제134회 석사학위논문
지도교수 임 창 원

이미지-텍스트 매칭을 위한 동적 어텐션 네트워크

Dynamic Attention Network for Image-Text Matching

중앙대학교 대학원
통계학과 통계학전공
김 영 동

2020년 12월

이미지-텍스트 매칭을 위한 동적 어텐션 네트워크

Dynamic Attention Network for Image-Text Matching

이 논문을 석사학위논문으로 제출함

2020년 12월

중앙대학교 대학원

통계학과 통계학전공

김 영 동

김영동의 석사학위논문으로 인정함

심사위원장 _____ ㊞

심 사 위 원 _____ ㊞

심 사 위 원 _____ ㊞

중앙대학교 대학원

2020년 12월

목 차

표 목 차

그 림 목 차

제1장. 서론

1.1. 연구 배경

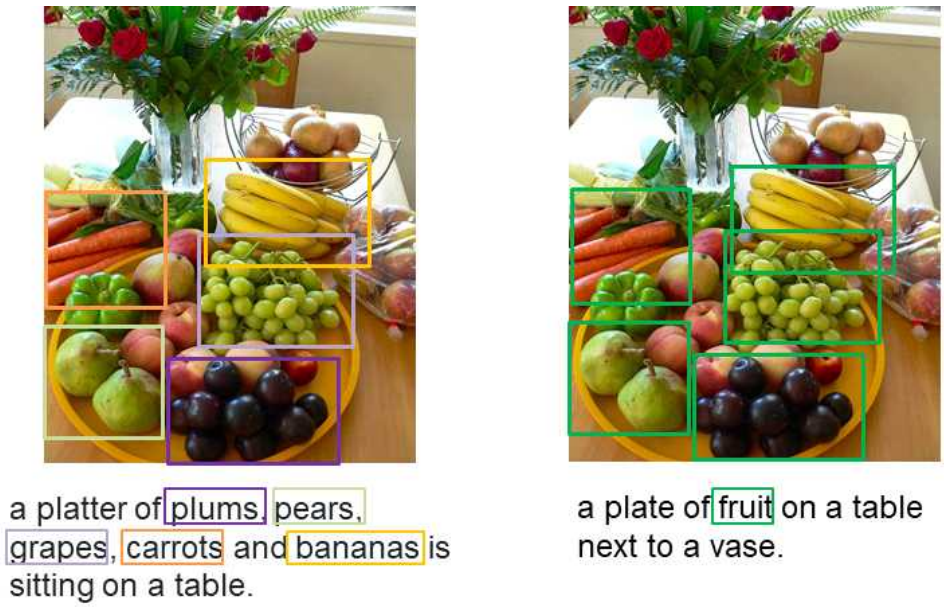
최근 모바일과 SNS(Social Network Service)의 인기와 함께 폭발적으로 늘어난 비정형 데이터의 다양성과 양으로 인해 교차 양식 검색(Cross Modal Retrieval)이 많은 관심을 끌고 있다. 이미지-문장 교차 양식 검색이란 주어진 문장과 관련된 이미지를 검색하거나 그 반대로 주어진 이미지와 관련된 문장을 검색하는 것을 말한다.

이미지-문장 교차 양식 검색의 핵심문제는 이미지와 문장 사이의 의미적 유사성을 어떻게 측정하는지에 있다. 최근 들어 딥 러닝 기술의 발달로 이미지와 문장 사이의 의미적 유사성을 측정하는 방법론들은 빠르게 발전하고 있지만 두 데이터의 큰 차이가 있기 때문에 여전히 큰 도전으로 남아있다.

초기의 연구에서는 이미지와 문장의 정보를 일반적인 잠재적 임베딩 공간 (latent embedding space)에 그대로 매핑하려고 시도했다. 예를 들어 Wang 등 (2016)은 이미지와 문장을 임베딩 공간에 각각 매핑하기 위해 두 개의 깊은 층의 가지를 채택하였다. 하지만 이러한 시도는 두 데이터의 대응을 거칠게 포착하여 이미지와 문장 사이의 미세한 상호작용을 묘사할 수 없었다.

이러한 미세 대응에 대한 더 깊은 이해를 얻기 위해 최근의 연구는 이미지-문장 교차 양식 검색을 위한 어텐션 메커니즘(attention mechanism)을 탐구했다. Karpathy 등 (2015)은 이미지내의 region의 특징과 문장 내의 단어의 특징을 추출하였고, region과 단어 쌍 간의 조밀한 정렬을 제안했다. Lee 등 (2018)은 region과 단어들 중 좀 더 두드러진 것들에 집중을 하는 Stacked Cross Attention Network(SCAN)을 제안하여 벤치마크 데이터세트에서 좋은 성능을 달성할 수 있었다.

1.2. 연구 목적



[그림 1-1] 하나의 이미지에 서로 다른 문맥의 설명이 있는 사진

[그림1-1]은 다른 문맥을 가진 검색 과정을 보여준다. 오른쪽의 녹색으로 표시된 region은 오른쪽 문장에서의 “fruit”에 해당한다. 하지만 그것들은 각각 왼쪽 그림에서 보라, 연두, 연보라, 주황, 노랑으로 표시된 왼쪽 문장의 “plums”, “pears”, “grapes”, “carrots”, “bananas”에 해당한다.

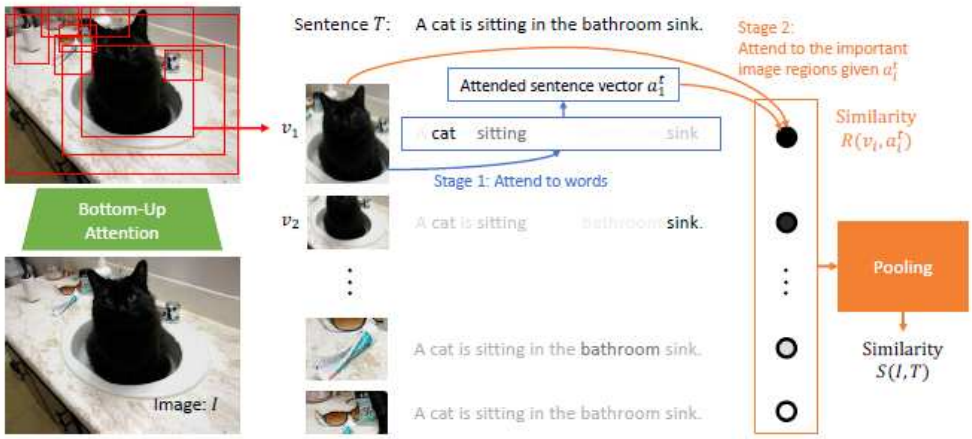
Zhang 등 (2020)에 의하면 사람들은 위의 그림처럼 region들을 하나씩 나열하여 설명할 수도 있고 하나로 묶어서 설명할 수도 있다. 그러나 이전의 어텐션 메커니즘 기반 이미지-문장 교차 양식 검색에서는 region 또는 단어가 다른 문맥에서 서로 다른 의미론을 가질 수 있다는 사실을 간과한다.

본 연구에서는 문맥적 정보를 활용하는 Intra modality 어텐션 메커니즘을 도입하는 새로운 모델을 제안하여 이미지-문장 교차 양식 검색의 성능 향상을 도모한다.

본 논문은 총 5장으로 구성되어 있다. 2장에서는 기존의 이미지-문장 교차 양식 검색 방법론에 대한 설명을 다루었으며, 3장에서는 본 연구에서 제안한 Dynamic Attention Network(DMAN) for Image-Text Matching을 설명하였고, 4장에서는 평가지표를 설명하고 제안한 방법과 기존 방법론들의 성능을 Flickr30K, MS-COCO 데이터를 적용하여 성능을 비교하였다. 5장에서는 본연구의 결론과 향후의 연구 방향에 대해 논하였다.

제2장. 선행연구

2.1. Stacked Cross Attention Network



[그림 2-1] Region에 대해 단어들을 어텐션하는 방법 (Lee 등, 2018)

Lee 등 (2018)은 이미지와 문장의 통합처리를 하는데 있어 기존의 방법들보다 크게 성능이 향상된 두 가지 방법을 제안하였다. 첫 번째 방법은 [그림 2-1]과 같이 각 이미지 안에 있는 region에 대해 문장의 단어들을 어텐션 하는 방법이다. 먼저 사전학습한 객체탐지기법인 Faster-RCNN (Girshick 등, 2015)과 ResNet-101 (He 등, 2016)을 이용해 이미지 내 region들의 특징들 $V=\{v_1,...,v_k\}$을 추출하고, Bi-directional Gated Recurrent Units(GRU) (Bahdanau 등, 2015; Schuster 등, 1997)를 이용하여 문장 내 단어들 $E=\{e_1,...,e_n\}$의 특징들을 추출한다. 그 다음 각각의 region에 대한 단어에 집중을 하고 어텐션된 문장벡터에 대한 region의 중요성을 결정한다.

주어진 이미지 I와 k개의 탐지된 region들, 주어진 문장 T와 n개의 단어들에 대해서 다음과 같이 모든 쌍들의 코사인 유사도를 구한 후 정규화를 한다.

$$s_{ij} = \frac{v_i^T e_j}{\| v_i \| \| e_j \|}, i \in [1, k], j \in [1, n], \quad (1)$$

$$\overline{s_{ij}} = \frac{[s_{ij}]_+}{\sqrt{\sum_{i=1}^{k} [s_{ij}]_+^2}}, [x]_+ = \max(x). \quad (2)$$

여기서 s_{ij}는 i번째 region과 j번째 단어 사이의 유사도를 의미한다. 그 다음 region에 대한 단어들을 어텐션하기 위해 단어벡터들의 가중치를 다음과 같이 구한다.

$$\alpha_{ij} = \frac{\exp(\lambda_1 \overline{s}_{ij})}{\sum_{j=1}^{n} \exp(\lambda_1 \overline{s}_{ij})}, \quad (3)$$

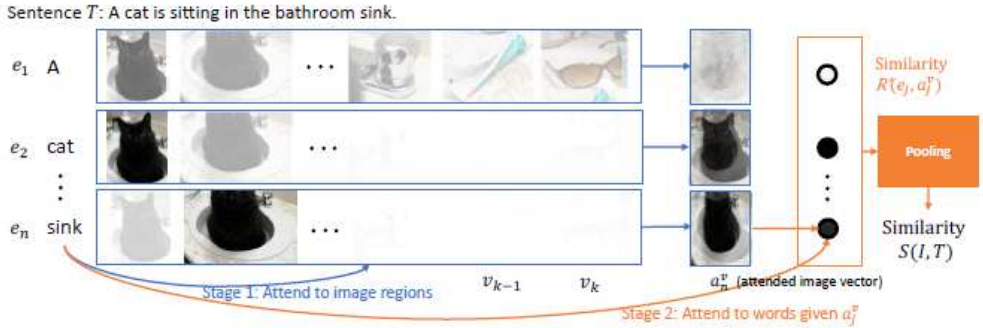
$$a_i^t = \sum_{j=1}^{n} \alpha_{ij} e_j. \quad (4)$$

그 다음, 주어진 문장에 대해 중요한 region을 결정하기 위해 i번째 region과 문장 사이의 관련성을 다음과 같이 region벡터인 v_i와 어텐션된 문장벡터인 a_i^t의 코사인 유사도로 구한다.

$$R(v_i, a_i^t) = \frac{v_i^T a_i^t}{\| v_i \| \| a_i^t \|}. \quad (5)$$

이미지 I와 문장 T사이의 유사도는 다음과 같이 He 등 (2008)에서 사용한 Log-SumExp(LSE) pooling으로 계산된다.

$$S_{LSE}(I, T) = \log\left(\sum_{i=1}^{k} \exp(\lambda_2 R(v_i, a_i^t))\right)^{(1/\lambda_2)}. \quad (6)$$



[그림 2-2] 단어에 대해 region들을 어텐션하는 방법

두 번째 방법은 반대로 [그림 2-2]과 같이 각 단어에 대한 이미지 region들을 어텐션하는 방법이다. 앞서 구한 s_{ij}의 정규화를 다음과 같이 구한다.

$$\bar{s}_{ij}^{'} = [s_{ij}]_+ / \sqrt{\sum_{j=1}^{n} [s_{ij}]_+^2} \tag{7}$$

그 다음 각 단어에 대한 region들을 어텐션하기 위해 다음과 같이 region벡터들의 가중치를 구한다.

$$\alpha_{ij}^{'} = \frac{\exp(\lambda_1 \bar{s}_{ij}^{'})}{\sum_{i=1}^{k} \exp(\lambda_1 \bar{s}_{ij}^{'})}, \tag{8}$$

$$a_j^v = \sum_{i=1}^{k} \alpha_{ij}^{'} v_i \tag{9}$$

주어진 이미지에 대해 중요한 단어를 결정하기 위해 j번째 단어와 이미지 사이의 관련성을 다음과 같이 문장벡터인 e_j와 어텐션된 region벡터인 a_j^v의 코사인유사도로 정의한다.

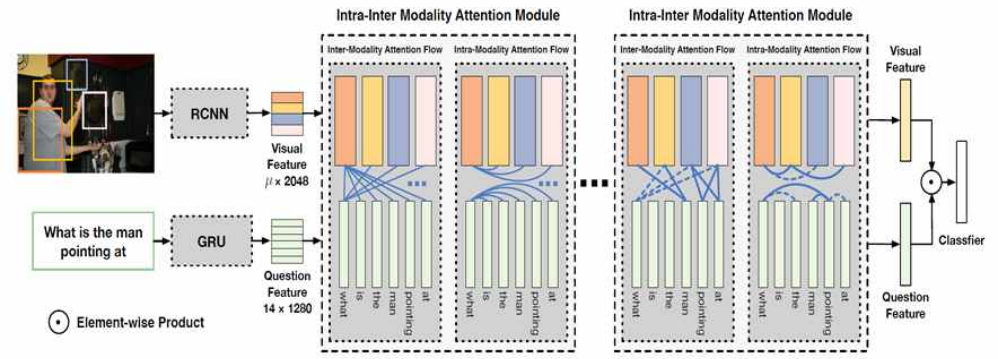
$$R^{'}(e_j, a_j^v) = \frac{e_j^T a_j^v}{\| e_j \| \, \| a_j^v \|}. \tag{10}$$

첫 번째 방법과 마찬가지로 이미지 I와 문장 T사이의 유사도는 LSE pooling으로 계산된다.

위와 같은 두 가지 방법으로 유사도를 학습하는 SCAN은 상대방의 정보를 활용하는 방법으로서 두드러지는 region 또는 단어를 잘 탐지할 수 있지만 문맥적 정보를 탐지하기에는 어려움이 있다.

2.2. Dynamic Fusion with Intra- and Inter-modality Attention Flow

Gao 등 (2019)은 이미지와 문장 데이터에 대해 내부와 외부 데이터의 정보흐름을 이용한 동적인 융합방법(Dynamic fusion with intra- and inter-modality attention flow, DFAF)을 제안한다. 이 방법은 이미지와 문장영역 간의 높은 수준의 상호작용을 강력하게 포착하여 Visual Question Answering(VQA) (Antol 등, 2015)의 성능을 크게 향상시켰다.



[그림 2-3] DFAF 구조 (Gao 등, 2019)

이미지의 특징을 $R \in \mathbb{R}^{\mu \times 2048}$, 문장의 특징을 $E \in \mathbb{R}^{14 \times 1280}$라 할 때, 먼저 Intra-modality attention flow과정으로 [그림 2-3]과 같이 region과 단어사이의 중요성을 포착하는 과정이다. 여기서 μ와 14는 이미지 내의 region의 개수와 문장 내의 단어의 최대 수를 의미한다. 학습된 가중치와 집계된 특징들을 전달하여 각 r

egion과 단어의 특징을 업데이트한다. 먼저, 주어진 이미지와 문장의 특징에 대하여 region과 단어 모든 쌍에 대한 가중치를 구한다. 그러기위해 각 이미지 특징과 단어 특징을 query, key, value로 나누고 차원을 맞춰준다. 각각을 $R_K, R_Q, R_V \in \mathbb{R}^{\mu \times \dim}$, $E_K, E_Q, E_V \in \mathbb{R}^{14 \times \dim}$로 정의한다. 여기서 dim은 region 벡터와 단어벡터의 차원을 의미한다.

$$R_K = Linear(R; \theta_{RK}), \;\; E_K = Linear(E; \theta_{EK}), \tag{11}$$

$$R_Q = Linear(R; \theta_{RQ}), \;\; E_Q = Linear(E; \theta_{EQ}), \tag{12}$$

$$R_V = Linear(R; \theta_{RV}), \;\; E_V = Linear(E; \theta_{EV}). \tag{13}$$

위 식에서 $Linear$는 각각 가중치 $\theta_{RK}, \theta_{RQ}, \theta_{RV}, \theta_{EK}, \theta_{EQ}, \theta_{EV}$를 가진 완전연결층을 의미한다. 이미지의 특징 중 query와 문장의 특징 중 key를 내적 하여 초기의 어텐션 가중치를 구하여 단어 특징에서 시각적 특징에 대한 정보를 수집하고 반대로도 진행한다. 그 다음 정규화를 하여 어텐션 가중치를 얻는다.

$$InterMAF_{R \leftarrow E} = softmax\left(\frac{R_Q E_K^T}{\sqrt{\dim}}\right), \tag{14}$$

$$InterMAF_{R \rightarrow E} = softmax\left(\frac{E_Q R_K^T}{\sqrt{\dim}}\right). \tag{15}$$

양방향의 InterMAF행렬을 통해 모든 region과 단어 쌍 사이의 중요성을 포착한다. 그 다음 InterMAF를 이용해 이미지영역과 문장영역을 업데이트하여 $R_{update} \in \mathbb{R}^{\mu \times \dim}$, $E_{update} \in \mathbb{R}^{14 \times \dim}$로 표현한다.

$$R_{update} = InterMAF_{R \leftarrow E} \times E_V, \tag{16}$$

$$E_{update} = InterMAF_{R \to E} \times R_V. \tag{17}$$

업데이트된 이미지와 문장 특징을 구한 후, 초기의 이미지 특징인 R과 문장 특징인 E와 각각 concatenate한 후 완전연결층을 이용하여 차원을 맞춰줘 최종적인 특징들을 만든다. 최종적으로 계산된 특징들은 이후 진행될 Dynamic Intra-modality Attention 과정의 입력값으로 사용된다.

$$R = Linear([R, R_{update}]^T; \theta_{RT}), \tag{18}$$

$$E = Linear([E, E_{update}]^T; \theta_{ET}). \tag{19}$$

여기에서, $Linear$는 각각 가중치 θ_{RT}, θ_{ET}를 가진 완전연결층을 의미한다. 그 다음으로 Dynamic Intra-modality Attention 과정으로써 다음과 같이 다른 모델의 조건부 정보를 요약하기 위해 조건부게이트를 만든 후, query와 key특징과 조건부게이트를 행렬 곱해준다 (Gehring 등, 2017).

$$G_{R \to E} = \sigma(Linear(AvgPool(R); \theta_{RP})), \tag{20}$$

$$G_{R \leftarrow E} = \sigma(Linear(AvgPool(E); \theta_{EP})). \tag{21}$$

$$\hat{R}_Q = (1 + G_{R \leftarrow E}) \odot R_Q, \ \hat{R}_K = (1 + G_{R \leftarrow E}) \odot R_K, \tag{22}$$

$$\hat{E}_Q = (1 + G_{R \to E}) \odot R_Q, \ \hat{R}_K = (1 + G_{R \to E}) \odot E_K. \tag{23}$$

여기서 $\sigma(\bullet)$는 시그모이드 함수를 뜻하고, $AvgPool$은 average pooling을, $\odot$은 element-wise 곱을 의미한다. 또한 $Linear$는 각각 가중치 θ_{RP}, θ_{EP}를 가진 완전연결층을 의미한다. Dynamic Intra-modality Attention 행렬인 $DyIntraMAF_{R \leftarrow R} \in \mathbb{R}^{\mu \times \mu}$와 $DyIntraMAF_{E \leftarrow E} \in \mathbb{R}^{14 \times 14}$은 다음과 같이 계산되고,

$$DyIntraMAF_{R\leftarrow R} = softmax\left(\frac{\hat{R}_Q\hat{R}_K^T}{\sqrt{\dim}}\right), \tag{24}$$

$$DyIntraMAF_{E\leftarrow E} = softmax\left(\frac{\hat{E}_Q\hat{E}_K^T}{\sqrt{\dim}}\right), \tag{25}$$

최종적인 이미지와 문장 특징들은 업데이트된 value특징인 R_V와 E_V로 다음과 같이 구해진다.

$$R_{update} = DyIntraMAF_{R\leftarrow R} \times R_V, \tag{26}$$

$$E_{update} = DyIntraMAF_{E\leftarrow E} \times E_V, \tag{27}$$

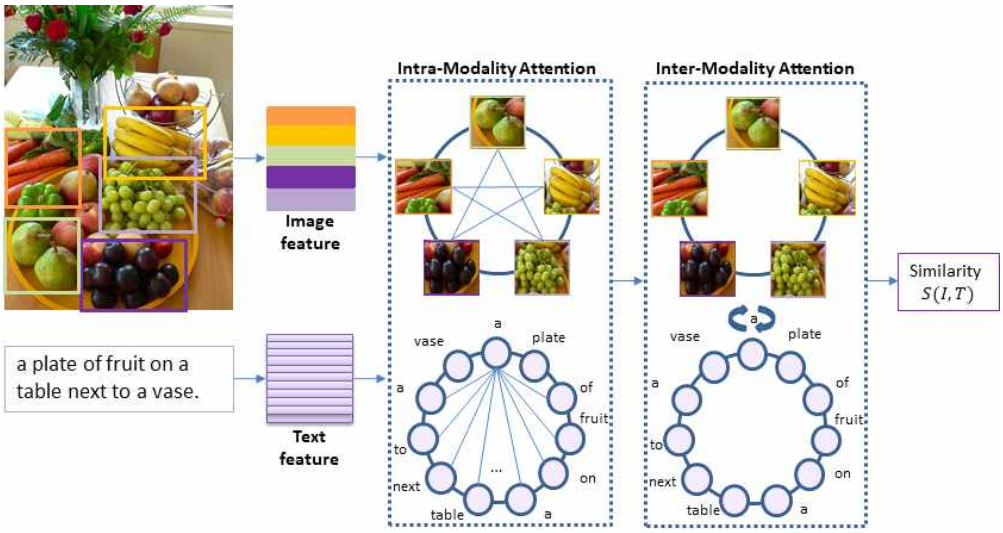
$$R = Linear(R + R_{update}; \theta_{RD}), \tag{28}$$

$$E = Linear(E + E_{update}; \theta_{ED}). \tag{29}$$

여기에서 $Linear$는 각각 가중치 θ_{RD}, θ_{ED}를 가진 완전연결층을 의미한다. 위와 같은 과정이 하나의 블록으로 위 논문에서는 블록을 8개 반복실행 하였을 때의 성능이 가장 좋았다고 한다. 하지만 이러한 반복과정은 학습시간이 오래 걸린다는 점과 메모리상의 문제를 야기할 수 있는 단점이 있다.

3장. 모델 제안

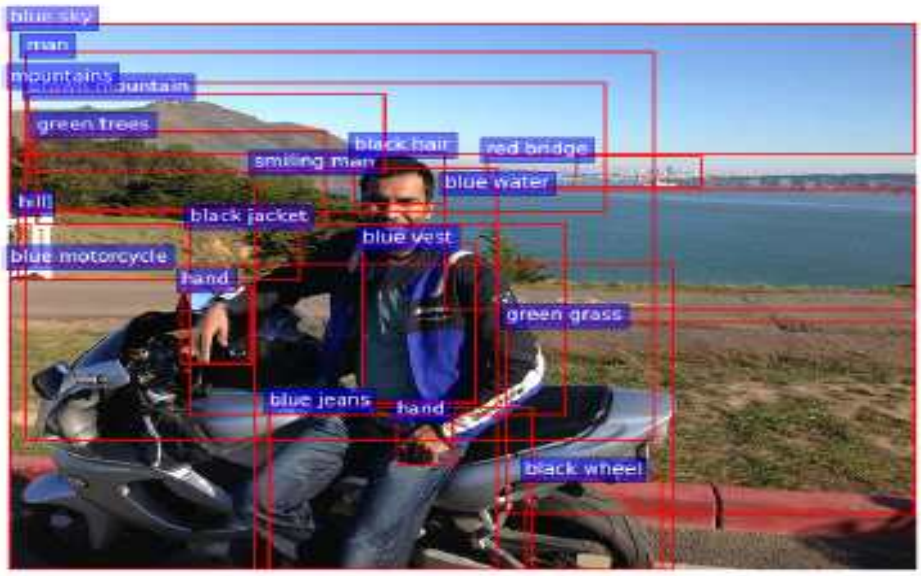
SCAN의 경우 region이나 단어가 다른 맥락에서는 다른 의미를 가질 수 있다는 사실을 간과하는 것에 대한 한계가 있기 때문에 본 논문에서는 DFAF의 Dynamic Intra-modality Attention의 아이디어를 착안하여 다른 데이터의 정보만 참고하는 것이 아닌 자기 자신의 정보도 참고하여 문맥적 정보파악을 통하여 region 및 단어 탐지가 가능한 개선된 Image-Text Matching모델인 Dynamic Attention Network(DMAN)을 제안한다. [그림 3-1]은 제안하는 모델의 대략적 구조이다.



[그림 3-1] DMAN 모델의 구조

3.1. Bottom-up attention을 이용한 이미지 특징 추출

주어진 이미지 I에 대해서, 이미지 특징집합인 $V=\{v_1,...,v_k\}, v_i \in \mathbb{R}^D$로 표현하기 위해 본 논문에서는 두드러진 region을 탐지하기 위해 Anderson 등 (2018)이 제안한 Bottom-up attention 방법과 Faster R-CNN (Girshick 등, 2015)을



[그림 3-2] Bottom-up Attention을 이용한 객체 (Anderson 등, 2018)

이용해 이미지 특징을 추출한다.

Faster R-CNN은 두 가지 단계가 있는 객체탐지기법이다. 첫 번째 단계는 Region Proposal Network(RPN)로 이미지를 입력받아 사각형 모양의 object proposal과 objectness score를 출력해준다. 두 번째 단계에서는 지역별 분류 및 bounding box 회귀를 위한 convolution feature map으로부터 Region Of Interests(ROIs)를 Pooling한다.

본 논문에서는 [그림 3-2]과 같이 의미가 풍부한 특징을 학습하기 위해 Andreson 등 (2018)이 Visual Genomes (Krishna 등, 2017) 데이터세트로 pre-trained한 Faster R-CNN과 ResNet-101 (He 등, 2016)을 채택하여 로컬화하기 어려운 region과 의미가 강한 region을 예측한다(예, 하늘, 풀, 빌딩 등).

각각의 선택된 region i에 대해, f_i는 이 객체의 2048차원인 mean-pooled convolutional feature를 의미한다. 그 다음 f_i를 다음과 같이 D차원의 벡터 v_i로 변환시킨다.

$$v_i = W_v f_i + b_v. \tag{30}$$

여기에서 W_v와 b_v는 각각 가중치와 편향이다. 따라서 최종적인 이미지의 임베딩표현은 다음과 같고 k는 이미지 내에 있는 region의 개수를 의미한다.

$$v = \{v_1, ..., v_k\}, v_i \in \mathbb{R}^D \tag{31}$$

3.2. 문장 특징 추출

이미지영역과 문장 영역을 연결하기 위해 이미지의 임베딩차원과 같게 문장도 D차원으로 매핑 해주어야 한다. T라는 문장이 주어졌을 때, 가장 간단한 접근법은 그 안에 있는 모든 단어를 개별적으로 매핑하는 것이다. 그러나 이 접근법은 문장 내에서 어떤 의미적 맥락도 고려하지 않는다. 때문에 본 논문에서는 Recurrent Neural Network(RNN)을 사용하여 단어의 맥락과 함께 단어를 임베딩한다.

문장 내의 i번째 단어에 대해, 먼저 단어의 index를 나타내는 one-hot 벡터로 나타내고, 임베딩 행렬 W_e을 통해 300차원으로 변환시킨다.

$$x_i = W_e w_i, i \in [1, \cdots, n] \tag{32}$$

여기서 n은 문장 내의 단어의 개수를 의미한다. 그 다음, bi-directional GRU을 이용하여 문장의 양방향 정보를 요약함으로써 문장의 문맥과 함께 최종적인 단어 벡터에 매핑한다. Bi-directional GRU는 첫 번째 단어인 w_1부터 마지막 단어인 w_n까지 읽는 forward GRU $\overrightarrow{h_i}$와 w_n부터 w_1까지 읽는 backward GRU $\overleftarrow{h_i}$로 구성되어 있다.

$$\overrightarrow{h_i} = \overrightarrow{GRU}(x_i), i \in [1, \cdots, n] \tag{33}$$

$$\overleftarrow{h_i} = \overleftarrow{GRU}(x_i), i \in [1, \cdots, n] \tag{34}$$

최종적인 단어 특징 t_i는 i번째 단어인 w_i 주위의 정보를 요약하는 $\overrightarrow{h_i}$와 $\overleftarrow{h_i}$의

평균으로 표현된다. 문장의 임베딩표현은 다음과 같다.

$$t = \{t_1, ..., t_n\}, t_i \in \mathbb{R}^D \tag{35}$$

3.3. Intra-modality attention

Intra-modality attention 과정에서는 자기 자신의 문맥적 정보를 얻기 위해 주어진 region벡터집합과 단어벡터집합 v, t에 대해, 다음과 같이 먼저 각각의 코사인 유사도를 구한 후 정규화를 한다.

$$x_{ij} = \frac{v_i^T v_j}{\| v_i \| \| v_j \|}, y_{ij} = \frac{t_i^T t_j}{\| t_i \| \| t_j \|}, \tag{36}$$

$$\overline{x_{ij}} = \frac{[x_{ij}]_+}{\sqrt{\sum_{i=1}^{k} [x_{ij}]_+^2}}, \overline{y_{ij}} = \frac{[y_{ij}]_+}{\sqrt{\sum_{i=1}^{n} [y_{ij}]_+^2}}, i \in [1, \cdots, k], j \in [1, \cdots, k] \tag{37}$$

여기서 $[a]_+ = \max(a)$를, x_{ij}는 i번째 region과 j번째 region 사이의 유사도를 의미하고, y_{ij}는 i번째 단어와 j번째 단어 사이의 유사도를 의미한다. 또한 정규화를 하는 이유는 실험적으로 정규화를 하였을 때 성능이 더욱 좋았기 때문이다. 그 다음 region에 대해 다른 region들을 어텐션하기 위해 다음과 같이 region벡터들의 가중치를 구한다.

$$\alpha_{ij} = \frac{\exp(\lambda_1 \overline{x}_{ij})}{\sum_{i=1}^{k} \exp(\lambda_1 \overline{x}_{ij})}. \tag{38}$$

그 다음 i번째 region과 어텐션을 해주어 문맥적 정보를 가지고 있는 region벡터로 업데이트 해준다.

$$v_i^{'} = \sum_{i=1}^{k} \alpha_{ij} v_i. \tag{39}$$

마찬가지로 단어에 대해 다른 단어들을 어텐션하기 위해 다음과 같이 단어벡터들의 가중치를 정의한 후 j번째 단어와 어텐션을 해주어 문맥적 정보를 가지고 있는 단어벡터로 업데이트 해준다.

$$\alpha_{ij}^{'} = \frac{\exp(\lambda_1 \bar{y}_{ij})}{\sum_{j=1}^{n} \exp(\lambda_1 \bar{y}_{ij})}, \tag{40}$$

$$t_j^{'} = \sum_{j=1}^{n} \alpha_{ij}^{'} t_j. \tag{41}$$

여기서 λ_1은 소프트맥스함수의 inversed temperature로서 소프트맥스를 계산하기 전에 logit을 나누는 양을 나타낸다 (Chorowski 등, 2015).

3.4. Inter-modality attention

Inter-modality attention과정에서는 상대방의 정보를 얻기 위해 주어진 region 벡터집합과 단어벡터집합 v, t에 대해 다음과 같이 region과 단어의 코사인유사도 s_{ij}를 구한다.

$$s_{ij} = \frac{v_i^T e_j}{\| v_i \| \ \| e_j \|}, i \in [1, \cdots, k], j \in [1, \cdots, n] \tag{42}$$

그 다음 문맥적 정보를 가지고 있는 region과 단어들에 대해 region에 대한 단어들을 어텐션하기 위해 단어벡터들의 가중치 β_{ij}를 구하고 마찬가지로 단어에 대한 region들을 어텐션하기 위해 region벡터들의 가중치 $\beta_{ij}^{'}$을 구한다.

$$\beta_{ij} = \frac{\exp(\lambda_2 s_{ij})}{\sum_{i=1}^{k} \exp(\lambda_2 s_{ij})},\ \beta_{ij}^{'} = \frac{\exp(\lambda_2 s_{ij})}{\sum_{j=1}^{n} \exp(\lambda_2 s_{ij})}. \tag{43}$$

여기서 λ_2도 마찬가지로 소프트맥스함수의 inversed temperature이다. 그 다음 각각을 intra-modality attention과정에서 업데이트된 region과 단어벡터에 어텐션을 해주어 문맥적 정보와 상대방의 정보를 동시에 갖고 있는 최종적인 region벡터 a_j^v와 단어벡터 a_i^t로 업데이트 해준다.

$$a_i^t = \sum_{j=1}^{n} \beta_{ij} t_j^{'},\ a_j^v = \sum_{i=1}^{k} \beta^{'ij} v_i^{'}\ . \tag{44}$$

3.5. 유사도 점수

마지막 과정으로서 주어진 이미지에 대해 중요한 단어를 결정하기 위해 j번째 단어와 이미지사이의 관련성을 다음과 같이 문장벡터인 e_j와 앞서 업데이트한 어텐션된 region벡터인 a_j^v의 코사인 유사도로 구한다.

$$R(e_j, a_j^v) = \frac{e_j^T a_j^v}{\| e_j \| \ \| a_j^v \|}. \tag{45}$$

마찬가지로 주어진 문장에 대해 중요한 region을 결정하기 위해 i번째 region과 문장사이의 관련성을 다음과 같이 region벡터인 v_i와 앞서 업데이트한 어텐션된 단어벡터인 a_i^t의 코사인 유사도로 구한다.

$$R^{'}(v_i, a_i^t) = \frac{v_i^T a_i^t}{\| v_i \| \ \| a_i^t \|}. \tag{46}$$

그리고 다음과 같이 각각을 LSE pooling해주어 더해주면 최종적인 이미지 I 와 문장 T 사이의 유사도 점수인 $S(I, T)$가 정의된다.

$$R(I, T) = \log\left[\sum_{j=1}^{n} \exp(\lambda_3 R(e_j, a_j^v))\right]^{(1/\lambda_3)}, \quad (47)$$

$$R'(I, T) = \log\left[\sum_{i=1}^{k} \exp(\lambda_3 R'(v_i, a_i^t))\right]^{(1/\lambda_3)}, \quad (48)$$

$$S(I, T) = R(I, T) + R'(I, T). \quad (49)$$

3.6. 손실 함수

Triplet loss는 이미지 문장 통합처리를 할 때 보통 사용되는 ranking objective이다. Karpathy 등 (2015)는 다음과 같이 hinge-based triplet ranking loss에 마진 α를 적용시켰다.

$$l(I, T) = \sum_{\hat{T}} [\alpha - S(I, T) + S(I, \hat{T})]_+ + \sum_{\hat{I}} [\alpha - S(I, T) + S(\hat{I}, T)]_+ \quad (50)$$

여기에서 $[x]_+ \equiv \max(x, 0)$, $S(I, T)$는 정답인 이미지-문장 쌍의 유사도점수이고, $\hat{T}$는 I와 매칭되지 않는 문장, $\hat{I}$는 T와 매칭되지 않는 이미지를 의미하며, $S(I, \hat{T})$와 $S(\hat{I}, T)$는 정답이 아닌 쌍의 유사도점수이다. 위의 손실함수는 정답인 이미지-문장 쌍은 가까이하고 부정적인 쌍은 거리를 멀게 하도록 한다. 하지만 이미지 문장 통합처리에 널리 사용되었음에도 불구하고, 샘플링의 무작위성으로 인해 높은 redundancy와 수렴이 느리다는 단점이 있다. 때문에 계산 효율을 위해 본 논문에서는 Faghri 등 (2018)이 제안한 방법인 미니배치에서 모든 부정적인 쌍을 합하기 보다는 부정적인 쌍 중 가장 유사도가 높은 것과 함께 손실함수를 계산하는 방법을 사용한다.

정답인 이미지-문장 쌍인 (I, T)에 대해, 부정적인 쌍 중 가장 유사도가 높은 것을 $\hat{I}_h = \arg max_{m \neq I} S(m, T)$, $\hat{T}_h = \arg max_{d \neq T} S(I, d)$로 정의한다. 최종적으로 본 논문에서 사용하는 손실함수는 다음과 같다.

$$l_{hard}(I,\ T) = \left[\alpha - S(I,\ T) + S(I,\ \hat{T}_h)\right]_+ + \left[\alpha - S(I,\ T) + S(\hat{I}_h,\ T)\right]_+ \quad (51)$$

제 4장. 실험

본 논문에서는 제안하는 모델인 DMAN을 평가하기 위해 이미지-문장 교차양식 검색에 많이 사용되는 Flickr30K (Plummer 등, 2015)와 MS-COCO (Young 등, 2014) 데이터셋을 이용하여 광범위한 실험을 수행하였고, 다른 다양한 이미지-문장 교차양식 검색 모델들과 비교하였다. 비교한 모델은 Deep visual-semantic alignments(DVSA) (Karpathy 등, 2015), selective multimodal LSTM(sm-LSTM) (Huang 등, 2017), 2WayNet (Eisenschtat 등, 2017), Visual-semantic embeddings(VSE++) (Faghri 등, 2018), Dual attention network(DAN) (Nam 등, 2017), Dual-path convolutional image-text embedding(DPC) (Zheng 등, 2017), Semantic concepts and order(SCO) (Huang 등, 2018), Generative cross-modal feature learning framework(GXN) (Gu 등, 2018), SCAN, BFAN이다.

4.1. 데이터 설명 및 평가지표

4.1.1. 데이터 설명

Flickr30K : 31,000개의 이미지들로 구성되어 있고, 각 이미지는 5개의 설명과 연결되어 있다. Karpathy 등 (2015)을 따라 1,000개의 이미지를 validation으로 사용하고 1,000개의 이미지를 test로 사용하고 나머지 이미지들을 train에 사용한다.

MS-COCO : 123,287개의 이미지들로 구성되어 있고, 각 이미지는 Flickr30K와 마찬가지로 5개의 설명과 연결되어 있다. 또한 Karpathy 등 (2015)을 따라 5,000개의 이미지를 validation으로 사용하고 5,000개의 이미지를 test로 사용하고 82,783개의 이미지를 train에 사용한다. 그리고 Faghri 등 (2018)을 따라 사용하지 않

은 30,504개의 이미지를 원래의 validation에 추가한다. [그림 4-1]은 MS-COCO의 예시를 보여준다.

- a weird looking blue bus in a field.
- a vintage blue and white bus displayed in a field.
- a fancy bus with multiple pictures and awards at a park.
- two different buses one is white and blue the other red and white.
- an old bus on show at an event parked next to motorcycles and another bus.

[그림 4-1] MS-COCO 데이터 예시 (https://cocodataset.org/)

4.1.2. 평가지표

본 논문에서는 이미지-문장 교차양식 검색에서 널리 사용되는 Recall at K (R@K)를 평가지표로 사용한다. R@K란 유사도점수를 기준으로 상위 K개의 후보 중 정답이 있는 비율을 의미한다. 본 실험에서는 R@1, R@5, R@10을 사용하였다.

4.2. 실험 결과

모델	Flickr30K Test					
	문장 검색			이미지 검색		
	R@1	R@5	R@10	R@1	R@5	R@10
(R-CNN, AlexNet)						
DVSA	22.2	48.2	61.4	15.2	37.7	50.5
(VGG)						
sm-LSTM	42.5	71.9	81.5	30.2	60.4	72.3
2WayNet	49.8	67.5	–	36.0	55.6	–
(ResNet)						
VSE++	52.9	80.5	87.2	39.6	70.1	79.5
DAN	55.0	81.8	89.0	39.4	69.2	79.1
DPC	55.6	81.9	89.5	39.1	69.2	80.9
SCO	55.5	82.0	89.3	41.1	70.5	81.1
(Faster-RCNN, ResNet)						
SCAN	67.4	90.3	**95.8**	48.6	77.7	85.2
BFAN	68.1	**91.4**	-	50.8	**78.4**	-
DMAN(ours)	**68.9**	90.5	**95.8**	**51.2**	77.9	**85.4**

[표 4-1] Flickr30K 데이터의 실험 결과

모델	MS-COCO Test					
	문장 검색			이미지 검색		
	R@1	R@5	R@10	R@1	R@5	R@10
(R-CNN, AlexNet)						
DVSA	38.4	69.9	80.5	27.4	60.2	74.8
(VGG)						
sm-LSTM	53.2	83.1	91.5	40.7	75.8	87.4
2WayNet	55.8	75.2	–	39.7	63.3	–
(ResNet)						
VSE++	64.6	90.0	95.7	52.0	84.3	92.0
DPC	65.6	89.8	95.5	47.1	79.9	90.0
GXN	68.5	–	97.9	56.6	–	94.5
SCO	69.9	92.9	97.5	56.7	87.5	94.8
(Faster-RCNN, ResNet)						
SCAN	72.7	94.8	**98.4**	58.8	88.4	94.8
BFAN	**74.9**	95.2	-	**61.6**	**89.6**	-
DMAN(ours)	73.5	**96.0**	**98.4**	57.9	88.6	**95.1**

[표 4-2] MS-COCO 데이터의 실험 결과

Flickr30K 데이터를 DMAN에 적용한 결과는 [표 4-1]과 같다. 분석결과 최신모델인 BFAN보다 문장검색의 R@1에서 0.8% 성능향상과 이미지검색의 R@1에서 0.4% 성능향상을 확인할 수 있었다. 또한 비교모델인 SCAN보다 문장검색의 R@1에서 1.5%, R@5에서 0.2% 성능향상을 보였고, 이미지검색의 R@1에서 2.6%, R@5에서 0.2%, R@10에서 0.2% 성능향상을 확인 할 수 있었다.

MS-COCO 데이터를 DMAN에 적용한 결과는 [표 4-2]와 같다. 분석결과 BFAN보다 문장검색의 R@5에서 0.8% 성능향상을 확인할 수 있었다. 또한 SCAN보다 문장검색의 R@1에서 0.8%, R@5에서 0.8% 성능향상을 보였고, 이미지검색의 R@5에서 0.2%, R@10에서 0.3% 성능향상을 확인 할 수 있었다.

실험결과 본 논문의 모델이 대부분의 지표에서 비교모델보다 향상된 성능을 보여주어 유의미한 결과를 도출한 것을 확인할 수 있었다.

제 5장. 결론

본 논문에서는 이미지-텍스트 매칭 문제에서 상대방의 정보를 고려하여 두드러진 region과 단어에 집중하는 SCAN의 한계점인 문맥적 정보를 간과한다는 사실에 착안하여 문맥적 정보를 고려할 수 있는 새로운 모델인 Dynamic Attention Network을 제안하였다.

문맥적 정보를 고려할 수 있는 Intra Attention과 상대방의 정보를 고려할 수 있는 Inter-Attention을 도입하여 비교모델인 SCAN보다 좋은 성능을 보임을 확인하였다. 때문에 이미지와 문장사이의 유사도를 고려할 때 맥락도 중요한 정보가 됨을 확인할 수 있었다. 때문에 본 연구는 향후 더 높은 성능을 보이는 이미지-문장 통합처리의 도움이 될 것이라고 생각한다.

다만, 본 연구는 Intra Attention과정에서 한 가지 데이터의 문맥만을 고려하고 다른 데이터의 문맥을 고려하지 못한다는 점에서 한계가 있을 수 있다. 때문에 향후 연구에서는 이러한 한계점을 극복하는 방향으로 진행하고자 한다.

참 고 문 헌

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and VQA. In: *CVPR.*

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, M., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In I*EEE International Conference on Computer Vision*, 2425-2433.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In: *ICLR.*

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. In: *NIPS*.

Eisenschtat, A., & Wolf, L. (2017). Linking image and text with 2-way nets. In: *CVPR*.

Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2017). VSE++: Improved visual-semantic embeddings. *arXiv preprint* arXiv:1707.05612.

Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S., Wang, X., & Li, H. (2019). Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6639-6648.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N,. (2017). Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, 1243-1252.

Gu, J., Cai, J., Joty, S., Niu, L. & Wang, G. (2018). Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In: *CVPR*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: *CVPR*.

Huang, Y., Wang, W., & Wang, L. (2017). Instance-aware image and sentence matching with selective multimodal LSTM. In: *CVPR*.

Huang, Y., Wu, Q., & Wang, L. (2018). Learning semantic concepts and order for image and sentence matching. In: *CVPR*.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In: *CVPR*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L. J., & Shamma, D. A., et al. (2017). Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1), 32-73.

Lee, K., Cehn, X., Hua, G., Hu, H., & He, X., (2018). Stacked cross attention for image-text matching. In *Proceedings of the European*

Conference on Computer Vision (ECCV), 201-216.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In: *ECCV*.

Liu, C., Mao, Z., Liu, A., Zhang, T., Wang, B., & Zhang Y. (2019). Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, 3-11.

Nam, H., Ha, J. W., & Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In: *CVPR*.

Plummer B, Wang L, Cervantes C, Caicedo J, Hockenmaier J, & Lazebnik S. (2015). Flickr30k entities: Collecting region to phrase correspondences for richer image to sentence models. In: *ICCV*, 2641-2649.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In: *NIPS*.

Wang, L., Li, Y., & Lazebnik, S. (2016). Learning deep structure-preserving image-text embeddings. In: *CVPR*.

Zhang, Q., Lei, Z., Zhang, Z., & Li, S. Z. (2020). Context Aware Attention Network for Image-Text Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Zheng, Z., Zheng, L., Garrett, M., Yang, Y., & Shen, Y. D. (2017). Dual-path convolutional image-text embedding. *arXiv preprint* arXiv:1711.05535.

국 문 초 록

이미지-텍스트 매칭을 위한 동적 어텐션 네트워크

김영동

통계학과 통계학전공

중앙대학교 대학원

Image-text matching은 이미지와 텍스트의 연결고리 역할을 하기 때문에 관심도가 증가하고 있다. 이러한 작업에는 cross-modal 검색(즉, 시각적 질의가 주어지면 해당되는 텍스트를 검색하거나, 반대로 언어적 질의가 주어지면 해당되는 이미지를 검색하는 작업)이 포함된다. 이 분야의 핵심은 이미지와 텍스트 사이의 유사성을 어떻게 학습하는지에 달려있다.

본 논문에서는 Image-text matching 작업에서 이미지 및 텍스트 사이의 동적 정보를 교대로 전달하는 intra-modal과 inter-modal 정보 흐름을 가진 Dynamic Attention Network(DMAN) 모델을 제안하고자 한다. 이 모델을 사용하면 이미지와 텍스트 사이에서 높은 수준의 문맥적 상호작용을 포착할 수 있기 때문에 image-text matching에서의 성능향상을 기대할 수 있다.

본 논문에서는 교차모달 검색을 통해 성능비교를 하기 위해 Flickr30K와 MS-COCO라는 두 가지 이미지, 텍스트 검색 데이터에 대해서 실험을 진행하였고, 그 결과 기존의 모델과 비교하여 성능 향상을 확인하였다.

주요용어: 딥 러닝, 멀티모달검색, 어텐션기법

ABSTRACT

Dynamic Attention Network for Image-Text Matching

Youngdong Kim
Major in Statistics
Department of Statistics
The Graduate School
Chung-Ang University

Image-text matching problem is a link between image and text, so it has attracted great interest in the past decades. Tasks in Image-text matching include cross-modal retrieval ($i.e.$, image search for given sentences with visual descriptions and the retrieval of sentences from image queries.). The key to this study depends on how we learn the similarity between images and texts.

In this paper, we propose a Dynamic Attention Network (DMAN) with Intra-modal and inter-modal information flow that alternately delivers dynamic information between images and texts in cross-modal task. DMAN can capture high level of contextual interaction between images and texts, so we expect to improve performance in cross-modal retrieval.

In this study, two data such as Flickr30K and MS-COCO were experimented to compare performance through cross-modal retrieval, and as a result, performance improvement was confirmed compared with other models.

Keyword : Deep learning, Multi-modal retrieval, Attention algorithm