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

모델 구조: 전반적인 모델 학습 구조



모델 구조: Loss Functions



Adversarial Loss

SG가 영상을 요약하고 SD가 영상을 구분할 때 발생하는 Loss

$$\mathcal{L}_{adv}(S_D, S_K) = \mathbb{E}_{s\sim p_{data}(s)}[\log S_D(s)] + \mathbb{E}_{v\sim p_{data}(v)}[\log(1 - S_D(S_K(v)))]$$

Reconstruction Loss

SG를 통해 생성된 Frame과 실제 Frame간의 차이를 최소화하기 위한 Loss Function

$$\mathcal{L}_{reconst}(S_K(v), v) = \frac{1}{k}\sum_{t=1}^{k} \| S_K(v)^t - v^{f_t} \|_2^2$$

Diversity Loss

SG를 거쳐 선택된 Frame들을 시각적으로 다양화하기 위한 Loss Function

$$\mathcal{L}_{div}(S_K(v)) = \frac{1}{k(k-1)}\sum_{t=1}^{k}\sum_{t'=1,t'\neq t}^{k} \frac{(S_K(v)^t)^T \cdot S_K(v)^{t'}}{\|S_K(v)^t\|_2 \|S_K(v)^{t'}\|_2}$$

참고문헌: Rochan, Mrigank, and Yang Wang. "Video summarization by learning from unpaired data." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

실험 Setup: Unpaired 데이터셋

전체 데이터셋 통계

	Train	Test	Total
TVSum	47	3	**50**
OVP	47	3	**50**
YouTube	36	3	**39**
SumMe	22	3	**25**
Total	**152**	**12**	**164**

- **TVSum**

  뉴스, 브이로그 등 다양한 장르의 영상

- **OVP**

  Open Video Project 에서 수집한 영상

- **YouTube**

  YouTube에서 수집한 영상 (뉴스, 스포츠 등)

- **SumMe**

  일반 사람들이 제작한 영상 (요리, 자전거 타기 등)

실험 Setup: Unpaired 데이터셋

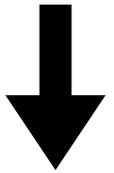
전체 데이터셋 통계

	Train	Test	Total
TVSum	47	3	**50**
OVP	47	3	**50**
YouTube	36	3	**39**
SumMe	22	3	**25**
Total	**152**	**12**	**164**

전체 Train Dataset: 152개

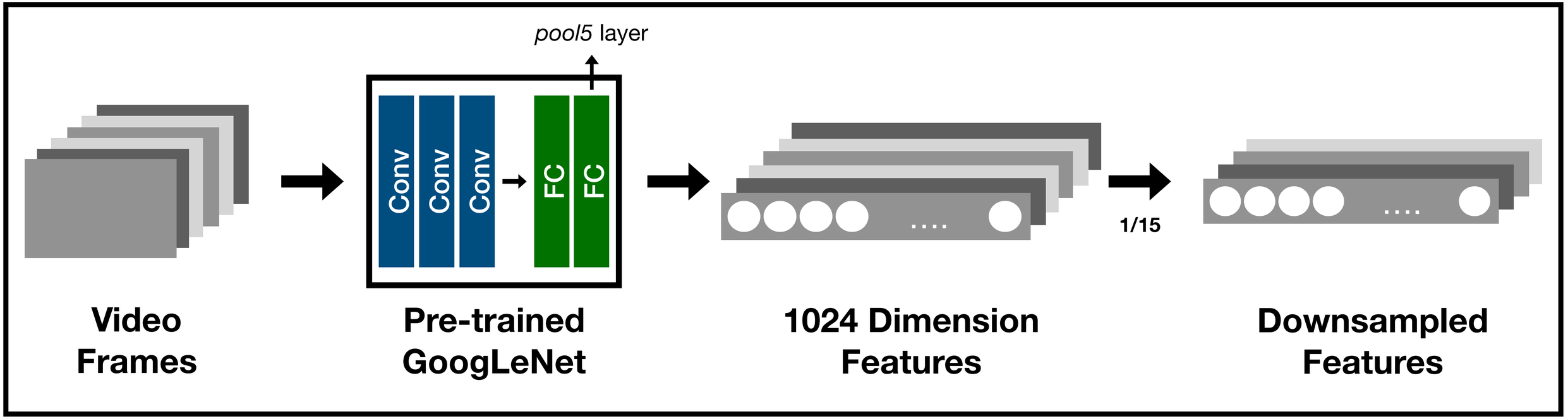
76개: SG input
원본 영상만 사용

76개: SD input
요약 영상만 사용



Unpaired 데이터셋 생성

실험 Setup: Feature Creation



Feature Creation 예시

(4500, 480, 360, 3) → Pre-trained GoogLeNet → (4500, 1024) → (1/15) → (300, 1024)

실제로 사용하는 데이터 Input

결과

F1 Score 계산 방법

$$P = \frac{\text{overlap in } X \text{ and } Y}{\text{duration of } X} \qquad F = \frac{2 \times P \times R}{P + R}$$

$$R = \frac{\text{overlap in } X \text{ and } Y}{\text{duration of } Y}$$

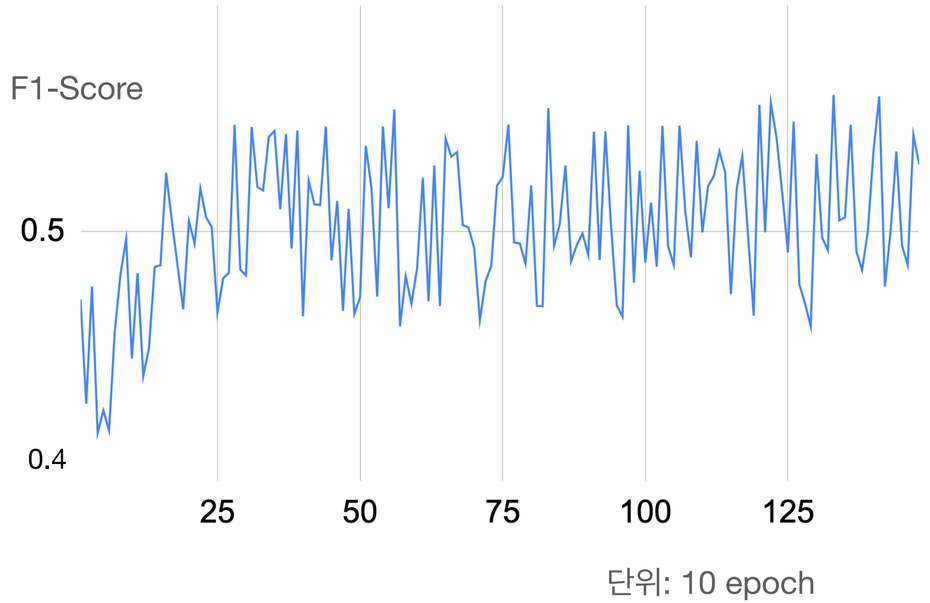
X : SG가 생성한 요약 영상

Y : Ground Truth 요약 영상

F1 Score 비교

#	F-Score
1. FCSN-(SumMe 테스트)	0.448
2. FCSN-(TVSum 테스트)	0.536
3. FCSN-(SumMe 테스트) advloss	0.465
4. FCSN-(TVSum 테스트) advloss	0.553
이번 실험 결과 **(12개 영상 테스트)**	**0.5296**

10 epoch 당 F1 Score



결론 및 향후 연구 방향

결론

- Unpaired 데이터셋과 Adversarial 방법을 사용하여, 영상 요약 모델 학습이 데이터에 종속되는 것을 줄이고자 했다.

향후 연구 방향

- 실제로 Unpaired 데이터셋이 영상 요약 모델 학습에 효과가 있는지 비교하기 위해 Paired 데이터셋으로 학습 후 성능을 비교하는 과정이 필요하다.
- 추가적으로 LSTM 등 영상 요약 성능이 검증된 모델을 사용하여 학습을 하여 결과 비교 후 졸업논문 내용에 추가할 계획이다.

발표 마치겠습니다.
들어주셔서 감사합니다.