Recurrent GNN은 입력과 출력이 아래와 같은 함수 f\_w를 정의하여 점의 상태를 업데이트 한다.

폰트, 타이포그래피, 서예, 친필이(가) 표시된 사진

자동 생성된 설명

GCN – CNN의 아이디어와 비슷, CNN은 학습 가능한 필터를 통해 중심 픽셀의 주변 픽셀을 합치는 것이다. Spatial Convolution Network는 이 아이디어에서 주변 픽셀 대신 연결된 점의 특징을 적용하는 것이다.

Scene Graph – (객체 검출 – CNN, RPN) 한 장면의 Contents를 그래프라는 데이터구조로 묘사한 Topological representation이라고 말할 수 있다. 그래프 이론의 관점으로 보았을 때 Scene Graph는 방향성이 있는 그래프(Directed graph)이고 Object, attributes, relationship 3종류의 Node를 가진다고 볼 수 있다. 혹은 그래프의 Node는 오직 object와 attributes를 같이 표현하고 Edge를 relationship을 표현할 수 있다. 어느 방식을 취할지는 연구마다 다르다.

이미지, 비디오, 3D mesh 등의 장면이 있을 때 Scene graph의 전체 구성 요소는 다음과 같은 triplet의 집합 RS로 표현된다. 이때 OS는 장면(Scene)에 있는 모든 물체의 집합이고 AS는 모든 특성들의 집합 이다. PS는 관계를 나타내는 서술어(Predicate)의 집합이다.

OS = {o1, ..., on}

AS = {a1,..., an}

RS OS PS (OS AS )

이때 triplet에서 두 개의 Object가 있는데 첫 번째 대상을 Subject라고 표현하고 두 번째 대상을 Object라고 표현하기도 하여 <s-p-o> triplet이라고 말한다.하나의 triplet은 rS, i, j = (oS, i , pS, i, j, oS, j) RS라고 표현할 수 있고 각각의 Object는 semantic label lS, k과 Bounding box bS, k 로 이용하여 oS, k = (lS, k, bS, k ) OS라고 표현할 수 있다.

활용 가능성

* Image Captioning은 이미지를 보고 자연어로 표현하는 작업이다. 기존에는 RNN이나 LSTM을 사용하였는데 물체간의 모든 관계를 사용할 수 없기 때문에 정확도를 많이 높이기 힘들었다.  Scene graph자체에 이미 무엇을 표현해야 할지에 대한 정보를 도출 할 수 있기 때문에 이를 바로 이용할 수 있다. 반대로 문장을 보고 이미지를 만드는 과정을 Image Generation이라고 하는데 여기에도 Scene graph가 활용될 수 있다.
* Visual Question Answering(VQA)는 이미지와 자연어로 된 질문을 넣었을 때 대답을 출력하는 작업이다.Scene Graph를 활용한 연구들에서 기존의 VQA 방법보다 성능 향상이 있음을 보여준다.  Image Retrieval은 내용을 묘사하는 설명을 보고 데이터베이스에서 이미지들을 뽑아내는 작업을 말하는데 이때 단순한 물체만 뽑아내는 것이 아니라 관계나 특성 또한 구조적으로 포함시켜서 얻어낼 수 있다. Referring Expression은 설명을 보고 이미지를 뽑아낸다는 점에선 Image Retrieval과 유사하지만 인풋 이미지에 대해 묘사와 일치되는 영역만 표시하는 것이다. 이렇듯 다양한 컴퓨터 비전 태스크에 Scene Graph가 핵심적인 중추역할을 할 수 있어서 많은 연구들이 진행되고 있다.

텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

Scene graph의 정의에 따라 우리가 찾아야 하는 것은 결국 물체와 물체의 관계이고 딥러닝 모델에서 처리하기 위해선 내부적으로 Feature라는 벡터의 형태가 구축되어야 한다.  그렇기 때문에 우선 물체와 관계자체를 검출하는 작업을 필요로 한다. 이를 위해 RPN(Region  proposal network)를 사용하여 물체와 관계의 ROI(Region of Interests)를 잡아내도록 한다. Bounding box를 통하여 검출된 영역만을 이용해 Proposal을 만들고 각각의 Proposal에 해당되는 Feature representation을 계산하도록 한다. 이런 Feature들은 Object, Subject, Predicate에 대한 외형, 공간 정보, 라벨, 깊이, 마스크 정보 등을 사용하여 표현할 수 있다.

이렇게 마련된 Feature들이 결국 그래프의 Node로 표현되어 최초의 Graph를 생성하도록 한다. 이 단계에서 일단 모든 물체끼리 쌍을 이루고 있다고 가정한다. 즉 Complete graph를 형성하도록 하는 것이다. 하지만 이런 초기 그래프는 아직 완벽한 장면내의 맥락을 반영하지 못한 상태이다. 예를 들어 어떤 triplet <s-p-o>가 있더라도 이와 연결된 주변의 노드들의 정보와 어떠한 의미론적인 의존구조가 있을 수 있다. 단순히 앞서 얻은 Proposal에만 의존하기 보다는 전체적인 시각을 고려해야 하는데 이를 Feature refinement 모듈에서 수행할 수 있다

영상에서 그래프 Generator, 시간에 따라 상관 관계를 가지는 관계들을 잡아내기 위해 Object detection과 Multi-object tracking 기법을 모두 사용한다.

가령 비디오를 오버랩을 포함하여 segment로 분리한 뒤 각각에 대해서 Relation Proposal과 Refinement를 수행한다. 그 뒤 짧은 기간에 대해 예측한 관계들을 병합시키면서 최종적인 <s-p-o> triplet을 얻게 된다.

Data Set - 다른 컴퓨터 비전 태스크와 마찬가지로 SGG에는 다량의 annotated data를 필요로 한다. Scene Graph Dataset는 최초의 Real-world scene graph 데이터셋이다. MSCOCO와 YFCC100m에서 공통되는 이미지 5,000장으로 구성하였고 사람이 직접 만든 scene graph가 들어 있다. [VRD](https://cs.stanford.edu/people/ranjaykrishna/vrd/) 또한 5,000장의 이미지로 구성되어 있고 SGG 태스크의 벤치마크로 사용되기 위해 구축되었다. Scene Graph Dataset은 object 당 평균 2.3개의 predicates가 있지만 VRD는 object당 평균 24.25개의 predicates를 가진다.

이 분야에서 가장 대표적으로 사용되는 공개 데이터셋은 [Visual Genome](https://visualgenome.org/)이다. 가장 다양한 object categories와 relation을 포함하고 있다. VRD는 컴퓨터 비전 전문가가 만든 반면에 Visual Genome을 많은 수의 데이터를 처리하기 위해 일반인 annotator가 참가해서 퀄리티가 떨어지는 단점이 있다. 그래서 연구자들마다 Visual Genome을 그대로 사용하기 보다는 나름대로 개선한 자신들만의 버전을 사용한다. 예를 들어 VrR-VG는 Visual Genome에서 중복되거나 시각적으로 덜 중요한 관계를 필터링하였다.

비디오를 기초로 하는 [Action Genome](https://www.actiongenome.org/index.html) 데이터셋도 존재한다. 기존의 비디오 데이터셋은 다량의 클립으로 구성하긴 했지만 물체와 관계가 어떻게 변화하는지에 대해 주목하진 않았다.

즉 단일 사건으로만 분류하였고 시간에 따라 관계가 어떻게 변하는지는 나타내지 않았다. Action Genome에선 시간에 따라 변하는 관계 또한 포함하기 때문에 Spatio-temporal SGG를 연구하는데 사용될 수 있다.

SGG를 수행하는 모델의 성능을 평가하는 태스크는 크게 4가지로 구성된다.

- Predicate Classification(PredCls) : 검출한 물체들에 대한 모든 pairwise relationship의 predicate를 예측한다. 다른 요소는 무시하고 predicate classifcaition의 성능만 검증한다.

- Scene Graph Classification(SGCls) : 검출한 물체들이 있을 때 모든 pairwise relationship에 대해서 predicate와 object, subject의 categories를 예측한다.

- Phrase Detection(PhrDets) : (s-p-o)의 label을 출력하고 전체 relationship을 하나의 bbox로 localize한다. 이때 ground truth box와 overlap이 0.5이상이어야 한다.

- Scene Graph Generation(SGGen) : object를 detection하는 것과 동시에 검출된 object의 pair간의 predicate를 예측한다. Object는 gt bbox와 0.5이상 IoU가 있을 때 성공적으로 검출한 것으로 본다.

일반적인 분류작업에서 평가지표에는 정확도(Accuracy), 정밀도(Precision) 등이 있으나 SGG에서 재현율(Recall)이 de-facto standard로 자리잡았다.

왜냐하면 불균등 데이터셋인 점을 감안해서 성능을 지표화하기엔 정확도가 완벽하지 않기 때문이다.

Recall@k라고 표현하는데 예측으로 나온 <s, p, o>의 Top-k개와 실제값을 비교하여 계산하게 된다. k는 보통 50, 100을 사용한다.

CNN

– Convolution Layer : 입력 데이터에 필터 적용 후 활성화 함수를 반영하는 필수 레이어

3차원의 텐서로 표현, RGB-> 채널의 크기 = 3.

높이X너비X채널이 4X4X1인 텐서 형태의 입력 이미지에 3X3 크기 필터 > 합성곱 계층에서 합성곱 연산 -> 2,2,1 텐서 형태의 이미지 생성

Stride : 필터의 이동 거리 (1,2,3 . . . ) -> 보통은 1이다.

합성곱 연산 후 이미지의 크기가 줄어들고, 픽셀의 정보가 사라지는 것을 방지하기 위해 패딩(Padding)을 사용 : zero-padding(0의 값을 갖는 픽셀추가)

–Pooling Layer: 선택적 레이어 -> 주로 Max Pooling 사용 -> 노이즈 감소, 속도 향상, 분별력이 좋아진다.

RPN – CNN으로부터 얻어진 특징 맵을 입력으로..

CrossEntorpyLoss 손실함수 – 분류 문제에 주로 적용, 모델의 예측과 실제 타깃 값 사이의 크로스 엔트로피(예측 값과 실제 값 사이의 차이, 정확할 수록 작아짐)를 계산, 소프트맥스 함수와 함께 사용되는 경우도 있다. 소프트맥스 함수는 모델의 출력을 클래스 확률로 변환해주는 함수.

Visual mask – 객체인식, 분할 작업에서 픽셀 수준으로 객체 영역을 나타내는 이진 마스크 -> 마스크의 픽셀 값은 객체에 해당하는 영역 : 1, 배경은 : 0

Point – 객체의 중심 좌표를 나타내는 정보, 객체 인식이나 검출 작업에서 객체 위치 표현에 사용

4

임베딩 – 주어진 데이터를 고차원 벡터로 변환하는 과정, 텍스트나 시퀀스 데이터의 형태를 처리할 때 사용.

Transformer attention – 쿼리, 키, 값으로 구성

Query – 어텐션을 수행하는 위치(단어)를 나타내는 벡터이다. 보통 현재 처리 중인 단어의 임베딩이 쿼리로 사용된다.

Key – 쿼리와 비교되어 어텐션 가중치를 계산하는 데 사용되는 위치(단어)들의 벡터이다. 보통 입력 시퀀스 내의 모든 위치의 임베딩이 키로 사용된다.

Value : 어텐션 가중치에 따라 가중 평균된 결과를 생성하기 위해 사용되는 위치들의 벡터이다. 주로 입력 시퀀스 내의 모든 위치의 임베딩이 값으로 사용된다.