Journal of Korea Multimedia Society Vol. 24, No. 1, January 2021(pp. 21-29)
https://doi.org/10.9717/kmms.2020.24.1.021

# 반려묘의 상황인지형 행동 캡셔닝 시스템

채희찬†, 최윤아††, 이종욱†††, 박대희††††, 정용화†††††

# Context-Awareness Cat Behavior Captioning System

Heechan Chae†, Yoona Choi††, Jonguk Lee†††, Daihee Park††††, Yongwha Chung†††††

## ABSTRACT

With the recent increase in the number of households raising pets, various engineering studies have been underway for pets. The final purpose of this study is to automatically generate situation-sensitive captions that can express implicit intentions based on the behavior and sound of cats by embedding the already mature behavioral detection technology of pets as basic element technology in the video capturing research. As a pilot project to this end, this paper proposes a high-level capturing system using optical-flow, RGB, and sound information of cat videos. That is, the proposed system uses video datasets collected in an actual breeding environment to extract feature vectors from the video and sound, then through hierarchical LSTM encoder and decoder, to identify the cat's behavior and its implicit intentions, and to perform learning to create context-sensitive captions. The performance of the proposed system was verified experimentally by utilizing video data collected in the environment where actual cats are raised.

**Key words:** Cat Behavior Monitoring, Cat Behavior Captioning System, Context-Awareness, Deep Learning

## 1. 서 론

2020년 반려동물 트렌드 리포트에 따르면 우리나라 전체 가구 중 23.7%가 반려동물을 양육하고 있다고 보고된다. 이는 전체 가구 중 1/4에 해당하는 많은 수의 가구가 반려동물과 생활하고 있음을 의미한다[1]. 특히 2012년 이후 반려묘의 연평균 증가율은 25.4%로 반려견(4.2%)에 비해 약 6배가량 높다. 이처럼 반려동물에 대한 관심이 높아지는 것과 비례하여 반려동물을 건강하게 키우고자 하는 관심 또한 매우 높다. 그러나 사람이 24시간 반려동물을 살펴볼 수는 없는 것이 현실이기 때문에, 반려동물을 자동으로 모니터링할 수 있는 시스템이 필요하다.

행동 탐지와 관련한 최근의 연구 동향에 의하면, IT 기반의 기술에 의하여 인간이나 움직이는 사물을 자동으로 탐지하고 분류하기 위한 연구들에 이어서, 반려동물의 행동까지도 파악하는 연구들이 현재 진행 중이다. 예를 들면, 가속도계 센서 데이터를 활용

※ Corresponding Author : Daihee Park; Jonguk Lee, Address: (30019) Sejong-ro 2511, Sejong City, Republic of Korea, TEL : +82-10-2299-1344, FAX : +82-44-860-1344, E-mail : dhpark@korea.ac.kr; eastwest9@korea.ac.kr
Receipt date : Oct. 19, 2020, Revision date : Nov. 20, 2020
Approval date : Dec. 17, 2020
† InfoValleyKorea
(E-mail : chay219@korea.ac.kr)
†† Dept of Computer and Information Science, Korea University
(E-mail : cyabc2@korea.ac.kr)
††† Dept of Computer and Convergence Software, Korea University
†††† Dept of Computer and Convergence Software, Korea University
††††† Dept of Computer and Convergence Software, Korea University (E-mail : ychungy@korea.ac.kr)
※ This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B07044938 and NRF-2020R1I1A3070835).

하여 강아지의 행동을 분류하는 연구[2], 아두이노와 소리 감지 센서를 활용한 반려견의 소리 분류 연구[3], 그리고 고양이의 울음소리를 기계학습을 통해 10가지 종류로 분류하는 연구[4] 등이 보고되고 있다.

한편, 동영상 속 객체의 탐지 및 행동 분류와 함께 해당 객체의 행동에 대한 캡셔닝 연구들이 최근 보고되고 있다. 예를 들면, 의미 특징 학습과 주의집중기술을 이용한 고밀도 비디오 캡션의 생성에 관한 연구[5], 시간정보를 활용하여 보다 긴 시간의 비디오 캡션을 가능하게 하는 계층적 RNN encoder를 제안한 연구[6] 등이 있다.

본 연구에서는 이미 충분히 성숙된 반려동물의 행동 탐지 및 소리 식별 기술을 동영상 캡셔닝 연구에 기본 요소기술로 임베딩하여, 반려묘의 행동과 소리에 따른 암묵적 의도를 표현할 수 있는 상황인지형 캡션의 자동 생성을 최종 목적으로 한다. 이를 위한 첫 번째 시도(pilot project)로써, 본 논문에서는 반려묘 동영상의 optical-flow, RGB 및 소리 정보를 활용한 캡셔닝 시스템을 제안한다. 즉, 제안하는 시스템은 동영상의 특징 벡터를 추출한 후, 계층적 LSTM encoder로 동영상의 지역적/전역적 상황을 파악하고, decoder를 거쳐 상황에 맞는 캡셔닝을 생성할 수 있도록 학습을 수행한다. 실제 반려묘를 양육하는 환경에서 수집한 동영상 데이터를 활용하여 제안하는 시스템의 성능을 실험적으로 검증한다. 이를 보다 구체적으로 설명하자면, 주인이 방으로 들어간 후, 반려묘가 뒤따르다 방문 앞에 앉아서 '야옹'하는 소리가 포함된 동영상(방문을 열어 달라는 반려묘의 암묵적 의도가 내포)을 예로 들 수 있다. 이때, 해당 동영상의 영상 정보만을 이용하는 캡셔닝의 결과는 영상의 시각적 정보만을 표현하는 문장("the cat is sitting at the door")이 생성되지만, 본 논문에서 제안하는 영상뿐만 아니라 소리 정보를 동시에 활용하는 모델의 캡셔닝 실험 결과에서는 반려묘의 암묵적 의도를 표현하는 캡셔닝 결과("the cat wants to open the door")를 확인할 수 있다. 이와 같은 실험 결과는 영상의 시각 정보(RGB)와 객체의 움직임 정보(optical-flow), 그리고 객체의 소리 정보('야옹': 무언가를 원한다는 반려묘의 감정)를 기반으로 고수준의 상황인지형 캡셔닝이 가능함을 시사한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 주제인 반려동물을 대상으로 진행된 행동 분류 연구들과 비디오 데이터를 대상으로 진행된 캡셔닝 연구들을 살펴보고, 3장에서는 본 논문에서 제안하는 반려묘의 행동을 캡셔닝하는 시스템을 소개한다. 4장에서는 제안하는 시스템의 성능을 정량적, 정성적으로 검증하고, 마지막으로 5장에서는 본 연구의 결론 및 향후 연구를 언급한다.

## 2. 관련 연구

본 논문의 주제인 반려동물에 대한 공학적 연구에 초점을 맞춰 최신의 선행 연구들을 살펴보면 다음과 같다. 먼저 반려견과 관련된 연구를 살펴보면, Ladha 등[7]은 웨어러블 가속도계 측정 플랫폼으로 반려견의 행동 데이터를 취득한 후, 통계 분류 프레임워크를 통해 강아지의 행동을 분류하는 연구를 진행하였다. Kumpulainen 등[2]은 3축 가속도계 센서를 활용하여 24마리의 반려견을 대상으로 7가지 동작을 분류하는 연구를 진행하였으며, Kim 등[3]은 아두이노와 소리 감지 센서를 활용해 반려견의 소리 분류 연구를 진행하였다. 한편, 반려묘에 대한 연구로는 Pandeya 등[4]이 제안한 고양이의 울음소리를 기계학습을 통해 10가지 종류로 분류한 연구와 Nanni 등[8]이 소개한 Convolutional Neural Networks(CNNs)와 데이터 확대기법(data augmentation)을 이용한 새와 고양이의 소리 분류 연구가 있다. 또한, Lee 등[9]은 반려동물 모니터링의 지역적 한계를 해결하고자 딥러닝을 이용하여 고양이를 검출하고 추적하는 이동식 시스템을 제안하였다. 이를 위하여, 객체 검출 신경망 모델의 하나인 You Look Only Once(YOLO)를 이용하여 데이터셋을 학습하고, 라즈베리파이와 노트북을 무선 랜으로 연결하여 고양이의 움직임과 상태를 실시간으로 확인할 수 있는 이동식 모니터링 시스템을 설계하였다. 이렇듯 최근 반려동물에 대한 관심과 비례하여, 반려동물을 대상으로 하는 다양한 공학적 시도들이 존재한다. 하지만 대다수의 연구들이 반려동물의 행동이나 소리에 대한 분류와 같이 단순히 반려동물의 상태를 식별하는 수준의 연구들이 주를 이루고 있으며, 그마저도 반려견을 대상으로 진행된 연구들이 대부분이고, 반려묘를 대상으로 진행된 학술적 연구는 매우 제한적이다.

한편, 비디오를 대상으로 한 최근의 캡셔닝 연구들은 살펴보면, 먼저, Lee 등[5]은 의미 특징과 어텐

션 기법을 활용한 비디오 캡션에 관한 연구를 수행하였으며, Pan 등[6]은 계층적 RNN 인코더를 활용하여 보다 긴 시간의 비디오 캡션을 가능하게 하는 연구를 진행하였다. 그리고 Wang 등[10]은 오디오와 시각적 특징을 어텐션 기법을 통해 결합한 멀티모달 방식의 비디오 캡션 생성 연구를 진행하였다. 이처럼, 최근 고수준의 비디오 캡셔닝 연구들이 다양하게 제안되고 있으며 일정 부분 성공적인 결과를 도출하고 있다. 하지만, 동영상의 공간적(spatial), 시간적(temporal) 정보와 객체의 암묵적 의도를 포함하는 소리(audio) 정보를 함께 아우르는 상황인지형(context-awareness) 캡션 생성에 관한 연구와 반려묘를 대상으로 하는 캡셔닝 연구는 아직까지 발견되지 않는다. 본 연구에서는 반려동물의 행동 탐지 및 소리 식별 기술을 동영상 캡셔닝 연구에 기본 요소기술로 임베딩하여 반려묘의 행동과 소리에 따른 암묵적 의도까지도 표현할 수 있는 상황인지형 캡션의 자동 생성에 적합한 딥러닝 신경망 모델을 제안한다.

## 3. 반려묘의 행동 캡셔닝 시스템

본 논문에서 제안하는 상황인지형 반려묘의 행동 캡셔닝 시스템의 구조는 Fig. 1과 같이 크게 특징 추출 모듈, 인코딩 모듈, 디코딩 모듈의 3단계로 구성된다. 즉, 제안된 시스템은 동영상의 optical-flow와 RGB정보를 특징 벡터로 출력하는 Google사의 Inflated 3D ConvNet(I3D)모델[11]과 소리 정보를 위

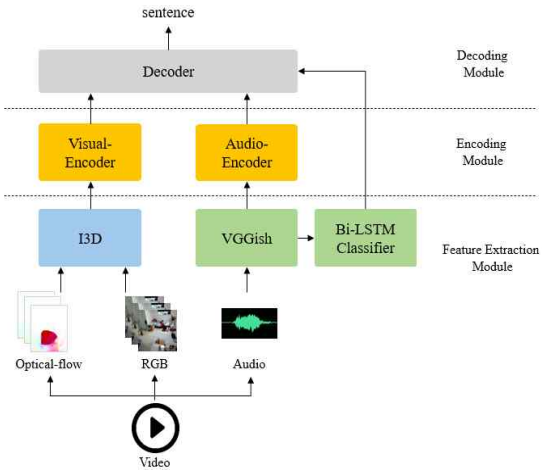



Fig. 1. Overall structure of cat behavior captioning system.

한 VGGish모델[12]에서 특징 벡터를 추출한 후, 어텐션 기법[13]이 적용된 계층적 LSTM 인코더[6]를 통해 문맥(context) 벡터를 추출한다. 이후, 문맥 벡터와 Bidirectional LSTM(Bi-LSTM) 모델[14-15]에서 생성된 소리 분류 특징 벡터가 영문 캡션을 생성하는 디코더에 전달됨으로써 최종 캡션 결과를 생성한다.

### 3.1 특징 추출 모듈

특징 추출과 관련된 기계 학습 패러다임의 변화는 사람이 수작업 특징(hand-crafted feature)을 구상하고 구현하는 과정이 특징 추출도 기계 학습하는 통째 학습(end-to-end learning)으로 진화했으며, 자동 추출된 특징이 보다 효과적임을 확인하는 다양한 응용 분야의 연구들이 보고되고 있다[16-22]. 본 연구의 특징 추출 모듈에서는 동영상 데이터에서 영상과 소리에 대한 특징을 효과적으로 추출하기 위하여 사전 학습된 CNN 모델을 사용하며, 그 구조는 Fig. 1의 하단부와 같다. 즉, 영상 특징은 Google사의 Deep Mind 팀에서 제공하는 I3D모델을 사용하여 optical-flow와 RGB에 해당하는 특징 벡터를 추출하고, 소리 특징은 VGG모델을 소리 데이터에 맞춰 변형한 VGGish 모델을 사용하여 소리 특징 벡터를 추출한다. 이처럼 추출된 영상/소리 특징 벡터는 각각 인코딩 모듈에 전달된다. 또한, VGGish 모델을 거쳐 생성된 소리 특징 벡터는 LSTM기반의 분류기를 통해 소리 분류 특징 벡터로 추출된 후, 디코딩 모듈에 전달된다(Fig. 2 참조). 이때, 반려묘의 소리는 야옹(meowing), 채터링(chattering), 하악질(hissing)과 그 외의 소리들(방에서 발생하는 다른 소리들)이며, 총 4가지의 클래스로 분류 학습한다.

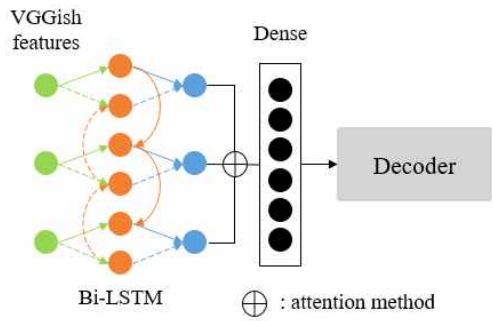



Fig. 2. Cat sound classification model.

### 3.2 인코딩 모듈

실험에 사용한 인코딩 모듈은 시각적(visual) 인코더와 소리(audio) 인코더로 구성되며, 두 인코더는 기본적으로 동일한 구조를 갖는다. 인코더는 Fig. 3과 같이 2단계의 계층적 LSTM으로 구성되며, 인코더 구조의 하단에 위치한 LSTM 계층은 동영상에서 나타나는 객체의 행동(지역적 특징)을 탐지하는 역할을, 상단에 위치한 LSTM 계층은 동영상의 객체 행동의 변화(전역적 특징)를 표현하는 역할을 수행한다. 여기서 지역적 특징이란 “고양이가 걷는다”, “고양이가 점프한다”와 같이 객체의 개별 행동에 대한 특징을 나타내며, 전역적 특징이란 “고양이가 걷다가 점프한다”와 같이 객체의 행동 변화에 초점을 맞춘 특징이다.

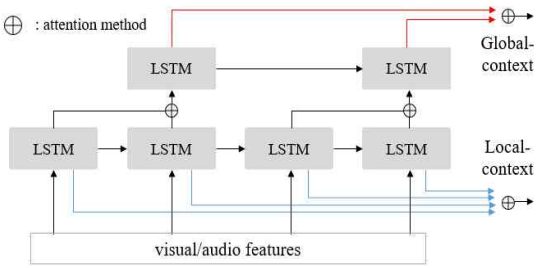



Fig. 3. Two-leveled hierarchical LSTM encoder.

### 3.3 디코딩 모듈

디코딩 모듈의 구조는 Fig. 4와 같으며, LSTM을 기반으로 크게 전역적 디코더와 지역적 디코더로 구성된다. 우선, 전역적 디코더는 어텐션 기법을 활용하여 소리와 영상 정보 중 보다 중요한 특징들을 강조하고, 객체의 전체적인 행동 변화의 순서를 이해한다. 다음으로, 지역적 인코더는 전역적인 행동 변화, 지역적인 문맥(local-context), 그리고 소리 분류 특징을 결합(concatenation)하여 지역적 디코더의 입력으로 사용한다. 이후 지역적 디코더의 출력값은 softmax를 통해 시간에 따라 가장 높은 확률의 단어를 생성하고, 최종적으로 생성된 단어를 시간 순서로 나열하여 캡션(문장)을 생성한다.

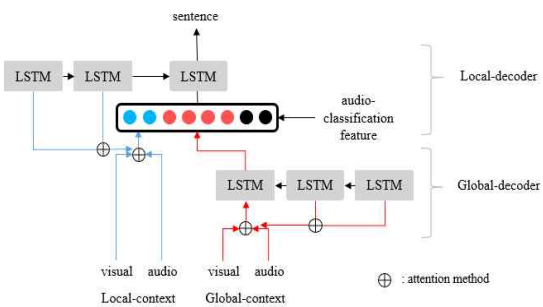



Fig. 4. Structure of the global/local decoder.

## 4. 실험 내용 및 결과

본 연구의 실험에서는 영상의 시각 정보(RGB)와 객체의 움직임 정보(optical-flow), 그리고 객체의 소리 정보를 사용한 고수준의 상황인지형 캡셔닝 결과를 정성적, 정량적 성능 평가로 확인한다.

### 4.1 실험 데이터 및 실험 내용

실험에 사용한 데이터는 실제 반려묘를 기르는 2.8m×2.7m 크기의 환경에서 취득한 약 2주 분량의 반려묘 동영상을 사용하였다. 영상 데이터는 라즈베리파이 카메라 모듈[23]을 사용하여 416×416 화질의 영상을 10fps로 촬영하였다. 촬영한 영상은 반려묘의 행동 및 소리를 기준으로 비디오 클립을 구성하였으며, Table 1에서 보듯이 5~30초 분량의 1,179개의 클립으로 구성된다. 여기서, 단일 동작(Single behavior)이란 반려묘의 걷기, 점프, 앉기, 그리고 먹기 등 한 가지 행동만을 포함한 데이터이며, 연속 동작(Continuous behavior)이란 단일 동작이 2개 이상 연속된 데이터(“The cat sits after walking”)를 말한다. 그리고 상호 동작(Interaction behavior)은 반려묘의 움직임뿐만 아니라 주인의 움직임도 포함된 데이터(“The cat walks with its owner”)이다. 실험 데이터는 학습 719개, 검증 150개, 테스트 310개로 구성했으며, 각 클립 당 5개의 정답지를 직접 생성하여 학습하였다. Fig. 5는 “고양이가 방안을 걷고 있다”에 해당하는 동영상으로 학습을 위한 캡션의 정답지 생성 예시이다.

Table 1. Configuration of cat video clips.

| Behavior | | Sound | |
|---|---|---|---|
| Single behavior | 793 | Meowing | 120 |
| Continuous behavior | 81 | Hissing | 45 |
| Interaction behavior | 95 | Chattering | 45 |
| Total | 1,179 | | |


Fig. 5. Example of video clip annotations.

본 논문에서 제안하는 반려묘의 행동 캡셔닝 모델을 검증하기 위한 학습 시스템은 1개의 NVIDIA TITAN X 그래픽 처리 장치, Intel(R) Core(TM) i7-8700K CPU 및 32GB RAM으로 구성하였으며, Ubuntu 18.04 운영체제에서 Python 3.6 및 Pytorch 0.4.1 프레임워크를 통해 수행하였다. 또한, 메인 시스템에서 사용되는 소리 분류 모듈(Bi-LSTM classifier)을 학습하기 위해 4개의 클래스별로 150개 총 600개의 소리 데이터를 사용하였으며, 사용된 데이터는 울음소리가 1회 이상 포함된다. 이때, 학습 하이퍼-파라미터(hyper-parameter)로는 50 epoch, 64 batch size, 128 units, 0.5 dropout 및 0.01의 learning rate를 사용하였고, 최종 실험 결과는 F-measure 0.82이다.

제안하는 반려묘의 행동 캡셔닝 모델을 두 가지 모델로 명명하여 실험하였다(1. RGB 정보와 optical-flow 정보를 사용한 제안 모델(Proposed-Op: Optical-flow); 2. Proposed-Op 모델에 소리 분류 모델을 추가한 제안 모델(Proposed-SCM: Sound Classification Model)). 제안한 모델들의 하이퍼-파라미터는 Table 2와 같다. 비교 실험을 위해서, 본 연구에서는 Wang 등[10]이 제안한 영상의 RGB 정보와 소리 정보만을 이용하는 캡션 생성 모델인 Hierarchically Aligned Cross-modal Attention (HACA) 모델을 선택하였다.

### 4.2 실험 결과의 정량적 평가

Table 3은 실험 결과의 정량적 평가로써, HACA, Proposed-Op, 그리고 Proposed-SCM 모델의 캡셔닝 결과와 실제 정답지 문장과의 유사도를 보여준다. 평가 지표로는 MS-COCO server[24]의 표준 평가 코드를 사용하여 계산된 BLEU, METEOR, 그리고 ROUGE-L 세 가지의 평가 지표를 사용하였다. 실험 결과, 본 논문에서 제안한 Proposed-SCM 모델의 성능이 다른 모델들에 비해서 우수함을 확인할 수 있다.

### 4.3 실험 결과의 정성적 평가

#### 4.3.1 optical-flow 유무에 따른 캡션 생성 결과

본 실험은 반려묘의 동영상에서 optical-flow 정보가 문장 생성에 미치는 영향력을 실험적으로 평가

Table 2. Hyper-parameters of the proposed model.

| | Visual | Audio |
|---|---|---|
| feature max len | 60 | 20 |
| low-level encoder LSTM hidden dim | 512 | 128 |
| high-level encoder LSTM hidden dim | 256 | 64 |
| global decoder LSTM hidden dim | 256 | |
| local decoder LSTM hidden dim | 1024 | |
| word embedding dim | 512 | |
| dropout | 0.5 | |
| batch size | 20 | |
| learning rate | 0.5 | |
| epochs | 50 | |
| adadelta optimizer | | |
| training data is shuffled at each epoch | | |

Table 3. Quantitative results of HACA and proposed models.

| | HACA | Proposed-Op | Proposed-SCM |
|---|---|---|---|
| BLEU_1 | 0.728 | 0.790 | 0.812 |
| BLEU_2 | 0.638 | 0.715 | 0.727 |
| BLEU_3 | 0.561 | 0.641 | 0.650 |
| BLEU_4 | 0.497 | 0.602 | 0.611 |
| METEOR | 0.452 | 0.490 | 0.510 |
| ROUGE_L | 0.746 | 0.798 | 0.801 |


Fig. 6. Result of cat behavior captioning with and without using optical-flow, respectively.

한다. Fig. 6("반려묘가 방에서 걷다가 점프"하는 영상)의 예시에 의하면, 제안된 시스템은 동영상에서 객체의 움직임 정보를 효과적으로 표현하는 optical-flow의 특징 벡터에 의해 "The cat jumps up after walking"이라는 시간 흐름에 따른 객체의 움직임을 순차적 복문의 형태로 표현함을 확인할 수 있다. 반면, 영상의 RGB 정보만(HACA 모델)을 활용하여 캡셔닝한 경우에는 시간에 따른 객체의 움직임 변화에 적절히 대응하지 못하는 캡션 결과("The cat is walking in the room")를 보여준다.

4.3.2 소리 유무에 따른 캡션 생성 결과

Fig. 7은 주인이 방으로 들어간 후, 반려묘가 뒤따르다 방문 앞에 앉아서 '야옹'하는 동영상을 캡쳐한 이미지로, 방문을 열어 달라는 반려묘의 암묵적 의도를 반영하는 상황으로 해석할 수 있는 시나리오이다[25]. 영상과 소리 정보를 동시에 고려하는 제안된 모델의 캡셔닝 실험 결과를 영상만 이용한 경우와 소리 정보까지 고려한 캡셔닝 실험 결과로 구분하여 Fig. 7과 Table 4에 정리하였다. 먼저, 영상 정보로만 판단하는 캡셔닝 결과는 영상의 시각적 정보를 표현하는 문장("the cat is sitting at the door")이 생성되지만, 울음소리 정보를 추가로 반영하는 본 시스템에서는 반려묘의 암묵적 의도를 표현하는 캡셔닝 결과


Fig. 7. Result of cat behavior captioning with and without using cat sounds, respectively.

("the cat wants to open the door")를 확인할 수 있다. 이와 같은 실험 결과는 영상의 시각 정보(RGB)와 객체의 행동 정보(optical-flow), 그리고 객체의 소리 정보('야옹': 뭔가를 원한다는 반려묘의 감정)을 기반으로 고수준의 상황인지형 캡셔닝이 가능함을 시사한다.

4.3.3 울음소리 종류에 따른 캡션 생성 결과

Fig. 8은 채터링(chattering)이 포함된 소리 데이

Table 4. Two levels of cat behavior captioning with and without using cat sounds, respectively.

| | | Cat sound X | | Cat sound O | |
|---|---|---|---|---|---|
| | | Visual | Audio | Visual | Audio |
| Caption | level 1 | sitting at the door | - | sitting at the door | want something |
| | level 2 | the cat is sitting at the door | | the cat wants to open the door | |

Sitting (visual) Chattering (sound)

**Caption results (chattering)**

HACA : 'the cat is sitting at the window <END>'

Proposed-Op : 'the cat is sitting at the window <END>'

Proposed-SCM
- level 1 : 'the cat is sitting at the window <END>' + chattering
- level 2 : 'the cat wants to hunting <END>'

Fig. 8. Captioning results of HACA and proposed models.

터가 포함된 경우에 대한 캡션 생성 결과이다. HACA 모델과 Proposed-Op 모델의 경우는 영상 정보에만 치우친 "the cat is sitting at the window"와 같은 캡션 결과를 생성한다. 반면, 소리 분류 모델이 추가된 제안된 Proposed-SCM 모델에서는 "the cat is sitting at the window"와 채터링이 합쳐져, "the cat wants to hunting"이라는 상황 인지적 의역이 생성됨을 확인할 수 있다(반려묘의 채터링 소리에는 사냥의 본능이 내제되어 있다[25-26]).

## 5. 결 론

본 논문에서는 실제 반려묘를 양육하는 환경에서 수집한 반려묘의 모습과 울음소리가 포함된 동영상을 이용하여 반려묘의 행동을 인지하고, 이를 기반으로 고수준의 캡션 결과를 생성하는 시스템을 제안하였다. 제안된 시스템은 이미 충분히 성숙된 반려동물의 행동 탐지 및 소리 식별 기술을 딥러닝 기반의 동영상 캡셔닝 연구에 기본 요소기술로 임베딩하여 반려묘의 행동과 소리에 따른 암묵적 의도까지도 표현할 수 있는 상황인지형 캡션의 자동 생성이 가능함을 예시를 통하여 실험적으로 확인하였다. 제안한 모델에서는 객체의 움직임 정보를 표현하는 optical-flow 스트림을 추가함으로써 RGB 정보만을 이용한 경우보다 해당 움직임 정보의 의미를 보다 정확하게 식별했으며, 이를 기반으로 고수준의 복문 형성이 가능함을 실험을 통해 입증하였다. 또한, 반려묘의 소리 정보를 부가적으로 활용하게 되면 고양이의 암묵적인 의도까지도 포함된 세밀한 캡션 결과가 생성됨을 확인하였다.

향후, 본 연구의 최종 목적인 주인과 반려묘 사이의 의사소통에 초점을 맞추어 주인과 반려묘의 행동에 대한 상호작용 및 상호관계에 대한 캡션 생성 시스템으로의 확장 작업을 진행할 예정이다.

## REFERENCE


[1] Pet Trend Report 2020. https://blog.opensurvey.co.kr/trendreport/companionanimal-2020 (accessed April 13, 2020).

[2] P. Kumpulainen, A. Valldeoriola, S. Somppi, H. Tornqvist, H. vaataja, P. Majaranta, et al., "Dog Activity Classification with Movement Sensor Placed on the Collar," *The Fifth International Conference on Animal-Computer Interaction*, Vol. 4, pp. 1-6, 2018.

[3] Y. Kim, J. Sa, Y, Chung, D. Park, and S. Lee, "Resource-Efficient Pet Dog Sound Events Classification Using LSTM-FCN Based on Time-Series Data," *Sensors*, Vol. 18, No. 11, 4019, 2018.

[4] Y. Pandeya, D. Kim, and J. Lee, "Domestic Cat Sound Classification Using Learned Features from Deep Neural Nets," *Applied Sciences*, Vol. 8, 1949, 2018.

[5] S. Lee and I. Kim, "Video Captioning with Visual and Semantic Features," *Journal of Information Processing Systems*, Vol. 14, No. 6, pp. 1318-1330, 2018.

[6] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning," *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1029-1038, 2016.

[7] C. Ladha, N. Hammerla, E. Hughes, P. Olivier, and T. Ploetz, "Dog's Life: Wearable Activity Recognition for Dogs," *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 415-418, 2013.

[8] L. Nanni, G. Maguolo, and M. Paci, "Data Augmentation Approaches for Improving Animal Audio Classification," *Ecological Informatics*, Vol. 57, 101084, 2020.

[9] M. Lee, J. Kang, and S. Lim, "Design of YOLO-Based Removable System for Pet Monitoring," *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 24, No. 1, pp. 22-27, 2020.

[10] X. Wang, Y. Wang, and W. Wang, "Watch, Listen, and Describe: Globally and Locally Aligned Cross-Modal Attentions for Video Captioning," *arXiv preprint arXiv:1804.05448*, 2018.

[11] J. Careira and Z. Andrew, "Quo Vadis, Action Recognition? a New Model and the Kinetics Dataset," *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 629-6308, 2017.

[12] S. Hershey, S. Chaudhuri, D.P. Elis, J.F. Gemmeke, A. Jansen, R.C. More, M. Plakal, D. Plat, et al., "CNN Architectures for Large-Scale Audio Classification," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 131-135, 2017.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv preprint arXiv:1409.0473*, 2014.

[14] Q. Zhou and H. Wu, "NLP at IEST 2018: BiLSTM-Attention and LSTM-Attention via Soft Voting in Emotion Classification," *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 189-194, 2018.

[15] Z. Huang, W. Yu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[16] M. Hong, H. Ahn, O. Atif, J. Lee, D. Park, and Y. Chung, "Field-Applicable Pig Anomaly Detection System Using Vocalization for Embedded Board Implementations," *Applied Sciences*, Vol. 10, No. 19, 6991, 2020.

[17] J. Seo, H. Ahn, D. Kim, S. Lee, Y. Chung, and D. Park, "EmbeddedPigDet: Fast and Accurate Pig Detection for Embedded Board Implementations," *Applied Sciences*, Vol. 10, No. 8, 2878, 2020.

[18] Y. Choi, O. Atif, J. Lee, D. Park, and Y. Chung, "Noise-Robust Sound-Event Classification System with Texture Analysis," *Symmetry*, Vol. 10, No. 9, 402, 2018.

[19] J. Lee, Y. Choi, Y. Chung, and D. Park, "Sound Noise-Robust Porcine Wasting Diseases Detection and Classification System Using Convolutional Neural Network," *Journal of Korean Institute of Information Technology*, Vol. 16, No. 5, pp. 1-13, 2018.

[20] S. Park, J. Kim, and D. Kim, "A Study on Classification Performance Analysis of Convolutional Neural Network Using Ensemble Learning Algorithm," *Journal of Korea Multimedia Society*, Vol. 22, No. 6, pp. 665-675, 2019.

[21] H. Park, S. Bhattacharjee, P. Deekshitha, C. Kim, and H. Choi, "A Study on Deep Learning Binary Classification of Prostate Pathological Images Using Multiple Image Enhancement Technique," *Journal of Korea Multimedia Society*, Vol. 23, No. 4, pp. 539-548, 2020.

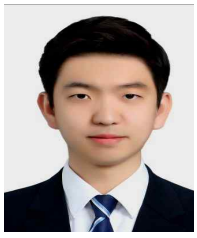
[22] K. Jo, S. Jung, and C. Sim, "A Study of Shiitake Disease and Pest Image Analysis Based on Deep Learning," *Journal of Korea Multimedia Society*, Vol. 23, No. 1, pp. 50-57, 2020.

[23] Raspberry Pi Official Website, https://projects.raspberrypi.org/en/projects/getting-started-with-picamera (accessed October 11, 2020).

[24] X. Chen, H. Fang, T.Y. Lin, R. Vedantam, and S. Gupta, "Microsoft COCO Captions: Data Collection and Evaluation Server," *arXiv preprint arXiv:1504.00325*, 2015.

[25] J.B Pam, "Think Like a Cat: How to Raise a Wel-Adjusted Cat-Not a Sour Pus," *Penguin*

*Books*, 2017.
[26] J.B Pam, "Cat Wise: America's Favorite Cat Expert Answers Your Cat Behavior Questions," *Penguin Books*, 2016.



**채 희 찬**

2018년 고려대학교 컴퓨터정보학과(학사)
2020년 고려대학교 컴퓨터정보학과(석사)
2020년~현재 ㈜인포벨리코리아 사원

관심분야 : 딥러닝, 텍스트마이닝, 영상처리, 빅데이터

**최 윤 아**

2019년 고려대학교 컴퓨터정보학과(학사)
2019년~현재 고려대학교 컴퓨터정보학과 석사과정
관심분야 : 빅데이터, 데이터마이닝, 딥러닝

**이 종 욱**

2002년 고려대학교 전산학과(학사)
2005년 고려대학교 전산학과(석사)
2014년 고려대학교 전산학과(박사)

2014년~현재 고려대학교 컴퓨터정보학과 초빙교수
관심분야 : 딥러닝, 데이터마이닝, 융합 IT, 음향분석

**박 대 희**

1982년 고려대학교 수학과(학사)
1984년 고려대학교 수학과(석사)
1989년 플로리다 주립대학 전산학과(석사)
1992년 플로리다 주립대학 전산학과(박사)

1993년~현재 고려대학교 컴퓨터정보학과 교수
관심분야 : 빅데이터, 데이터마이닝, 인공지능, 융합 IT

**정 용 화**

1984년 한양대학교 전자통신공학과(학사)
1986년 한양대학교 전자통신공학과(석사)
1997년 U. of Southern California(박사)

1986년~2003년 한국전자통신연구원 생체인식기술연구팀(팀장)
2003년~현재 고려대학교 컴퓨터정보학과 교수
관심분야 : 병렬처리, 영상처리, 융합 IT