Textual Explanations for Self-Driving Vehicles Paper Review

Jinkyu Kim et al. ECCV 2018 07

C.E 18011573 이장후 / Sejong Artificial Intelligence : S team

https://arxiv.org/abs/1807.11546

Topics

1. Attention
2. Abstract, Introduction
3. Related Work
4. Explainable Driving Model
5. BDD-X Dataset
6. Results & Evaluations
7. Conclusion

18011573
Janghoo Lee
Department of Computer Engineering

Attention

Attention Model Overview

18011573
Janghoo Lee
Department of Computer Engineering

Key Reference

- http://dmqm.korea.ac.kr/activity/seminar/280
  - Korea University
- https://youtu.be/W2rWgXJBZhU
  - CodeEmporium
- https://youtu.be/Sb3b0ocD8mI?t=2799
  - Stanford university
- https://youtu.be/SysgYptB198
  - Andrew Ng

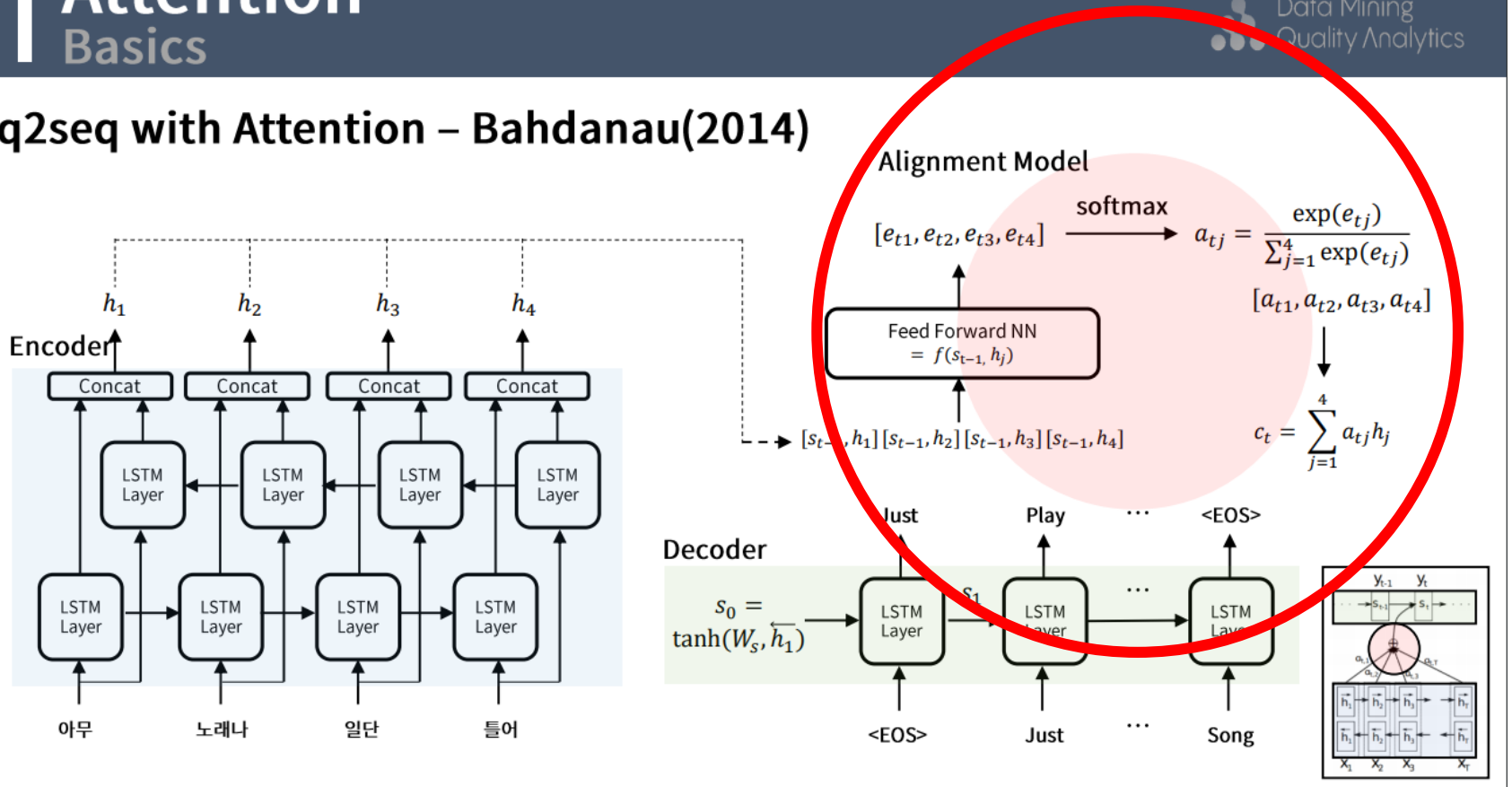
- https://ardino.tistory.com/59
  - https://medium.com/@sunwoopark/show-attend-and-tell-with-pytorch-e45b1600a749

Attention : Machine Translation

02 | Attention Basics

Data Mining Quality Analytics

❖ **Seq2seq with Attention – Bahdanau(2014)**



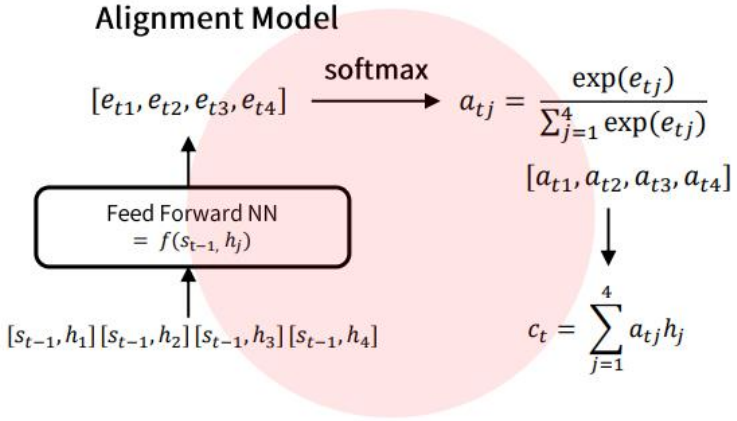
Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

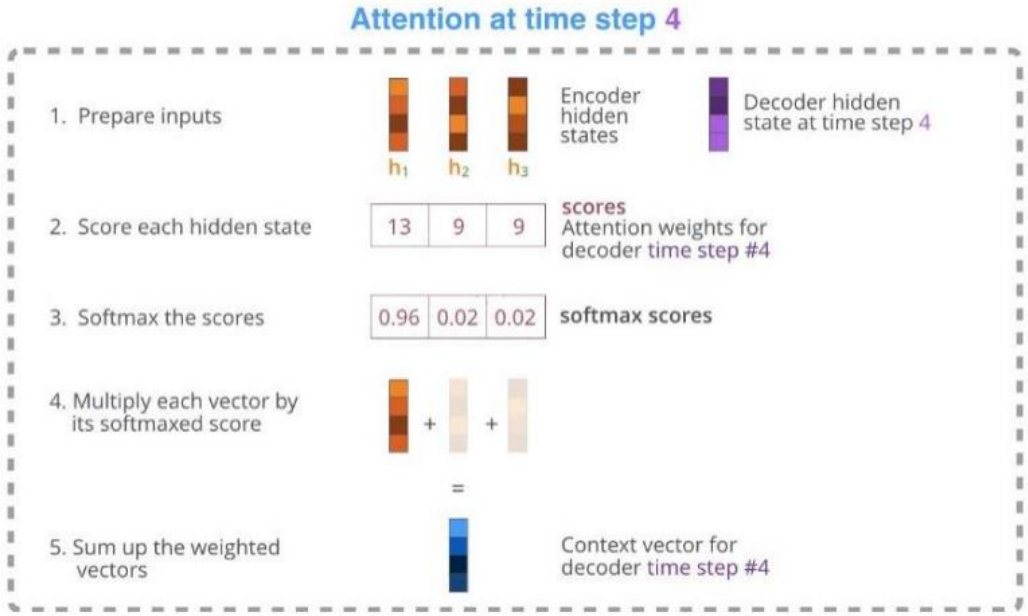
Attention : Machine Translation (2)

02 | Attention Basics

Data Mining Quality Analytics

❖ **Seq2seq with Attention – Bahdanau(2014)**





https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

200214 Open HCAI Seminar – Visual Attention 10/55

Attention : Machine Translation (3)

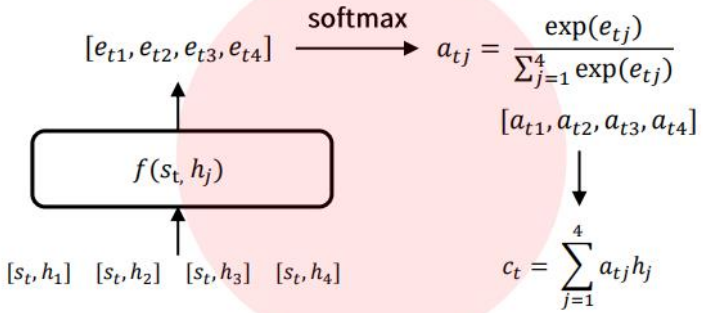
02 | Attention Basics

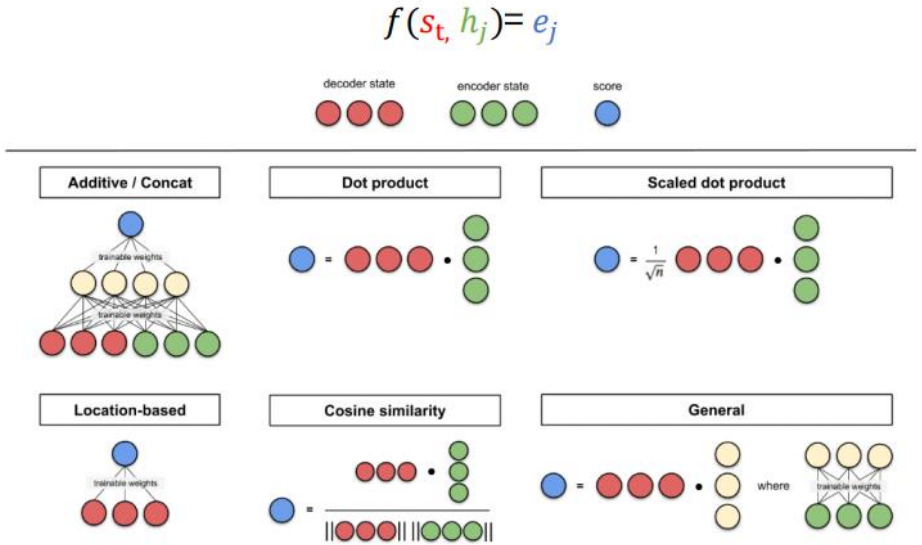
Data Mining Quality Analytics

❖ Seq2seq with Attention – Luong(2015)

- Alignment model의 input으로 s_{t-1}가 아닌 s_t을 사용
- 다양한 similarity function $f(s_t, h_j)$ 제시

Alignment Model



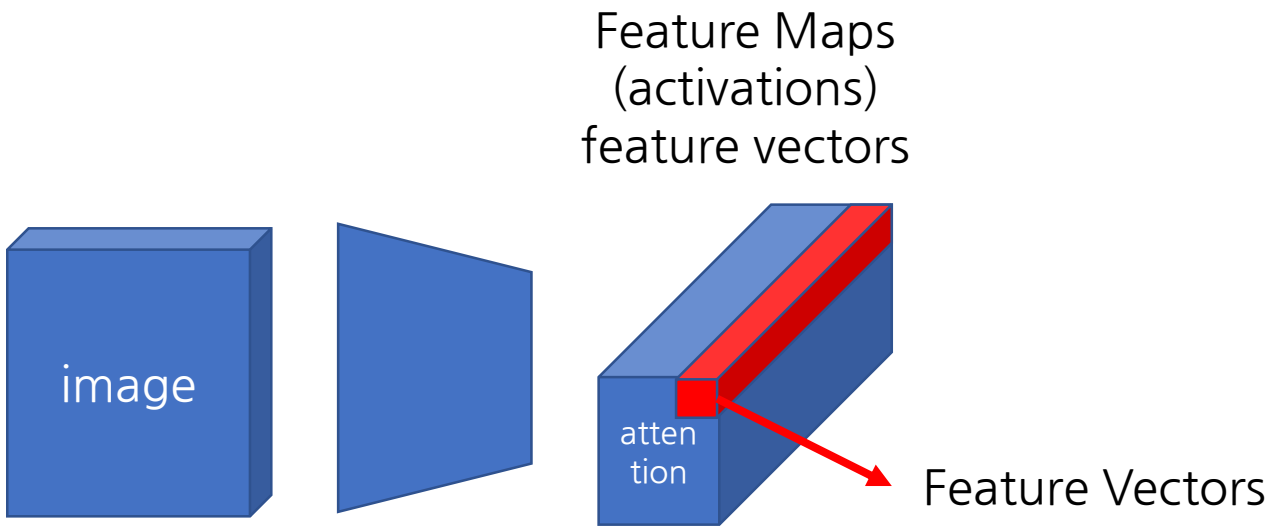


https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025* (2015).

Attention : Computer Vision (Captioning)

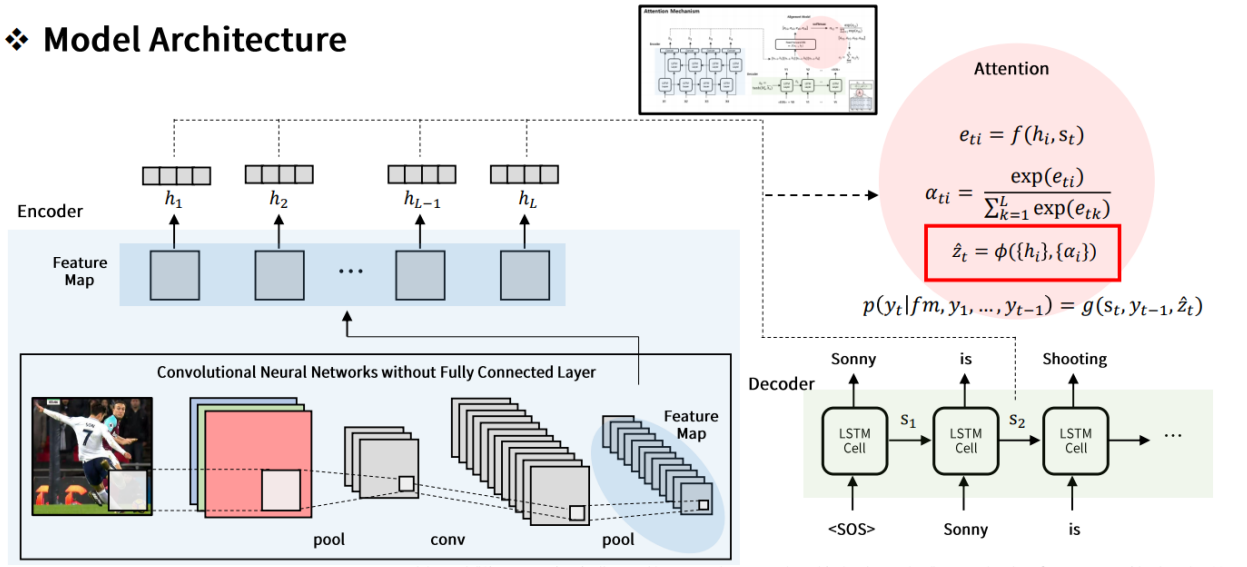
- 아이디어의 태동은 자연어처리 분야의, **Machine Translation**.
  - “Which sequence elements are most important to constructing an accurate output sequence.”
- Vision 쪽에서 attention 의 keynote paper : **“Show, Attend and Tell”**



https://medium.com/@shairozsohail/a-survey-of-visual-attention-mechanisms-in-deep-learning-1043eb25f343

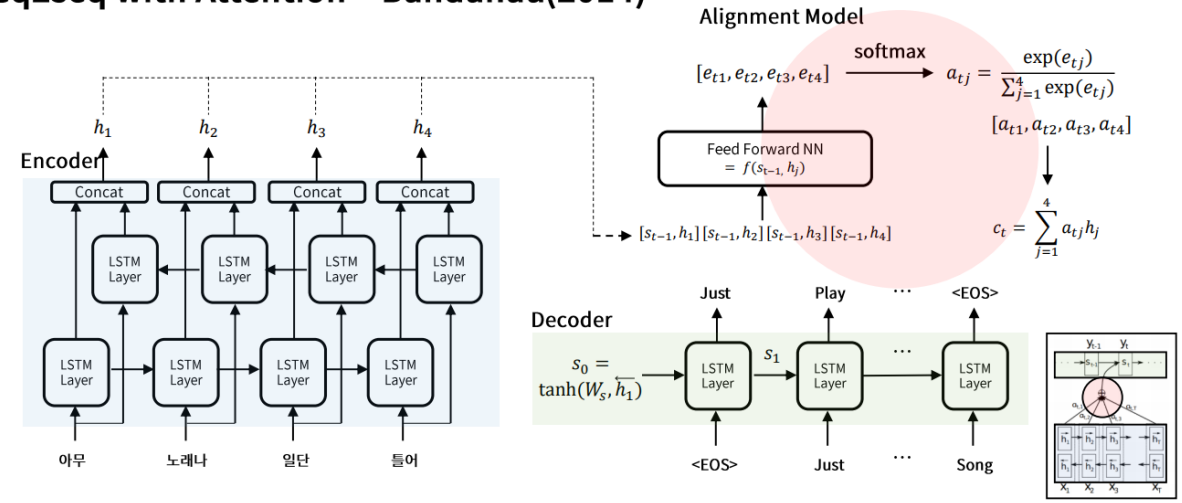
Attention : Computer Vision (Captioning) (1)

❖ **Model Architecture**



Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning*. 2015.

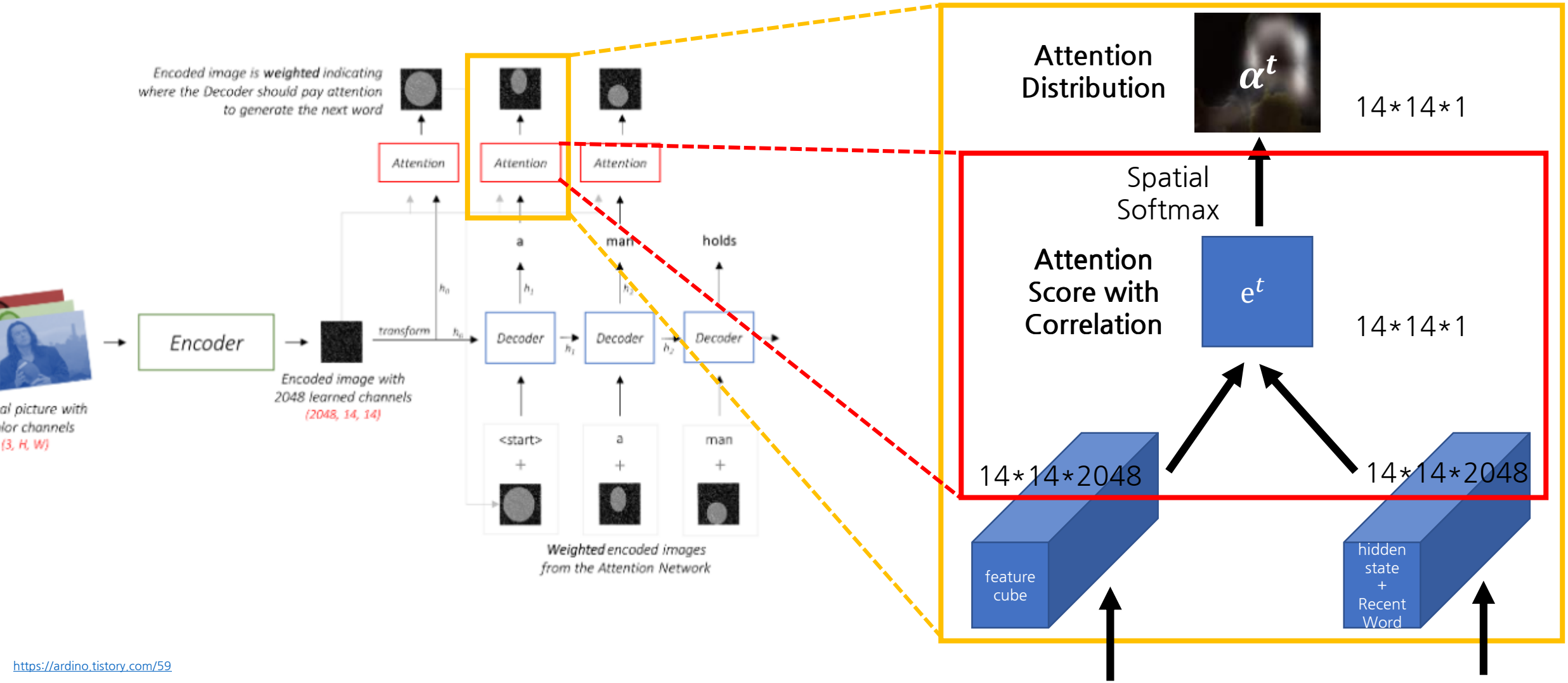
❖ **Seq2seq with Attention – Bahdanau(2014)**



Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

18011573
Janghoo Lee
Department of Computer Engineering

Attention : Computer Vision (Captioning) (2)



https://ardino.tistory.com/59

18011573
Janghoo Lee
Department of Computer Engineering

Attention? Localization?

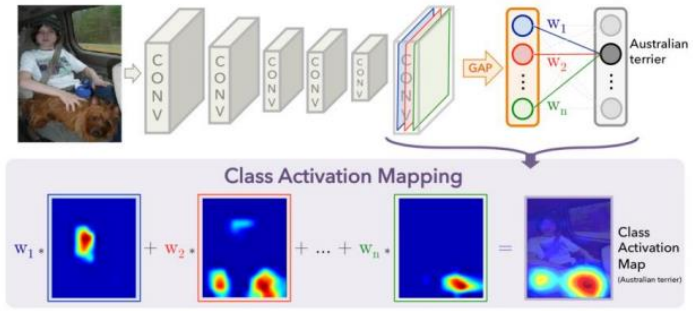
- Attention : Interpretability (Localization) + Performance
  - ref : DMQA Lab (Korea University)

02 | Attention Conclusion

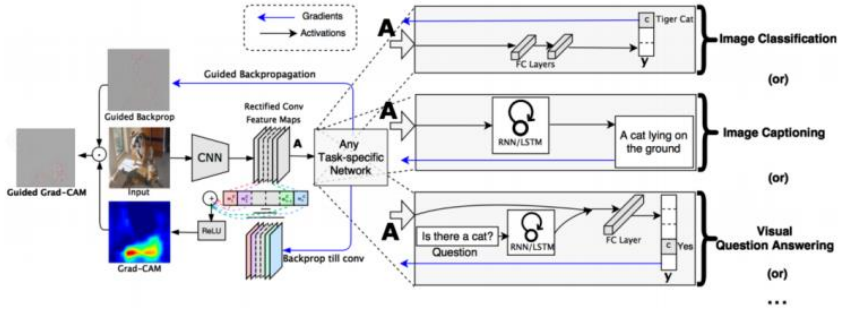
Data Mining Quality Analytics

❖ **Localization vs Attention**

- Localization: Interpretability
- Attention: Interpretability + Model performance



Class Activation Map(CAM)



Grad_CAM

Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

http://dmqm.korea.ac.kr/activity/seminar/280

18011573
Janghoo Lee
Department of Computer Engineering

+ Naming

- 그… 그만하라굿!

arxiv.org › cs ▾ 이 페이지 번역하기

Show and Tell: A Neural Image Caption Generator

2014. 11. 17. - In this **paper**, we present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and ...

O Vinyals 저술 - 2014 - 3713회 인용 - 관련 학술자료

arxiv.org › cs ▾ 이 페이지 번역하기

Show, Attend and Tell: Neural Image Caption Generation with ...

2015. 2. 10. - Title:**Show**, **Attend and Tell**: Neural Image Caption Generation with Visual Attention ... We also show through visualization how the model is able to automatically learn to fix its ... Which authors of this **paper** are endorsers?

K Xu 저술 - 2015 - 5445회 인용 - 관련 학술자료

2014, Image Captioning

2015, Image Captioning

2015, Speech Recognition

2017

2015

arxiv.org › cs ▾ 이 페이지 번역하기

Show, Ask, Attend, and Answer: A Strong Baseline For Visual ...

2017. 4. 11. - Abstract: This **paper** presents a new baseline for visual question **answering** task. Given an image and a question in natural language, our ...

V Kazemi 저술 - 2017 - 85회 인용 - 관련 학술자료

arxiv.org › cs ▾ 이 페이지 번역하기

Listen, Attend and Spell

2015. 8. 20. - Abstract: We present **Listen**, **Attend and Spell** (LAS), a neural network that learns to transcribe ... Which authors of this **paper** are endorsers?

W Chan 저술 - 2015 - 288회 인용 - 관련 학술자료

arxiv.org › cs ▾ 이 페이지 번역하기

Ask, Attend and Answer: Exploring Question-Guided Spatial ...

2015. 11. 17. - Title:**Ask**, **Attend and Answer**: Exploring Question-Guided Spatial Attention for Visual Question **Answering** ... We evaluate our model on two published visual question **answering** ... Which authors of this **paper** are endorsers?

H Xu 저술 - 2015 - 513회 인용 - 관련 학술자료

Abstract, Introduction

18011573
Janghoo Lee
Department of Computer Engineering

Abstract, Introduction

DNN 자율주행차를 이용자가 받아들이기 위해서는 당연히 높은 신뢰가 요구.

- 동작에 대한 설명이 필요.
- 왜 이런 동작을 했는지를 이해할 필요.
- 더 효율적으로 소통할 필요.

그렇다면 어떤 정보를 소통해야 하나?

- rationalization : 악셀을 밟거나 제동을 하거나 등 이러한 뉴럴넷 동작의 결정을 주고 동작에 대한 설명 (post-hoc manner)
- introspective explanation : 악셀을 밟거나 제동을 하거나 등의 결정이 처리되는 그 이유. 그 동작에 대한 운전자 입장에서의 설명

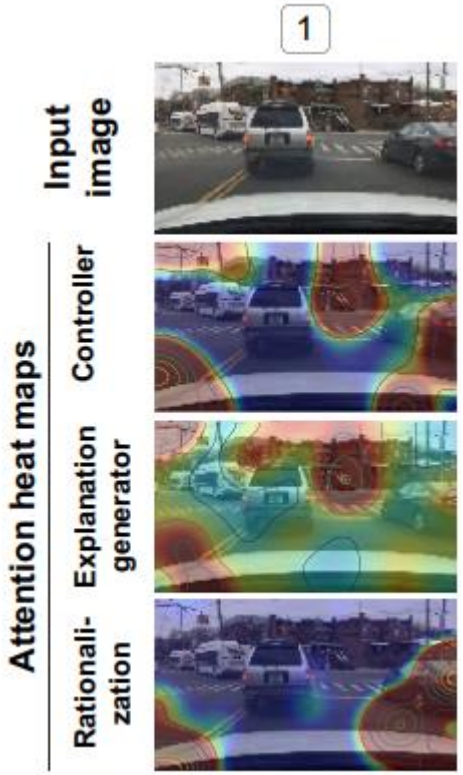
Abstract, Introduction

- 기존에는 Activation map, Visual Attention 을 통해서 자율주행 분야에서 설명가능성을 얻고자 했음.
- 그런데 그러한 방법들은, 뉴럴넷의 동작과 바로 연결이 안 됨. 또한 이용자가 그걸 일일이 생각하면서 다시 돌려보아야 함. 예를 들어, “신호등이 빨간색이므로 차의 브레이크를 밟습니다.” 에 비해 정면을 촬영한 이미지에 신호등이 Activation 된 것은 번거로울 수 있음.
- 따라서 이용자에게 generating textual description 을 통해 rationalization 과 introspective explanation 을 제공하는 것에 집중함.

Abstract, Introduction

3 Contributions

- **Contribution 1**. textual description 을 통한 rationalization 과 introspective explanation 을 제공하는 방법 제안
- **Contribution 2**. 다른 정보를 주지 않고 vehicle controller 의 동작 결과와 이미지만을 잘 연결해서 textual explanation 을 생성해낸 것 (rationalization) 과, vehicle controller 의 attention context 와 이미지를 연결해서 textual explanation 을 생성해낸 것 (introspective explanation) 을 비교하고, 어떠한 결과보다 어떤 결과를 도출하는 네트워크의 상태 (state) 를 활용해서 학습하는 것이 더 좋다는 것을 밝힘.
- **Contribution 3**. 77시간어치 Berkeley Deep Drive 데이터셋에 eXplanation (BDD-X) Dataset 을 추가함. (Human-Annotated)



Human: The car steadily driving + now that the cars are moving.

Ours (WAA): The car is driving forward + because traffic is moving freely.

Ours (SAA): The car heads down the road + because traffic is moving at a steady pace.

Rationalization: The car slows down + because it's getting ready to a stop sign.

Textual Explanations for Self-Driving Vehicles

사람이 평가한 결과입니다.

Type	Model	Control inputs	λ_a	λ_c	Correctness rate	
					Explanations	Descriptions
Rationalization	Ours (no constraints)	Y	0	0	64.0%	92.8%
Introspective explanation	Ours (with SAA)	Y	-	100	62.4%	90.8%
	Ours (with WAA)	Y	10	100	**66.0%**	**93.5%**

Related Work

1. End to End Learning for AV
2. Visual and Textual Explanations

18011573
Janghoo Lee
Department of Computer Engineering

Related Work - End to End Learning for AV

자율주행차 controller 을 만드는 방법은 크게 두 가지로 나뉨.

이 논문은 end-to-end 방법론. (최근 이쪽 연구는 대부분 end-to-end 로 진행)

- 여러가지를 인지한 결과를 조립해서 만드는 방법 (mediated)
  - 신호등 인식 모델, 보행자 인식 모델, 차선 인식 모델을 모두 활용해서 컨트롤러 생성
- 그냥 사진이나 영상을 주고 통째로 학습시켜버리는 방법 (end-to-end)
  - input : 최근 몇 타임동안의 영상, 조향, 악셀 … etc, output : 조향, 악셀 … etc

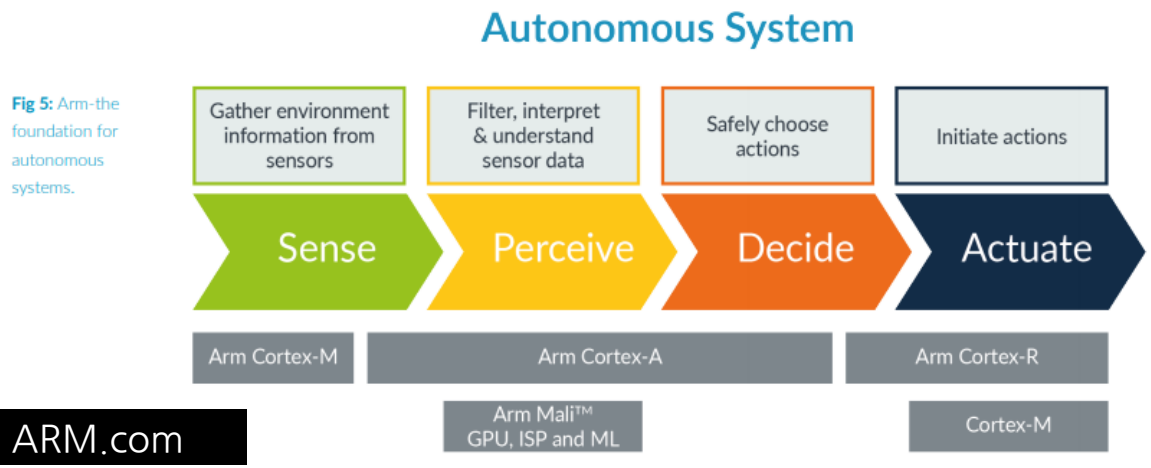
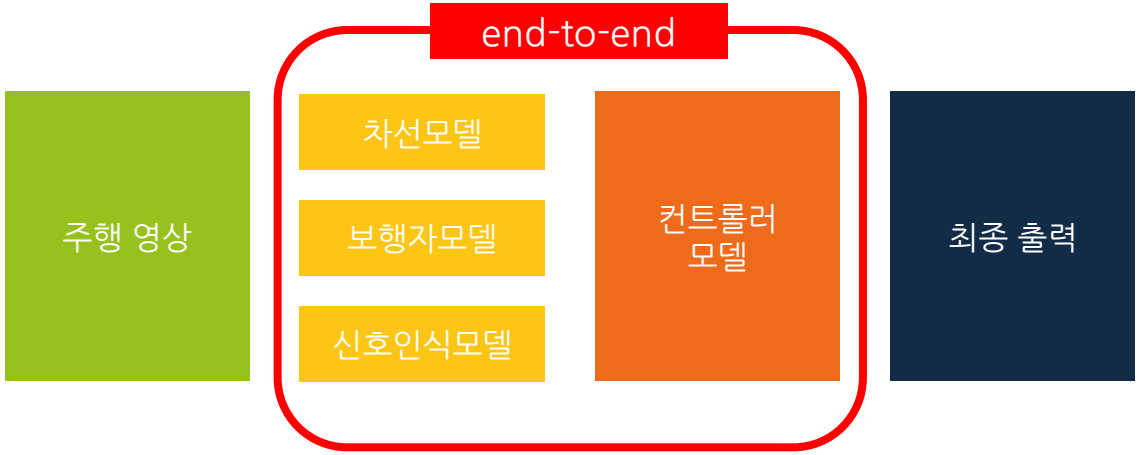


Fig 5: Arm-the foundation for autonomous systems.

ARM.com



18011573
Janghoo Lee
Department of Computer Engineering

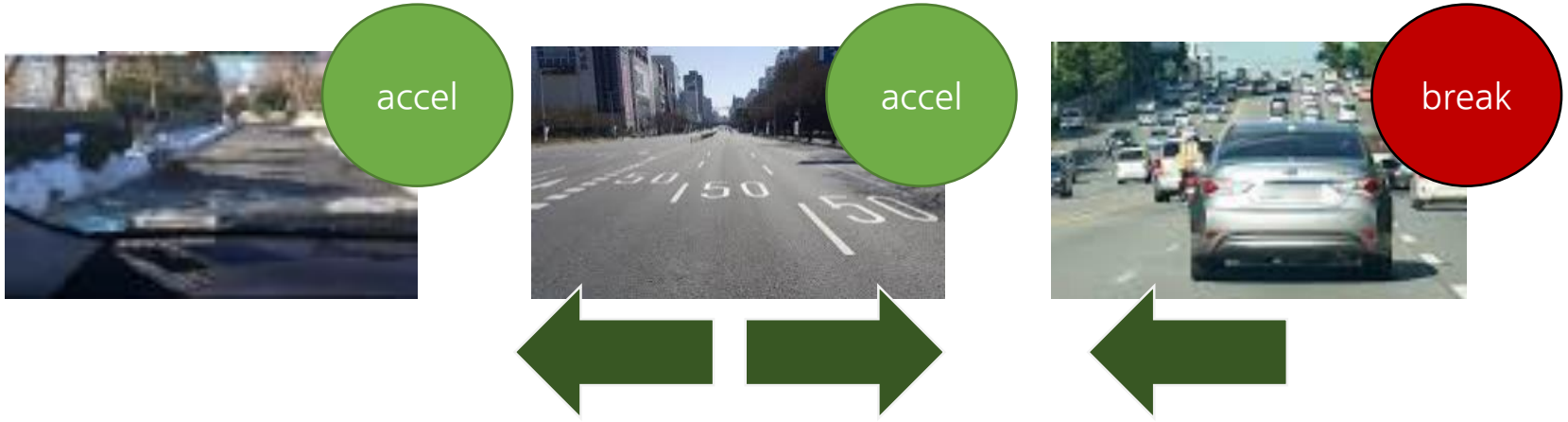
Related Work - End to End Learning for AV

기존의 end to end 시도들, (그리고 한계 지적 : 설명력 부족)

- (1) : dashcam image -> control
- (2) : 이전차량상태 + raw image pixel -> control

end to end 방법론 자율주행데이터셋 만드는 방법

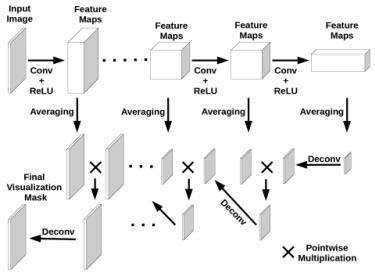
- Behavior cloning : 사람 실주행으로부터 조향, 엑셀, 브레이크 로그와 영상 페어 데이터를 만들고, 지도학습.



Related Work - Visual and Textual Explanations

기존의 rationalization 과 introspective explanation 을 제공하는 방법들

- (1) : CNN Deconvolution
- (2) : Image Captioning, CNN Justification with LSTM
- (3) : Deconvolution Based Pixelwise Activation Map
  - VisualBackProp: visualizing CNNs for autonomous driving, 2016
- **(4) : Attention Based Explanations with Causal filtering**
  - 의의 : 동작에 영향을 미치는 부분을 콕 집어서 표시하는 방법을 제안함.
  - Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention 2017
    - https://openaccess.thecvf.com/content_ICCV_2017/papers/Kim_Interpretable_Learning_for_ICCV_2017_paper.pdf



이러한 설명들을 가장 빨리 평가(justify)

하는 방법은 textual justification 이라고 주장.

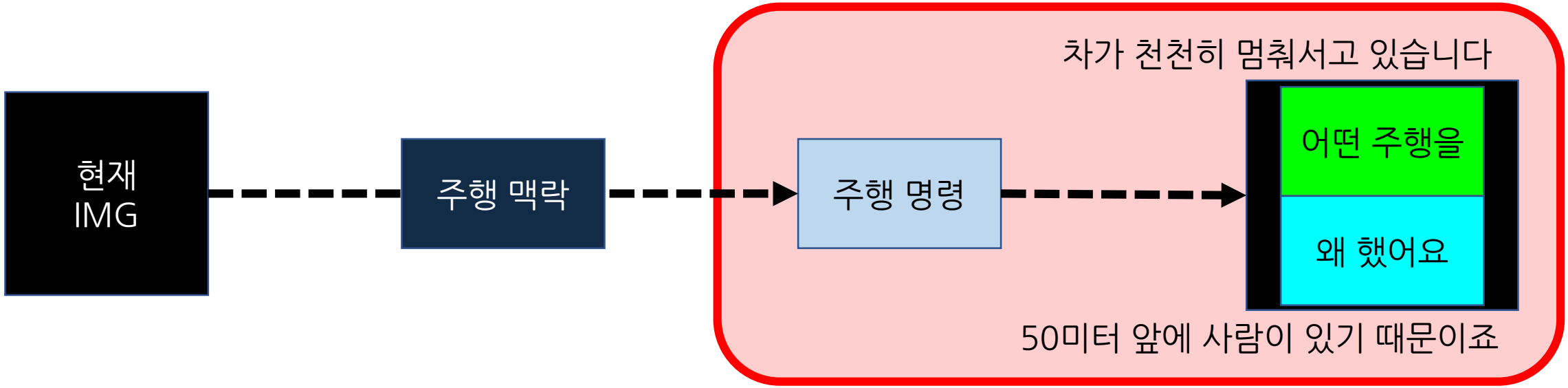
http://www.columbia.edu/~aec2163/NonFlash/Papers/VisualBackProp.pdf

Explainable Driving Model

18011573
Janghoo Lee
Department of Computer Engineering

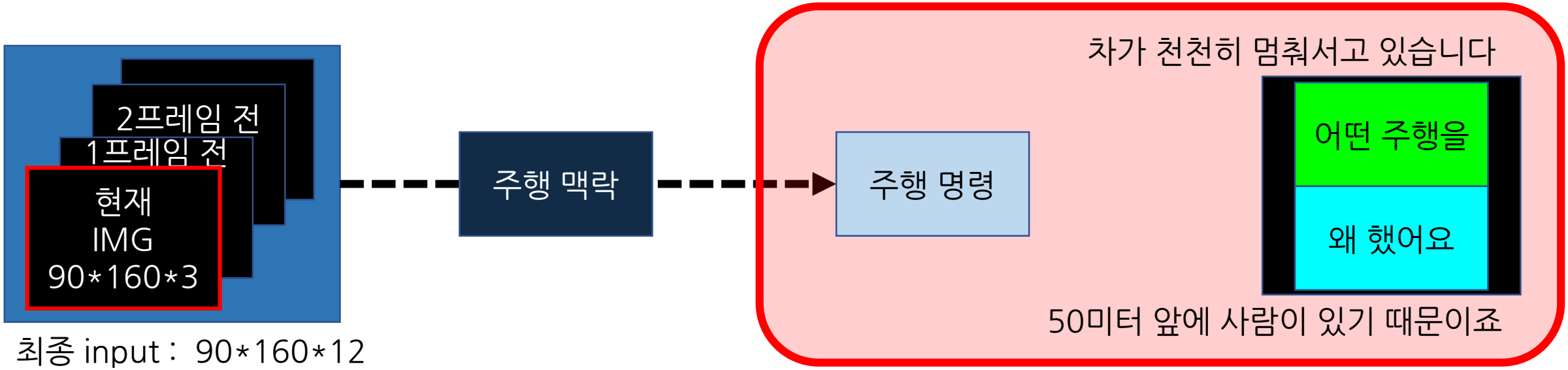
Purposed Model

- 이미지를 바탕으로, 아래와 같은 output 을 내는 end-to-end 모델을 구상
- Vehicle Controller
- Vehicle Controller description
- justification (explanation)



Purposed Model

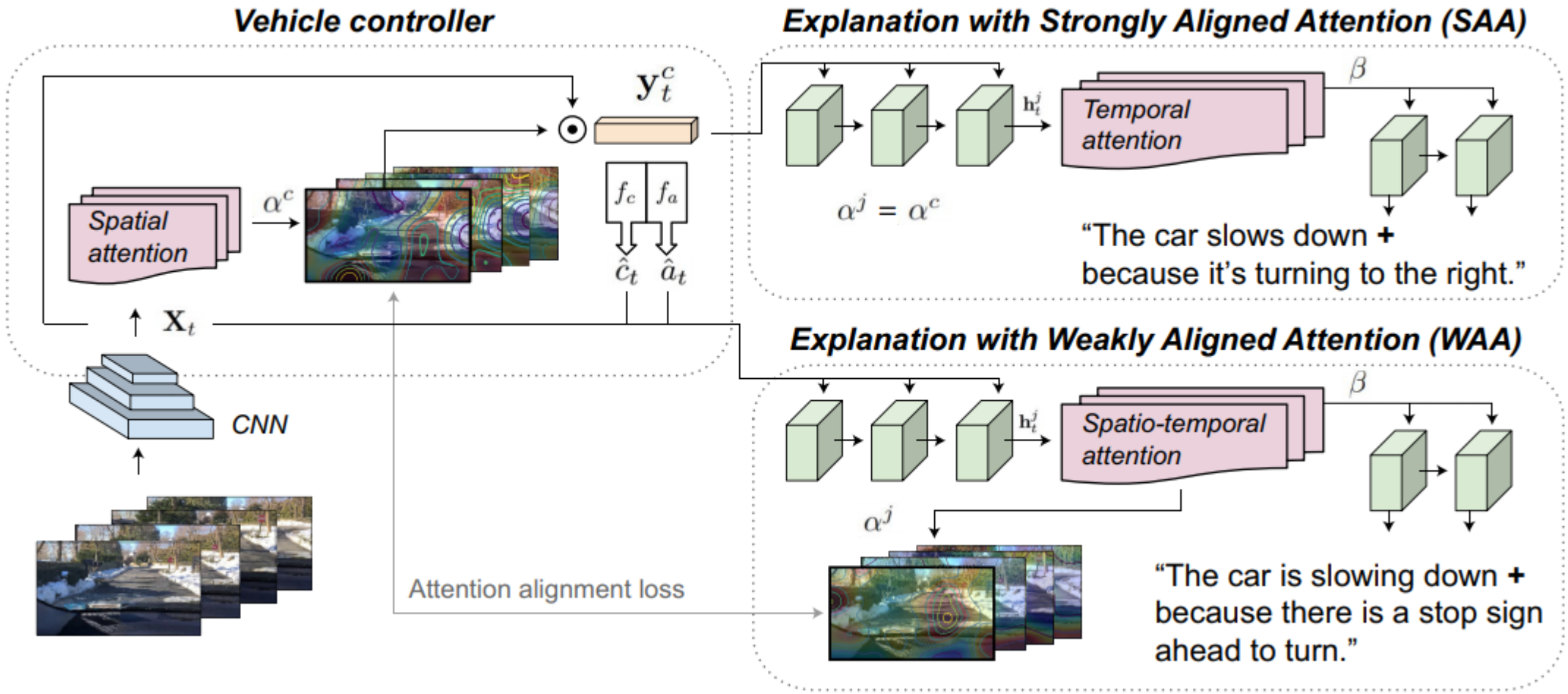
- input 으로 IMG 를 넣어줄 때, 최근 4 frame 의 input 들을 쌓아서 넣어줌.
  - resize : nearest neighbor 로 downsampling
  - normalize : resize 된 각 이미지 frame 마다 (x-m)/std 정규화



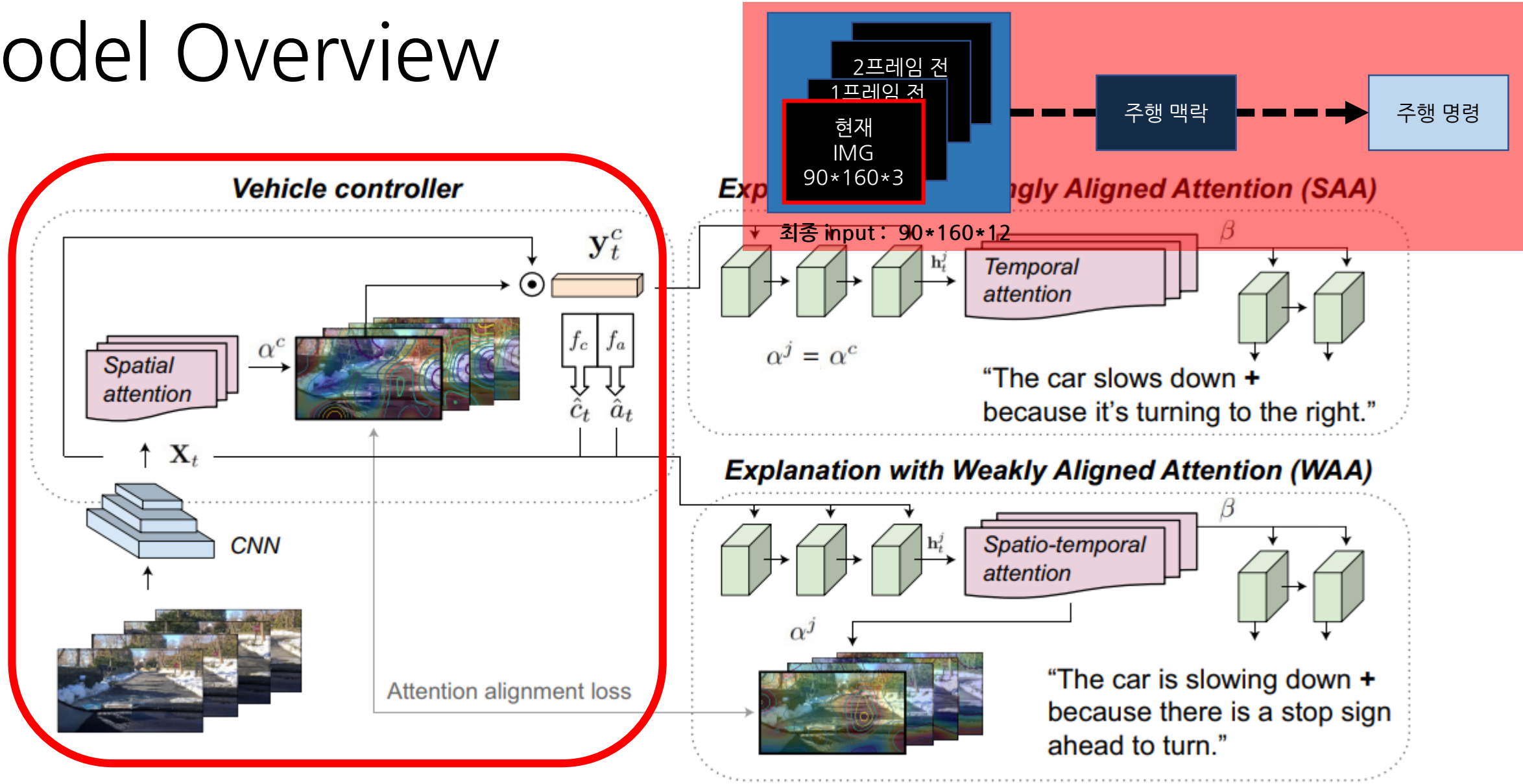
최종 input : 90*160*12

18011573
Janghoo Lee
Department of Computer Engineering

Model Overview



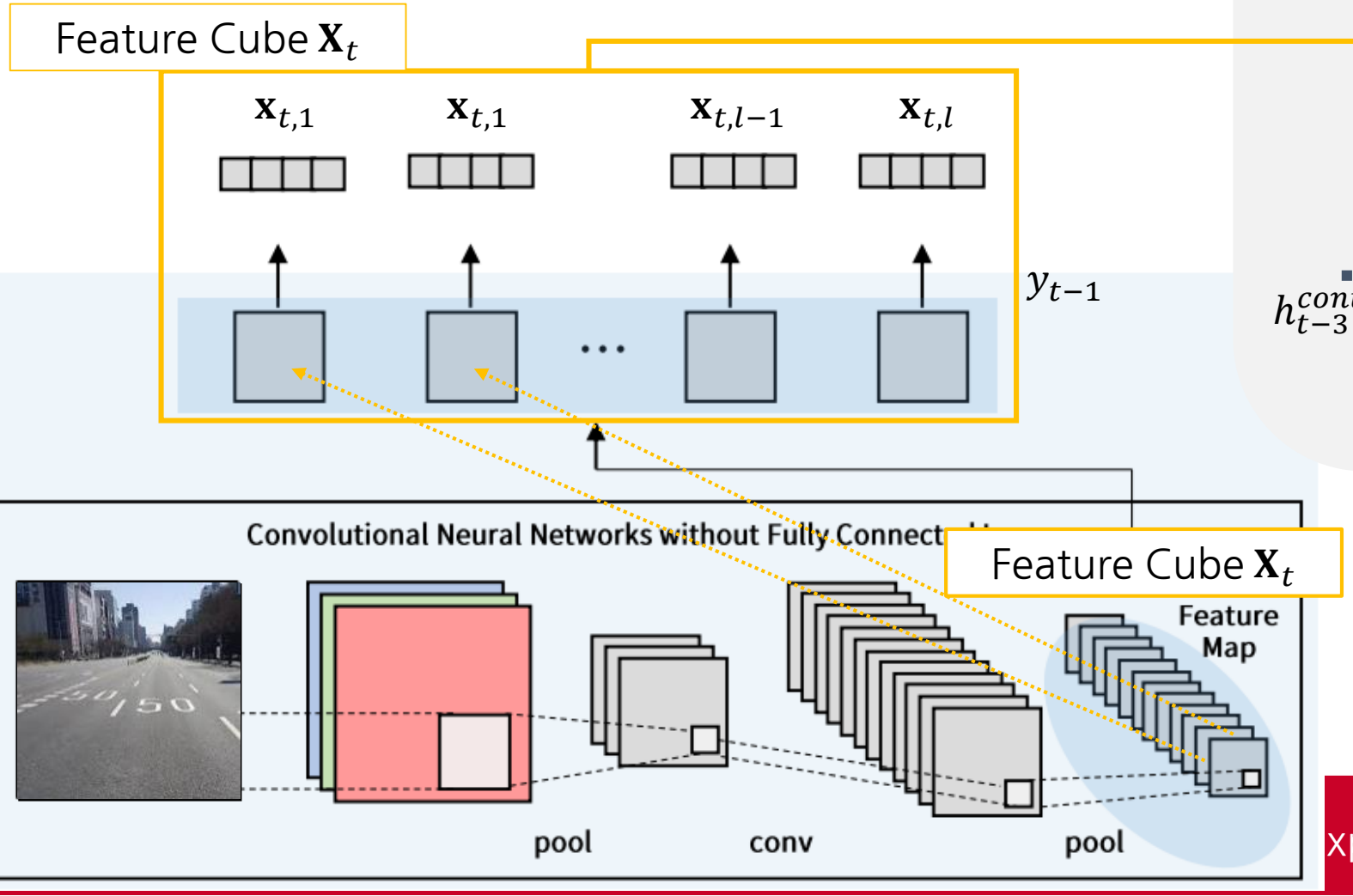
Model Overview

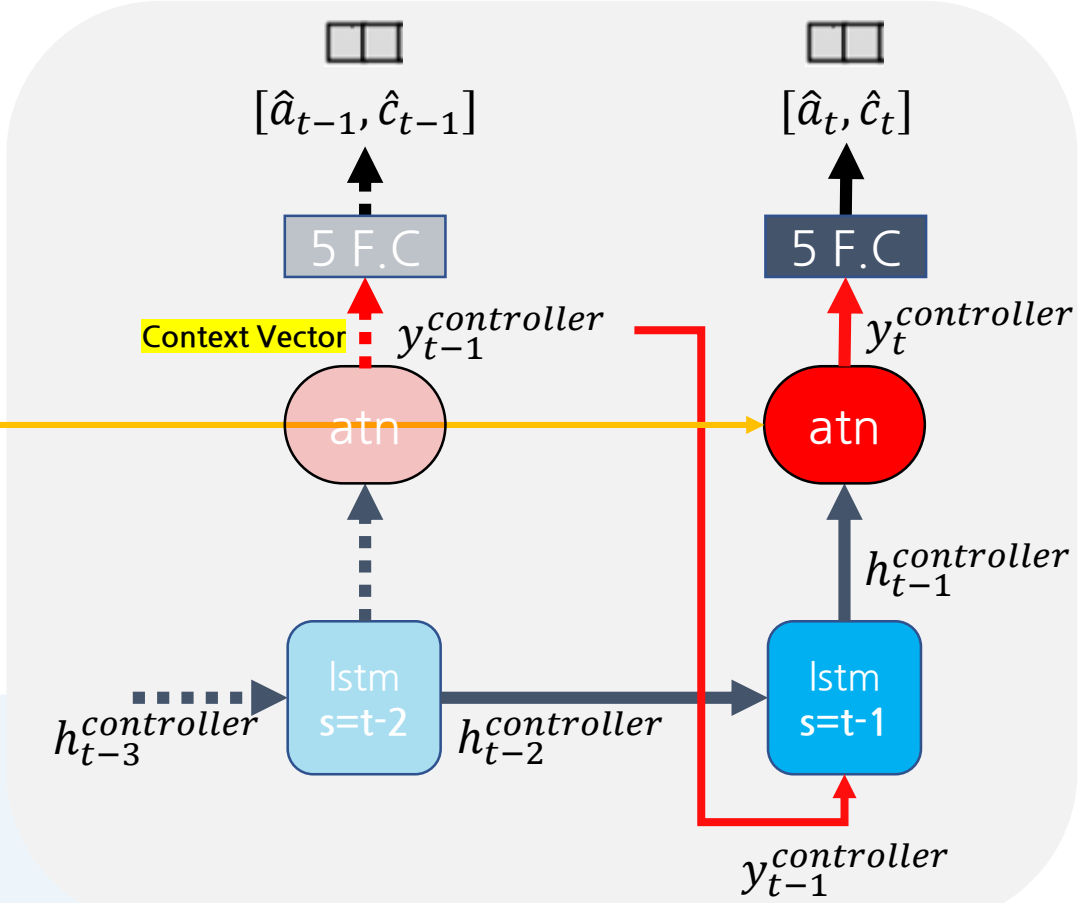


18011573
Janghoo Lee
Department of Computer Engineering

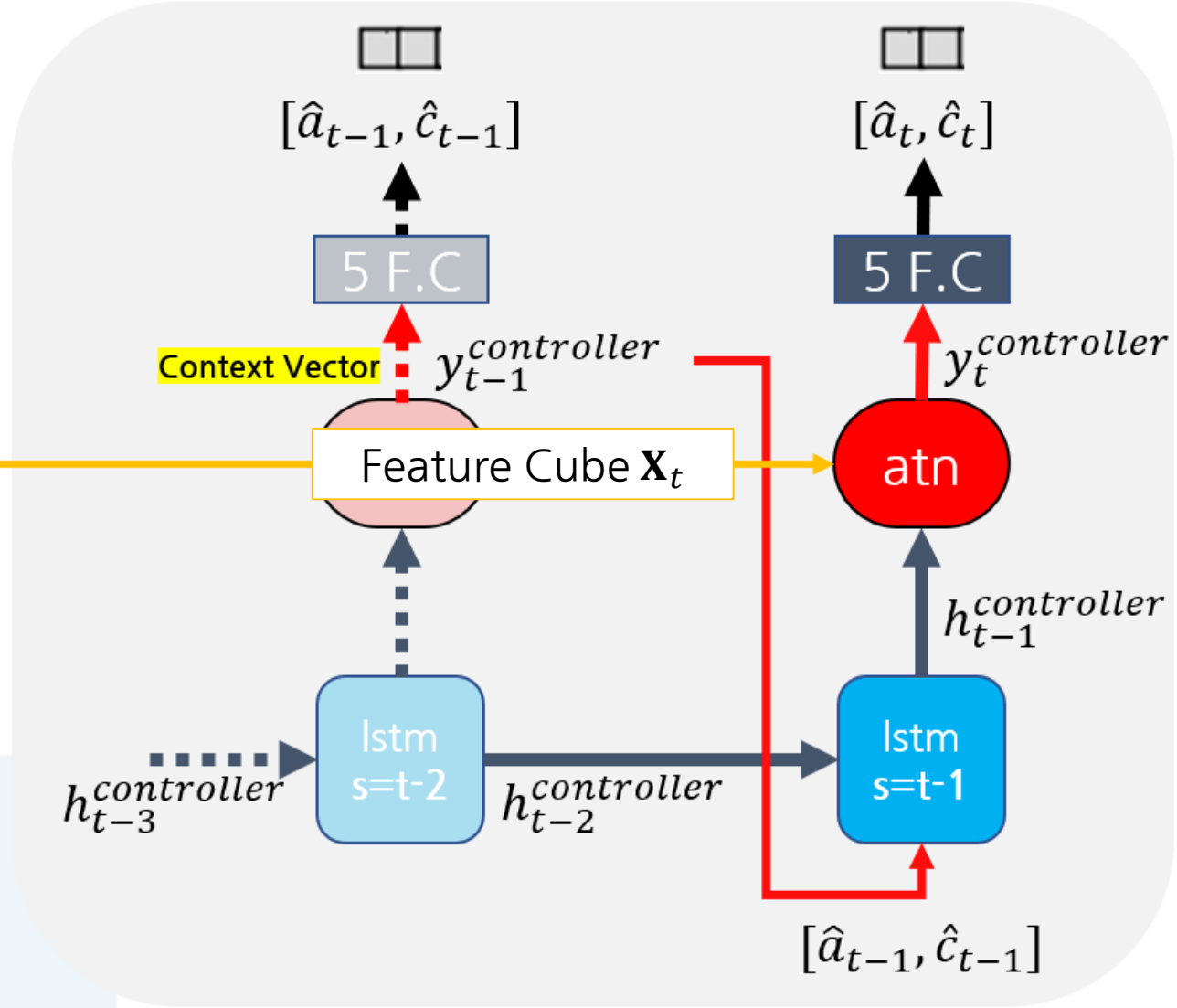
Model - Controller

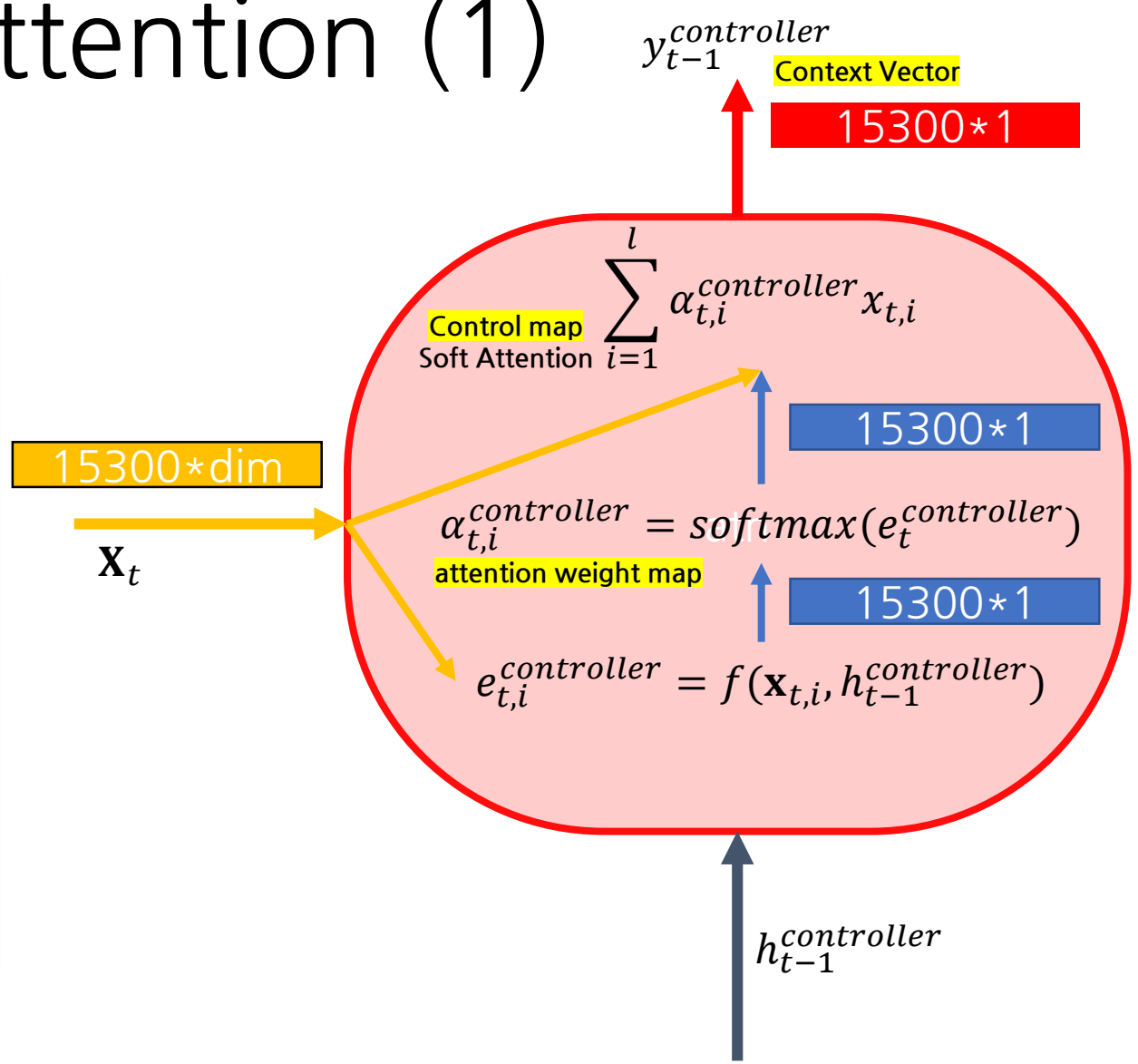
- $\hat{a}_{t-1}$: 이번엔 엑셀(=브레이크) 를 얼마나 밟을까?
- $\hat{c}_{t-1}$: 이번엔 차선을 어떻게 바꿀까?





Model - Controller - Attention (1)

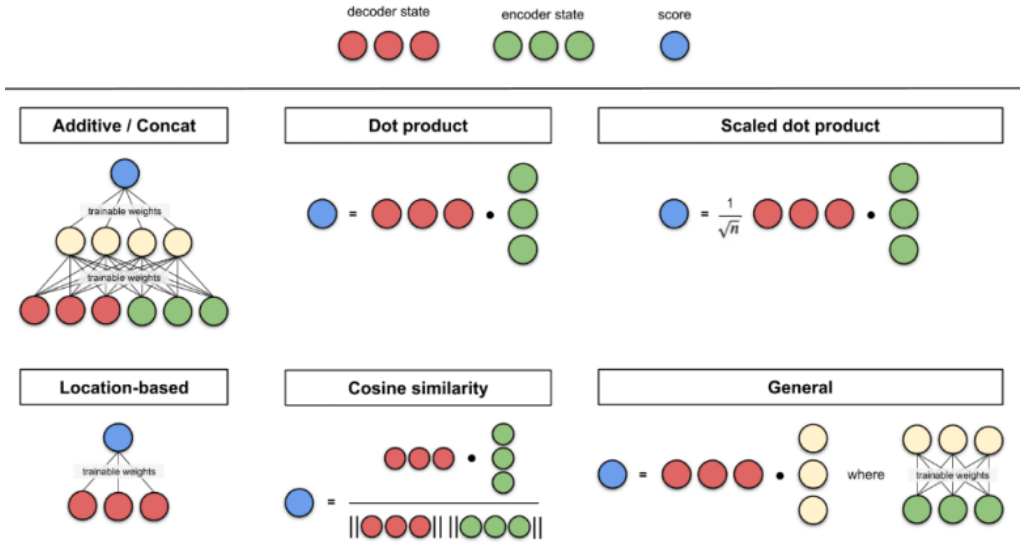




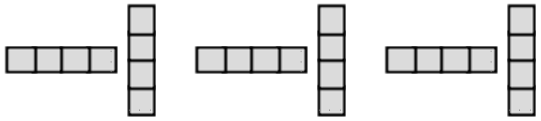
18011573
Janghoo Lee
Department of Computer Engineering

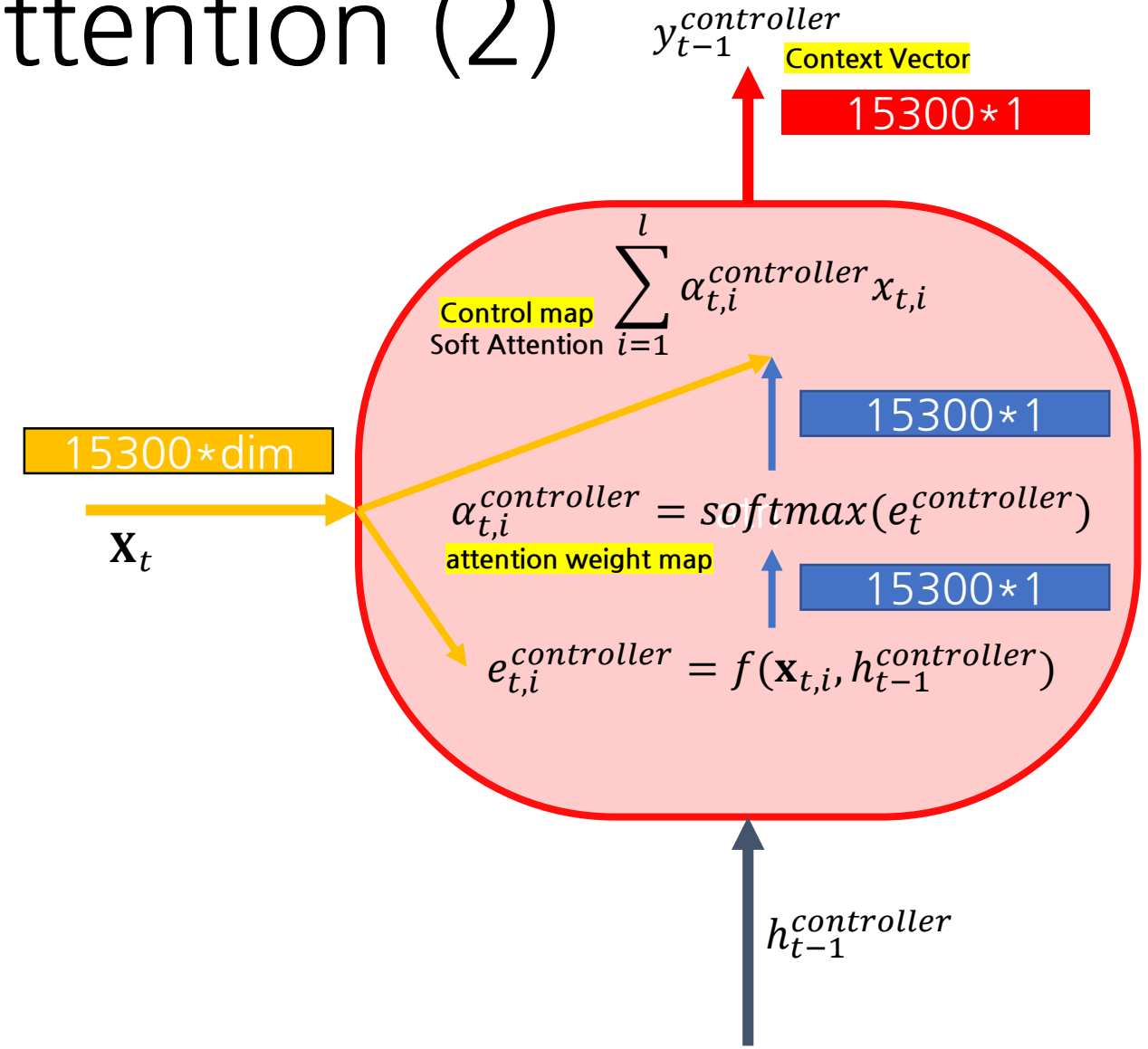
Model - Controller - Attention (2)

$f(s_t, h_j) = e_j$



$$e_{t,i}^{controller} = f(\mathbf{x}_{t,i}, h_{t-1}^{controller}) = Corr(\mathbf{x}_{t,i}, h_{t-1}^{controller})$$



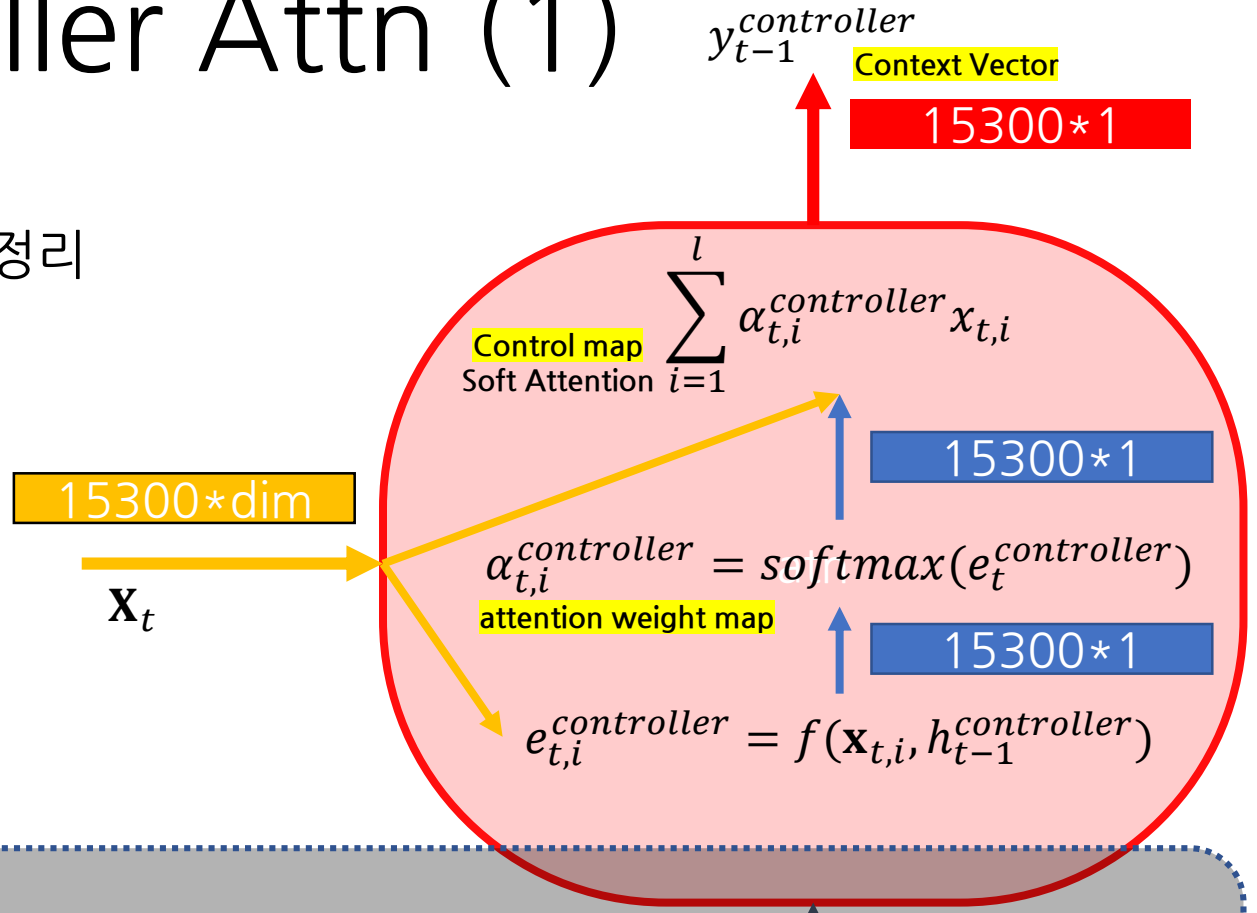


18011573
Janghoo Lee
Department of Computer Engineering

Understanding Controller Attn (1)

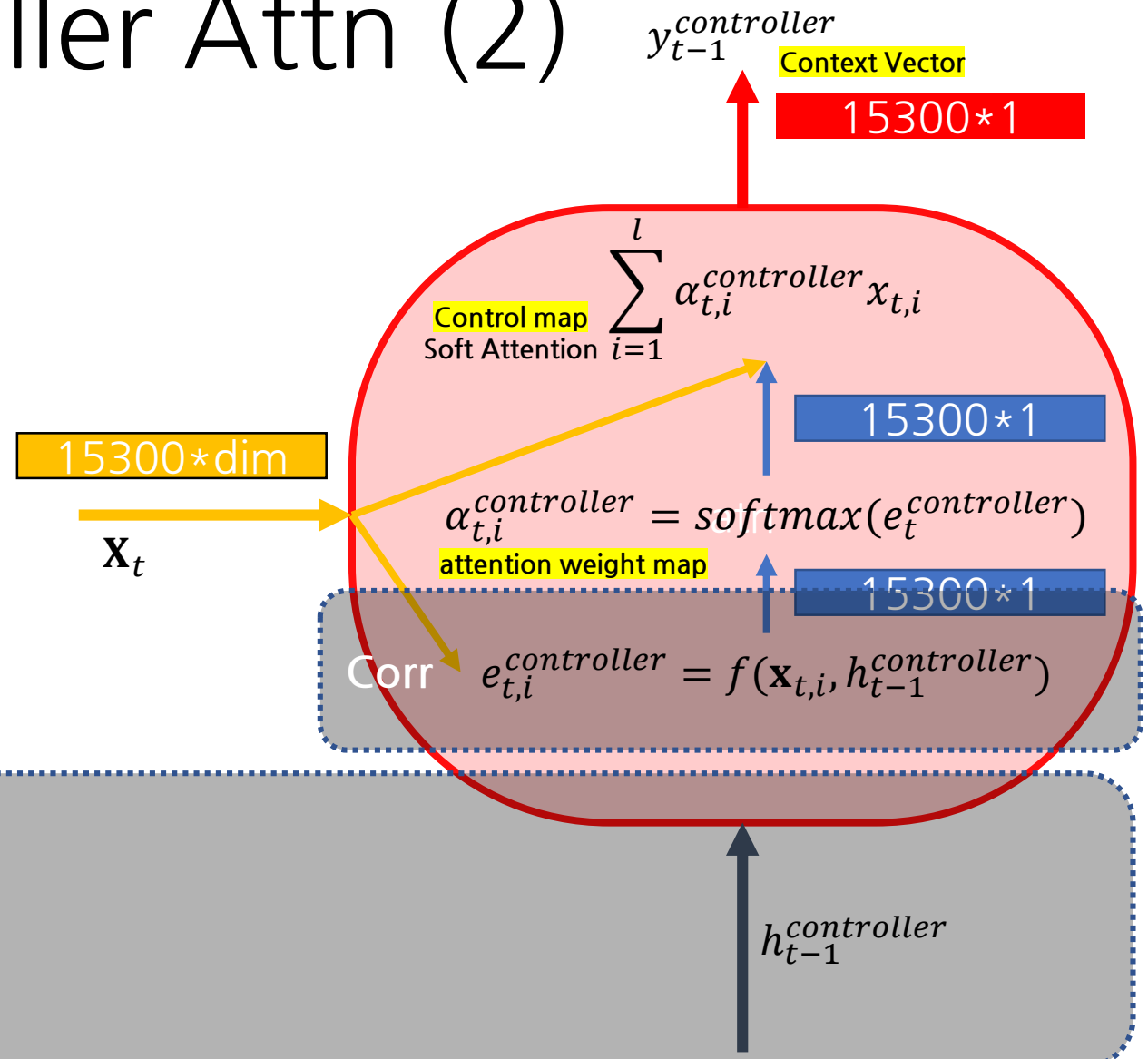
나같이 Spatial Attention 을 처음 접해서

직관적 이해가 안 되는 사람을 돕기 위해 한 번만 다시 정리



이전 attention 의 context 와
이전 상태에 대한 정보를 받음.

Understanding Controller Attn (2)

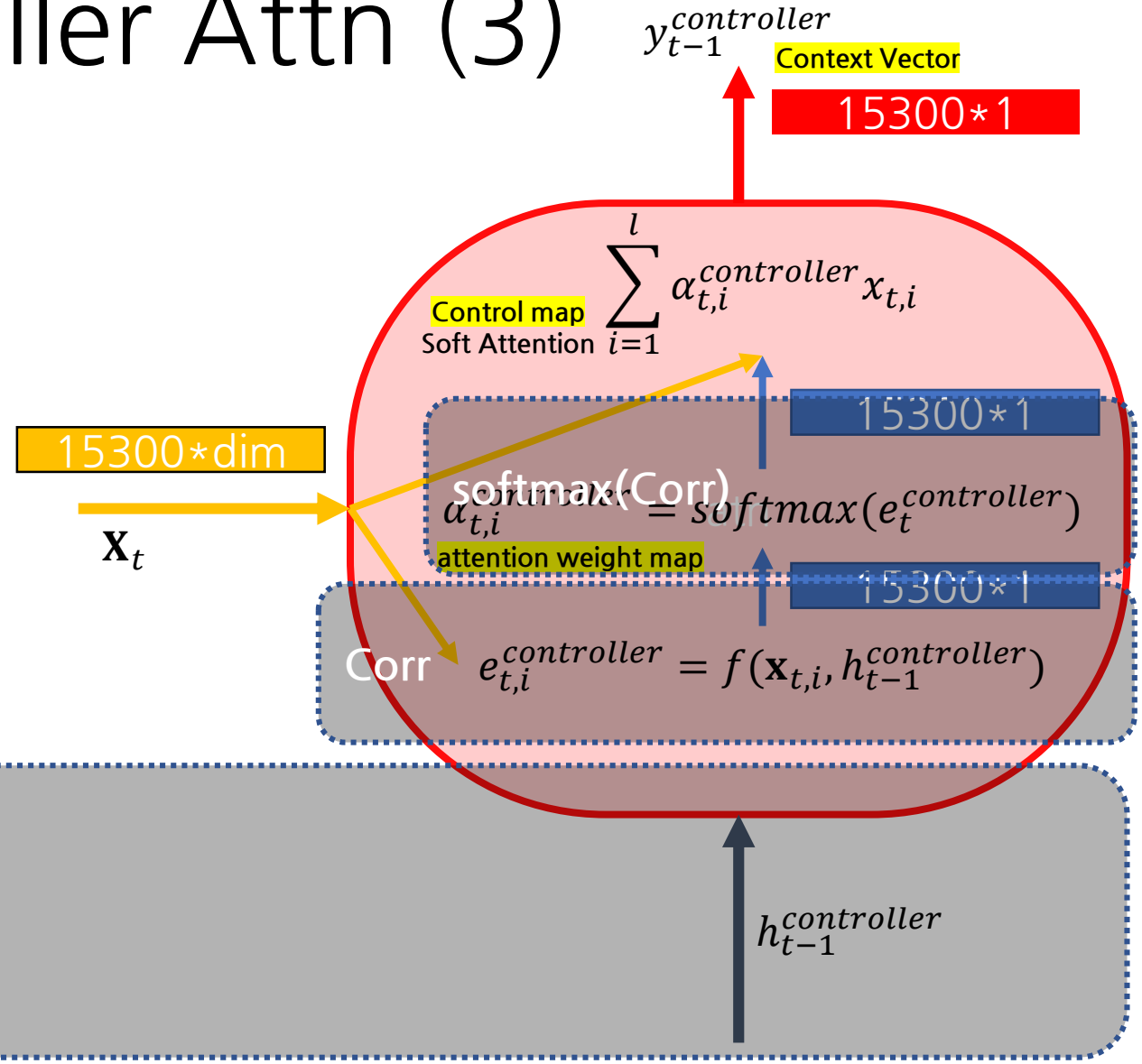


방금 받은 정보와, 4개 frame 으로부터 조합한 image $\mathbf{x}_t$ 의 correlation 을 구함.

이전 attention 의 context 와
이전 상태에 대한 정보를 받음.

18011573
Janghoo Lee
Department of Computer Engineering

Understanding Controller Attn (3)



Attention Weight Map : 각 Correlation 값을 Softmax 에 넣음. 즉 상관 관계 값을, 이미지상에서 현재 집중할 위치와의 상관관계 확률 값으로 바꾸어 줌.

방금 받은 정보와, 4개 frame 으로부터 조합한 image $\mathbf{x}_t$ 의 correlation 을 구함.

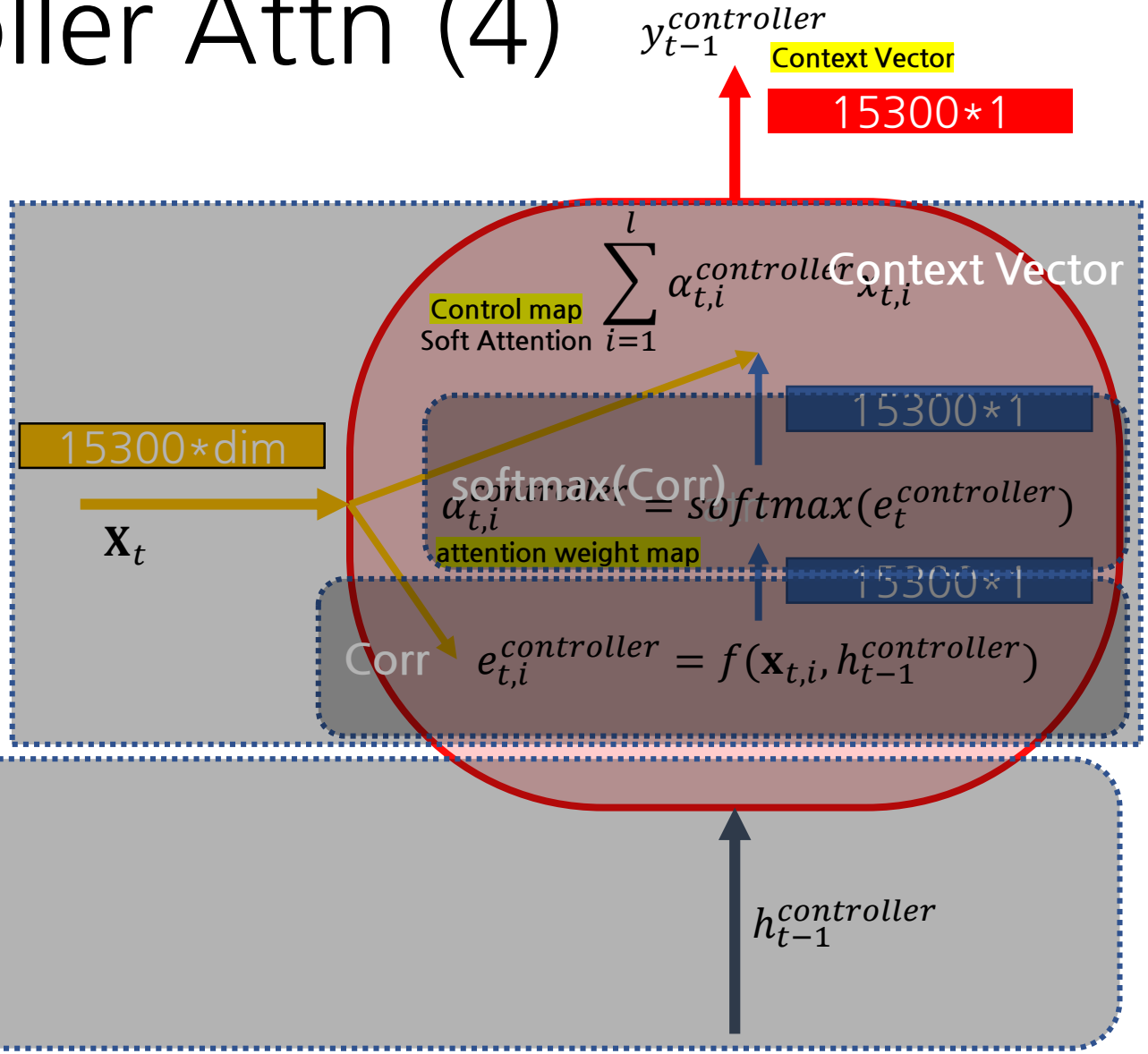
이전 attention 의 context 와
이전 상태에 대한 정보를 받음.

Understanding Controller Attn (4)

Context Vector : 방금 Attention Weight Map 을 실제 특징을 담은 공간적 cube $\mathbf{X}_t$ 에 가중치처럼 곱해서 공간적 특징에 중요도를 매겨서 다시 만든 피처값

Attention Weight Map : 각 Correlation 값을 Softmax 에 넣음. 즉 상관 관계 값을, 이미지상에서 현재 집중할 위치와의 상관관계 확률 값으로 바꾸어 줌.

방금 받은 정보와, 4개 frame 으로부터 조합한 image $\mathbf{X}_t$ 의 correlation 을 구함.



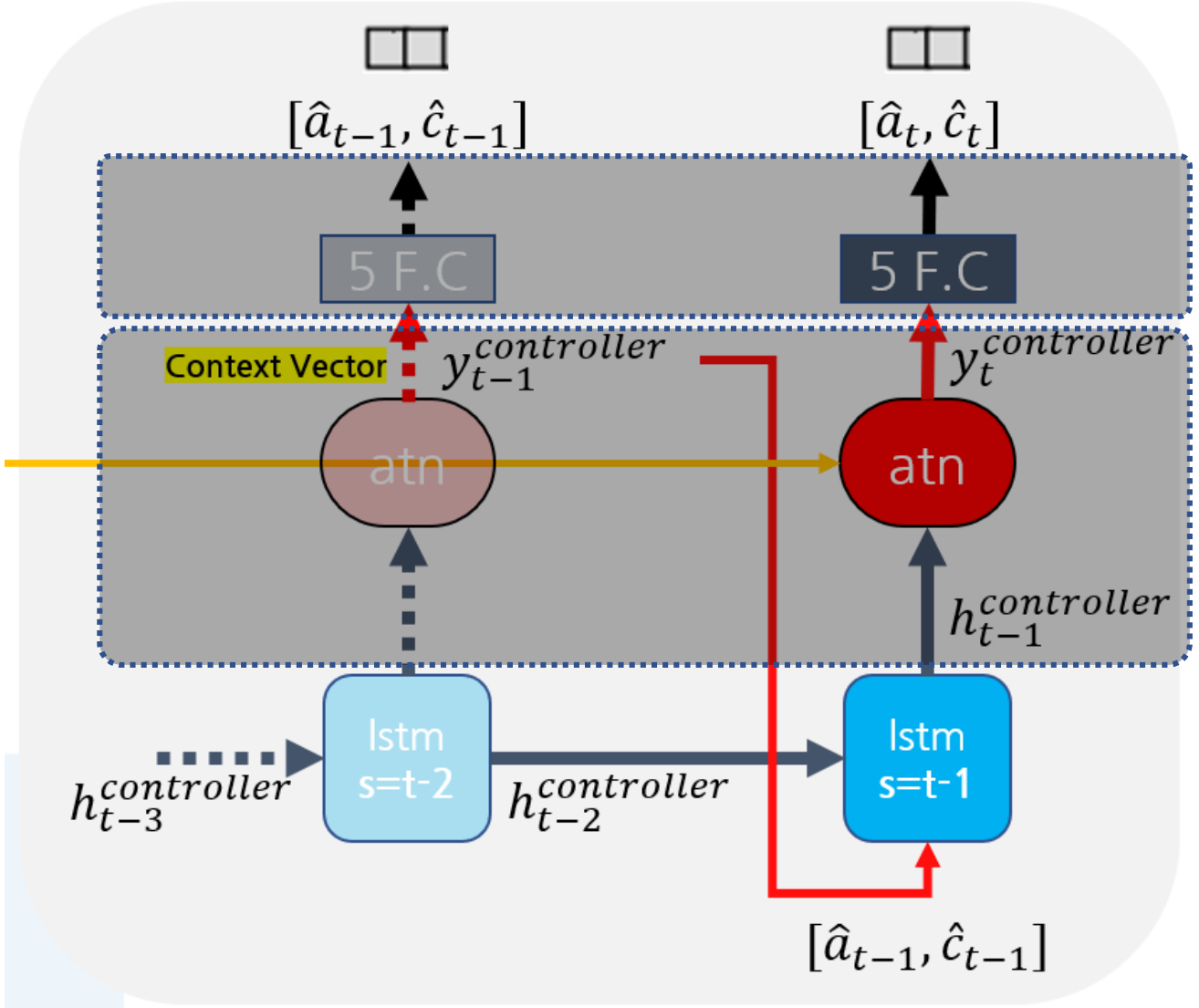
이전 attention 의 context 와
이전 상태에 대한 정보를 받음.

18011573
Janghoo Lee
Department of Computer Engineering

Understanding Controller Attn (5)

이를 FC Layer (논문에서는 5 개의 FC Layer) 을 거치며 Context Vector 가 가속/감속, 차선 전환 값으로 변환

Context Vector : 방금 Attention Weight Map 을 실제 특징을 담은 공간적 cube $\mathbf{x}_t$ 에 가중치처럼 곱해서 공간적 특징에 중요도를 매겨서 다시 만든 피처값



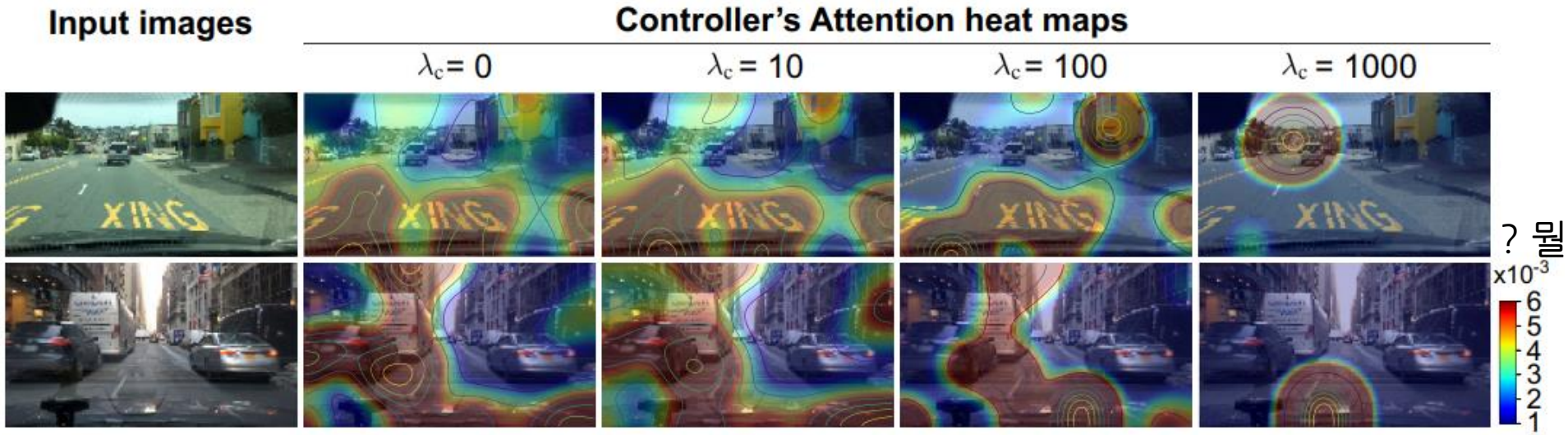
18011573
Janghoo Lee
Department of Computer Engineering

Controller Loss

가속/감속 차선전환 정보량이 많으면 패널티 (entropy)

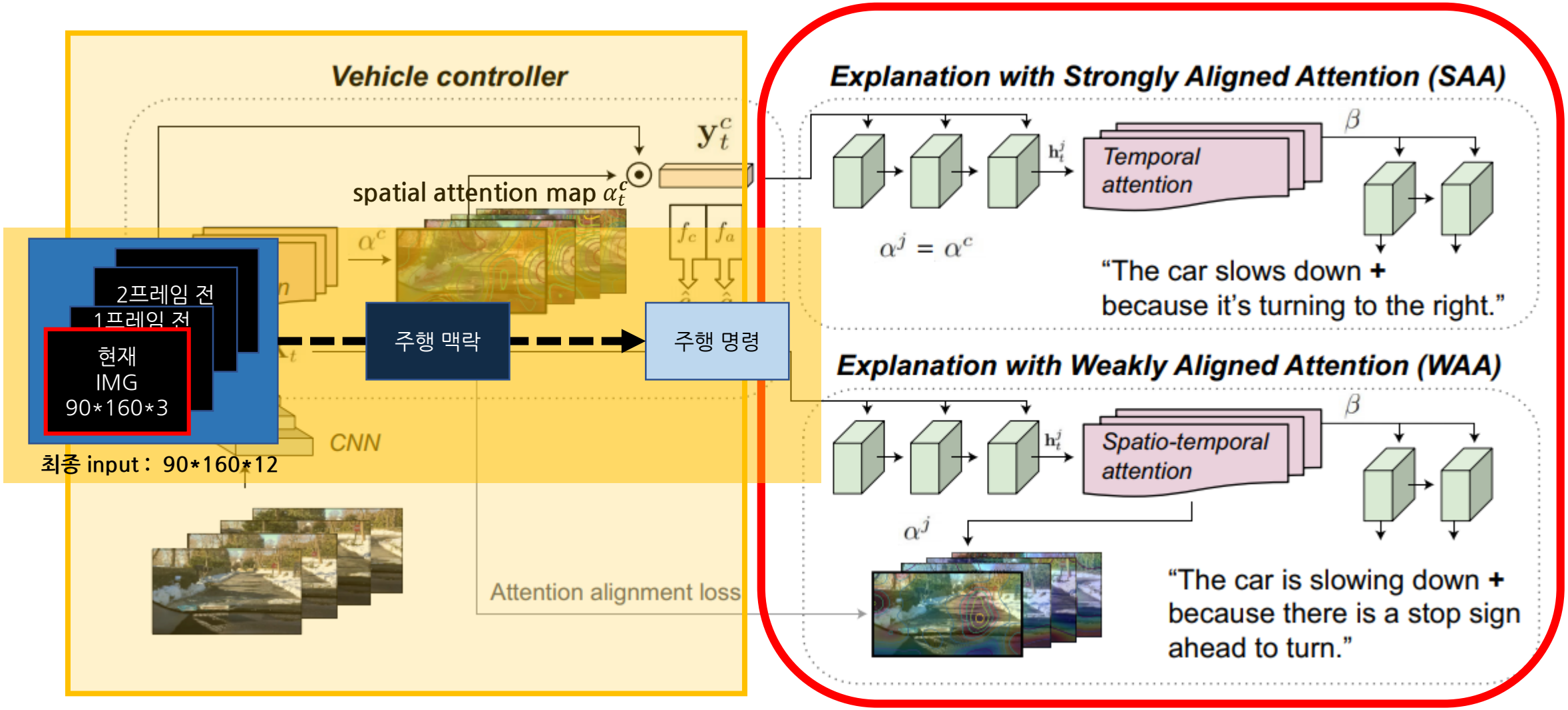
$$\mathcal{L}_c = \sum_t \left((a_t - \hat{a}_t)^2 + (c_t - \hat{c}_t)^2 + \lambda_c H(\alpha_t^c) \right) \quad (2)$$

The entropy is computed on the attention map as though it were a probability distribution. Minimizing loss corresponds to minimizing entropy. Low entropy attention maps are sparse and emphasize relatively few regions. We use a hyperparameter λ_c to control the strength of the entropy regularization term.



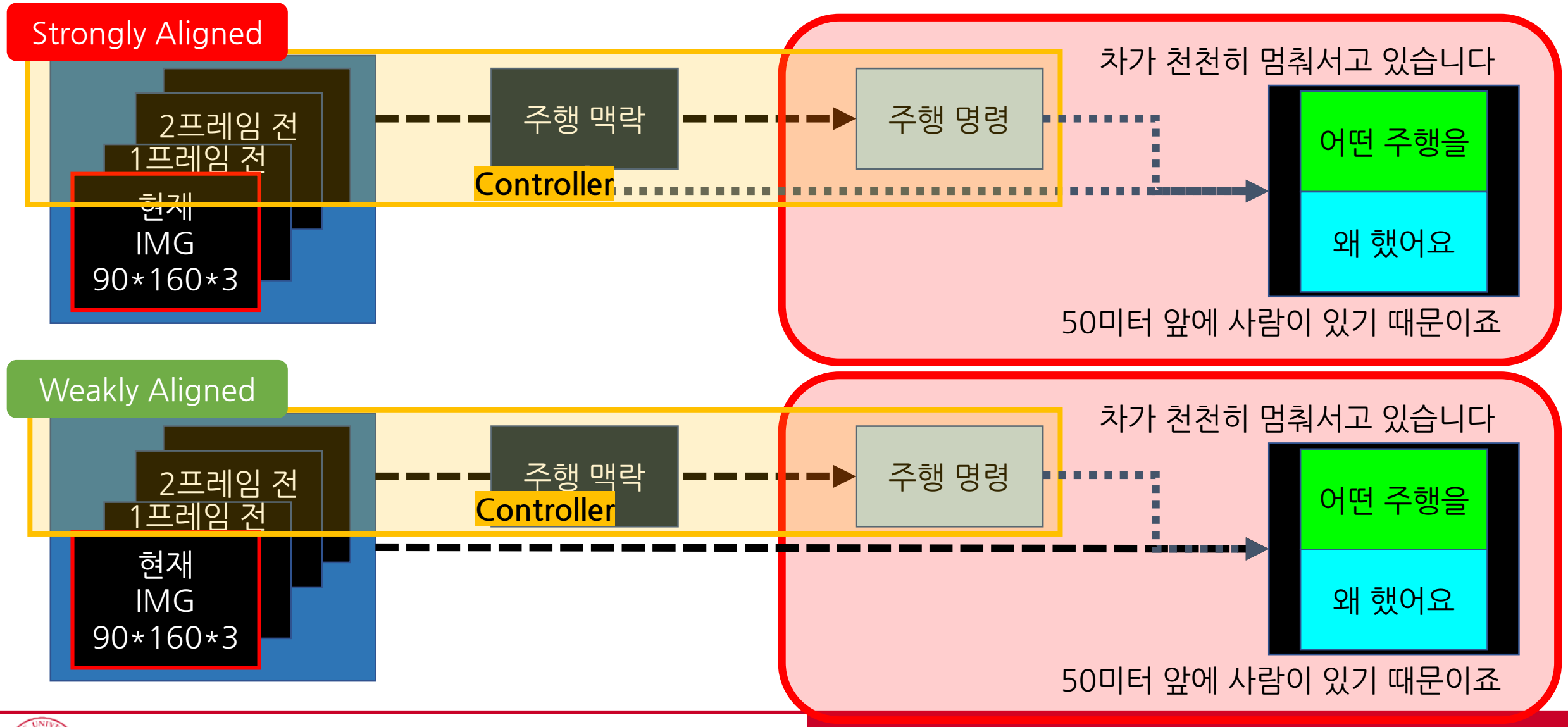
? 뭘 보고있누?

Model Overview



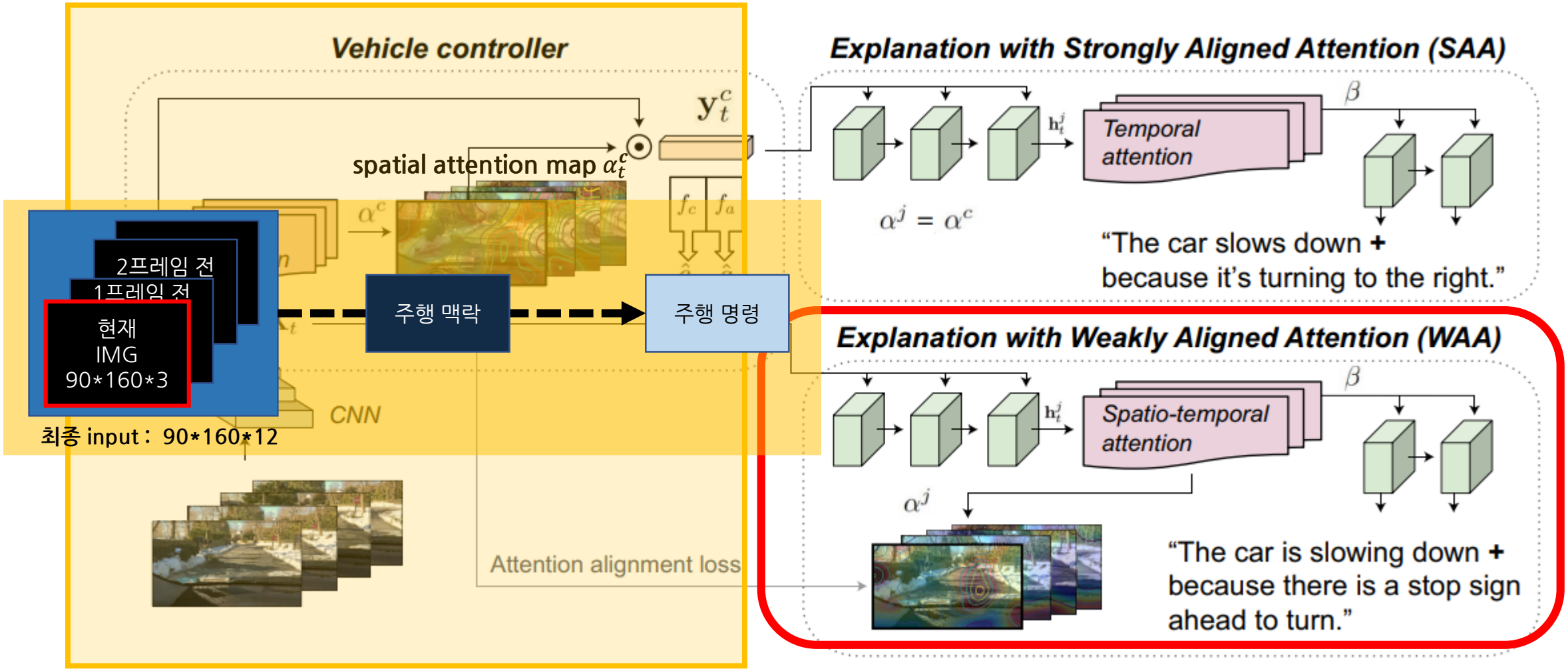
18011573
Janghoo Lee
Department of Computer Engineering

Attention Alignment



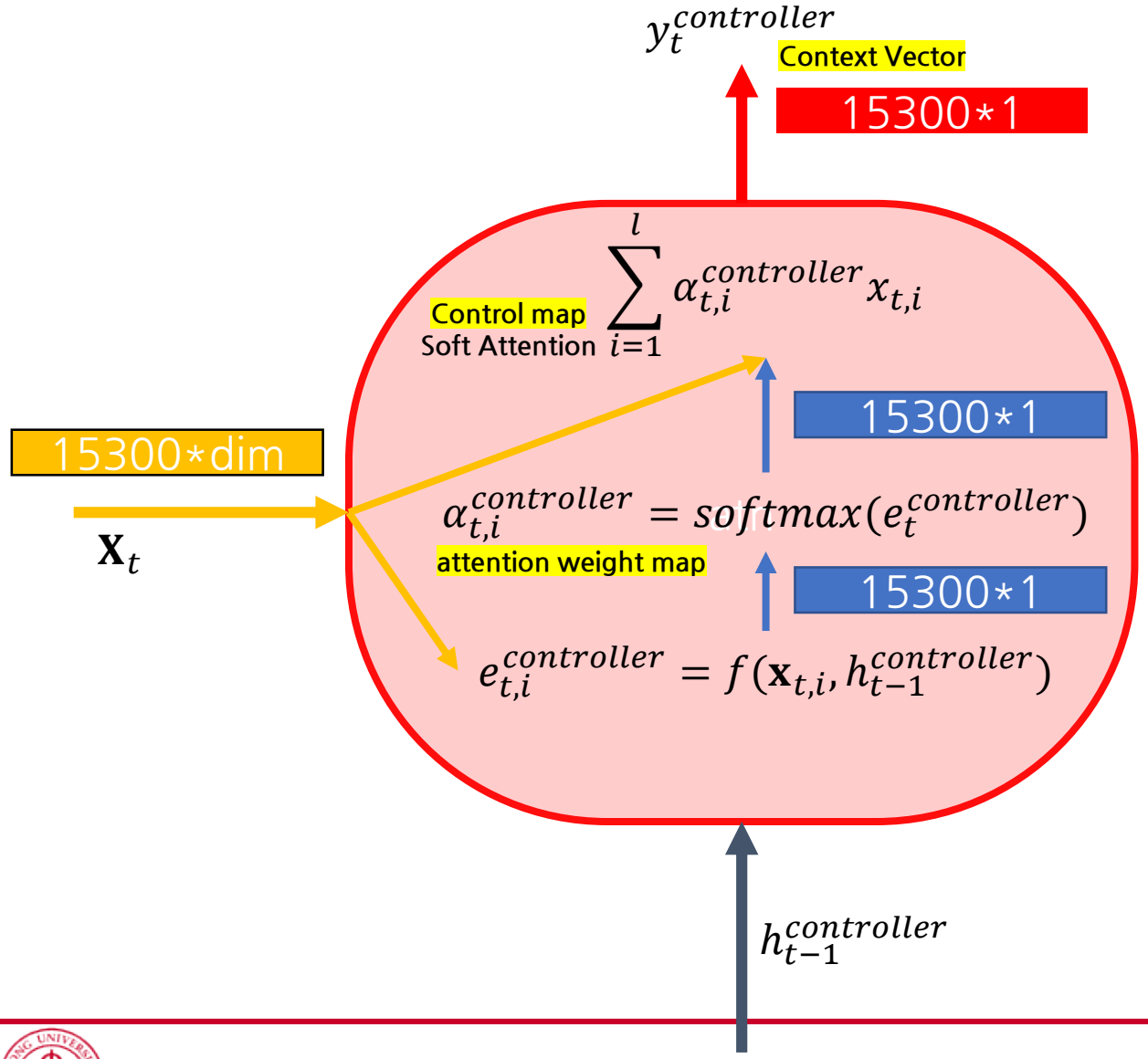
18011573
Janghoo Lee
Department of Computer Engineering

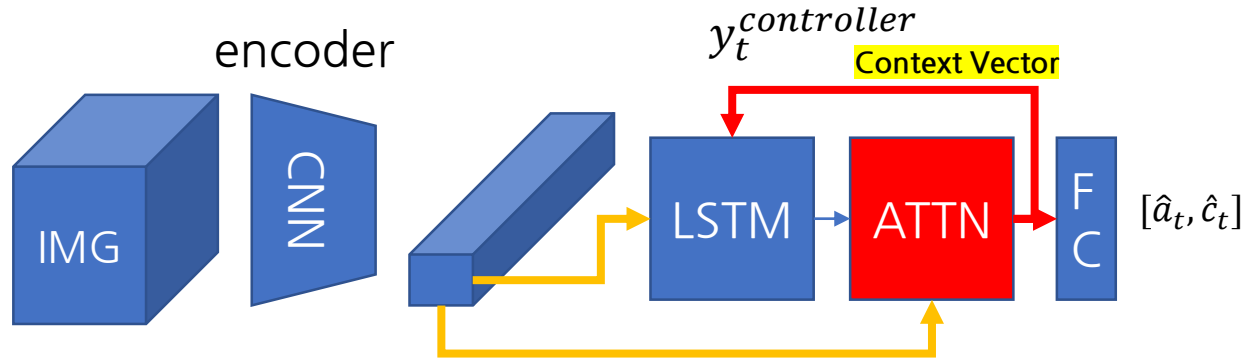
Model Overview



18011573
Janghoo Lee
Department of Computer Engineering

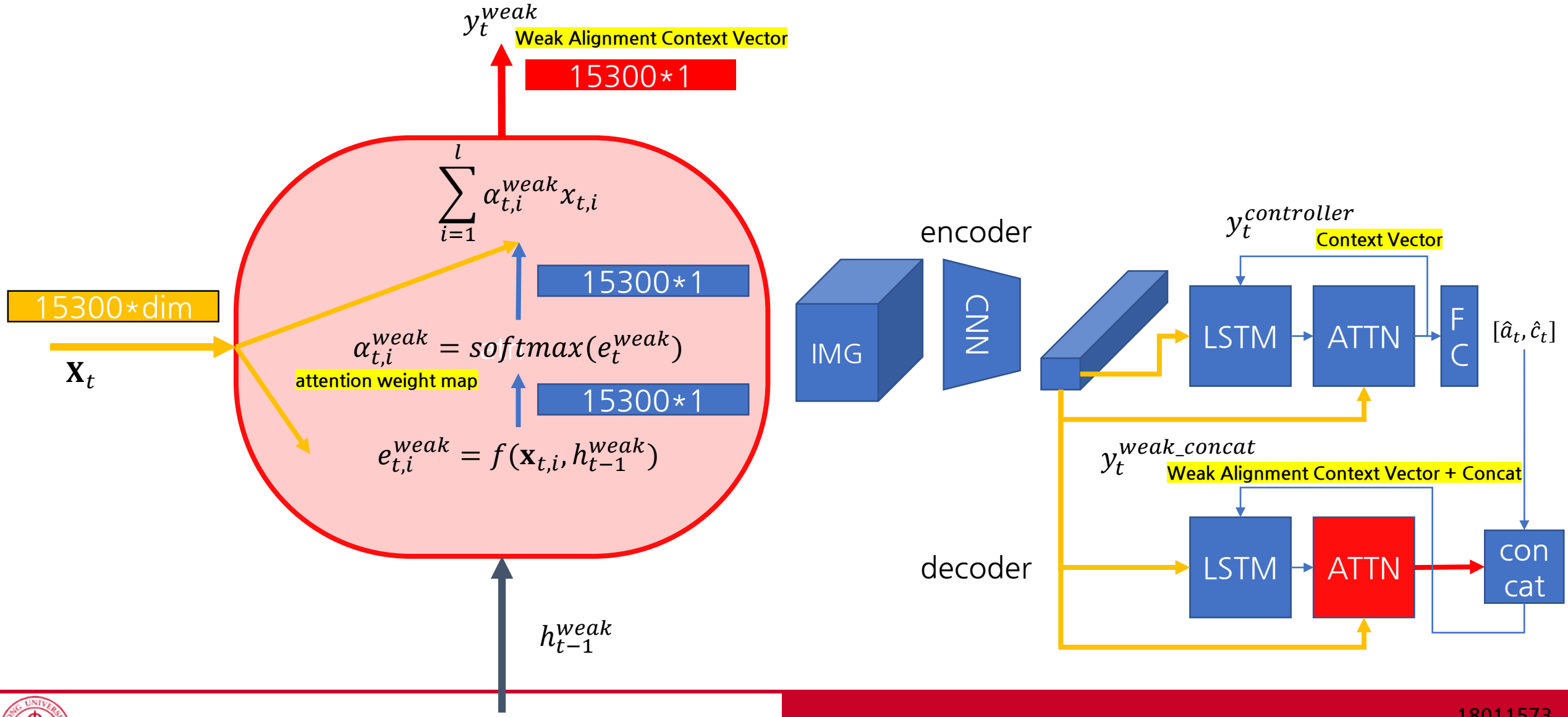
Controller





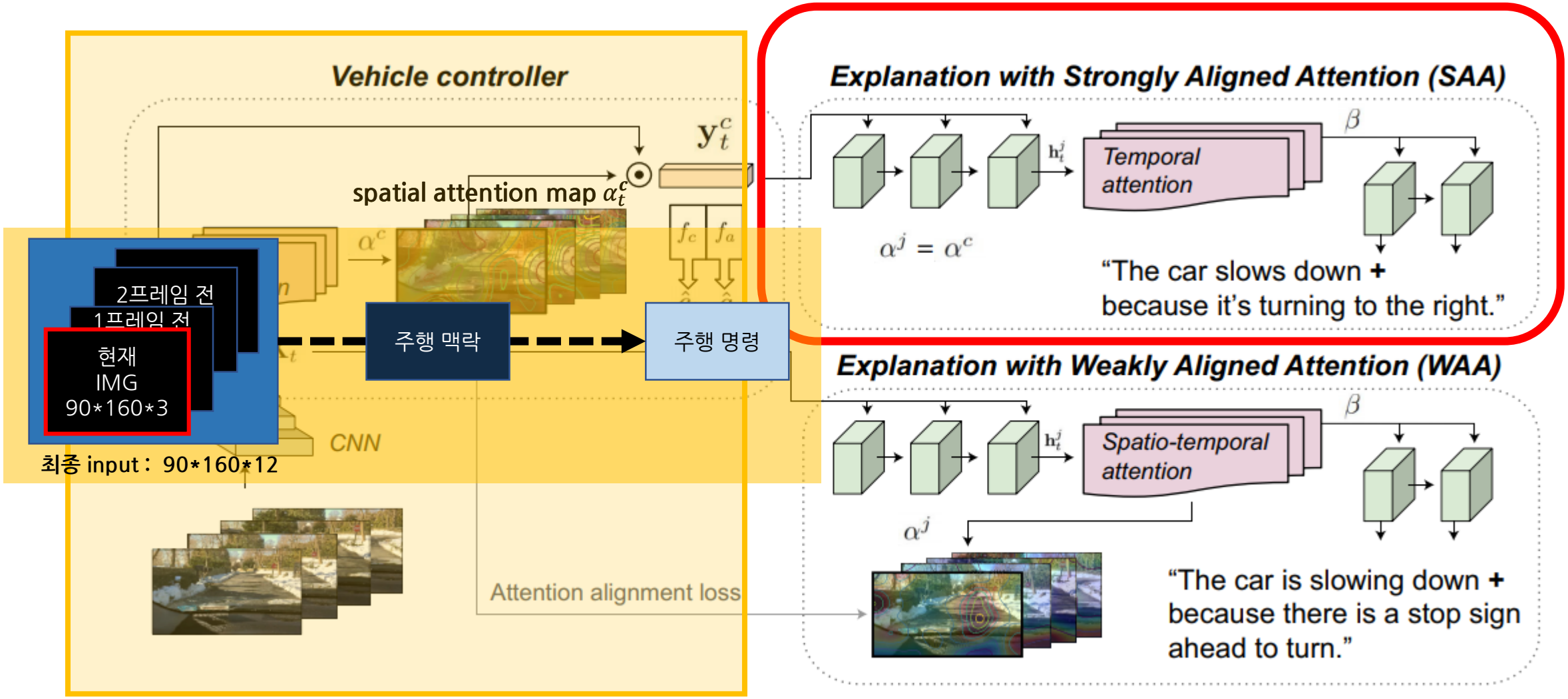
18011573
Janghoo Lee
Department of Computer Engineering

Weakly Aligned Attention Explanation

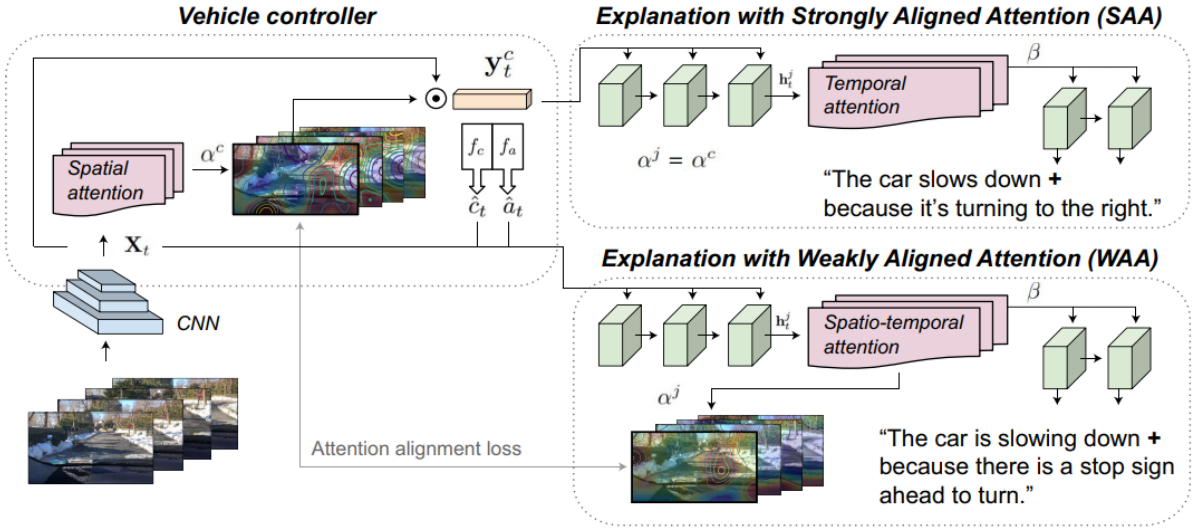


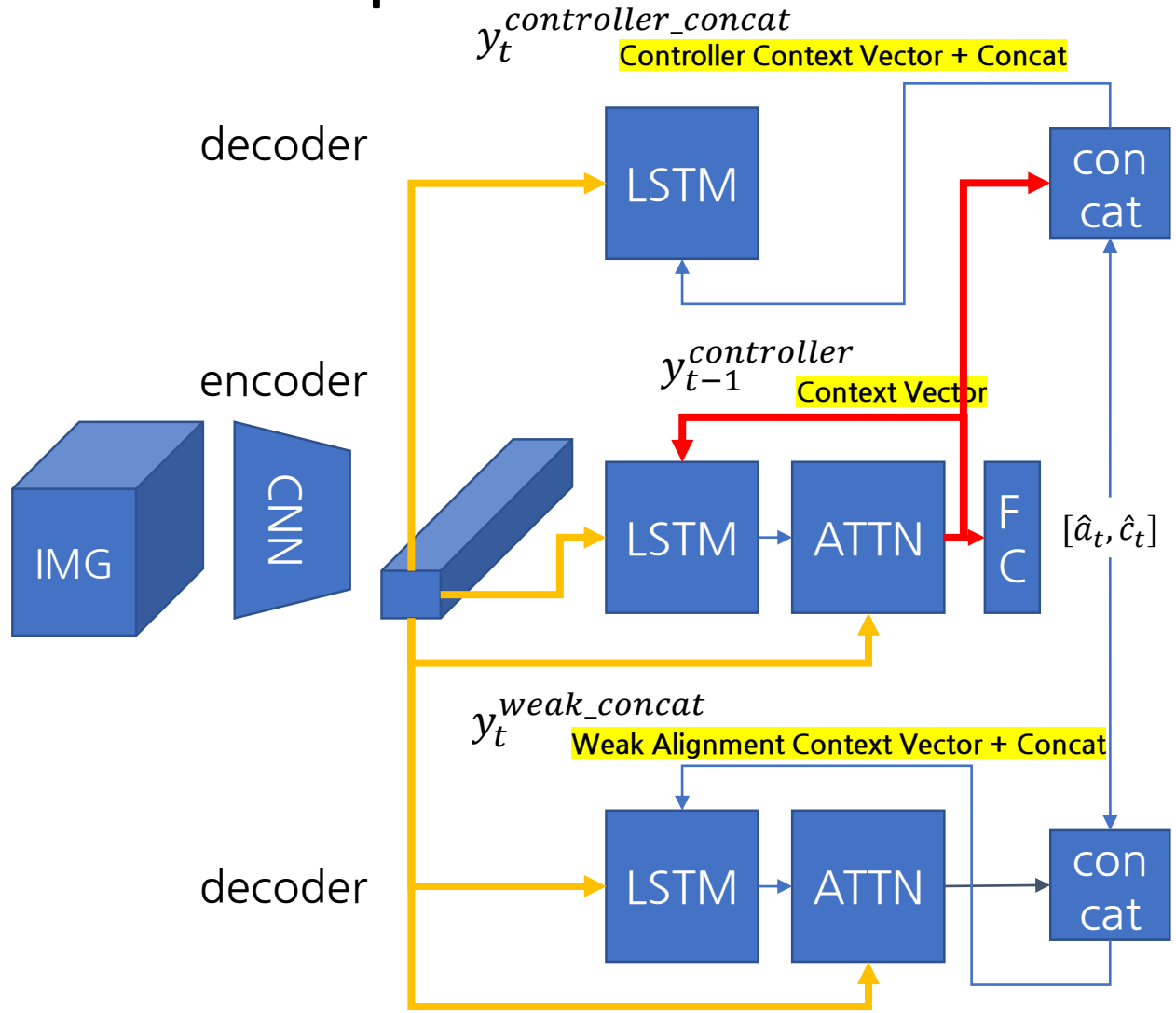
18011573
Janghoo Lee
Department of Computer Engineering

Model Overview



Strongly Aligned Attention Explanation



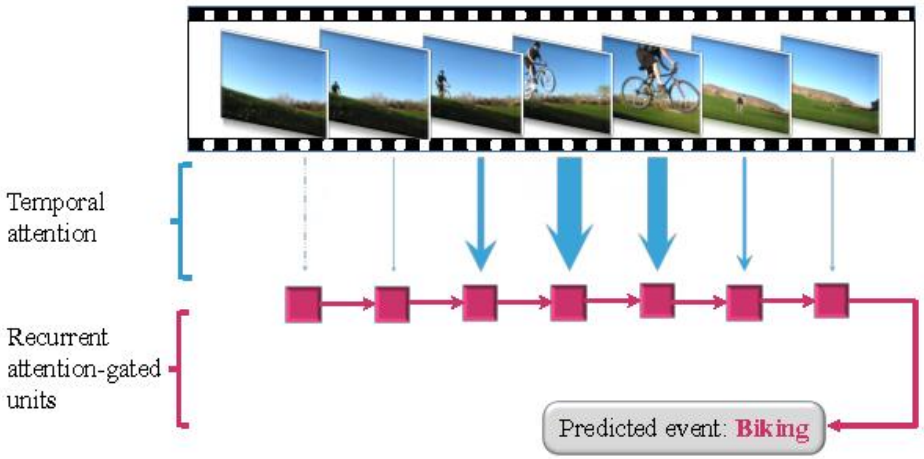


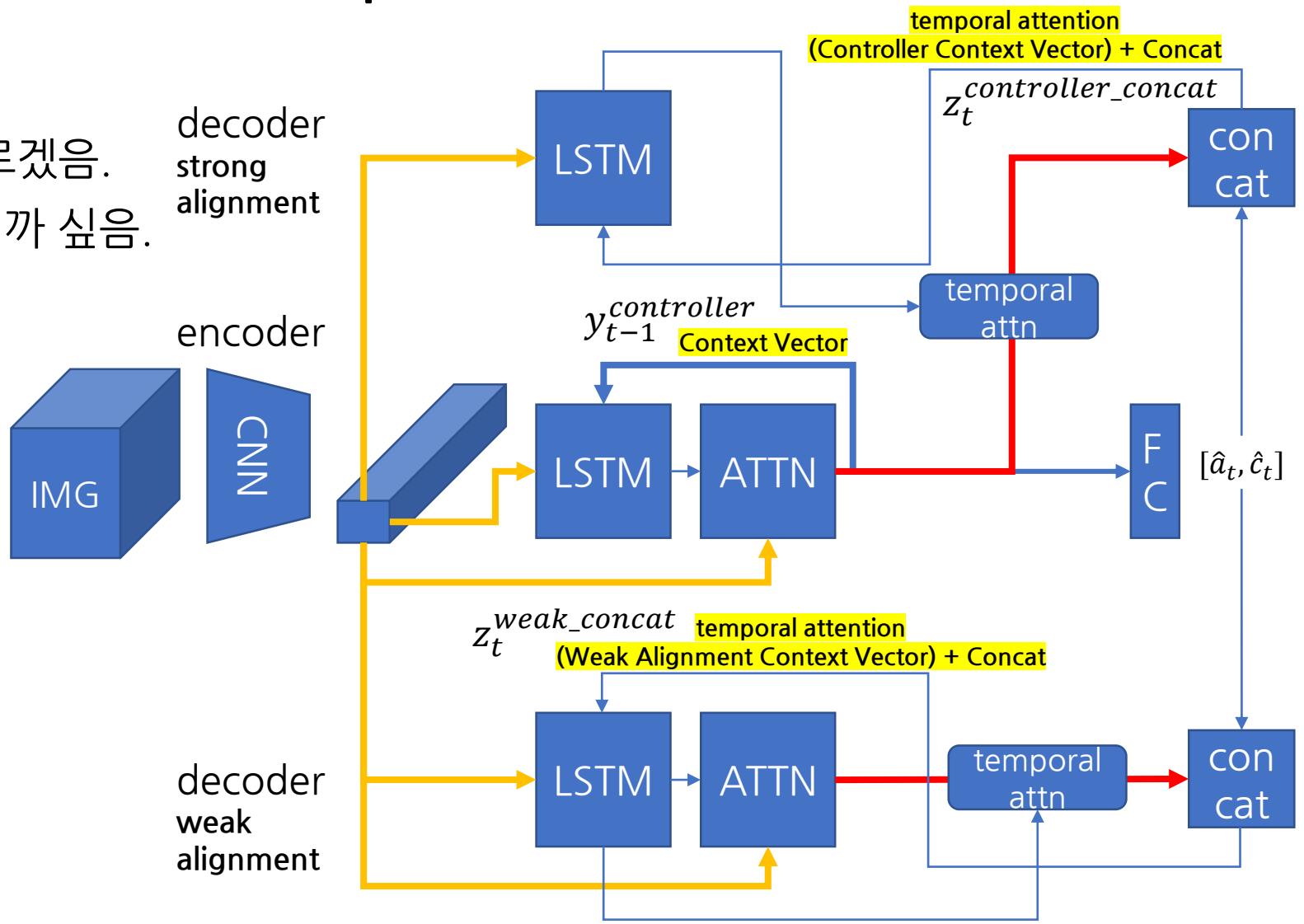
18011573
Janghoo Lee
Department of Computer Engineering

Aligned Attention + Temporal Attention

Temporal Attention

- 이쯤 되니 어떻게 동작하는지는 잘 모르겠음.
- 그냥 일반적인 기계번역 어텐션이 아닐까 싶음.



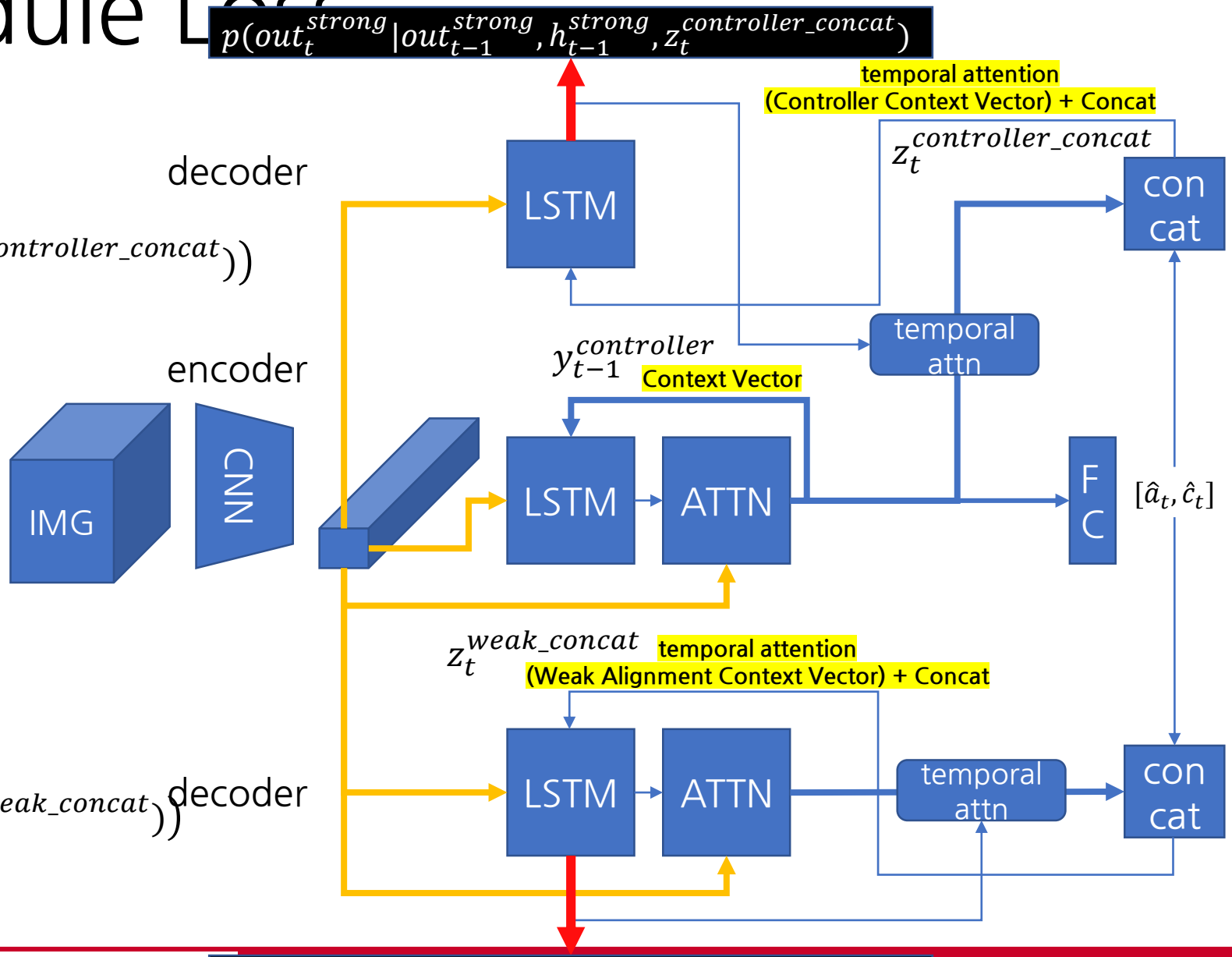


18011573
Janghoo Lee
Department of Computer Engineering

Explanation Module Loss

$$L_{weak\_explanation} = \sum_t log\left(p(out_t^{strong} | out_t^{strong}, h_{t-1}^{strong}, z_t^{controller\_concat})\right)$$

$$L_{strong\_explanation} = \sum_t log\left(p(out_t^{weak} | out_{t-1}^{weak}, h_{t-1}^{weak}, z_t^{weak\_concat})\right)$$



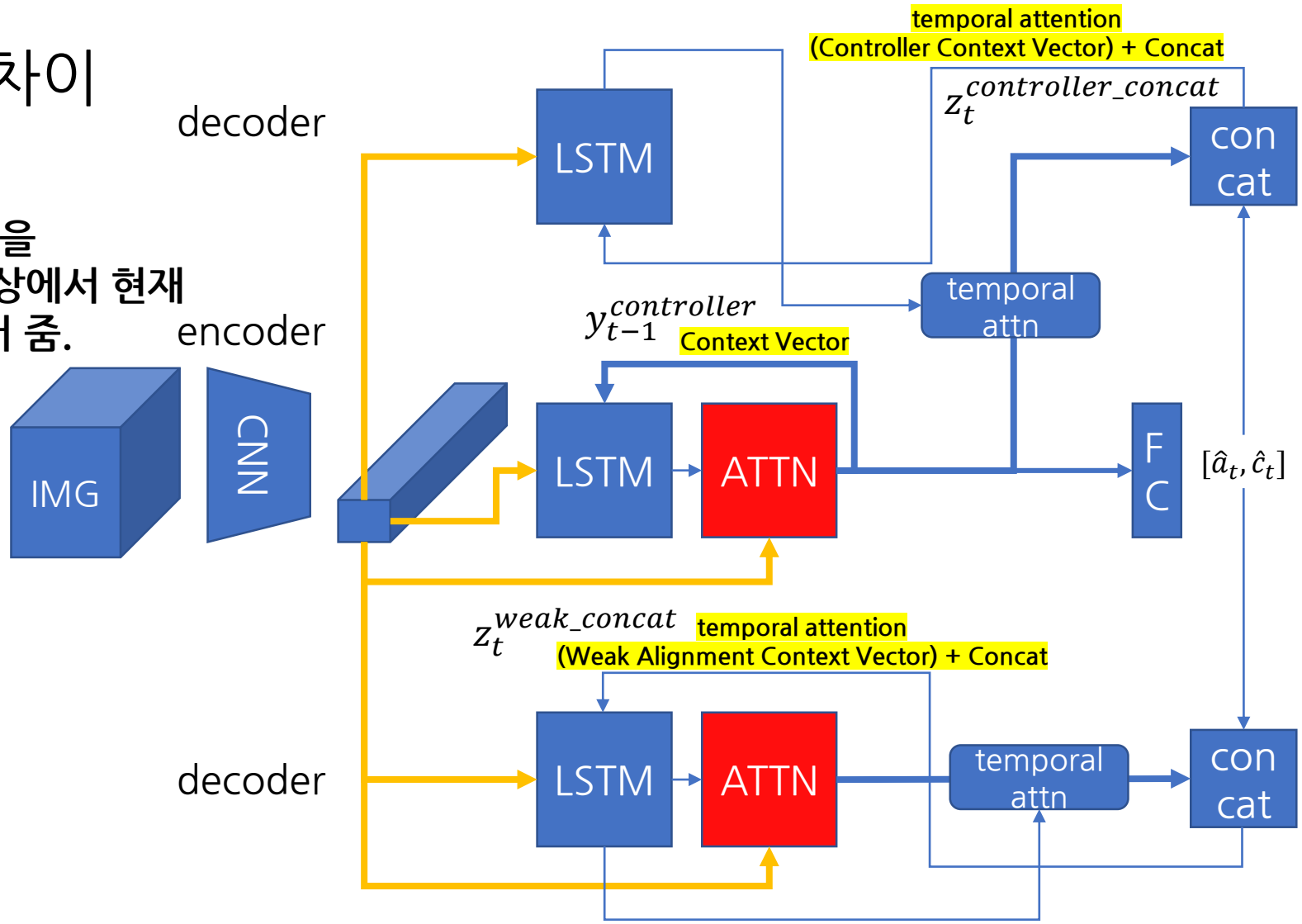
18011573
Janghoo Lee
Department of Computer Engineering

Attention Model Loss

- 두 attention map 사이의 차이
- $\alpha_t^{controller} || \alpha_t^{weak}$

Attention Weight Map : 각 Correlation 값을 Softmax 에 넣음. 즉 상관 관계 값을, 이미지상에서 현재 집중할 위치와의 상관관계 확률 값으로 바꾸어 줌.

$$L_{attn} = \lambda_{attn} \sum_t D_{KL}\left(\alpha_t^{controller} || \alpha_t^{weak}\right)$$



18011573
Janghoo Lee
Department of Computer Engineering

Total Loss

$$\mathcal{L}_c = \sum_t \left((a_t - \hat{a}_t)^2 + (c_t - \hat{c}_t)^2 + \lambda_c H(\alpha_t^c) \right) \quad (2)$$

여기서 c 는 controller

$$L_{weak\_explanation} = \sum_t log\left(p(out_t^{strong} | out_t^{strong}, h_{t-1}^{strong}, z_t^{controller\_concat})\right)$$

$$L_{strong\_explanation} = \sum_t log\left(p(out_t^{weak} | out_{t-1}^{weak}, h_{t-1}^{weak}, z_t^{weak\_concat})\right)$$

$$L_{attn} = \lambda_{attn} \sum_t D_{KL}\left(\alpha_t^{controller} || \alpha_t^{weak}\right)$$

$$L = L_{controller} + L_{attn} + L_{weak\_explanation} + L_{strong\_explanation}$$

이게 도대체 왜되는걸까.

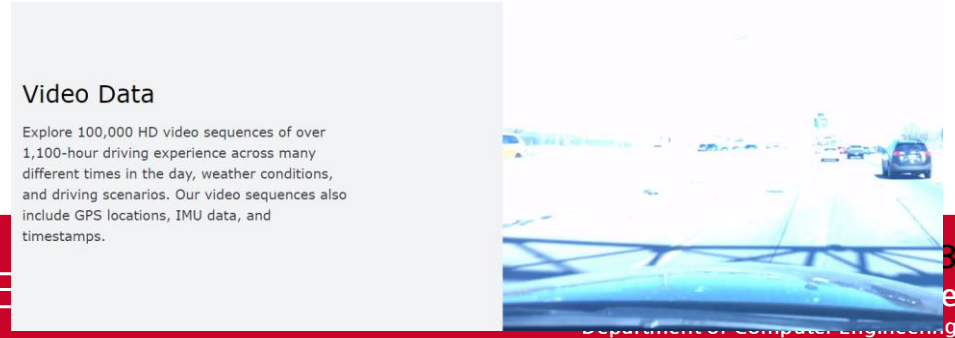
About Dataset : BDD-X

18011573
Janghoo Lee
Department of Computer Engineering

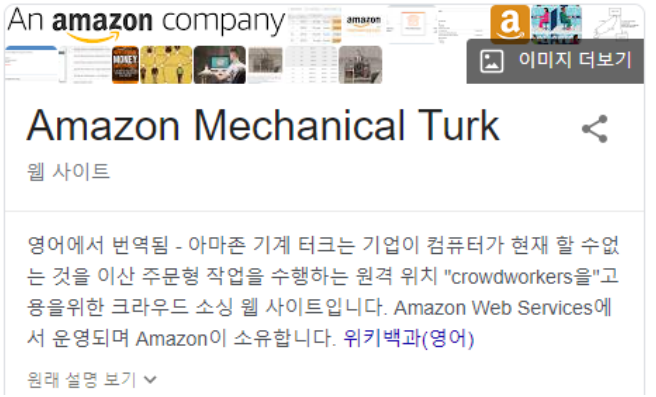
Dataset : BBD-X?

- 논문의 네 번째 단락은 BBD-X : Berkeley DeepDrive eXplanation 이야기.
- 전부 만든 것은 아니고 이미 Berkeley DeepDrive 데이터셋은 존재했음.
  - 차량에 장착된 카메라로 촬영된 영상의 집합
  - 40초정도 길이
  - 도시에서 촬영된 영상들로 구성
  - 다양한 기후에서 녹화됨, 낮과 밤에 모두 녹화됨.
  - lane marking 이 없는 도로에서도 녹화됨.
- 이 논문에서 textual explanation model 을 학습시키기 위해 Berkeley DeepDrive 데이터셋 영상에, 주행영상에 대한 설명을 추가한 것.
- 이걸 BBD-X 라고 이름붙인듯.

https://bdd-data.berkeley.edu/



Dataset : Annotation



- 라벨링 크라우드소싱을 함. *와 뭐 저런게 다있냐*
- 믿을만한 라벨링 결과를 얻기 위해, 다양한 노력을 함
  - 미국 도로에 익숙한 사람들만 선발함.
  - 한번 시험을 봐서 양질의 라벨링 인력을 뽑았음. (“qualified workers”, pass a test)
  - annotation 하는 방법을 미리 강의함.
- 라벨링한 것들은 다음과 같음
  - description : “운전자 입장에서” 운전자가 무엇을 하는지
  - justification (explanation) : “운전자 입장에서” 운전자가 왜 그러한 동작을 하는지
  - time stamp : 이러한 동작들의 시작점, 끝점.
  - 라벨링할 때 세심한 배려 : description 칸과 justification 칸이 따로 있어서, 안 헷갈리도록 함. 하지만 데이터를 실제로 모델에 활용할 때는 두 칸에 입력이 들어온 내용을 합쳐서 하나의 문장으로 학습시킴.

Dataset : Statics

- 데이터 특성
  - 40초짜리 영상에서 동작 3개정도를 입력함.
  - 데이터셋의 총 description/justification (explanation) 2600개
  - description/justification (explanation) 포함되는 동영상 프레임 8.4Million 개

- 당연하게도, 흔한 단어들과 흔한 추론들이 자주 등장하는 경향.
  - description : “앞으로 가고있음.” “멈춰있음.” “가속 중임” …
  - justification (explanation) : “신호등이 ~~해서..” “차가 ~~해서..” “횡단보도 ~~”

Dataset : Inter-Human Agreement

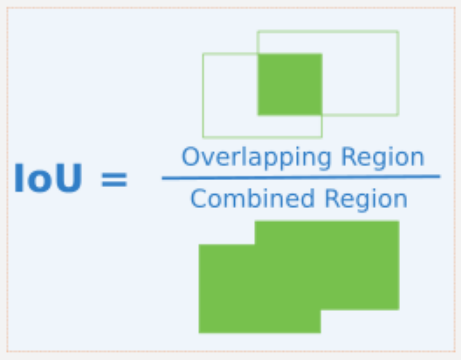
- 데이터 수집과정에서 특히 재밌는 부분
  - 사람들끼리 데이터셋 일부를 두고 교차검증을 함 (inter-human agreement)
  - 총 6948 개의 영상 중, 998 개의 영상에 대해서 2명이 라벨링하고 비교함.
    - 전체 영상에 대해서 두 번씩 시킬 수는 없으니까.

- 수집한 데이터 타당성 평가
  - 라벨링된 부분의 개수는 비슷한가?
  - 라벨링된 부분은 시간적으로 얼마나 겹치는가? (temporal IoU)
  - 이미지 설명이 얼마나 합당한지를 보는 정량적 지표에 따른 점수는 충분히 높은가? (CIDEr Socre)

18011573
Janghoo Lee
Department of Computer Engineering

Dataset : Inter-Human Agreement - Result

- 라벨링된 부분의 개수는 비슷한가?
  - 총 998 개의 영상 중 72% 의 영상은 두 명의 라벨링 개수 차이가 3개 이내였다.
  - 예를 들어 48초짜리 영상에서 A는 4개, B는 1개 라벨을 달았다면 차이는 3개.



- 라벨링된 부분은 시간적으로 얼마나 겹치는가? (temporal IoU)
  - 하나의 영상을 평가한 두 명은 평균적으로 63% 동일한 구간에 라벨링했다.
  - 예를 들어 A가 1초~3초 에 라벨링, B가 1초~2초에 라벨링하면 IoU 50%

- 이미지 설명이 얼마나 합당한지를 보는 정량적 지표에 따른 점수는 충분히 높은가? (CIDEr Socre)
  - 이게 어떻게 돌아가는 것인지는 모르겠음.

www.cv-foundation.org › papers ▾ PDF 이 페이지 번역하기

CIDEr: Consensus-Based Image Description Evaluation - The ...

We propose a new automatic consensus metric of image description quality – **CIDEr** (Consensus-based Image De- scription Evaluation). Our metric measures the similarity of a generated sentence against a set of ground truth sentences written by humans. Our metric shows high agreement with consensus as assessed by humans.

R Vedantam 저술 - 2015 - 1287회 인용 - 관련 학술자료

Results & Evaluation

18011573
Janghoo Lee
Department of Computer Engineering

Training Details (1)

encoder

- CNN 5 layer
- CNN 을 거치며 추출된 Feature cube 의 shape 는 12x20x64

controller

- FC 5 layer

training

- 우선 vehicle controller 을 training 시킴.
- 그리고 controller 을 얼리고 explanation generator 을 학습시킴.

Training Details (2)

- Adam, dropout 0.5, Xavier init.
- 8:1:1 dataset split
- Titan X 로 1일 이내 학습.

18011573
Janghoo Lee
Department of Computer Engineering

Evaluation Details - Controller

- Controller 평가
  - 이 논문에서 제안하는 Controller 의 경우, 브레이크와 엑셀에 연속적인 변화가 존재하는 output 임. 이미지마다 독립적인 output 을 가지는 방식보다, 논문에서 제안한 네트워크가, “시간적인 정보들을 기억” 하고 있기 때문에 (사람도 그러긴 하니까.) 특히 도시환경에서 더 정확할 수 있다고 주장함.
  - 독립적인 output 만을 사용하는 다른 컨트롤러와 차이가 있어서 유의미한 평가를 하기 위해 독립적 output 을 가지는 controller 모델을 올렸음. 단, CNN 은 동일한 것을 사용함.
    - 그런데, CNN 도 이 논문에서 제안한 controller 에 맞게 from scratch 로 학습됐는데 CNN 이 동일해도 되나 싶음.

Model	λ_c	Mean of absolute error (MAE)		Mean of distance correlation	
		Acceleration (m/s^2)	Course (degree)	Acceleration (m/s^2)	Course (degree)
CNN+FC [3]†	-	6.92 [7.50]	12.1 [19.7]	0.17 [0.15]	0.16 [0.14]
CNN+FC [3]+P	-	6.09 [7.73]	6.74 [14.9]	0.21 [0.18]	0.39 [0.33]
CNN+LSTM+Attention [11]†	-	6.87 [7.44]	10.2 [18.4]	0.19 [0.16]	0.22 [0.18]
CNN+LSTM+Attention+P (Ours)	1000	5.02 [6.32]	6.94 [15.4]	0.65 [0.25]	0.43 [0.33]
CNN+LSTM+Attention+P (Ours)	100	2.68 [3.73]	6.17 [14.7]	0.78 [0.28]	0.43 [0.34]
CNN+LSTM+Attention+P (Ours)	10	2.33 [3.38]	6.10 [14.7]	0.81 [0.27]	0.46 [0.35]
CNN+LSTM+Attention+P (Ours)	0	**2.29 [3.33]**	**6.06 [14.7]**	**0.82 [0.26]**	**0.47 [0.35]**

Evaluation Details - Explanation (1)

- description/justification (explanation) 평가 지표로는
  - METEOR
  - CIDEr-D
  - BLEU

- 객관적인 비교를 위해 SOTA S2V2 를 논문의 5 Layer CNN 을 활용하여 구현하고, 논문에서 제안하는 모델들을 비교.

Evaluation Details - Explanation (1)

- S2VT 모델보다 Attention 이 결합된 Alignment 모델이 현재 우리가 하고자 하는 task 에서 더 좋은 모습을 보인다.

Type	Model	Control inputs	λ_a	λ_c	Explanations (*e.g., "because the light is red"*)			Descriptions (*e.g., "the car stops"*)		
					BLEU-4	METEOR	CIDEr-D	BLEU-4	METEOR	CIDEr-D
	S2VT [25]	N	-	-	6.332	11.19	53.35	30.21	27.53	179.8
	S2VT [25]+SA	N	-	-	5.668	10.96	51.37	28.94	26.91	171.3
	S2VT [25]+SA+TA	N	-	-	5.847	10.91	52.74	27.11	26.41	157.0
Rationalization	Ours (no constraints)	Y	0	0	6.515	12.04	61.99	31.01	28.64	205.0
Introspective explanation	Ours (with SAA)	Y	-	0	6.998	12.08	62.24	**32.44**	29.13	213.6
	Ours (with SAA)	Y	-	10	6.760	12.23	63.36	29.99	28.26	203.6
	Ours (with SAA)	Y	-	100	7.074	12.23	66.09	31.84	29.11	214.8
	Ours (with WAA)	Y	10	0	6.967	12.14	64.19	32.24	29.00	**219.7**
	Ours (with WAA)	Y	10	10	6.951	**12.34**	68.56	30.40	28.57	206.6
	Ours (with WAA)	Y	10	100	**7.281**	12.24	**69.52**	32.34	**29.22**	215.8

Evaluation Details - Explanation (1)

- 250개를 무작위로 뽑아서 사람이 직접 네트워크의 output 을 평가
- 완벽, 맞음, 약간 오류, 치명적 오류 4지 선다.
- 성공 기준 : 적어도 2명 이상이 완벽, 맞음에 투표해야 True

Type	Model	Control inputs	λ_a	λ_c	Correctness rate	
					Explanations	Descriptions
Rationalization	Ours (no constraints)	Y	0	0	64.0%	92.8%
Introspective explanation	Ours (with SAA)	Y	-	100	62.4%	90.8%
	Ours (with WAA)	Y	10	100	**66.0%**	**93.5%**

Conclusion

18011573
Janghoo Lee
Department of Computer Engineering

Conclusion

- Controller 로 먼저 학습시켜서 뽑은 CNN Feature, 게다가 attention 으로 인해 더욱 제한된 피처들만으로도 충분히 성공적으로 Explanation 을 만들어낼 수 있다.
  - 이 Feature 은, image 에서 유의미한 정보들을 사람이 제시한 것이 아니고, 단지 image 로부터 브레이크와 엑셀 결과를 뽑아내면서 자동적으로 학습한 제한된 (grounded) Feature 임. 이런 것으로부터
  - Attention 을 결합하면 더 좋은 성능을 낸다.
  - 논문에서 제안한 두 가지 attention 배열 모두 준수한 성능을 보여준다.

Suggestion

- Causal Filtering (?) 을 이용하면 detection 에서 더 좋은 성능을 낼 수 있지 않을까 하고 툭 던짐.
- 운전자의 시선과 협력하는 모델을 만들면 어텐션을 더 재밌게 만들 수 있지 않을까 하고 제안함.

18011573
Janghoo Lee
Department of Computer Engineering

thank you!

- 스터디가 필요한 곳
  - Causal Filtering?
  - Temporal Attention?
  - CIDEr-D score?
  - Attention Back-Prop, How it works?
  - KL divergence / Cross Entropy Detail

18011573
Janghoo Lee
Department of Computer Engineering