자연어 처리와 영상 처리를 이용한 조인트임베딩 기반 영상 검색 기술

□ 함경준 / 한국전자통신연구원

Ⅰ. 서 론

하루 일과 중에 적어도 한 번 정도는 포털 사이트에서 검색을 하는 것이 우리의 일상 생활이 되어 버렸다. 기존에는 웹 문서를 대상으로 검색을 주로 하였다면, 지금은 그 검색의 대상이 오디오, 비디오 등의 다양한 멀티미디어로 확대되었다. 특별히 유튜브나 포털에서 제공하는 영상 공유 플랫폼을 통해 개인 영상 미디어가 폭발적으로 증가하고 있으며 기존의 방송 및 케이블 사업자들도 영상을 재편집하여 영상을 업로드하고 있는 상황이다. 불과 4~5년 전에는 N사의 지식 검색과 같은 텍스트 검색 서비스가 주류를 이루었는데, 이제는 블로그나 뉴스기사 보다는 유튜브에서 관련 영상을 검색하는 모습을 많이 볼 수 있게 되었다. 실제로 2019년에 미디어 조사업체에서 인터넷 이용자의 검색 이용 채널을 설문조사 하였을 때 10명중 6명이 유튜브를 통해 검색을 하는 것으로 조사되었다.

본 고에서는 영상 검색 서비스를 위해 필요한 기술을 소개하고자 한다. 특히 기존의 키워드 기반 검색 기술을 영상 검색에 활용하였을 때 겪게 되는 문제점과 이를 해결하기 위해 몇 년 전부터 두각을 나타내고 있는 심층학습 기반의 영상 검색 기술을 살펴보기로 한다.

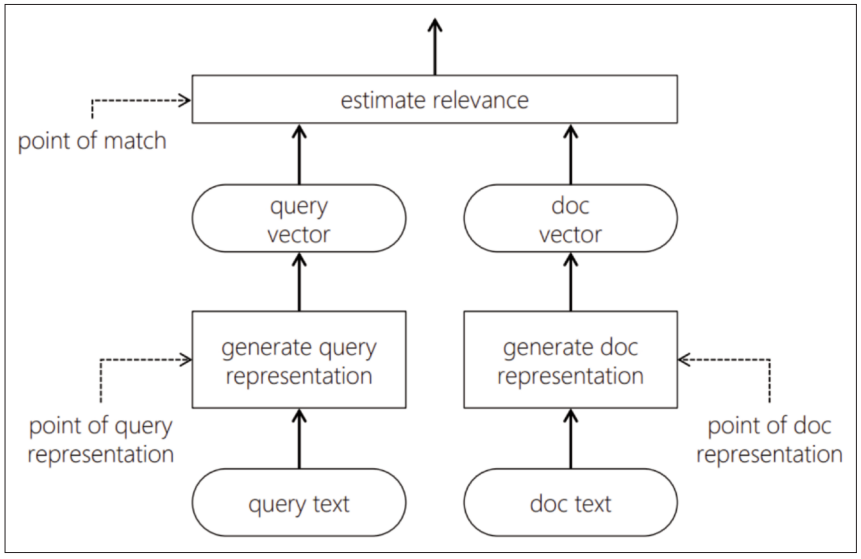
Ⅱ. 영상 검색 기술

사용자의 입장에서 보았을 때 어떠한 검색 기술이 사용되었더라도 UI를 통해 접하는 서비스 형태는 사용자가 입력한 질의어(Query)에 대하여 검색 결과를 보여주는 것이다. 본 고에서는 영상 검색 서비스의 내부 과정에서 사용되는 기술, 특히 키워드 기반의 영상 검색과 심층학습을 이용한 영상 검색

기술에 대하여 좀 더 자세히 알아보려고 한다.

1. 키워드 기반 영상검색

키워드 기반의 영상 검색은 웹 문서나 뉴스 기사처럼 기존의 텍스트 문서를 대상으로 검색하기 위해 고안한 기술을 이용하여 영상 검색 기능을 제공한다. 다만, 영상 자체에는 텍스트 정보가 없으므로 영상에 대한 메타데이터, 즉 영상의 주제나 주요 사건 등의 정보를 텍스트 형태로 구축하고 검색을 수행하게 된다. 사용자 입장에서는 마치 영상 검색이 되는 것처럼 느껴지지만, 내부적으로는 영상에 부여된 텍스트 메타데이터를 대상으로 키워드 기반 검색 기술을 이용하여 문서검색을 수행하는 것과 다를 바 없다. 이러한 키워드 기반 영상 검색은 충분히 성숙되어 있는 기술이며, 초기 도입이 쉽고 구현이 용이하기 때문에 소규모의 영상 풀(Pool)을 가진 영상 검색 서비스에 적합하다. 하지만 텍스트 기반 검색 기술이 스케일이 큰 경우에도 빠른 검색 속도를 자랑하기 때문에, 유튜브도 영상 검색 과정에서 업로더가 작성한 설명 텍스트를 기반으로 영상 검색 결과를 제공한다. 대부분의 키워드 기반 검색을 하기 위한 절차는 〈그림 1〉과 같다. 검색이라는 문제는 결국 사용자가 입력한 질의문과 검색 대상이 되는 문서 간에 유사도를 어떻게 정확하게 계산할 것인가로 정리할 수 있다. 이 문제를 세분화하면, 사용자가 입력한 불완전한 혹은 불분명한 검색 의도를 가진 질의문을 구체화하여 표현하는 문제와 문서의 중요한 의미를 담고 있는 단어와 그렇지 못한 단어가 마구잡이로 섞여 있는 검색 대상 문서를 정제하여 표현하는 문제로 나눌 수 있다. 질의문과 문서가 적절히 정확하게 표현되었다면 이들 간의 유사도 측정은 쉽사리 이루어질 수 있게 되고, 성능 좋은 검색엔진을 만들 수 있게 된다.



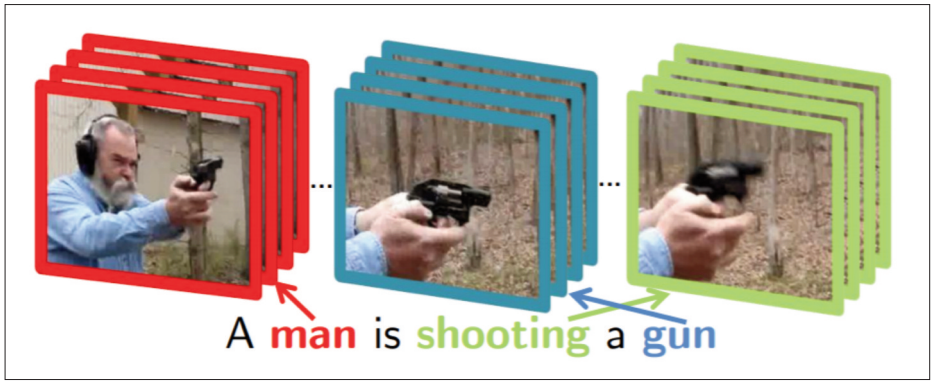
<그림 1> 키워드 기반 검색 과정 도식화

키워드 기반 검색 관련 연구에서 주로 이슈가 되는 부분은 사용자마다 서로 다른 의미와 형태로 입력되는 질의어를 의미적으로 분석하고 보강하여 검색 대상과 잘 매칭이 되도록 하는 것이다. 이에 따라, 질의어 정보를 보충하기 위해 관련 키워드를 추가하여 검색을 시도하는 질의어 확장 기술이 대안으로 제시되었다. 이러한 기술은 온톨로지나 워드넷과 같이 개념에 대한 명세를 체계적으로 구축한 지식베이스를 도입하여 해당 키워드의 의미를 구체화할 수 있는 단어를 추가하는 방법으로 구현할 수 있다. 예를 들어 '간식'이라는 키워드가 질의어로 주어졌을 때 해당 용어의 의미를 구체화하여 과자, 빵, 떡, 아이스크림 등의 관련 단어를 추가하여 검색을 수행하는 것이다. 이때, 검색 정확도를 높이기 위해서는 지식베이스의 구축과 업데이트가 매우 중요한 요소로 작용한다. 하지만 지식베이스 구축을 위한 시간과 비용이 발생하고, 지식베이스가 모든 용어에 대한 정보를 담고 있기에는 사실상 불가능하기 때문에 단어에 인접하여 출현하는 통계정보를 분석하여 관련 단어를 추가하려는 'Word2Vec'과 같은 연구가 대안으로 제시된다.

키워드 기반 검색 기술은 상용 검색 엔진이 있을 정도로 성숙된 기술 분야이기 때문에 영상 검색 분야에서도 많이 사용되고 있는 기술이다. 하지만 지식베이스 구축의 비용 및 한계가 있는 것처럼 영상에 대한 텍스트 형태의 메타데이터를 구축하는 것 또한 비용과 한계가 존재한다. 이러한 문제점을 극복하기 위하여 최근에는 심층 학습을 이용하여 영상과 질의어 이해를 통해 영상 검색을 하려는 연구가 활발히 이루어지고 있다.

2. 심층학습 기반 영상 검색

심층학습 기반 영상 검색 기술은 영상에 대한 추가적인 텍스트 메타데이터를 구축하지 않아도 영상 콘텐츠 검색이 가능하도록 영상과 그에 해당하는 캡션을 학습데이터로 구축하고 심층신경망을 학습시켜 영상 검색기능을 제공하는 기술이다. 본 고에서는 해당 기술에 대한 상세한 기술적 명세보다는 비전문가도 이해할 수 있는 기술의 개념적 설명과 기능 위주로 다루고자 한다. 영상에 대한 심층 학습 기반 영상 검색 기술을 소개하기 전에 이해를 돕기 위하여 〈그림 2〉를 이용하여 어떤 부분이 주요 이



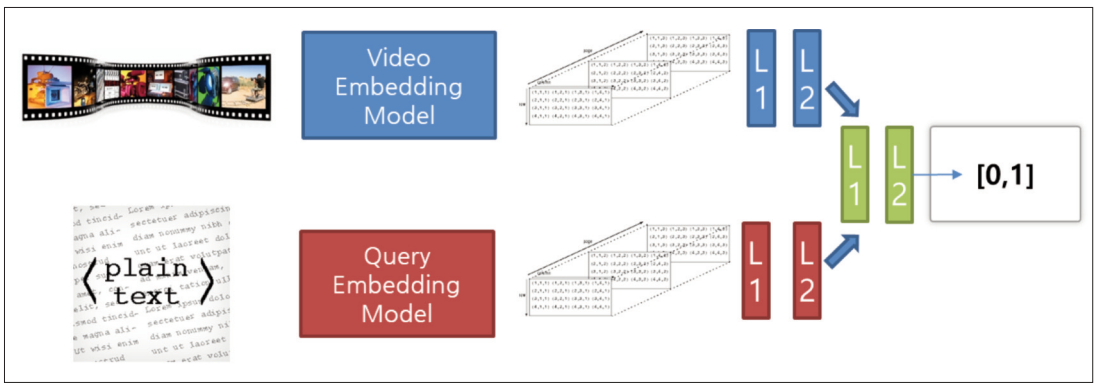
<그림 2> 심층학습 기반 영상 검색 개념도

슈인지 파악해 보기로 한다.

심층학습 기반의 영상검색 기술이라 할지라도 키워드 기반 영상 검색 부분에서 언급했던 검색의 기본 골격은 변하지 않는다. 즉 영상 자체에서 특징(feature)을 추출하여 해당 영상을 이해하고 그것을 바탕으로 적절한 형태의 정보로 표현을 해야 한다. 문장(질의어)에 대해서도 문장의 전체 의미를 파악하고 적절한 형태의 정보로 표현하는 과정이 필요하다. 적절히 표현을 하였다면 키워드 기반 검색에서도 그러했듯이 유사도를 계산할 수 있게 되고 주어진 질의어에 따라 적합도(Relevance) 점수가 계산되어 사용자에게 검색 결과로 제공된다. 여기서 주로 살펴봐야 할 것은 영상이 연속된 프레임 이미지로 구성된 집합체라는 것과 문장의 경우 단어의 출현 순서와 맥락에 따라 의미가 달라지는 점을 주의하여 표현을 해야 한다는 것이다. 결국 이미지 프레임 시퀀스와 단어 시퀀스에 대한 분석을 토대로 각각의 정보를 표현하고 서로의 정보가 의미적으로 유사하다면 유사도 값이 높아지도록 심층학습망을 학습하는 과정이 필요하다. 심층학습 기반 검색에서는 일련의 숫자로 구성되어 있어야 학습망에 넣어 학습 가능하므로 영상과 문장을 고차원의 벡터로 임베딩(embedding)하여 표현하게 된다. 〈그림 3〉은 심층학습 기반 영상 검색 학습 네트워크 개념도를 보여주고 있다. 문장을 임베딩 하기 위한 학습 네트워크를 구축하고 동시에 영상을 임베딩 하기 위한 학습 네트워크를 구축하여 이를 같은 벡터 공간에 표현하는 조인트 임베딩 과정을 거치고 유사도 값을 기반으로 하는 손실함수를 이용하여 네트워크 학습 파라미터를 업데이트하게 된다. 즉, 주어진 영상과 관련된 문장에 대해서는 높은 유사도가 계산되도록 학습 파라미터를 업데이트하고 관련 없는 문장에 대해서는 낮은 유사도가 계산되도록 학습을 시키게 되면 어느 정도 만족할만한 영상 검색 결과를 얻게 된다. 3절에서는 질의문 임베딩 과정에 대한 소개를 하고, 4절에서는 영상 임베딩하는 과정을 소개하며, 5절에서는 각각 임베딩한 결과를 조합(Joint)하여 학습을 하는 과정과 최신 연구를 소개하고자 한다.

3. 질의문 임베딩

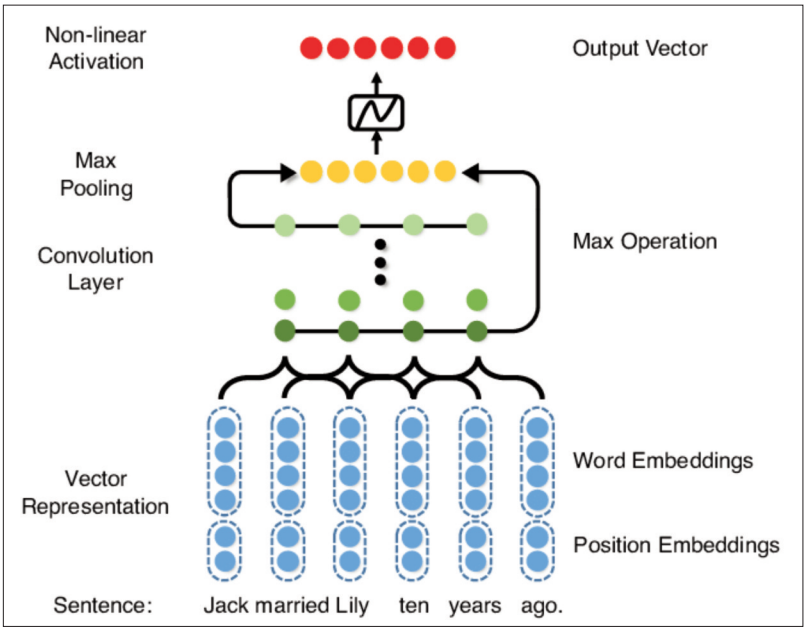
질의문을 학습 네트워크의 입력으로 사용하기 위



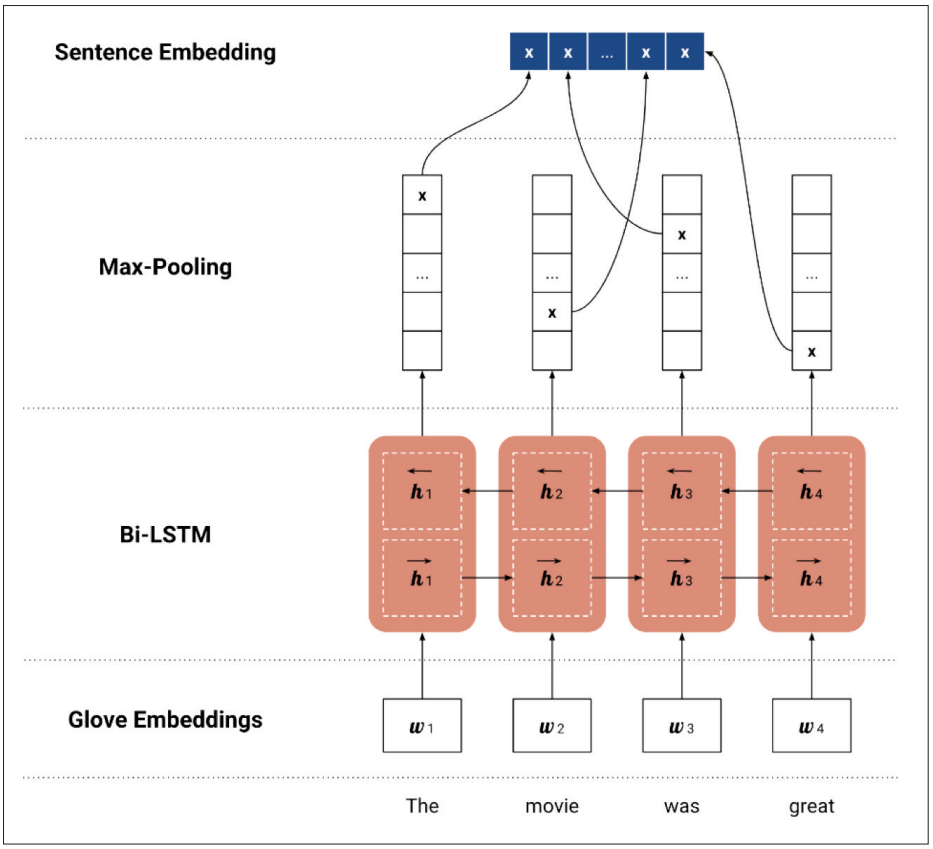
<그림 3> 심층학습 기반 영상 검색 학습 네트워크 개념도

해서는 문자형태의 정보를 숫자 형태의 정보로 표현하는 임베딩 과정이 필요하다. 이를 위해서는 형태소 분석과 같은 기본적인 전처리 과정이 필요할 수 있으며, 단어 혹은 문장 단위의 정보를 벡터로 표현해야 한다. 앞부분에서 언급한 Word2Vec와 같은 알고리즘을 사용하면 각각의 단어를 해당 단어의 의미정보를 포함하고 있는 벡터로 표현할 수 있게 된다. 개별 단어 벡터를 기반으로 문장을 임베딩하는 과정은 최근에도 빈번하게 사용되고 있다. 〈그림 4〉는 단어 벡터를 기반으로 문장을 임베딩하는 과정을 도식화한 그림이다. 각 단어의 벡터를 구한 후 이를 그대로 붙여서(Concatenation) 조합하여 사용을 하게 되면 문장의 벡터 차원 수가 너무 커지게 되어, 이후 신경망에서 제대로 학습이 안되거나 학습에 필요한 리소스가 부족하게 되므로 적절한 차원수로 압축통합(aggregation)하여 최종 문장 임베딩 벡터를 구하게 된다. 〈그림 5〉도 단어 벡터를 기반으로 문장 임베딩을 수행하는 모델의 구조인데, 단어 간의 순서를 고려하여 문장의 의미를 보다 정확히 함축하여 벡터를 구하도록 중간 레이어에 LSTM(Long Short-Term Memory) 모델 구조를 도입한 것이 특징이다.

최근에는 이러한 접근 방식에서 벗어나 정말로 방대한 텍스트 데이터를 비지도 학습데이터로 사용하여 언어표현의 여러가지 패턴을 사전 학습시킨 범용 목적의 언어이해 모델을 이용하기도 한다. 구글에서 발표한 Bert(Bidirectional Encoder Representations from Transformers)가 대표적인 모델이다. Bert와 같은 사전학습 언어이해 모델을 이용하여 문장 임베딩을 전이학습(Fine-tuning)하게 되면 기존의 접근 방식보다 정확한 문장 임베딩 결과를 기대할 수 있게 된다. Bert는 주로 영어를 대상으



<그림 4> 단어 벡터 기반 문장 임베딩 모델



<그림 5> 단어 벡터 기반 문장 임베딩 모델 - LSTM도입

로 다양한 딥러닝 프레임워크(Tensorflow, Torch 등)로 공개되어 있으며, 최근에는 한국어를 보다 잘 처리하고 이해할 수 있도록 추가 한글 학습 데이터 셋을 이용하여 사전학습을 수행한 한글 Bert도 공개되어 있어서 한글문장 임베딩에 사용할 수 있다.

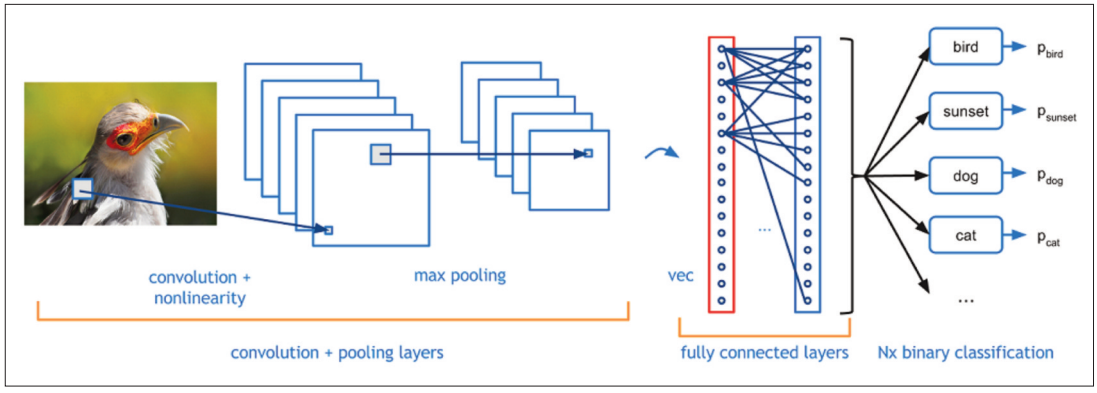
4. 영상 임베딩

일반적인 영상은 초당 24장 혹은 30장의 프레임 이미지로 구성되어 있으므로, 영상을 임베딩하기 위해서는 각 프레임 이미지에 대한 분석이 필요하다. 각 이미지를 통해 얻어지는 정보가 하위 정보라고 한다면 이미지 시퀀스 분석을 통한 상위 정보를 추출해야 영상을 온전히 이해하였다고 볼 수 있다. 또한 영상에는 오디오를 포함하고 있기 때문에 오디오를 통한 정보 추출도 필요하다.

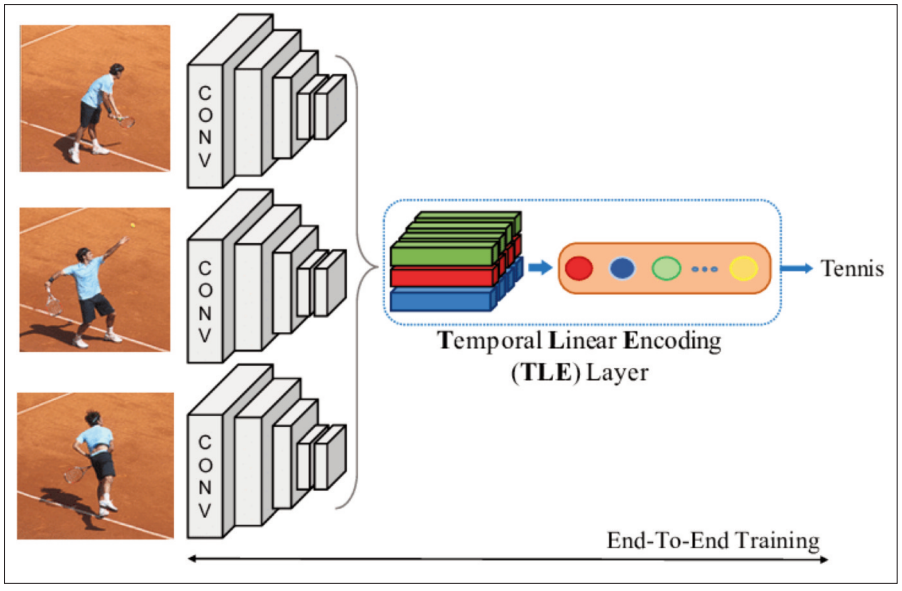
우선 이미지에 대한 이해는 주로 이미지 내에 포함되어 있는 객체를 식별하는 것을 의미한다. 〈그림 6〉은 이미지 내에 객체를 식별하기 위한 신경망 개념도이다. 심층학습이 세상에 널리 알려지게 된 계기가 합성곱 신경망을 통한 객체 식별이기 때문인데, 대량의 학습 이미지를 통해 사전학습된

객체 식별 모델을 사용하여 정확한 객체 정보를 추출할 수 있다. 많이 사용되는 객체 식별 공개 모델로는 ResNet, Yolo 등이 있다. 이러한 사전학습 모델을 도입하면 영상 내에 출현한 객체 정보를 포함하고 있는 임베딩 벡터를 얻을 수 있게 된다.

그렇다면 이미지 시퀀스 분석을 통한 정보 추출에 대하여 알아보기로 하자. 이미지 시퀀스를 통해 추출되는 상위 정보는 일반적으로 행위(Activity) 정보이다. 예를 들어 각 프레임 이미지에서 사람과 라켓, 공 객체가 연속으로 검출되는 영상일때, 이미지 시퀀스로부터 얻을 수 있는 행위정보는 '테니스'일 가능성이 높다. 〈그림 7〉은 행위 정보를 추출하기 위한 신경망 개념도이다. 각 프레임 이미지로부터 추출된 객체정보를 시간적인 순서를 고려하여



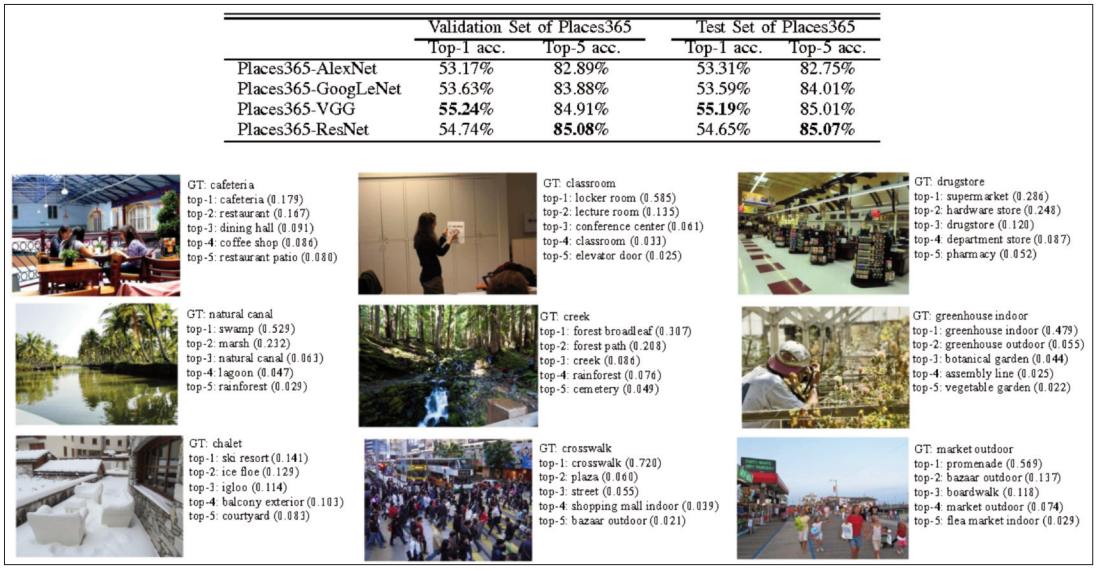
<그림 6> 이미지 객체 식별 신경망 개념도



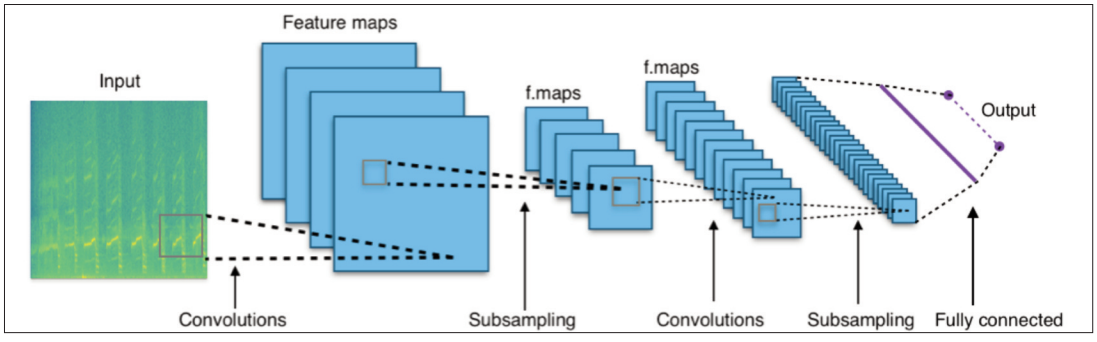
<그림 7> 행위(Activity) 정보를 추출하기 위한 신경망 개념도

고차원의 벡터로 인코딩하게 되면 영상이 포함하고 있는 행위 정보를 유추할 수 있는 임베딩 벡터를 구할 수 있게 된다. 행위 정보 추출 모델은 앞에서 살펴본 객체 식별 신경망을 기반으로 확장하여 사용하는 경우가 대부분이며 대표적으로 C3D, I3D 등의 모델이 있다. 현재까지도 행위 정보 식별 정확도를 높이기 위한 새로운 구조를 도입한 모델들의 논문과 소스코드, 그리고 학습 모델이 공개되고 있는데, 최근에는 언어 이해 분야에 많이 쓰이는 Bert 구조를 행위 식별 신경망에 도입하여 정확도를 높인 연구도 발표되었다.

영상으로부터 추출할 수 있는 정보는 객체와 행위 외에도 많은 종류가 있다. 영상에 있는 그래픽 문자를 식별하여 텍스트 정보를 추출할 수 있고, 시간(낮, 밤 등), 장소, 소리, 대사 등이 존재한다. 〈그림 8〉과 〈그림 9〉는 각각 장소와 소리를 식별하기 위한 모델



<그림 8> 영상으로부터의 장소 정보 추출



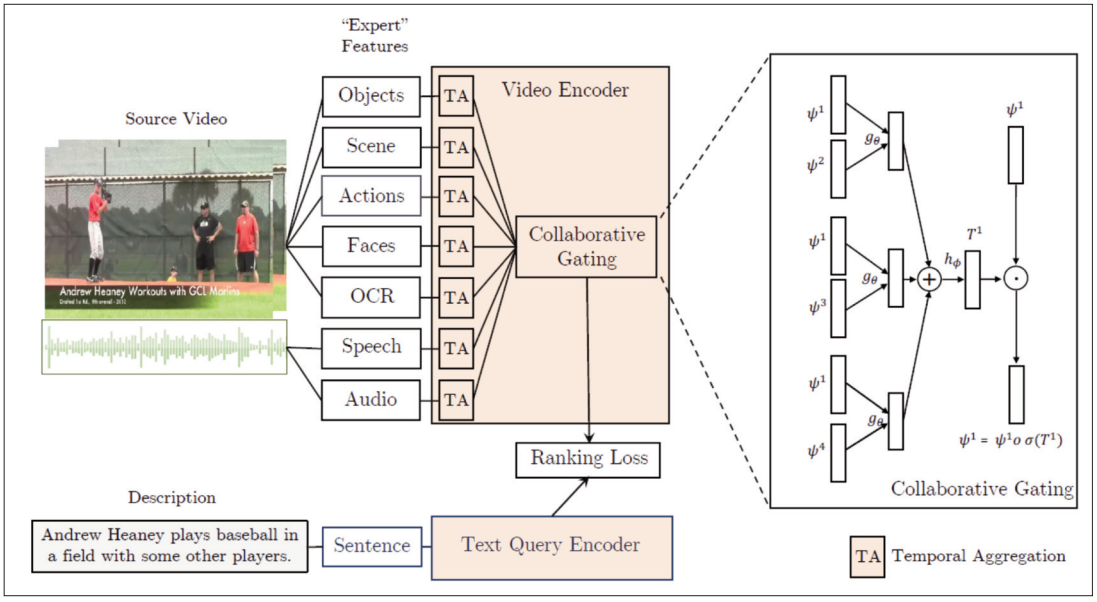
<그림 9> 영상의 오디오 스펙트로그램으로부터 소리 정보 추출

과 관련된 것이다. 장소의 경우 기본적으로 객체 검출 신경망을 기반으로 Place365와 같은 장소 관련 학습데이터를 이용하여 장소 정보를 식별할 수 있는 학습 모델을 구축하여 사용할 수 있다. 소리 정보를 추출하기 위해서는 오디오를 스펙트로그램(Spectrogram)의 이미지로 표현하고 이미지 처리에 많이 사용되는 합성곱 신경망을 이용하여 영상에 포함되어 있는 소리 정보(고양이 소리, 새 소리, 군중 소리, 피아노 소리 등)를 추출할 수 있다.

5. 조인트 임베딩

질의문 임베딩과 영상 임베딩 과정을 통해 고차원의 벡터를 얻을 수 있음을 알아보았다. 사람이 이해할 수 없는 숫자로만 구성된 각각의 벡터에는 문장과 영상에 대한 이해 정보가 압축되어 포함되어 있을 것으로 가정을 하고, 동일한 벡터 공간에 두 개의 벡터를 놓고 의미적으로 연관된 벡터 간에는 코사인 유사도 값이 높게 계산되도록 신경망을 학습하는 조인트 임베딩 모델을 구축하여야 한다. 〈그림 3〉의 개념도에서 문장과 영상이 각각 임베딩되어 만나는 지점이라고 할 수 있다. 영상 검색을 위한 조인트 임베딩 모델 학습에서 많이 쓰이는 손실함수는 Max-margin 랭킹 손실함수이다. 이 함수는 주어진 문장에 대한 정답 영상의 유사도 값이 높게 나오면 손실 값이 작아지고, 반대의 경우 손실 값이 커지게 된다. 한편, 주어진 문장과 관련이 없는 Negative 샘플링 영상에 대해서는 유사도 값이 높아지면 손실 값도 커지게 된다. 이러한 손실함수로 학습을 진행하면서 질의문과 영상 임베딩 신경망의 학습 파라미터를 업데이트하여 손실값을 최소화하게 되면 주어진 문장, 즉 질의문에 대하여 가장 적합한 영상을 찾아낼 수 있게 된다.

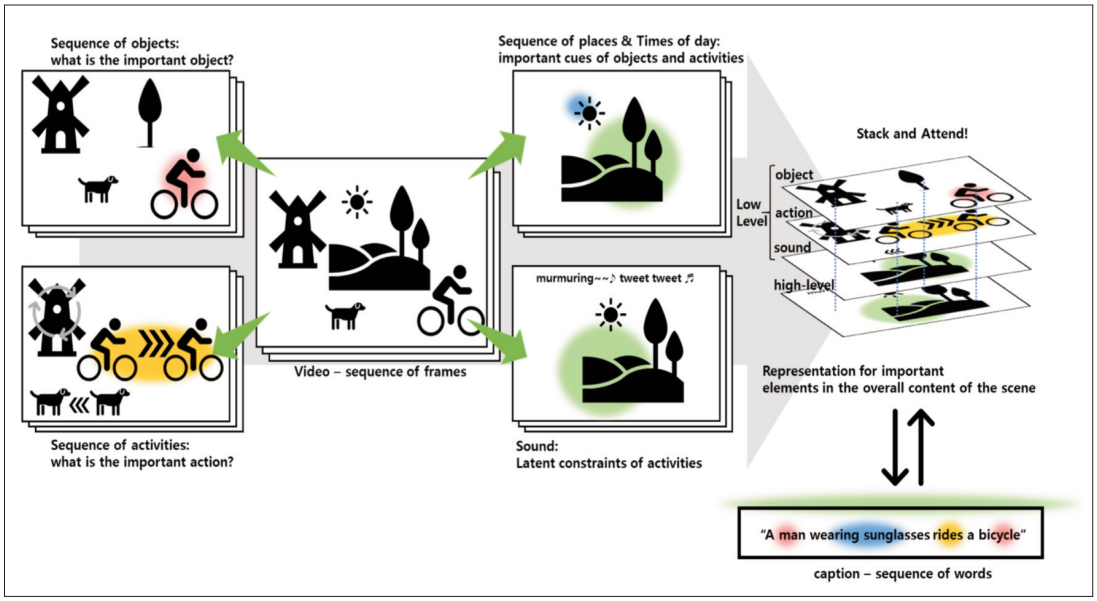
조인트 임베딩 기술을 이용하여 영상 검색을 수



<그림 10> 영상검색을 위한 신경망 구조 - 멀티모달 기반 영상 검색

행하는 연구는 몇 년 전부터 활발히 이루어지고 있다. 〈그림 10〉은 옥스포드대학교에서 발표한 영상 검색을 위한 신경망 구조이다. 영상으로부터 객체, 씬 유형, 행위, 얼굴, 그래픽 문자, 대사, 소리 정보를 추출하여 임베딩을 한 것이 인상적이다. 영상 검색 결과의 품질, 즉 검색 정확도를 높이기 위해서는 장면에 대한 정확한 이해가 필수 요소이며, 이를 위해 가능한 모든 정보를 추출하여 사용하는 접근방법이 효과적이다. 물론 이렇게 추출한 멀티모달 정보를 적절히 필터링하고 취합하여 단일 임베딩 벡터로 인코딩하는 과정도 중요하다.

저자가 속한 ETRI 미디어지능화연구실에서도 영상 검색을 위한 연구를 수행하고 있다. 공개되는 대부분의 연구는 MSR-VTT, LSMDC와 같은 공개 학습 데이터셋을 이용하여 성능을 검증하는데, 이러한 공개 데이터셋은 영어로 되어있기 때문에 한글 기반의 영상 검색 엔진을 구축하기에는 한계가 있다. 따라서 한글 영상 검색이 가능하도록 250편의 한국영화를 20초 길이로 분할하고 각 클립에 대한 명세문을 수동으로 작성하여 영상-문장 8만여 개 쌍을 학습 데이터로 구축하여 장면-문장 조인트 임베딩 연구를 수행하였다. 또한 영상과 문장에 대한 보다 정확한 이해가 가능하도록 영상의 멀티모달 추출 정보를 입체적으로 재구성 및 어텐션 기법으로 선별하여 영상의 전반적인 의미를 파악할 수 있는 네트워크 구조를 도입하였으며, 한글 Bert 사전학습 모델을 전이 학습시켜 문장에 대한 정확한 임베딩이 수행되도록 하였다. 〈그림 11〉은 ETRI에서 제시한 조인트 임베딩 학습 모델의 개념도이다. 〈그림 12〉는 ETRI에서 개발한 영상검색 엔진을 사용하여 검색을 수행해 본 예제 케이스이다. 전반적인 검색결과에서 주어진 질의문에 적합한 영상이 상위에 랭크되어 있음을 확인할 수 있었으며, 정량적 수치로 10개의 검색 결과 안에서 정답 영상이 포



<그림 11> ETRI에서 개발한 조인트 임베딩 기반 영상 검색 학습 모델의 개념도

<그림 12> 심층 학습기반 영상 검색 엔진을 이용한 검색 결과 예

함되어 있을 확률(Recall@10)은 평균 50.8%로 영상에 대한 수동 태깅이 없더라도 상당히 높은 정확도의 검색 성능을 제공함을 확인할 수 있었다.

Ⅲ. 결 론

본 고에서는 영상에 대한 메타데이터 수동 태깅 입력이 없어도 온전히 영상을 이해하고 질의문을 분석하여 영상 검색을 수행하는 조인트 임베딩 기반 영상 검색 기술에 대하여 소개하였다. 이를 위해 기존의 키워드 기반 영상 검색 접근방법에 대해서도 가볍게 다루었으며, 장면-문장 조인트 임베딩을 하기 위해 필요한 문장 임베딩 기술과 영상 임베딩 기술에 대해서도 소개하였다. 또한, 최신의 관련 연구 소개 및 저자가 속한 기관에서 개발한 영상 검색 엔진을 소개하고 검색 결과의 예제를 살펴 보았다.

심층 학습 기반의 영상 검색 기술은 앞으로 더욱 발전하여 보다 높은 검색 정확도를 제공할 것으로 예상되며 다양한 상용 영상 검색엔진이 출시될 것으로 보인다. 고화질 영상이 넘쳐나고 360도 영상과 같은 VR 미디어도 활발하게 공유될 것으로 보이기 때문에 새로운 유형의 미디어를 적정시간 내에 분석하는 모델 최적화 연구가 필요하며, 영상검색 학습 모델을 경량화하여 스마트폰에서도 자체적인 리소스만으로도 영상 검색을 수행할 수 있는 연구도 본격적으로 이루어질 것으로 보인다.

참 고 문 헌

[1] Describing Videos by Exploiting Temporal Structure, 10.1109/ICCV.2015.512
[2] Incorporating Relation Paths in Neural Relation Extraction, 10.18653/v1/D17-1186
[3] https://blog.dataiku.com/how-deep-does-your-sentence-embedding-model-need-to-be
[4] https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks/
[5] Deep Temporal Linear Encoding Networks, 10.1109/CVPR.2017.168
[6] https://github.com/gojibjib/jibjib-model
[7] https://github.com/CSAILVision/places365
[8] Use What You Have: Video Retrieval Using Representations from Collaborative Experts, (http://arxiv.org/abs/1907.13487)
[9] Neural Models for Information Retrieval (https://arxiv.org/pdf/1705.01509.pdf)
[10] Learning a Video-Text Joint Embedding using Korean Tagged Movie Clips, ICTC 2020

필 자 소 개

함 경 준

– 2014년 : 카이스트 산업공학과 박사
– 2015년 ~ : 한국전자통신연구원 미디어지능화연구실
– 주관심분야 : 검색, 뉴럴검색, 영상검색, 자연어처리