Computer Vision

Multi-modal learning: Captioning and Speaking

Tae-Hyun Oh (오태현)
전자전기공학과
POSTECH

Slide by Juyong Lee (이주용)
TAs: {Dongmin Choi , Jongha Kim, Juyong Lee, Sungbin Kim} (in alphabetic order)

boostcamp AI Tech

1. Overview of multi-modal learning

2. Multi-modal tasks (1) – Visual data & Text

- 2.1 Text embedding
- 2.2 Joint embedding
- 2.3 Cross modal translation
- 2.4 Cross modal reasoning

3. Multi-modal tasks (2) – Visual data & Audio

- 3.1 Sound representation
- 3.2 Joint embedding
- 3.3 Cross modal translation
- 3.4 Cross modal reasoning

1.

Overview of multi-modal learning

Toward the world beyond images

1. Overview of multi-modal learning

Modalities in multi-modal learning

Social

Odor

Texture

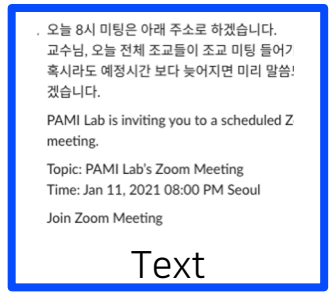
Taste

Unimodal

Audio

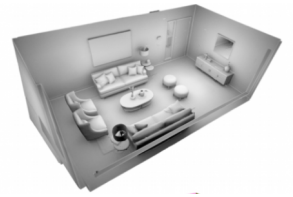
Vision

Multi-modal

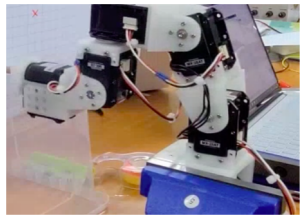


Text

Unimodal



Depth



Force

×

[https://arxiv.org/pdf/1906.05797.pdf]

1. Overview of multi-modal learning

Challenge (1) - Different representations between modalities

Audio

Image

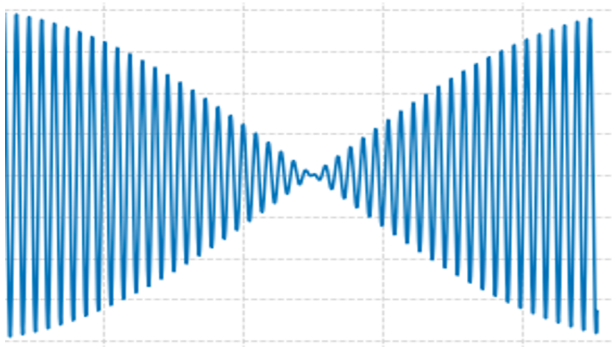
오늘 8시 미팅은 아래 주소로 하겠습니다.
교수님, 오늘 전체 조교들이 조교 미팅 들어가
혹시라도 예정시간 보다 늦어지면 미리 말씀!
겠습니다.

PAMI Lab is inviting you to a scheduled Z meeting.

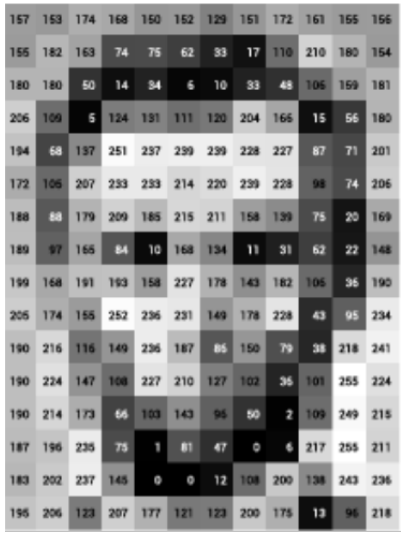
Topic: PAMI Lab's Zoom Meeting
Time: Jan 11, 2021 08:00 PM Seoul

Join Zoom Meeting

Text



×

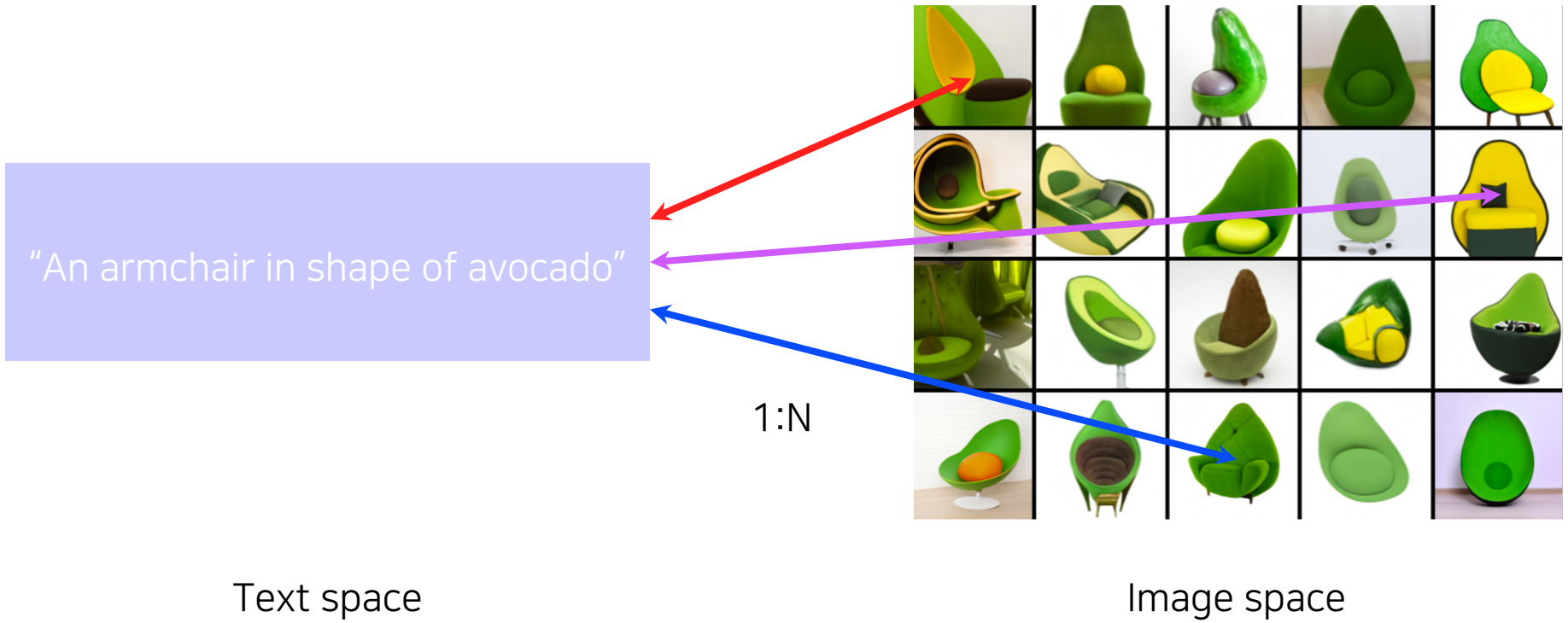


	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8

[https://ai.stanford.edu/~syyeung/cvweb/tutorial1.html, https://medium.com/@hari4om/word-embedding-d816f643140]

1. Overview of multi-modal learning

Challenge (2) - Unbalance between heterogeneous feature spaces

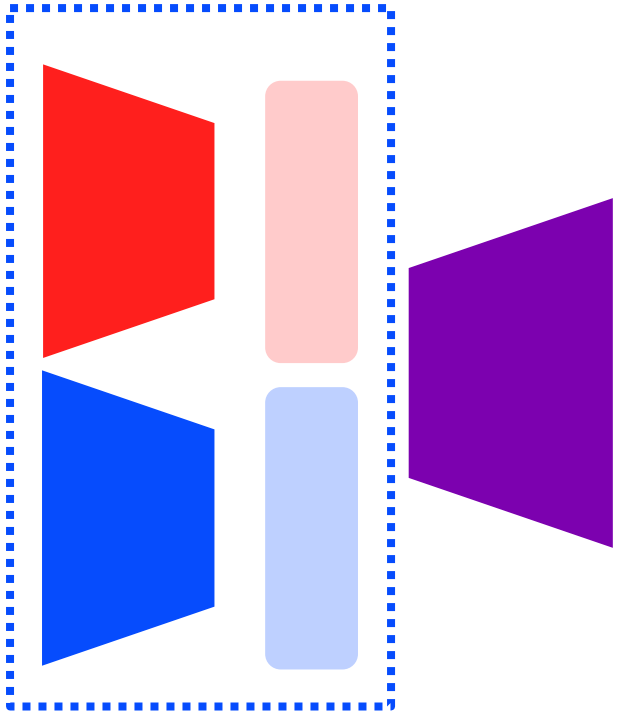


[https://openai.com/blog/dall-e]

1. Overview of multi-modal learning

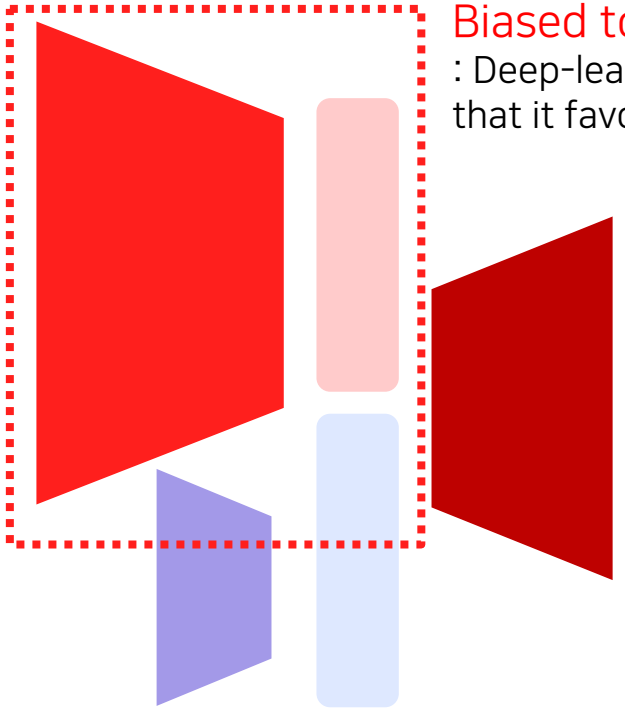
Challenge (3) - May a model be biased on a specific modality

[Wang et al., CVPR 2020]



Want to get benefits from multiple modalities

Poor training

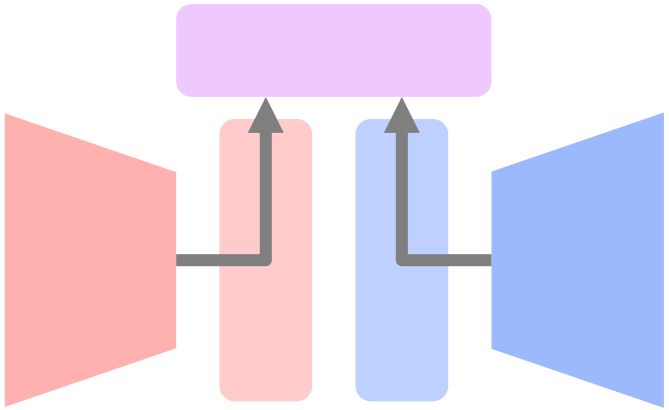


Biased to one modality
: Deep-learning is so lazy that it favors an easy path

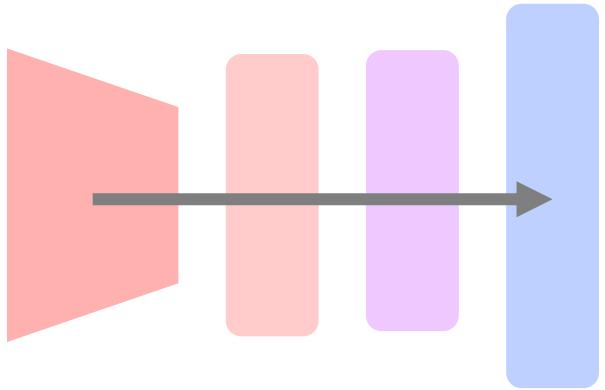
Model focuses only on an easy modality

1. Overview of multi-modal learning

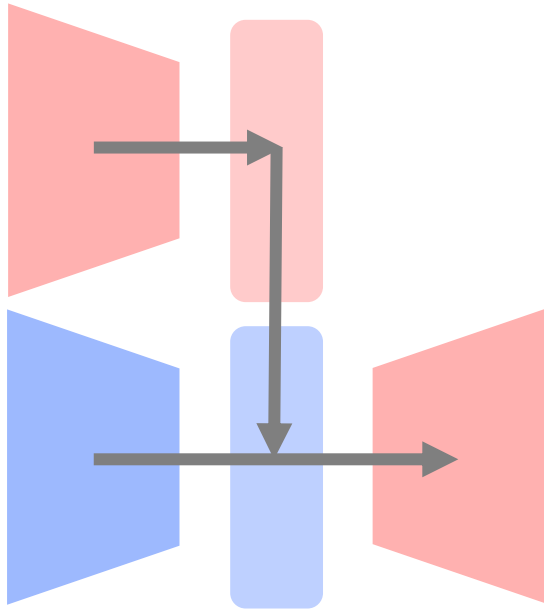
Despite the challenges, multi-modal learning is fruitful and important



Matching



Translating



Referencing

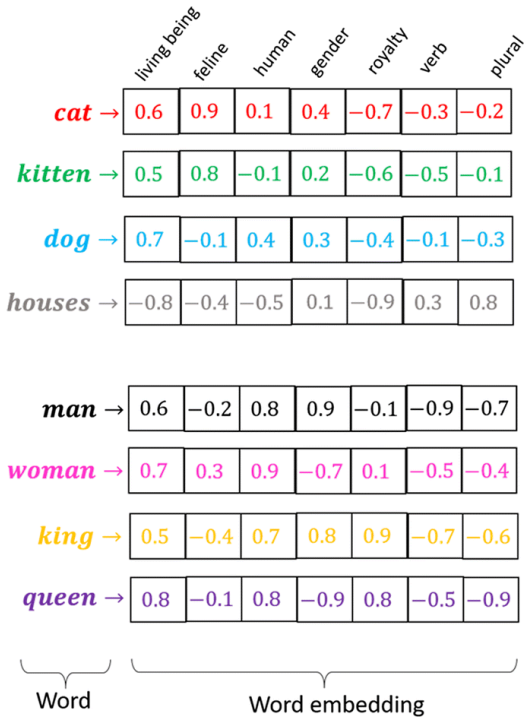
2.

Multi-modal tasks (1) - Visual data & Text

2.1 Text embedding

Text embedding – Example

- Characters are hard to use in machine learning
- Map to dense vectors

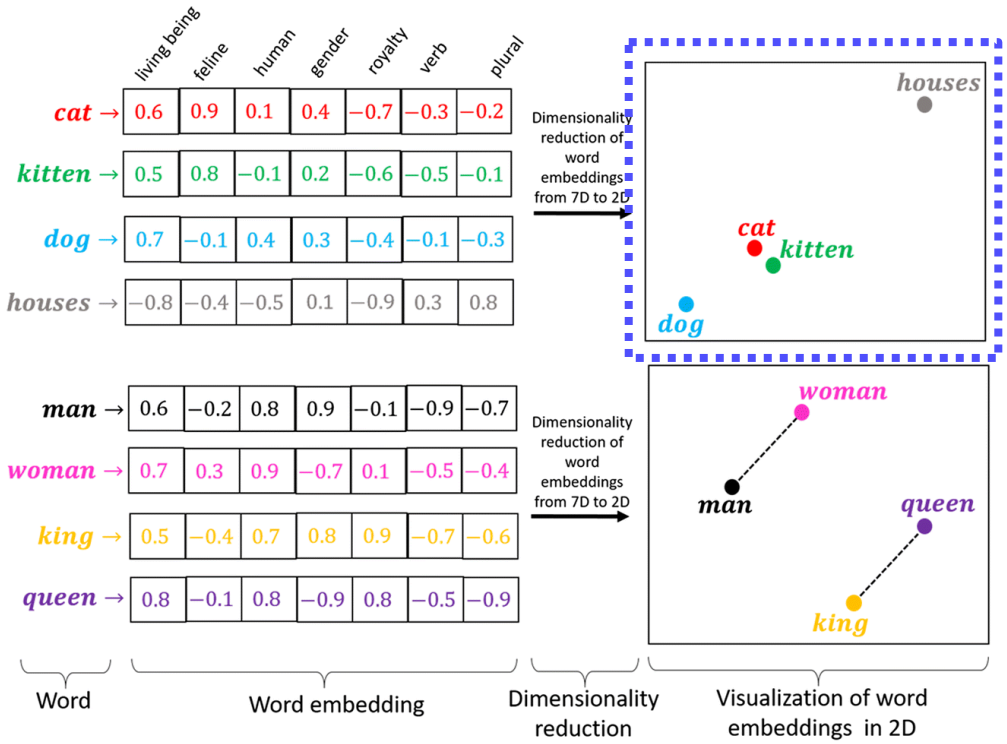


[https://medium.com/@hari4om/word-embedding-d816f643140]

2.1 Text embedding

Text embedding – Example

- Surprisingly, generalization power is obtained by learning dense representation
  - E.g., cat vs. kitten

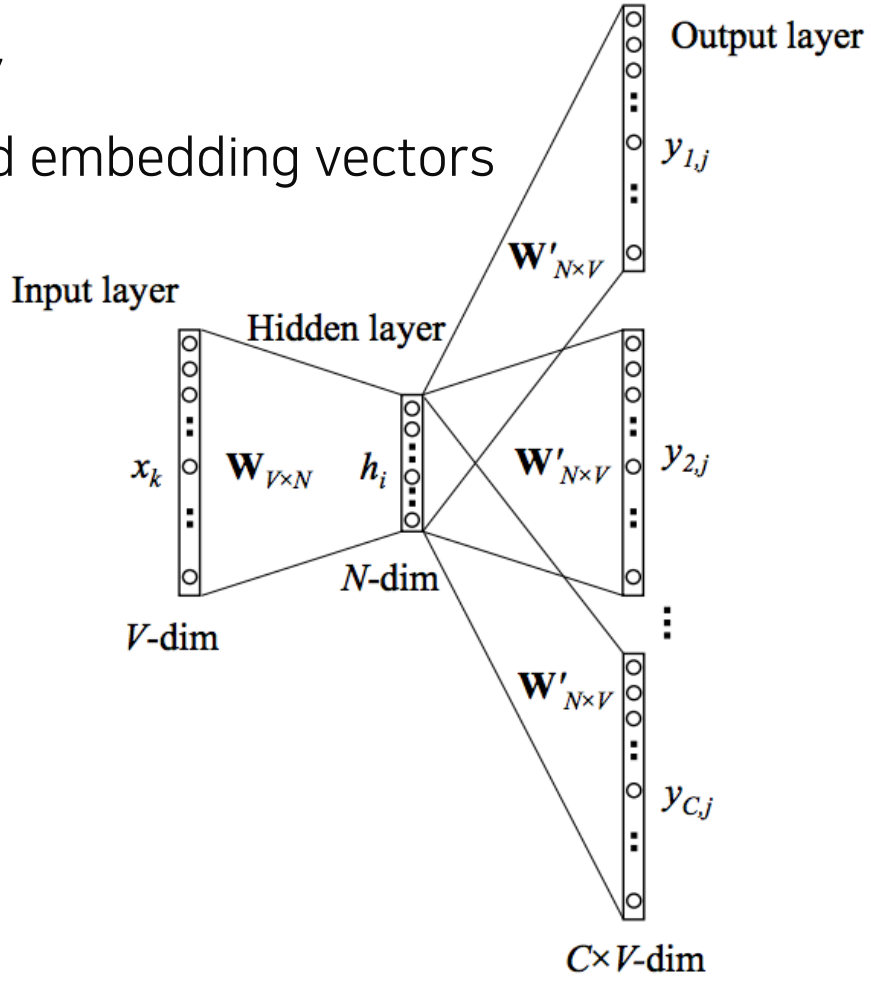


[https://medium.com/@hari4om/word-embedding-d816f643140]

2.1 Text embedding

word2vec - Skip-gram model

- Trained to learn W and W'
- Rows in W represent word embedding vectors

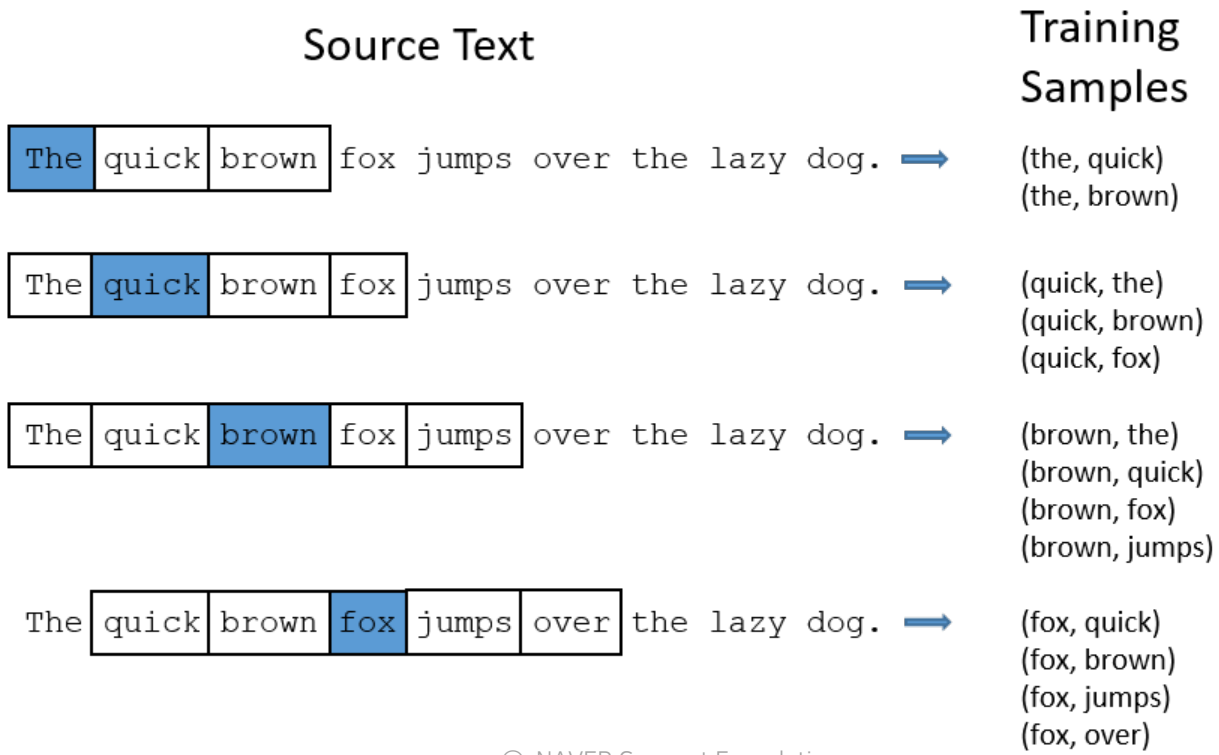


[https://imgur.com/TupGxMl]

2.1 Text embedding

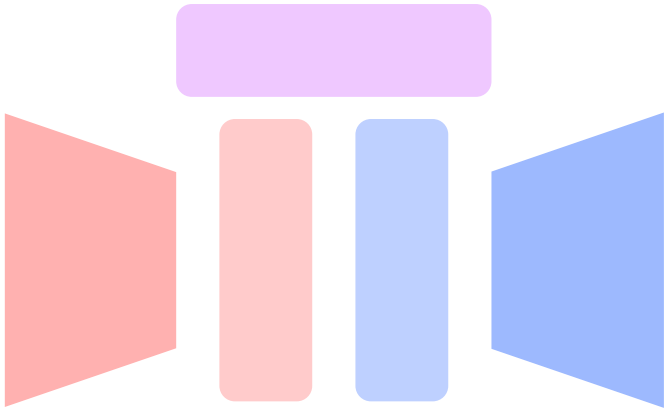
word2vec - Skip-gram model

- Learning to predict neighboring N words for understanding relationships between words
  - Given a model with a window of size 5, the center words depend on 4 words



[https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b]

2.2 Joint embedding



Matching

Translating

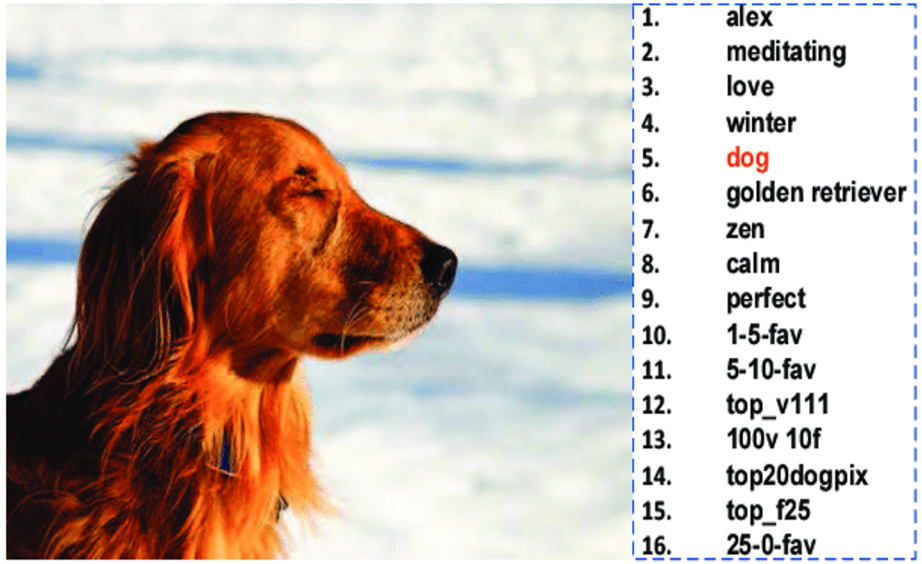
Referencing

2.2 Joint embedding

Application – Image tagging

[Srivastava and Slakhutdinov, JMLR 2014]

- Can generate tags of a given image, and retrieve images by a tag keyword as well



Tag generation of an image

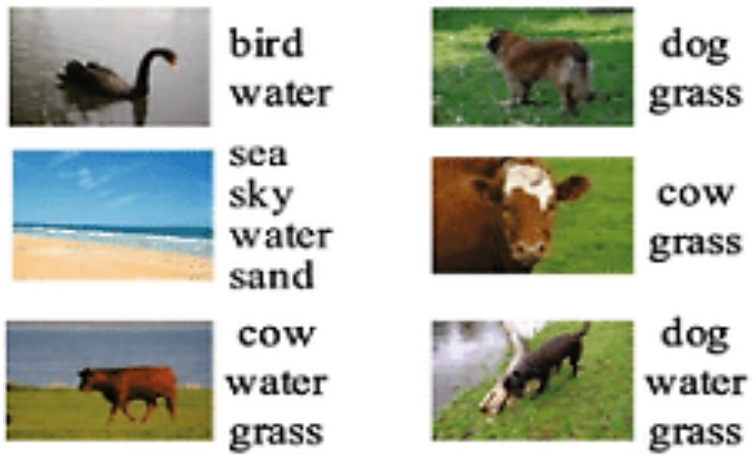


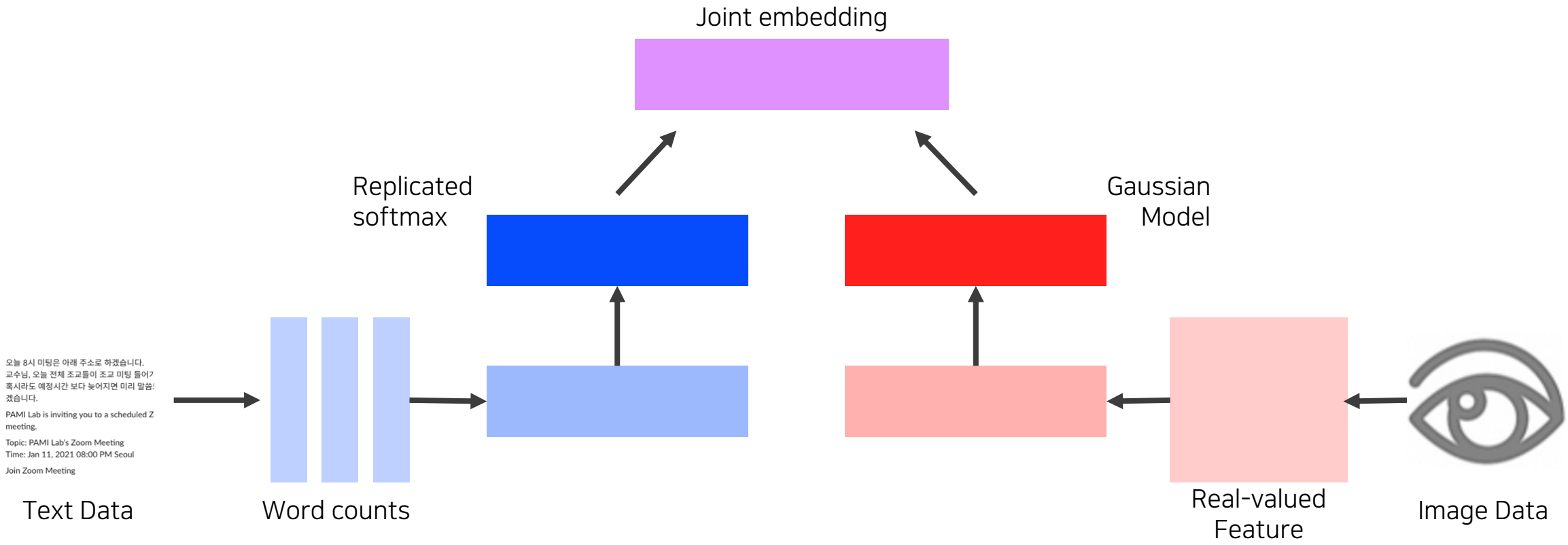
Image retrieval by a tag keyword

[Liu et al., ICWWW, 2009], [Li et al., IJCV 2020]

2.2 Joint embedding

Image tagging – Combining pre-trained unimodal models

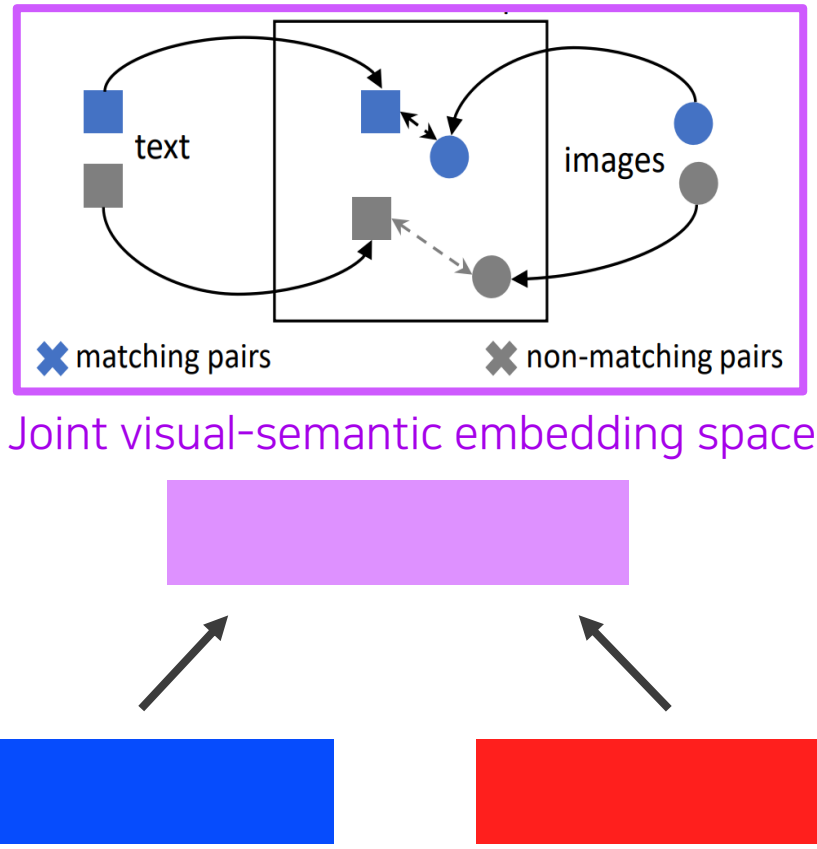
[Srivastava and Slakhutdinov, JMLR 2014]



2.2 Joint embedding

Image tagging – Metric learning in visual-semantic space

[Srivastava and Slakhutdinov, JMLR 2014]



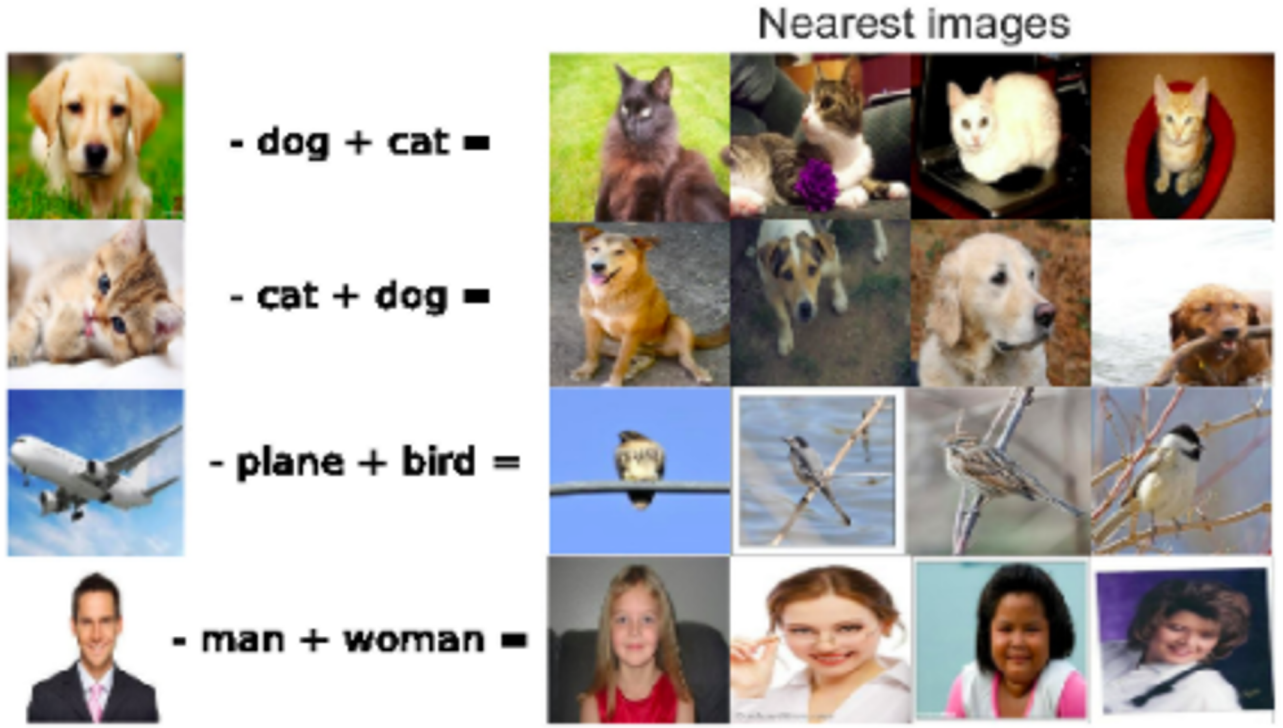
[Cornia et al., ECCV 2018]

2.2 Joint embedding

Image tagging – Interesting property

[Srivastava and Slakhutdinov, JMLR 2014]

- Surprisingly, learned embeddings hold analogy relationships between visual and text data



Multi-modal analogy

[Kiros et al., NIPS 2014]

2.2 Joint embedding

Application – Image & food recipe retrieval

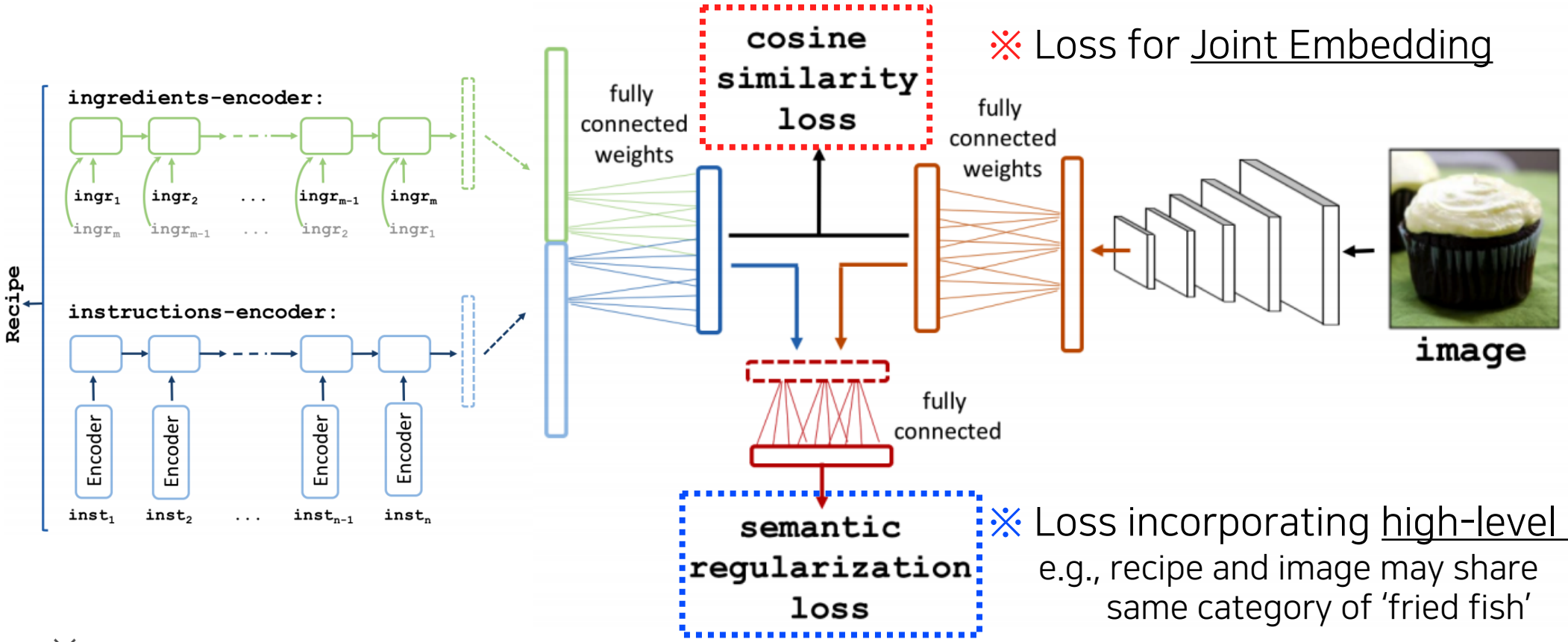
[Marin et al., TPAMI 2019]

Retrieved ingredients and images given the query images

2.2 Joint embedding

Recipe text (sentence) vs. food image

[Marin et al., TPAMI 2019]



※ Loss for Joint Embedding

※ Loss incorporating high-level semantics

e.g., recipe and image may share same category of 'fried fish'

2.3 Cross modal translation

Matching

Translating

Referencing

2.3 Cross modal translation

Application – Image captioning

a cat sitting on a suitcase on the floor

a cat is sitting on a tree branch

a dog is running in the grass with a frisbee

a white teddy bear sitting in the grass

two people walking on the beach with surfboards

a tennis player in action on the court

two giraffes standing in a grassy field

a man riding a dirt bike on a dirt track

2.3 Cross modal translation

Captioning as image-to-sentence - CNN for image & RNN for sentence

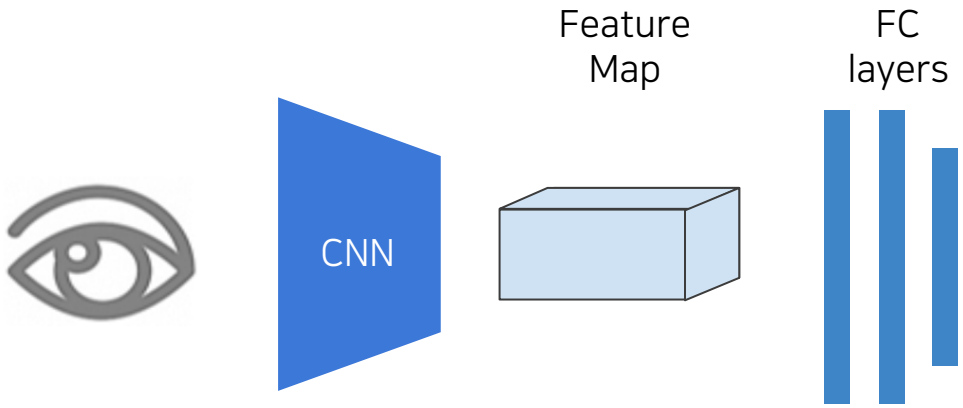
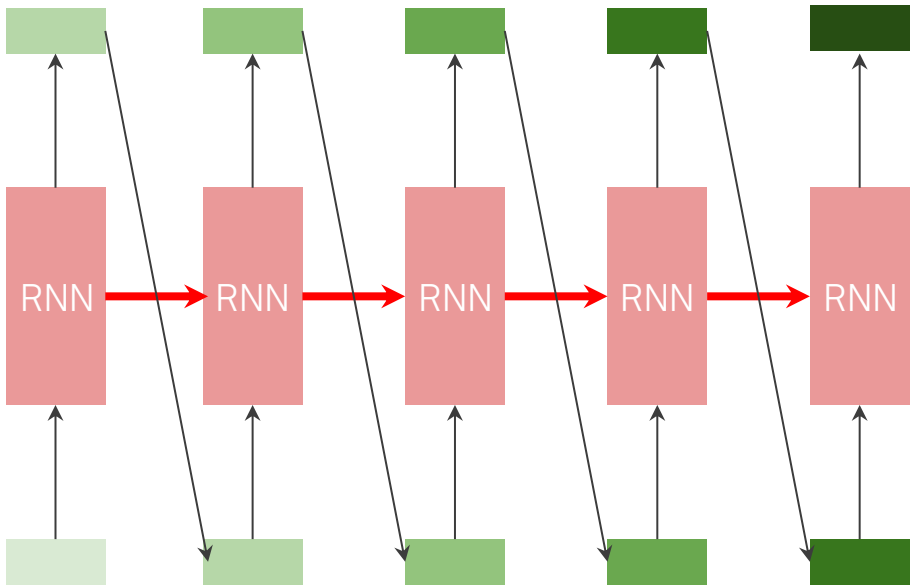


Image representation by CNN



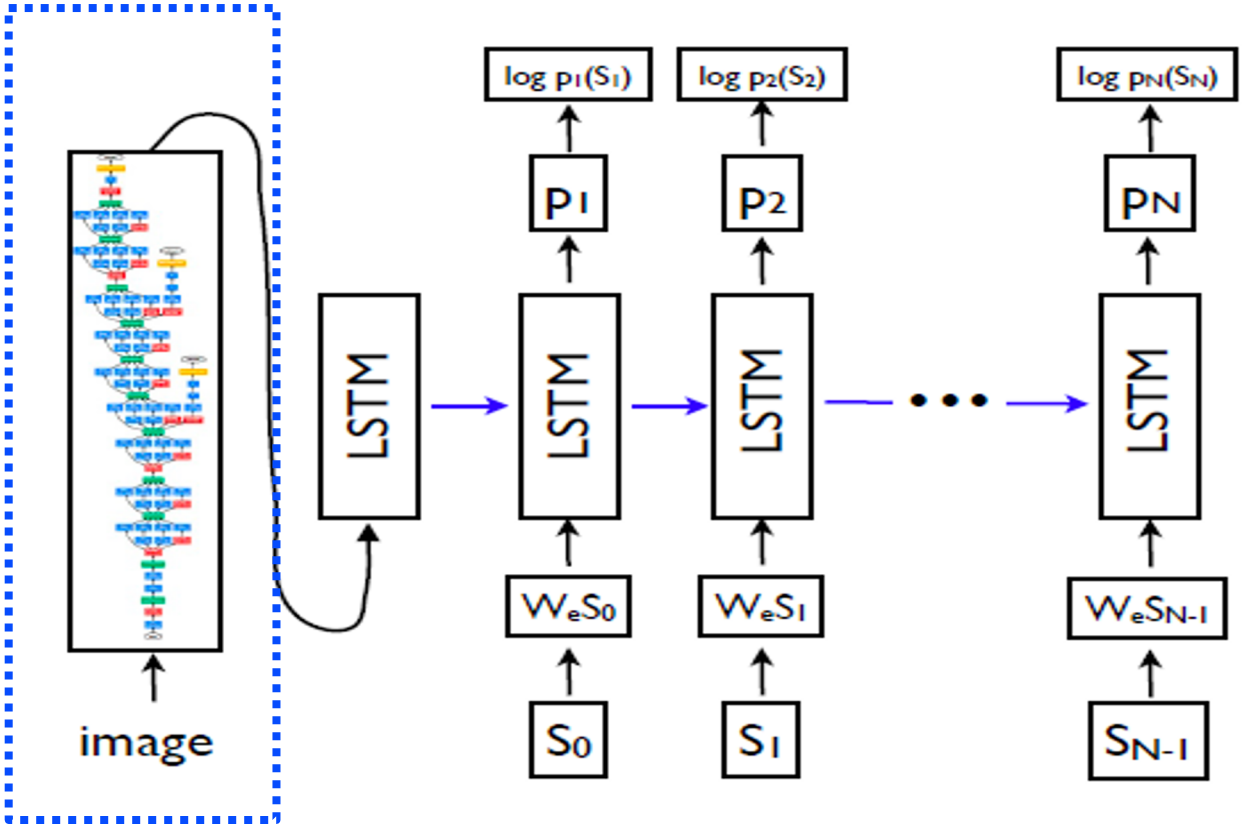
Sentence prediction by RNN

2.3 Cross modal translation

Show and tell

[Vinyals et al., CVPR 2015]

- Encoder: CNN model pre-trained on ImageNet
- Decoder: LSTM module

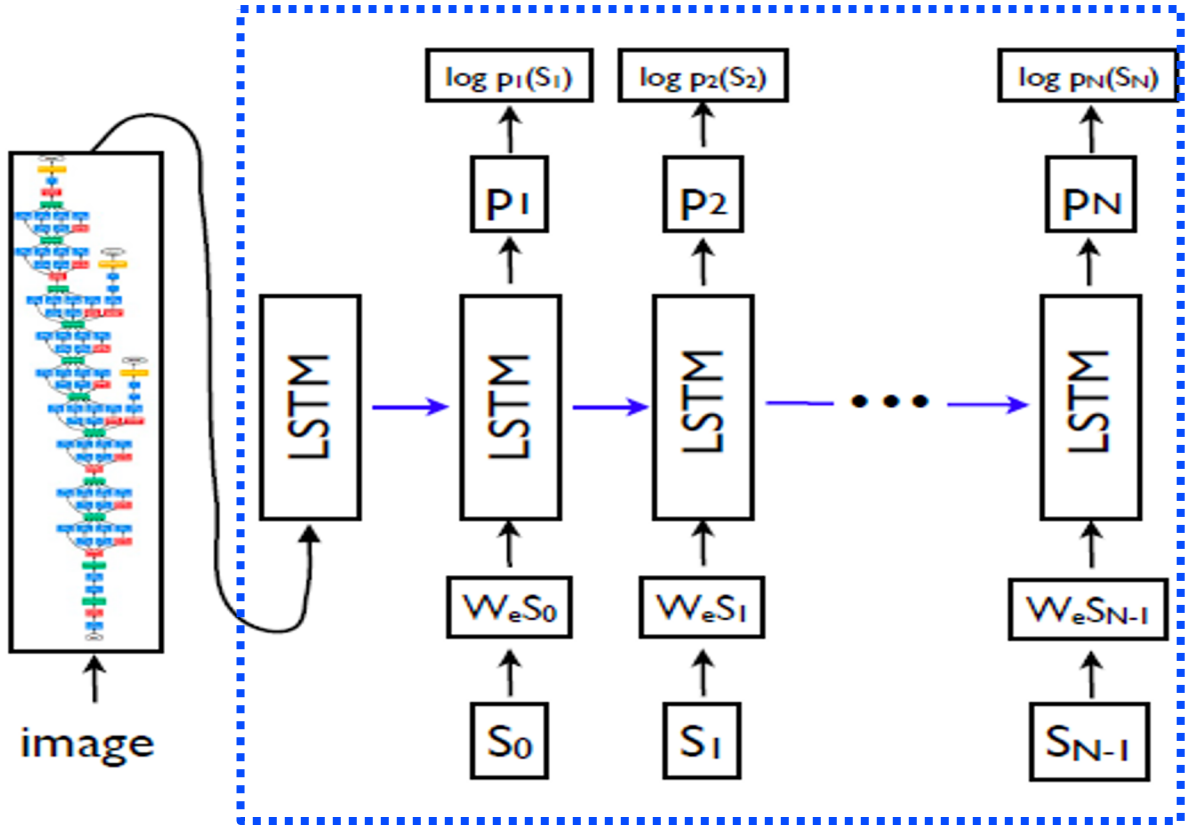


2.3 Cross modal translation

Show and tell

[Vinyals et al., CVPR 2015]

- Encoder: CNN model pre-trained on ImageNet
- Decoder: LSTM module



2.3 Cross modal translation

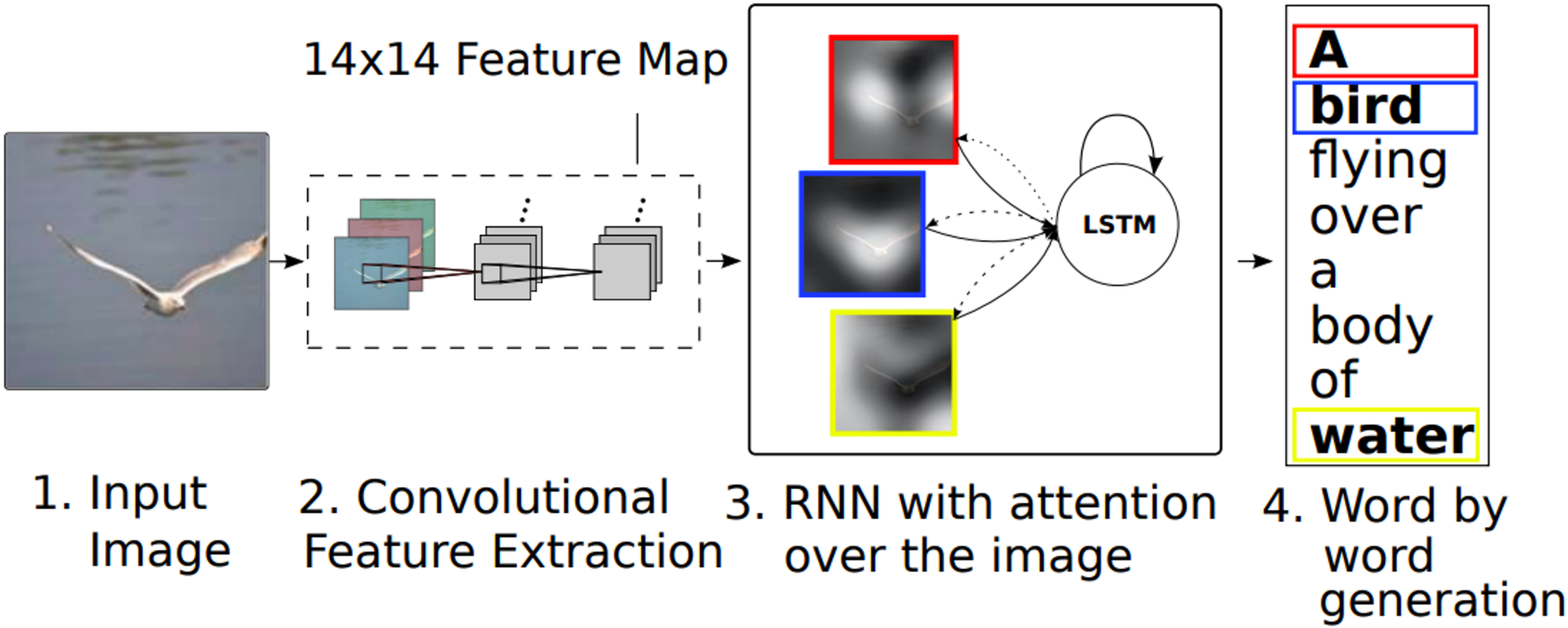
Show, attend, and tell – Example results

[Xu et al., ICML 2015]

2.3 Cross modal translation

Show, attend, and tell

[Xu et al., ICML 2015]



2.3 Cross modal translation

Show, attend, and tell - Attention

[Xu et al., ICML 2015]

CC BY-SA 2.0 licensed image

Human attention shifts on face image

×

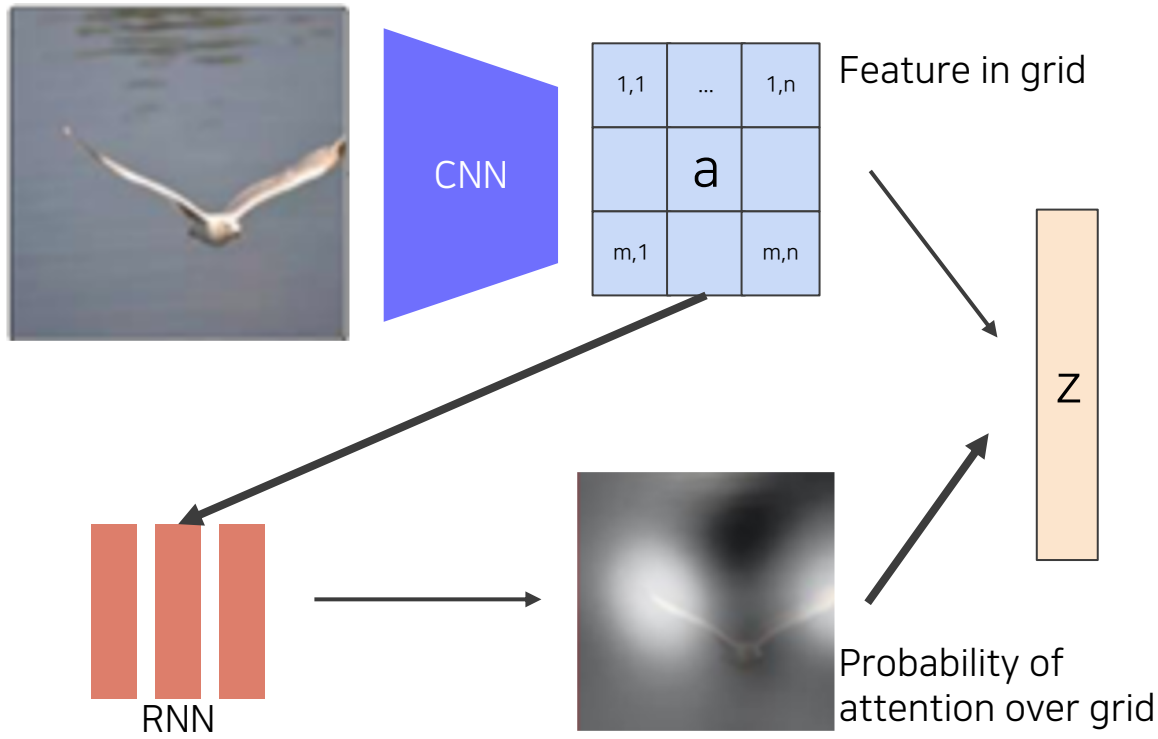
2.3 Cross modal translation

Show, attend, and tell - Soft Attention

[Xu et al., ICML 2015]

CC BY-SA 2.0 licensed image

Human attention shifts on face image

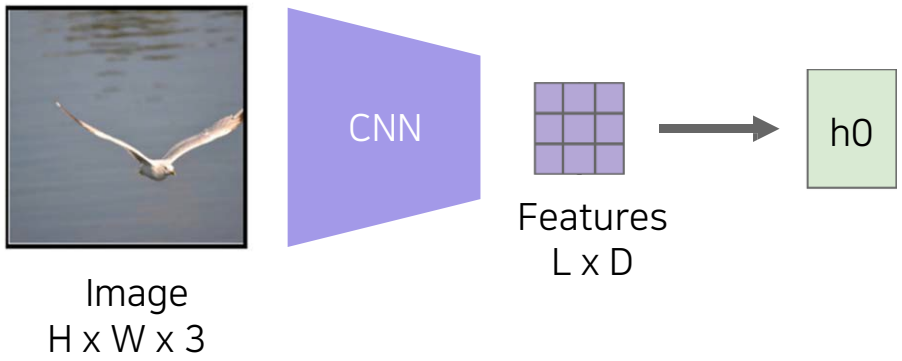


Soft attention embedding: z = a ⊙ s

2.3 Cross modal translation

Show, attend, and tell - Inference

[Xu et al., ICML 2015]

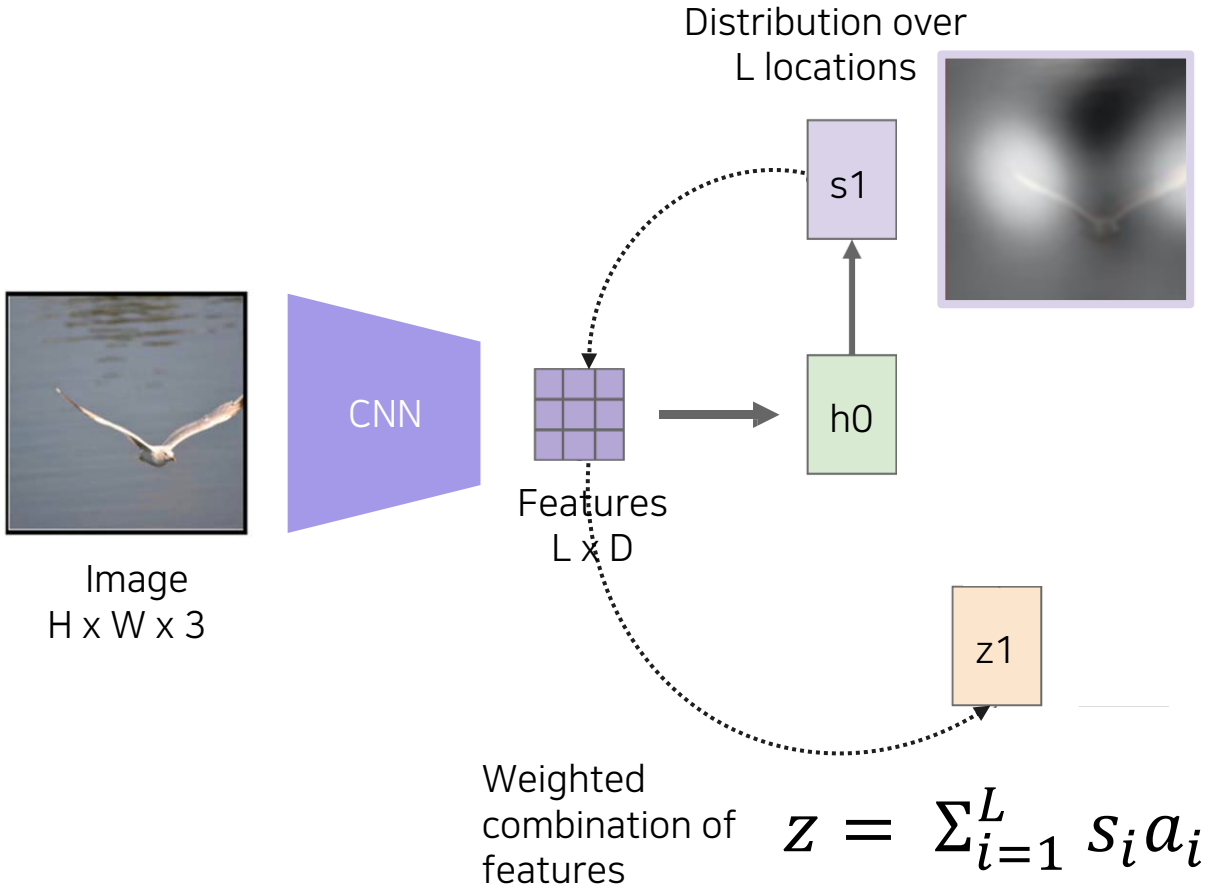


[Slide credit: Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n]

2.3 Cross modal translation

Show, attend, and tell - Inference

[Xu et al., ICML 2015]

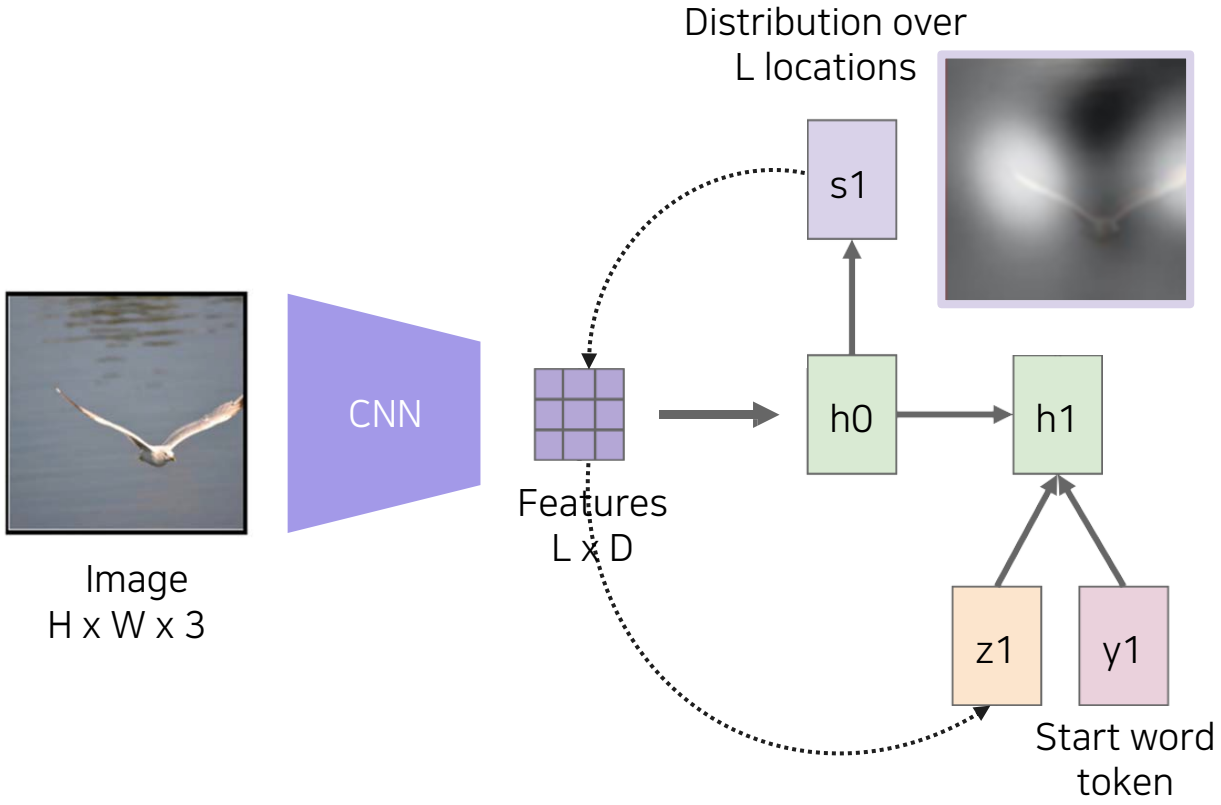


[Slide credit: Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n]

2.3 Cross modal translation

Show, attend, and tell - Inference

[Xu et al., ICML 2015]

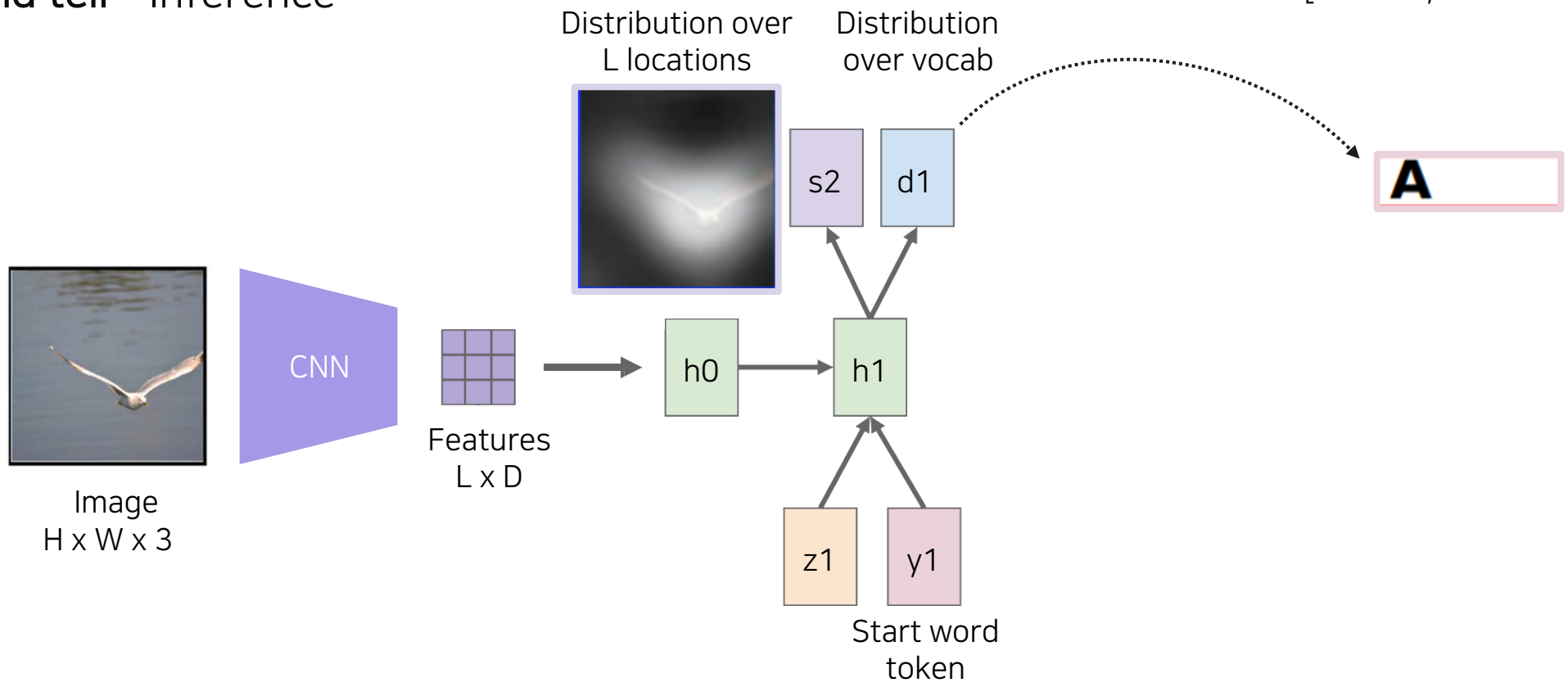


[Slide credit: Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n]

2.3 Cross modal translation

Show, attend, and tell - Inference

[Xu et al., ICML 2015]

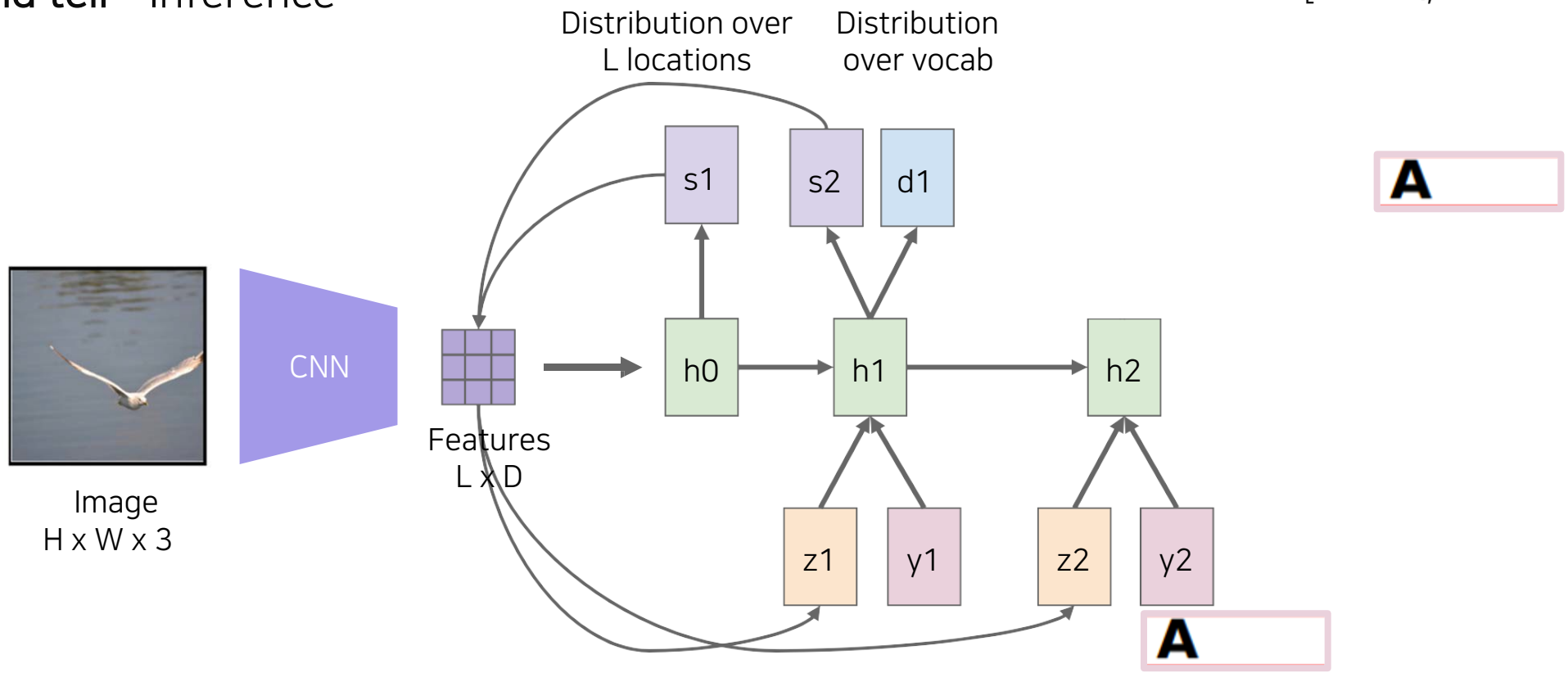


[Slide credit: Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n]

2.3 Cross modal translation

Show, attend, and tell - Inference

[Xu et al., ICML 2015]

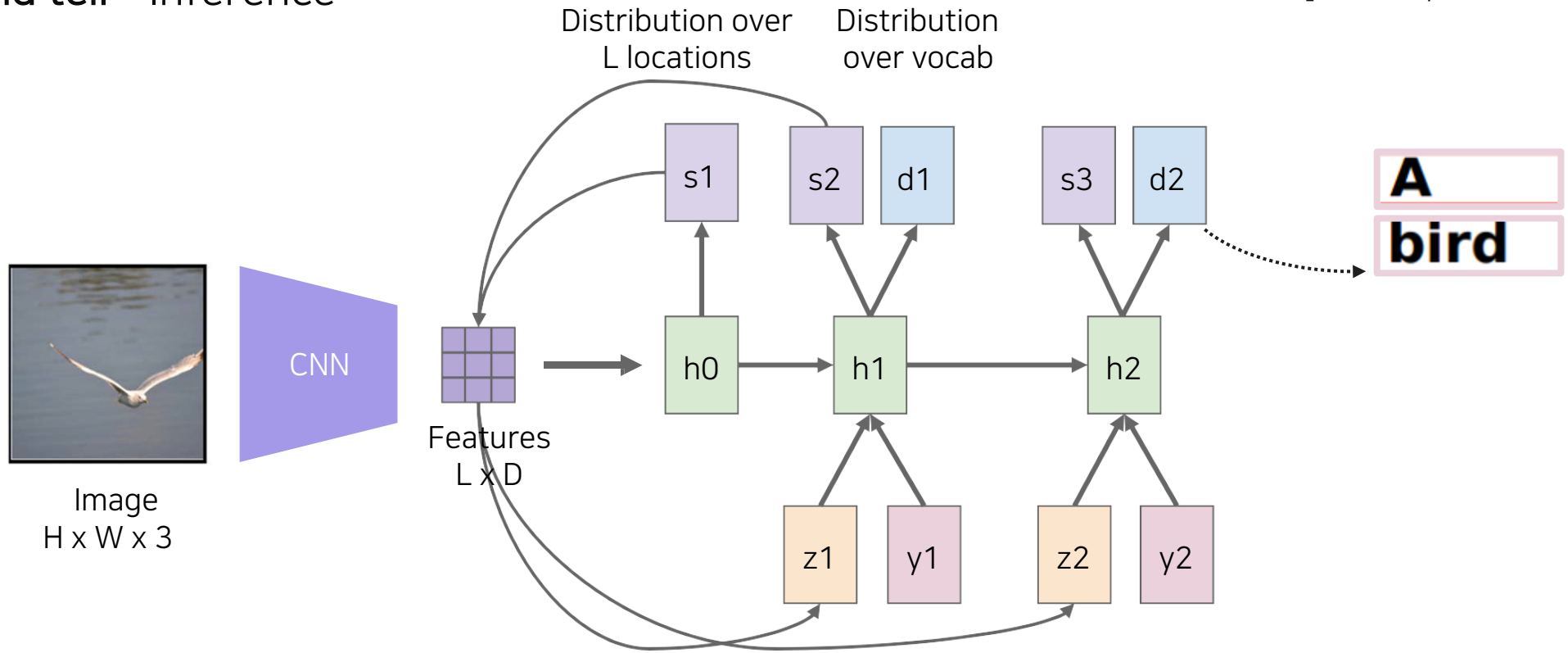


[Slide credit: Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n]

2.3 Cross modal translation

Show, attend, and tell - Inference

[Xu et al., ICML 2015]

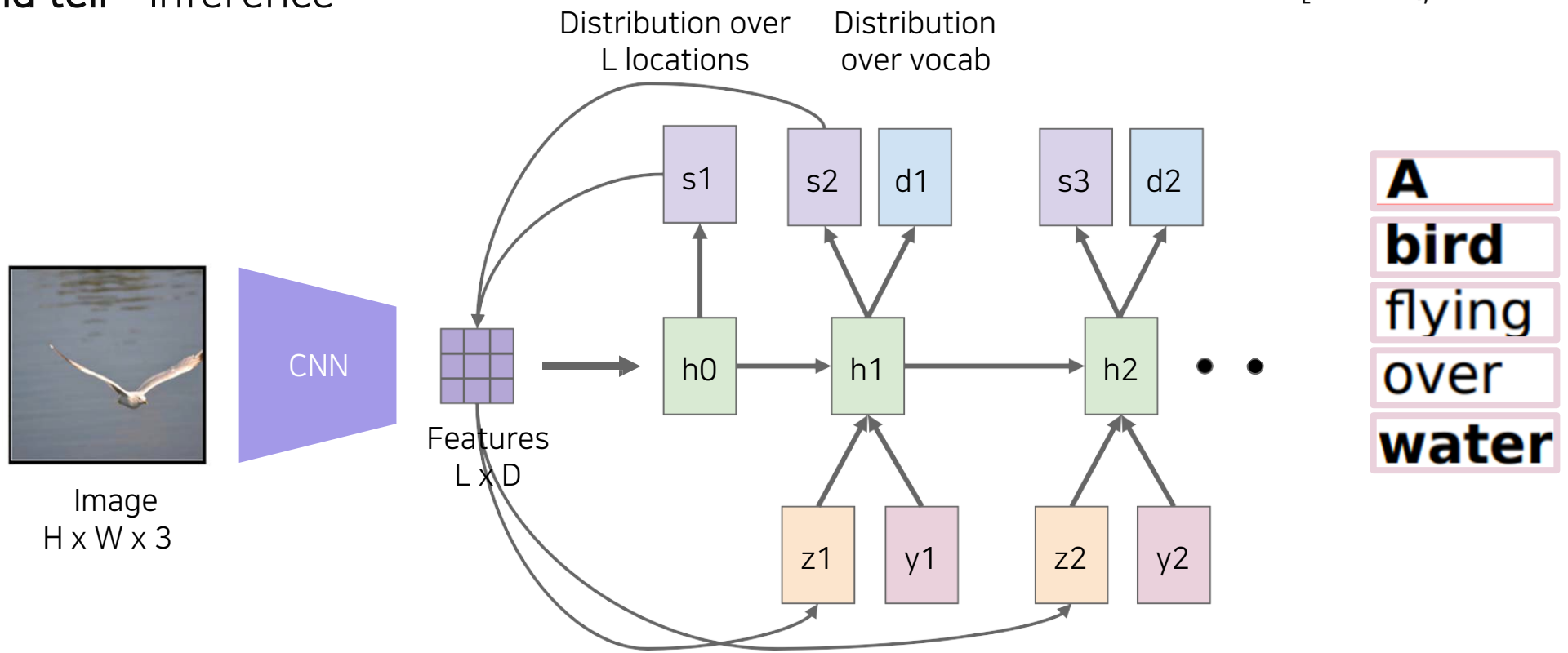


[Slide credit: Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n]

2.3 Cross modal translation

Show, attend, and tell - Inference

[Xu et al., ICML 2015]



[Slide credit: Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n]

2.3 Cross modal translation

Text-to-image by generative model - Example

[Reed et al., ICML 2016]

this small bird has a pink breast and crown, and black primaries and secondaries.

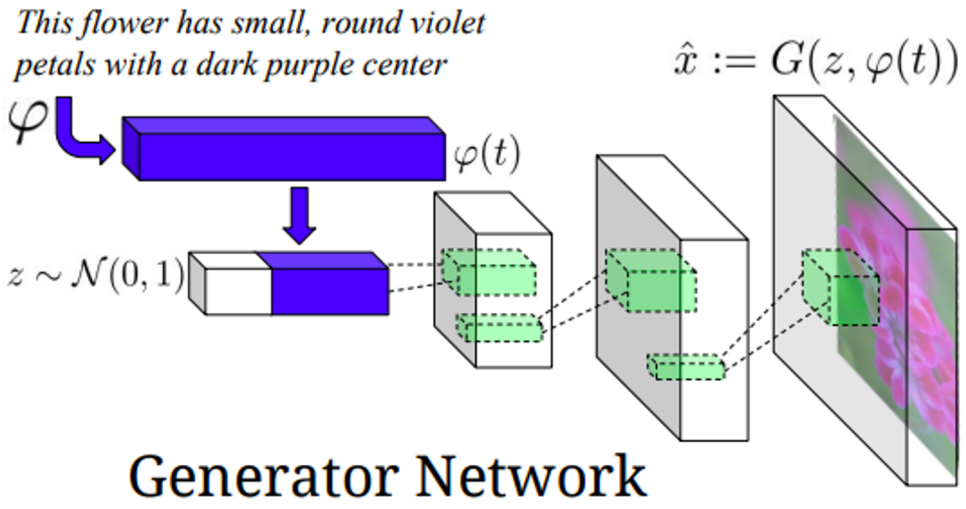
this magnificent fellow is almost all black with a red crest, and white cheek patch.

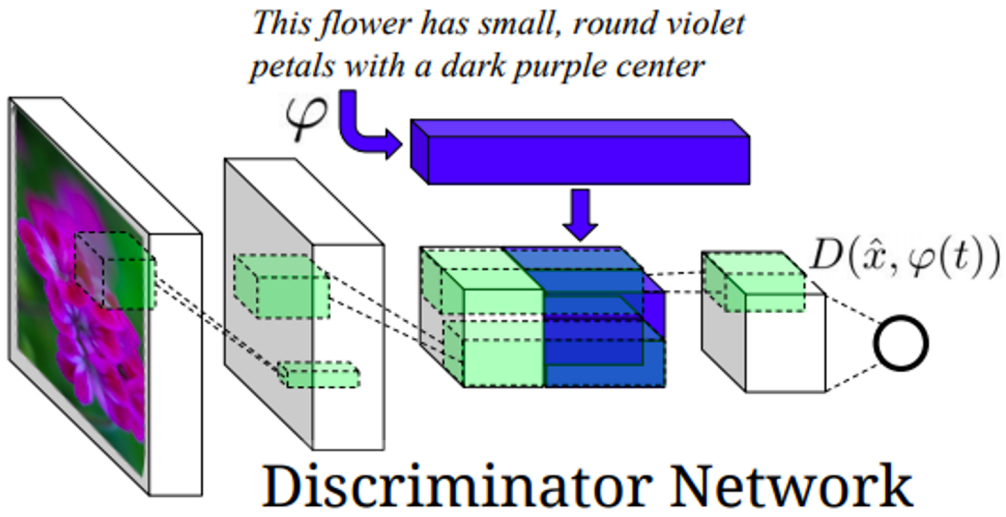
(Top) Query sentences and (Bottom) corresponding 6 generated images

2.3 Cross modal translation

Text-to-image by generative model

[Reed et al., ICML 2016]

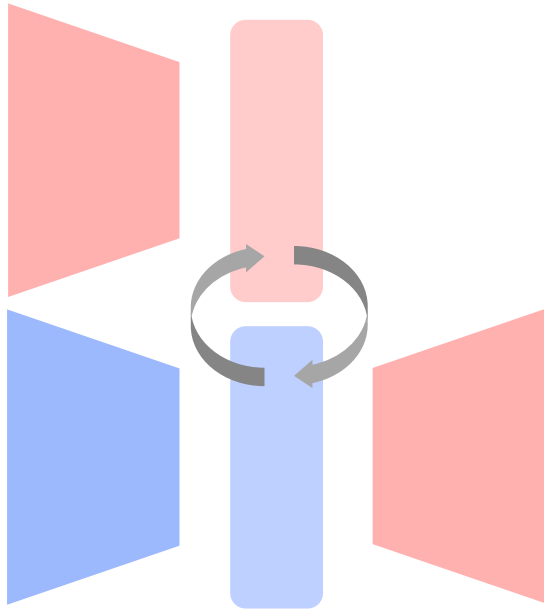




2.4 Cross modal reasoning

Matching

Translating

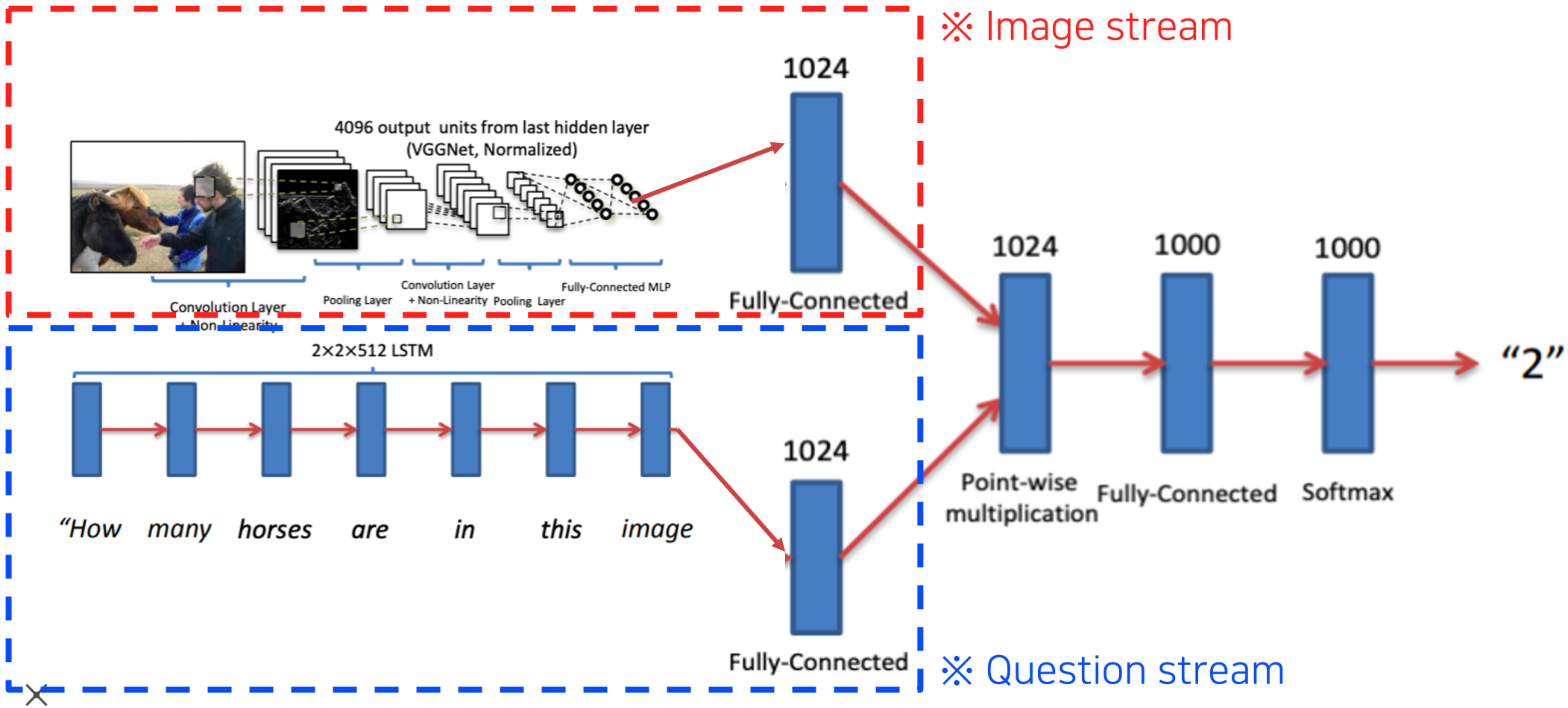


Referencing

2.4 Cross modal reasoning

Visual question answering – Multiple streams

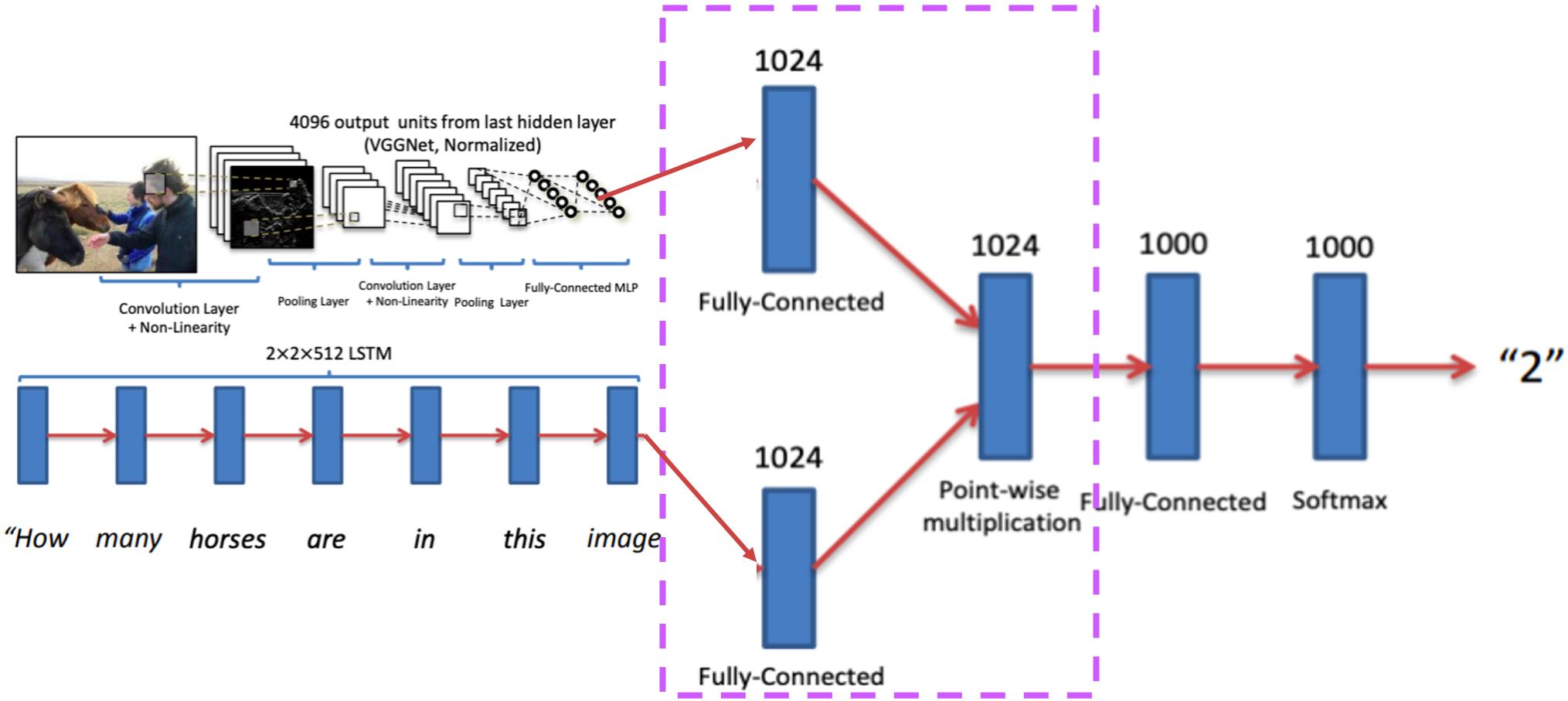
[Antol et al., ICCV 2015]



2.4 Cross modal reasoning

Visual question answering – Joint embedding

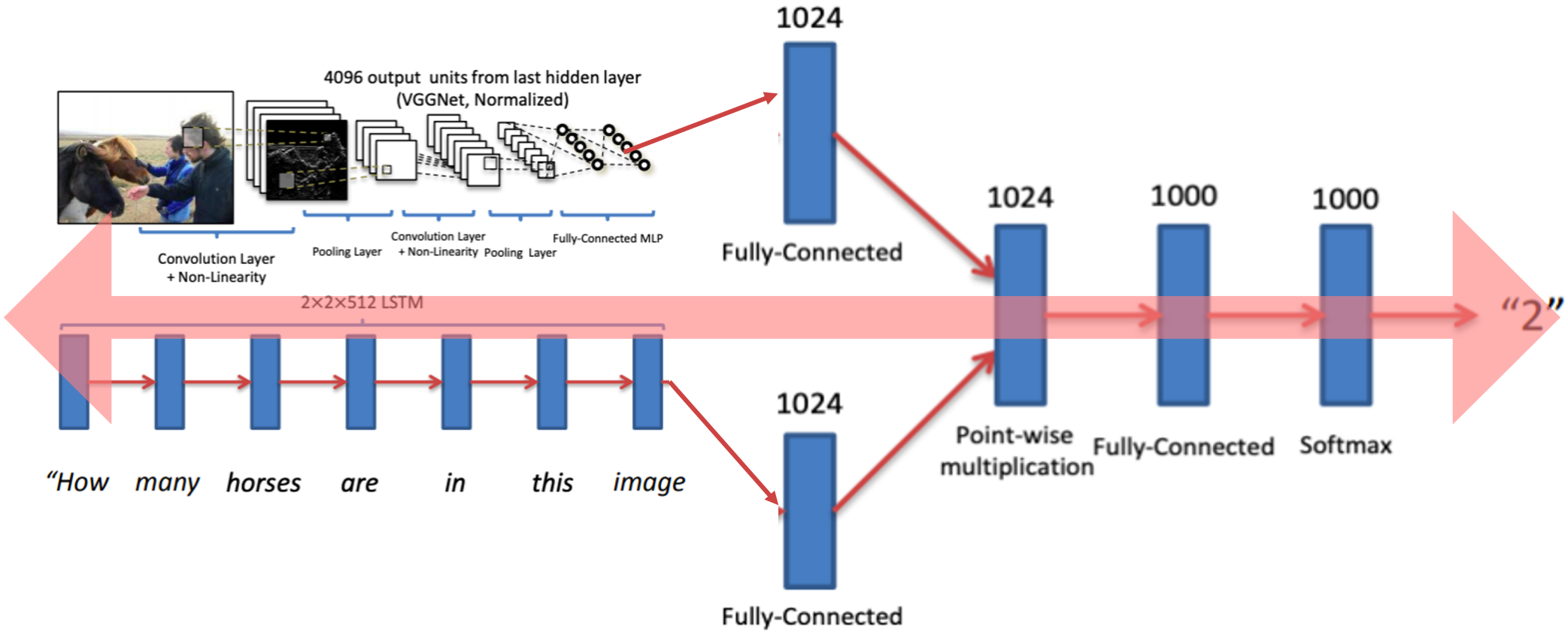
[Antol et al., ICCV 2015]



2.4 Cross modal reasoning

Visual question answering – End-to-end training

[Antol et al., ICCV 2015]



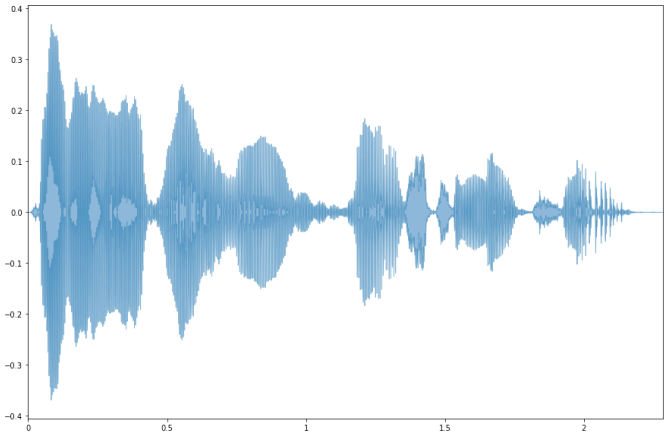
3.

Multi-modal tasks (2) - Visual data & Audio

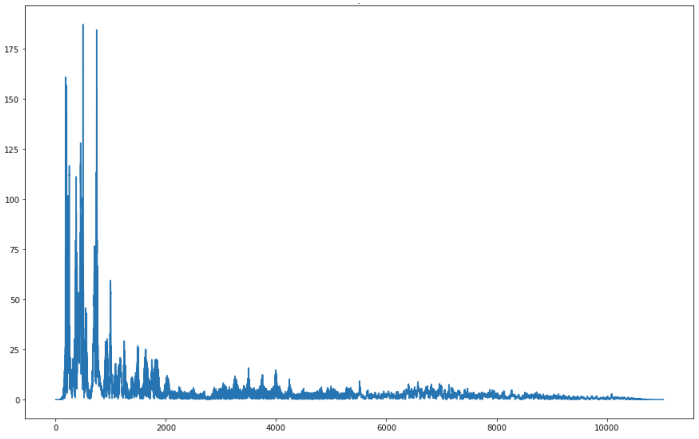
3.1 Sound representation

Sound representation

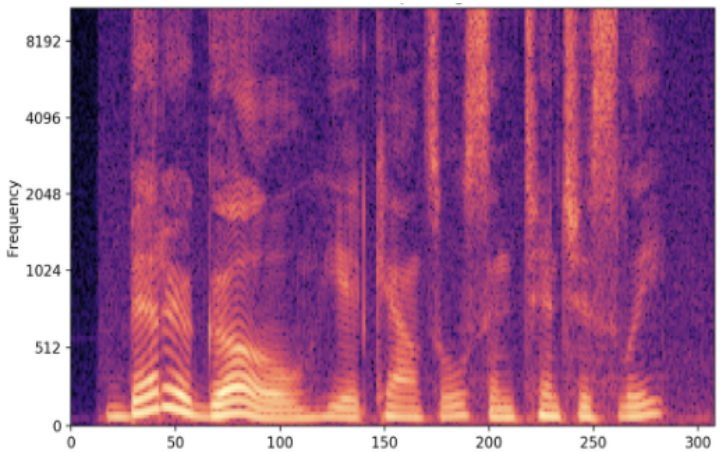
- Acoustic feature extraction from waveform to spectrogram



Waveform



Power spectrum



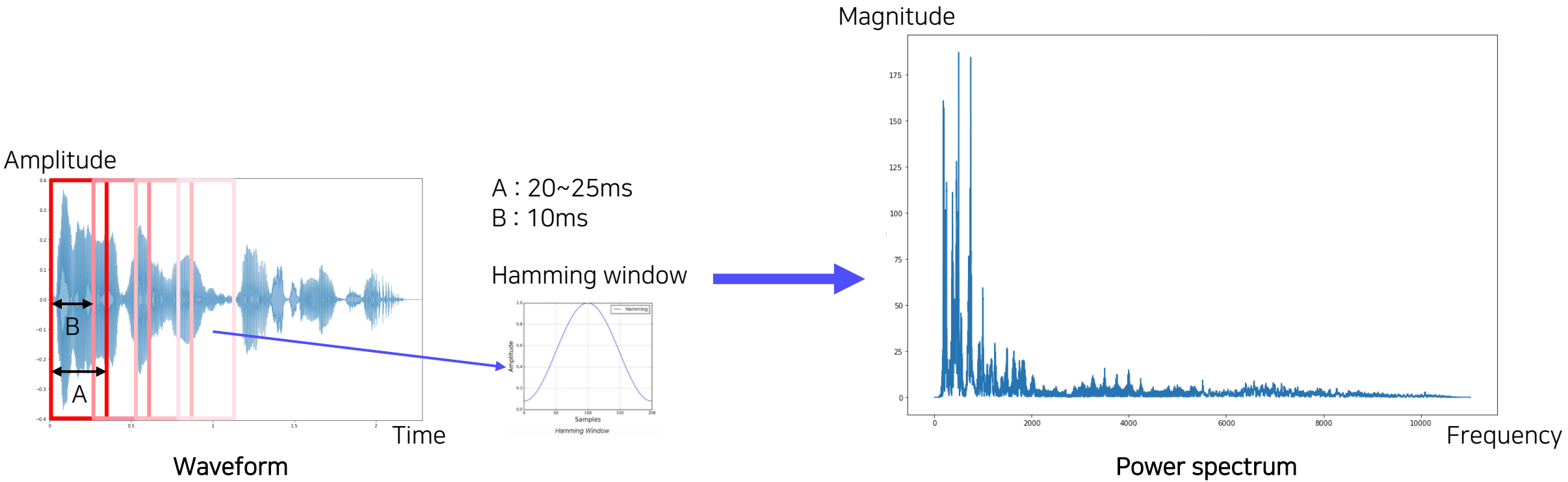
Spectrogram

[https://hyunlee103.tistory.com]

3.1 Sound representation

Sound representation – Fourier transform

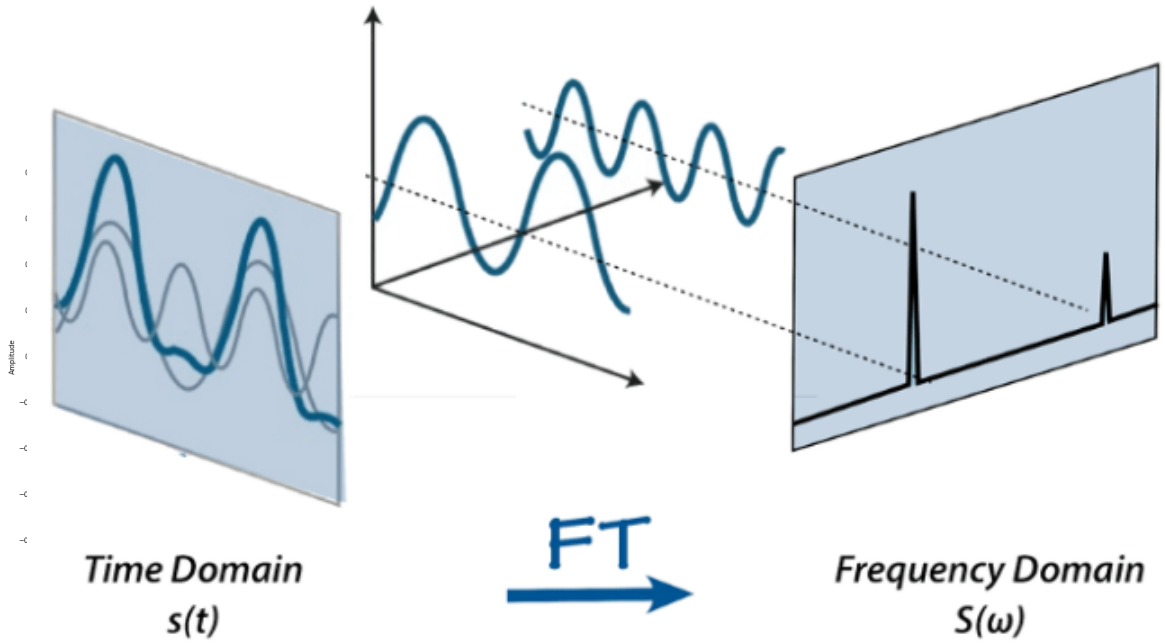
- Short-time Fourier transform (STFT)
  : Fourier transform (FT) on windowed waveform results in frequency-magnitude graph

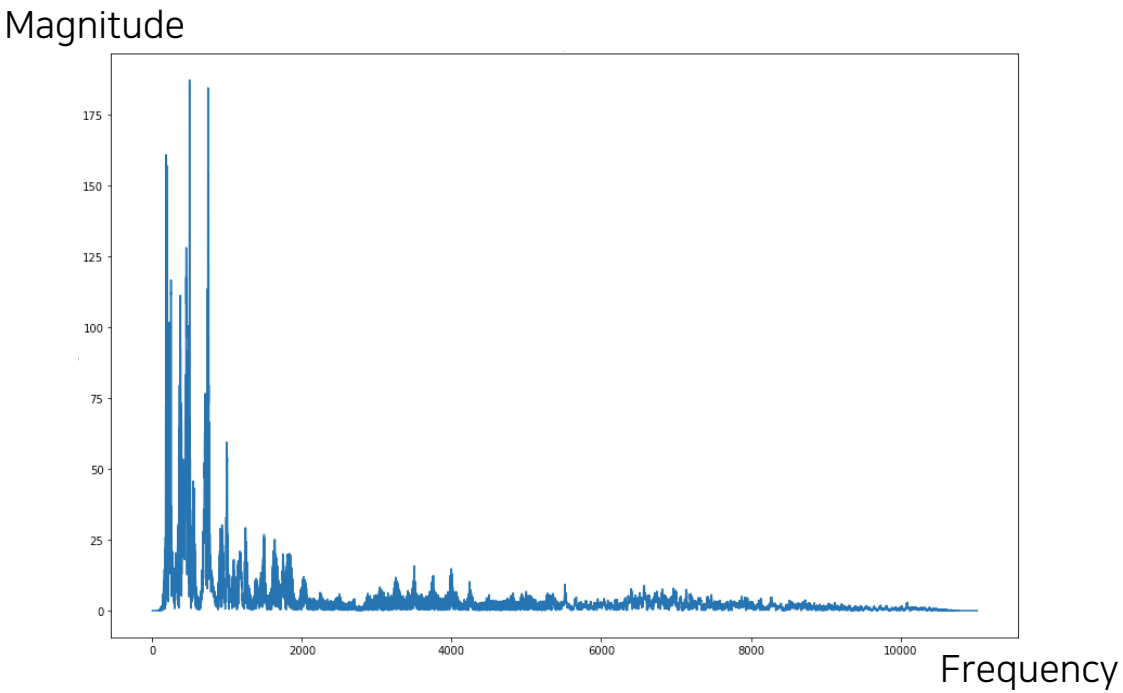


[https://hyunlee103.tistory.com]

Sound representation – Fourier transform

- FT decomposes an input signal into constituent frequencies





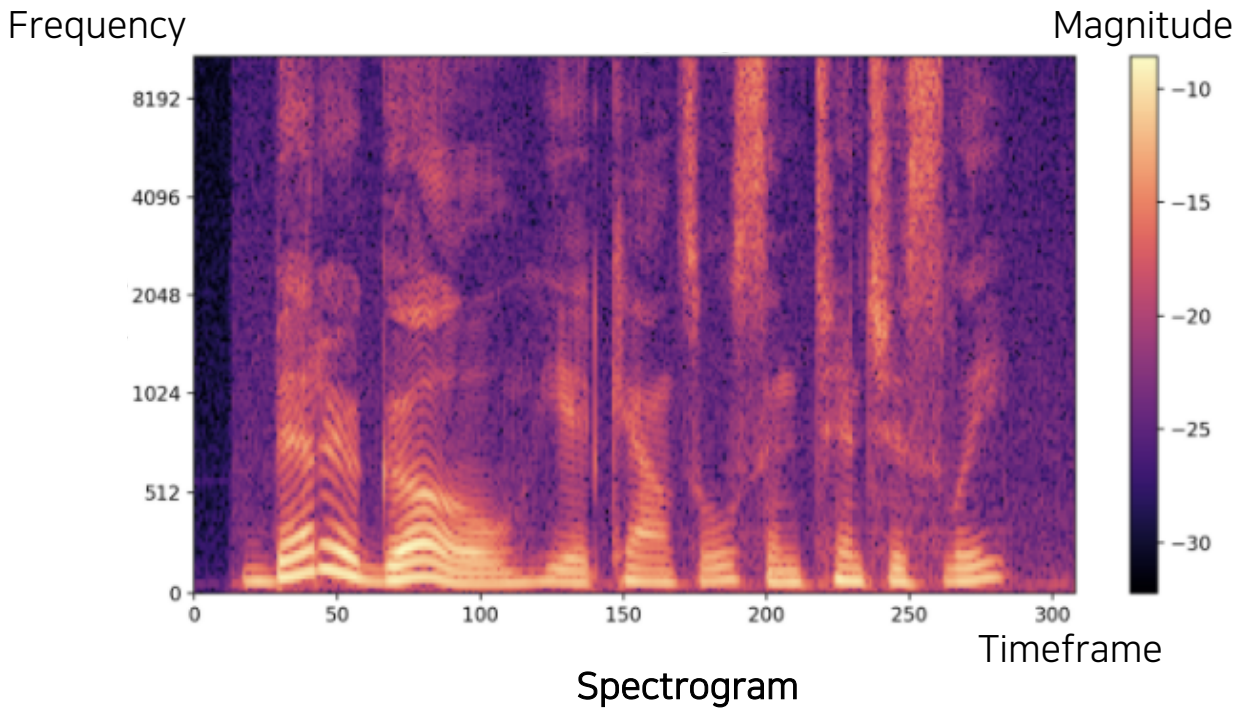
Power spectrum

[https://towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520]

3.1 Sound representation

Sound representation - Spectrogram

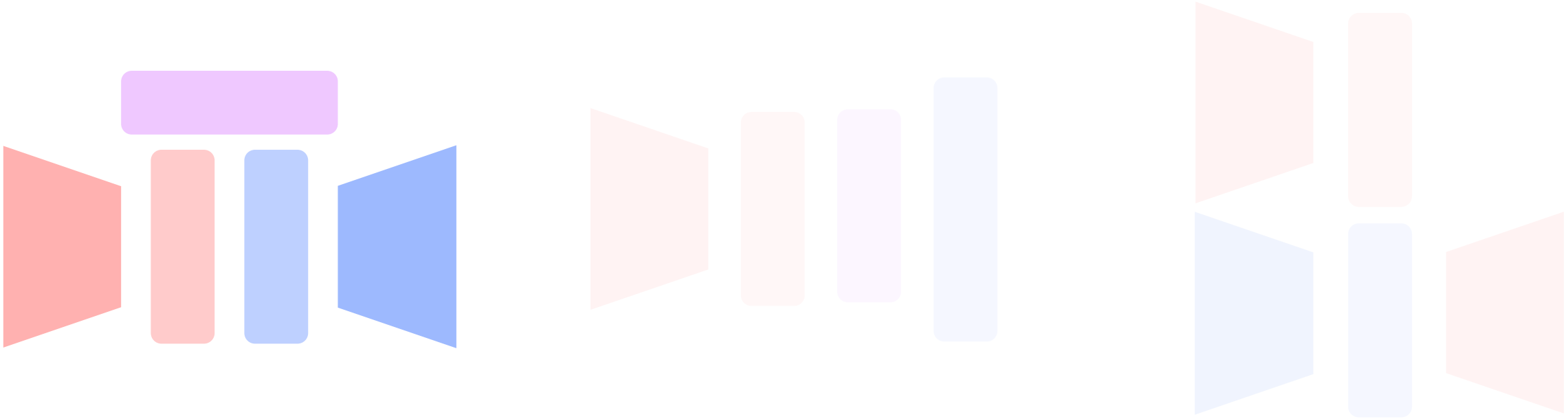
- Spectrogram: A stack of spectrums along the time axis



Spectrogram

[https://towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520]

3.2 Joint embedding



Matching

Translating

Referencing

3.2 Joint embedding

Application - Scene recognition by sound

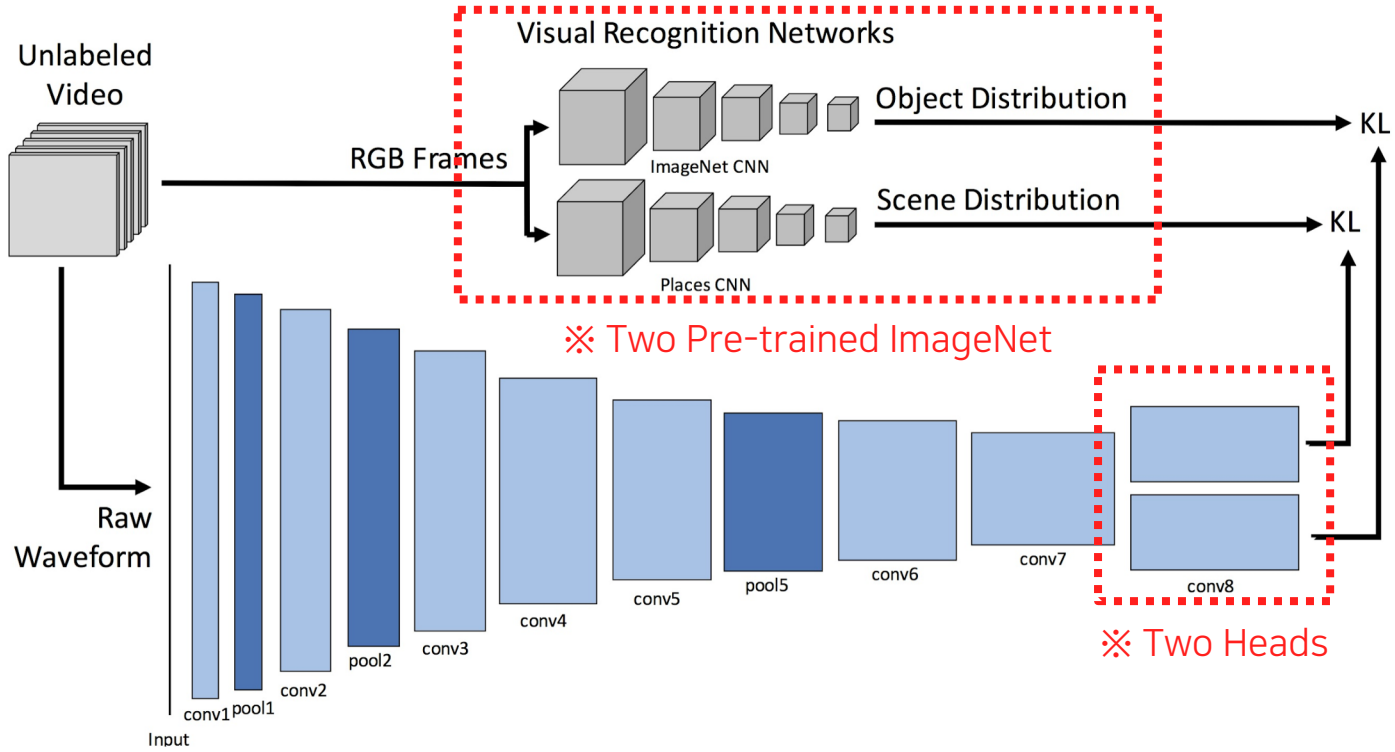
[Aytar et al., NIPS 2016]

3.2 Joint embedding

SoundNet

[Aytar et al., NIPS 2016]

- Learn audio representation from synchronized RGB frames in the same videos
- Train by the teacher-student manner
  - Transfer visual knowledge from pre-trained visual recognition models into sound modality

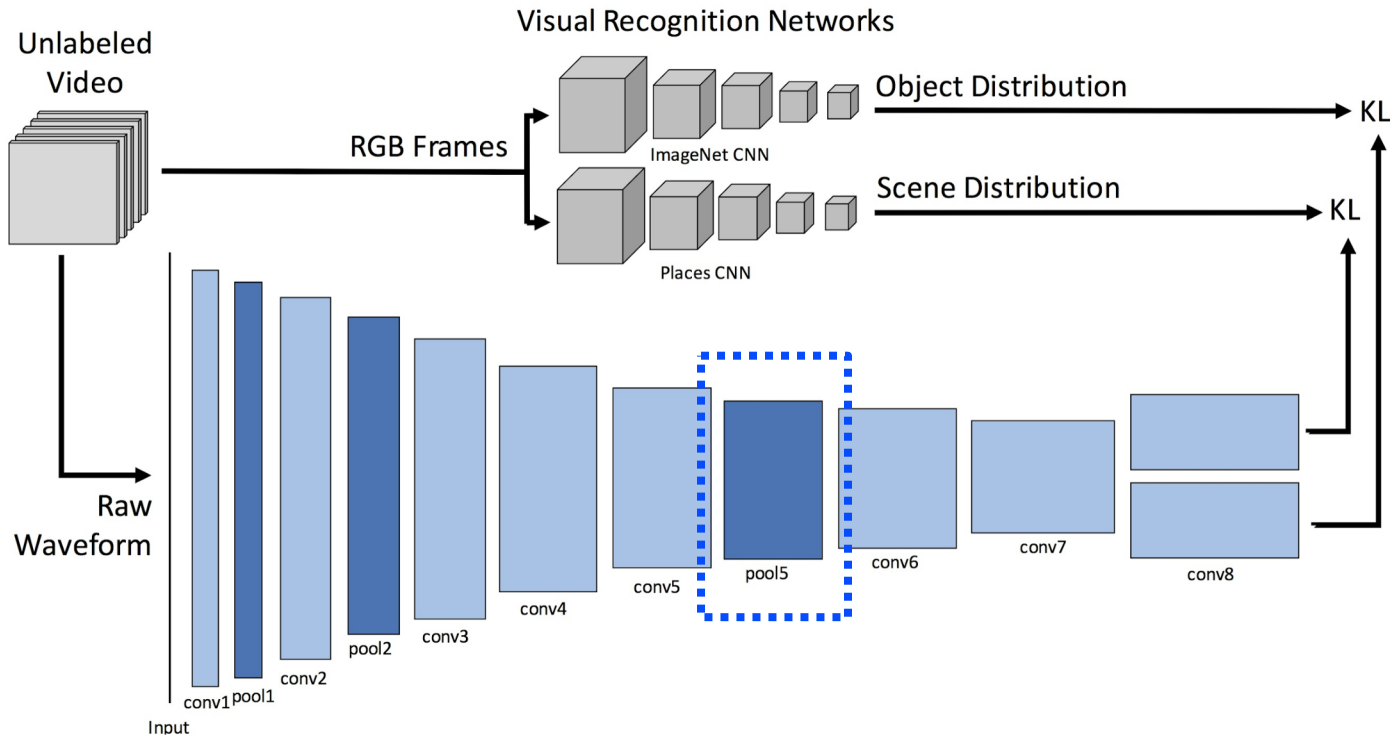


3.2 Joint embedding

SoundNet

[Aytar et al., NIPS 2016]

- For a target task, the pre-trained internal representation (pool5) is used as features
- Training a classifier with the pool5 feature
  - Instead of the output layer, the pool5 feature posses more generalizable semantic info.



3.3 Cross modal translation

Matching

Translating

Referencing

3.3 Cross modal translation

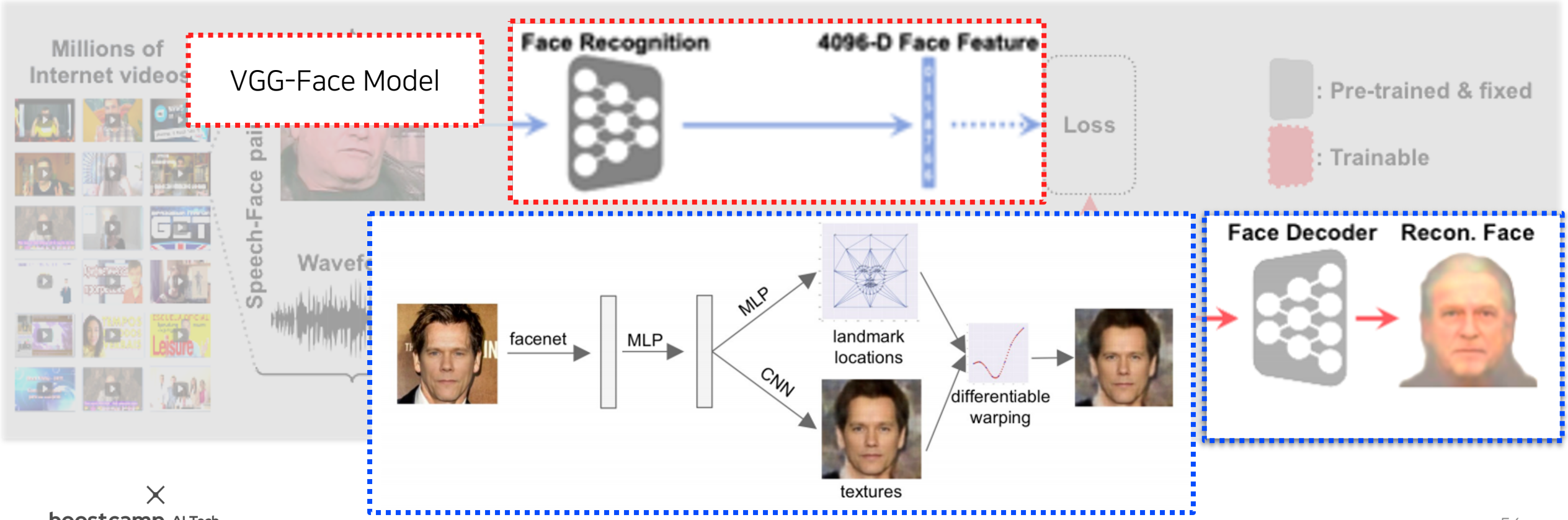
Speech2Face

[Oh et al., CVPR 2019]

3.3 Cross modal translation

Speech2Face – Module networks

[Oh et al., CVPR 2019]



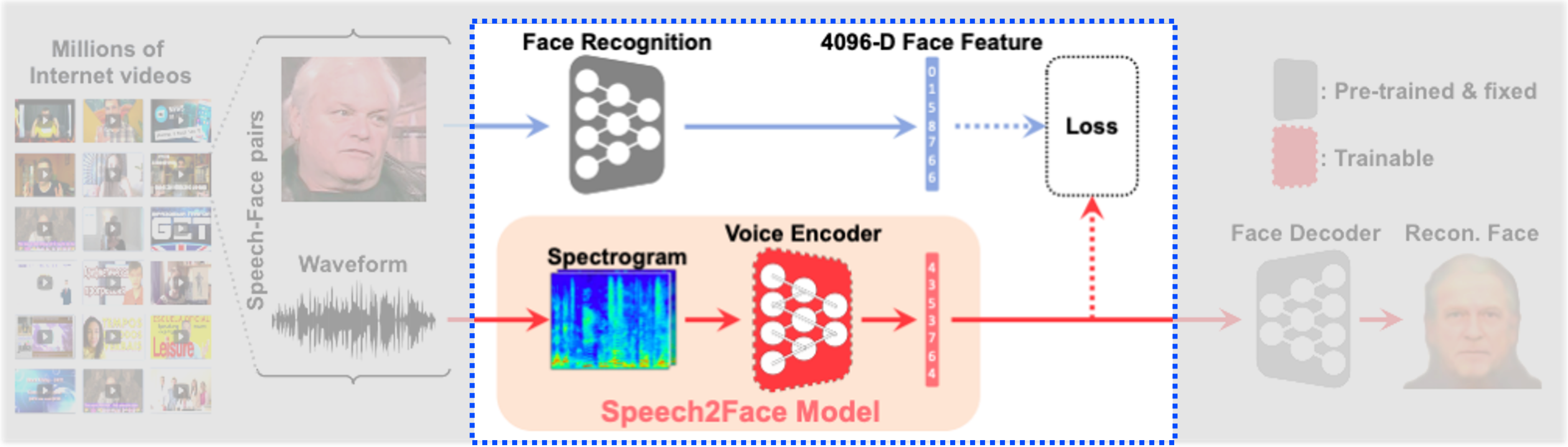
[Cole et al., CVRP 2017]

3.3 Cross modal translation

Speech2Face – Training

[Oh et al., CVPR 2019]

- Training by feature matching loss (self-supervised manner) for making features compatible
  - Natural co-occurrence of speaker's speech and facial images

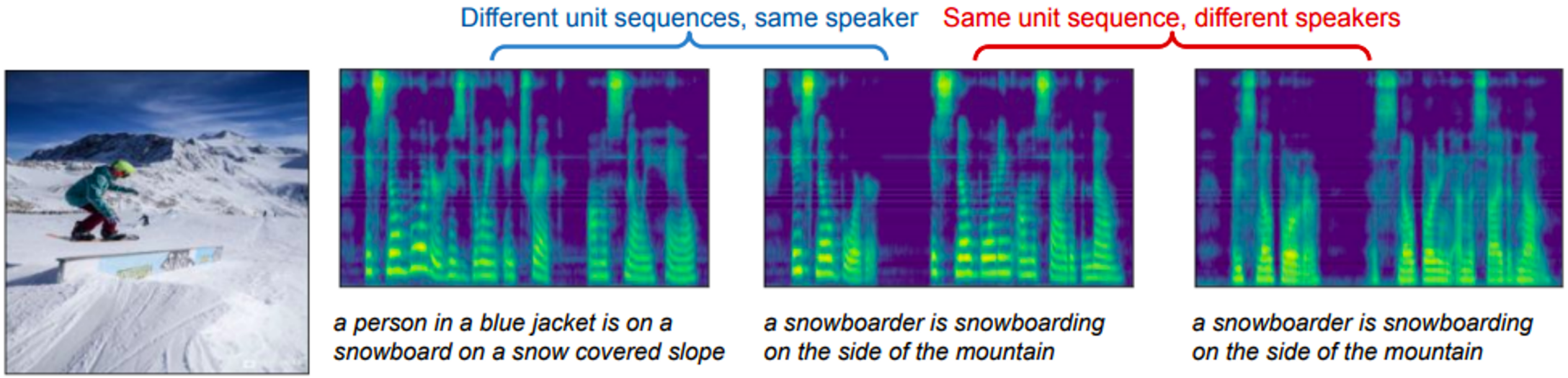


[F. , Cole et al., Synthesizing Normalized Faces from Facial Identity Features, CVRP 2017]

3.3 Cross modal translation

Application - Image-to-speech synthesis

[Hsu et al., arXiv 2020]



Succeed with discrete unit representation and exhibit robustness to different speakers

3.3 Cross modal translation

Image-to-speech synthesis – Module networks

[Hsu et al., arXiv 2020]

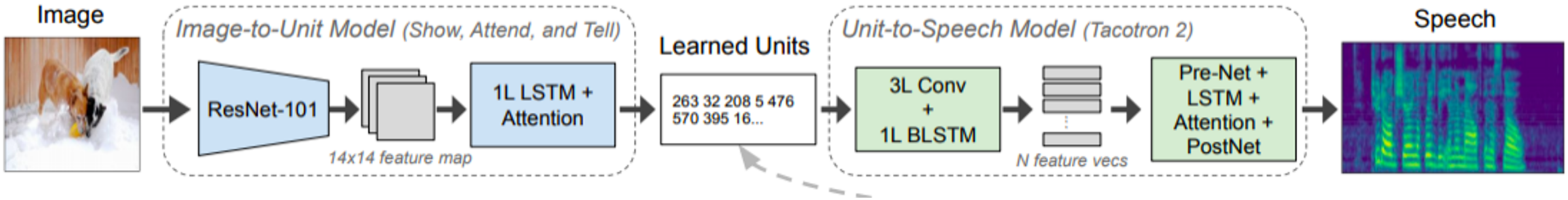


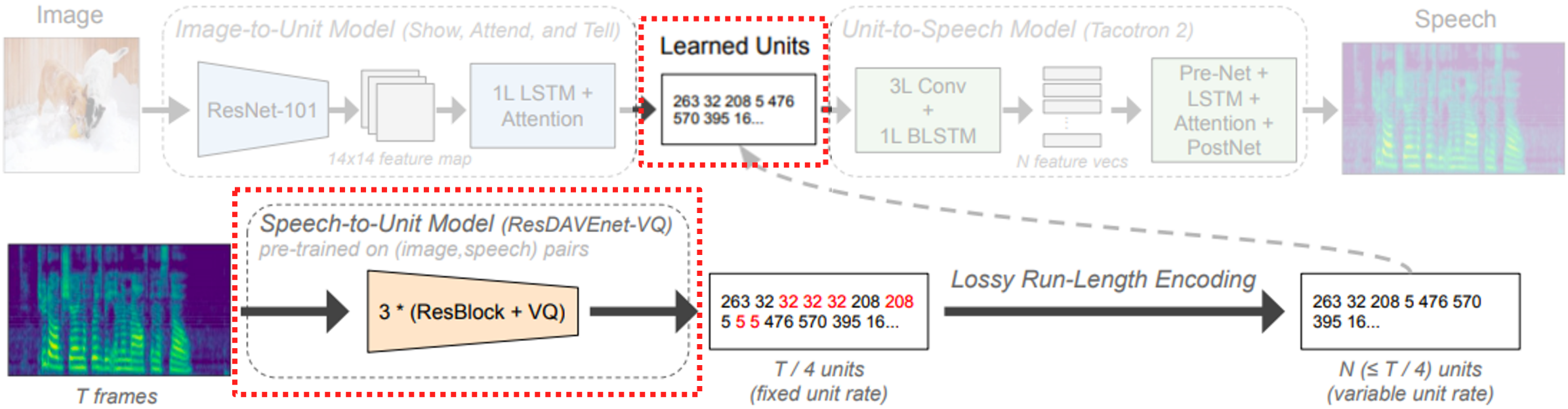
Image captioning,
but <u>sub-word units</u> not natural language

Speech synthesis

3.3 Cross modal translation

Image-to-speech synthesis – Module networks

[Hsu et al., arXiv 2020]

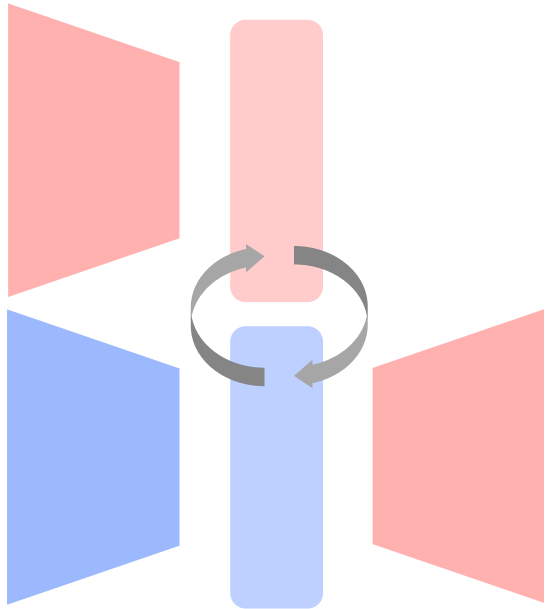


Associate speech with contextually relevant visual inputs using a triplet loss

3.4 Cross modal reasoning

Matching

Translating

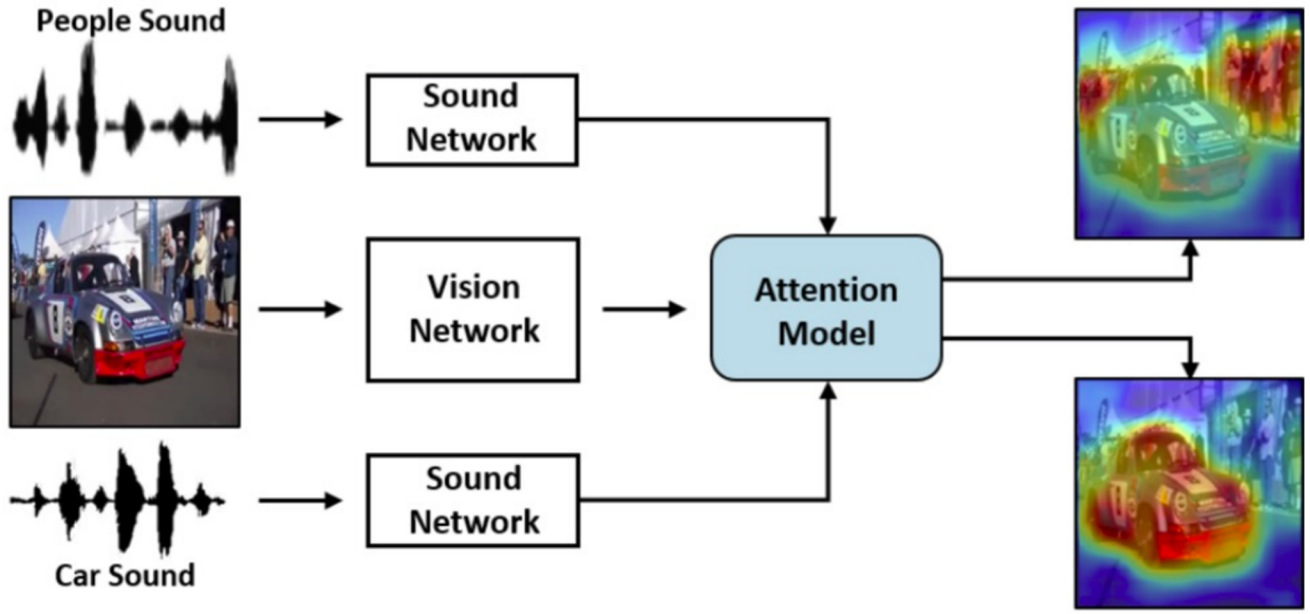


Referencing

3.4 Cross modal reasoning

Application - Sound source localization

[Senocak et al. CVPR 2018]



3.4 Cross modal reasoning

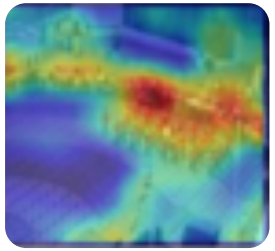
Sound source localization

[Senocak et al. CVPR 2018]

Image

Audio

×

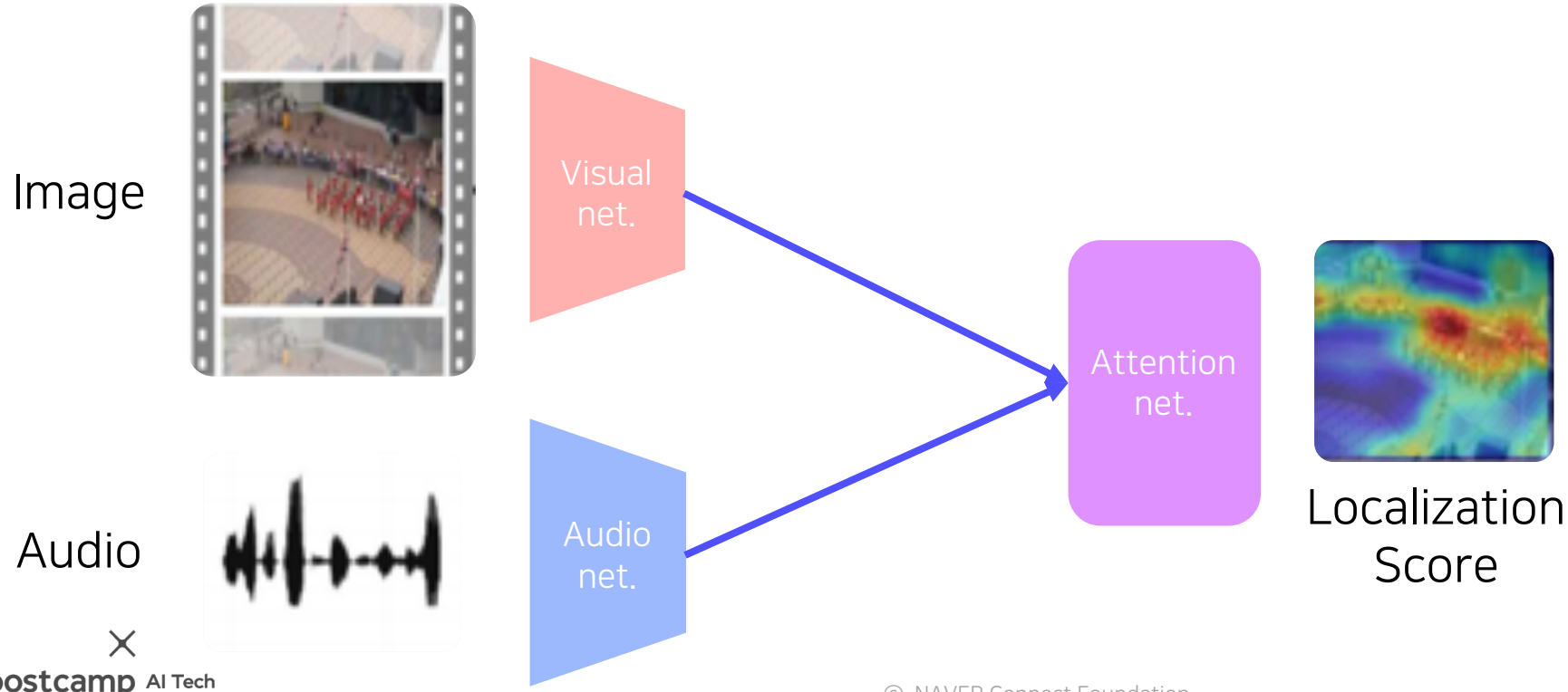


Localization Score

3.4 Cross modal reasoning

Sound source localization

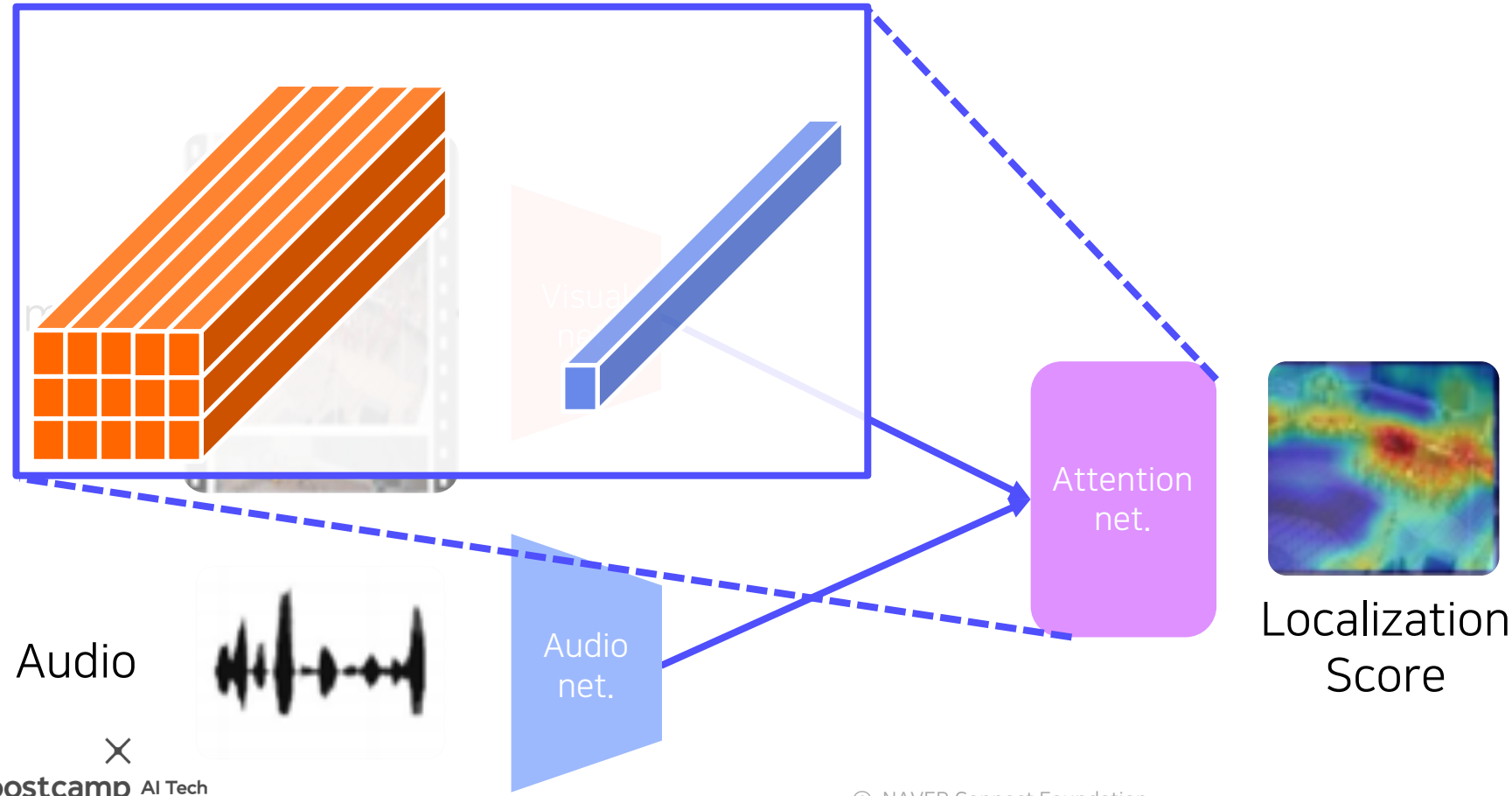
[Senocak et al. CVPR 2018]



3.4 Cross modal reasoning

Sound source localization

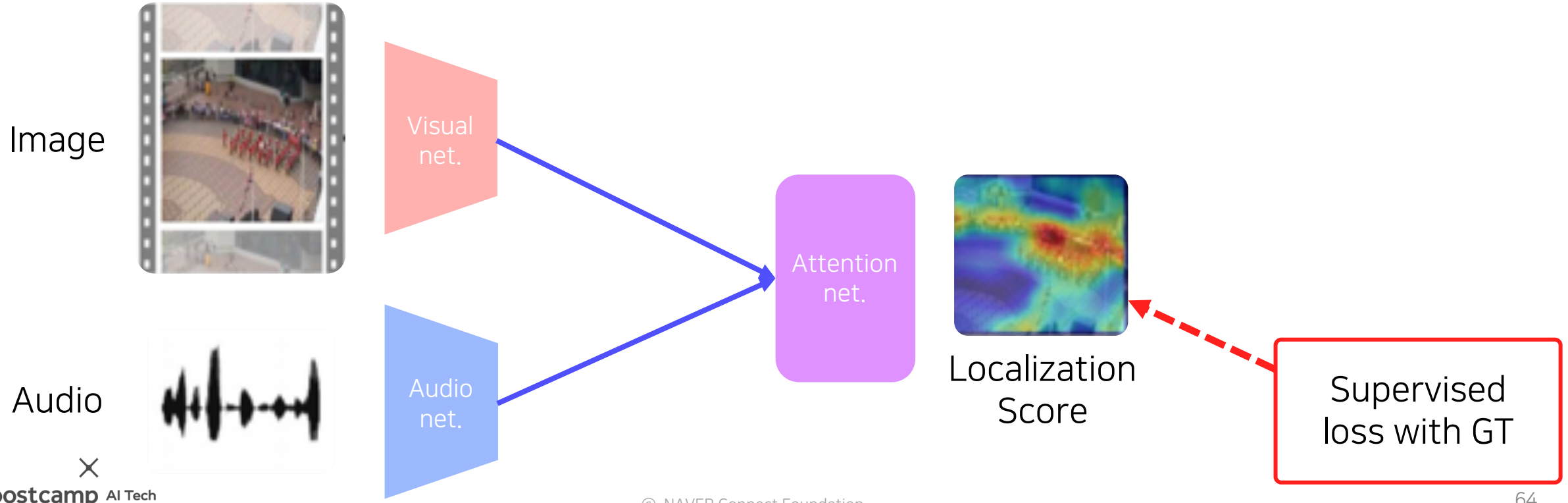
[Senocak et al. CVPR 2018]



3.4 Cross modal reasoning

Sound source localization – Fully supervised version

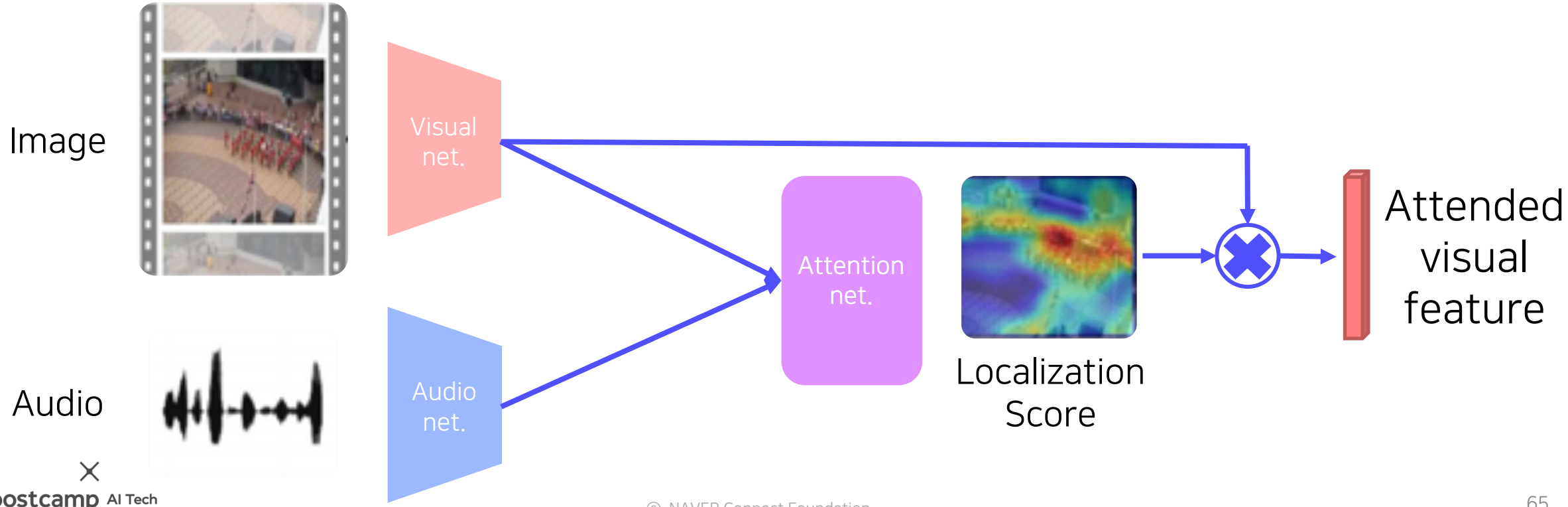
[Senocak et al. CVPR 2018]



3.4 Cross modal reasoning

Sound source localization – Unsupervised version

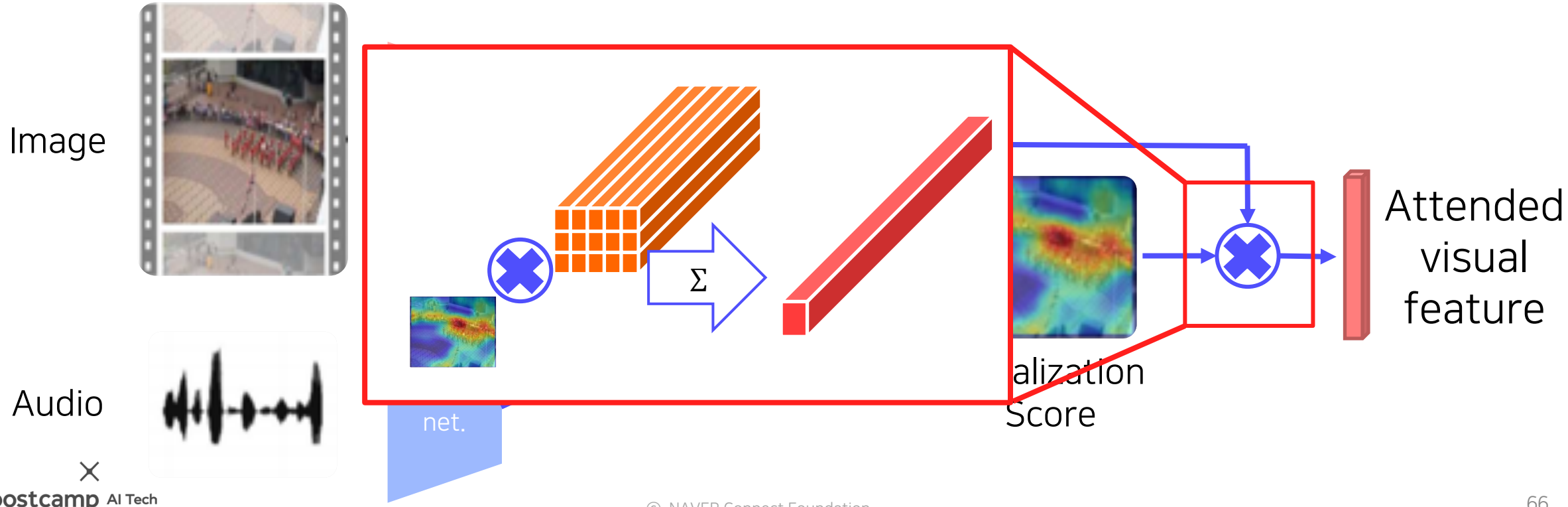
[Senocak et al. CVPR 2018]



3.4 Cross modal reasoning

Sound source localization – Unsupervised version

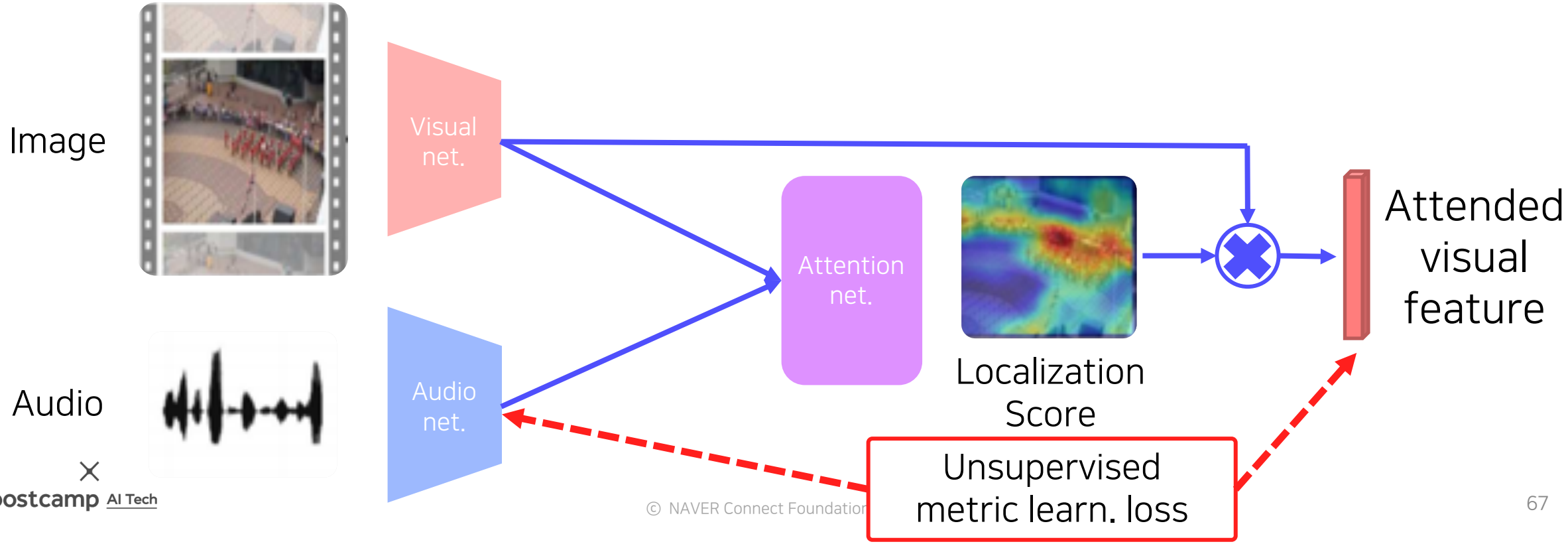
[Senocak et al. CVPR 2018]



3.4 Cross modal reasoning

Sound source localization – Unsupervised version

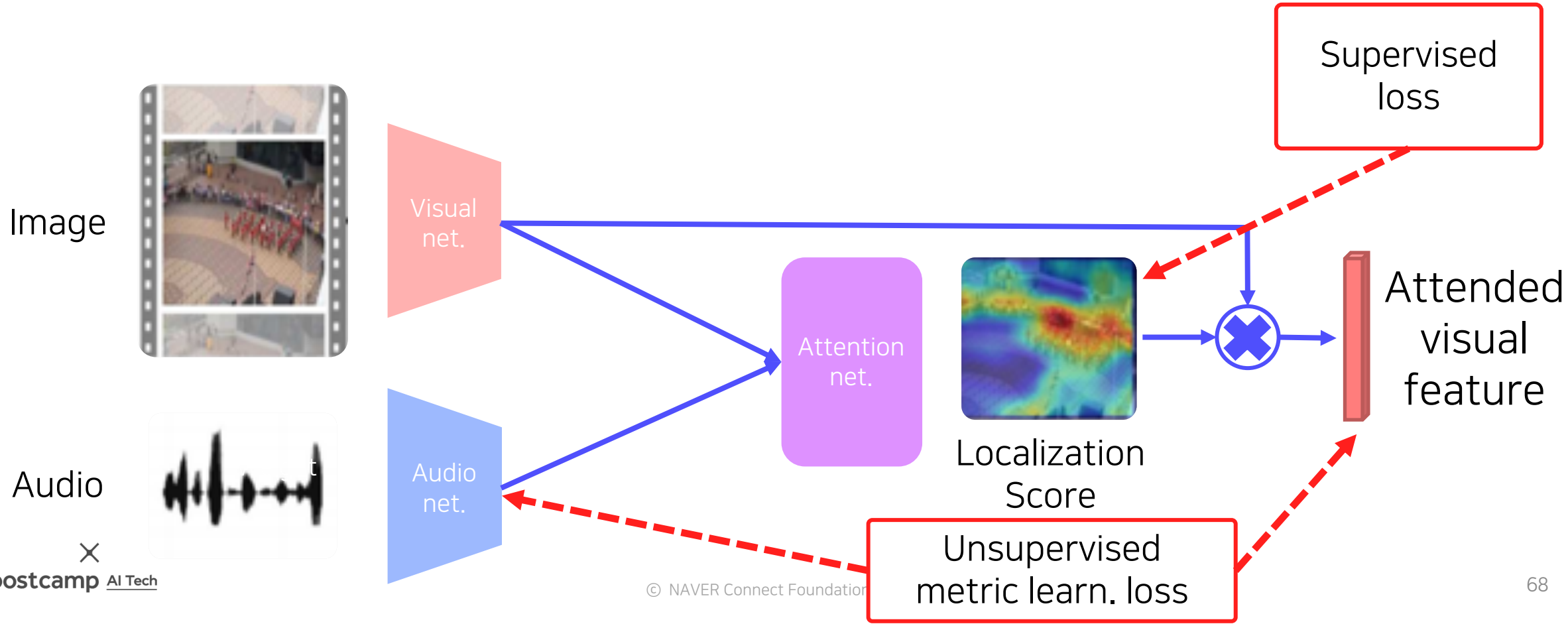
[Senocak et al. CVPR 2018]



3.4 Cross modal reasoning

Sound source localization – Semi-supervised version

[Senocak et al. CVPR 2018]



3.4 Cross modal reasoning

Application – Speech separation: Looking to listen at the cocktail party

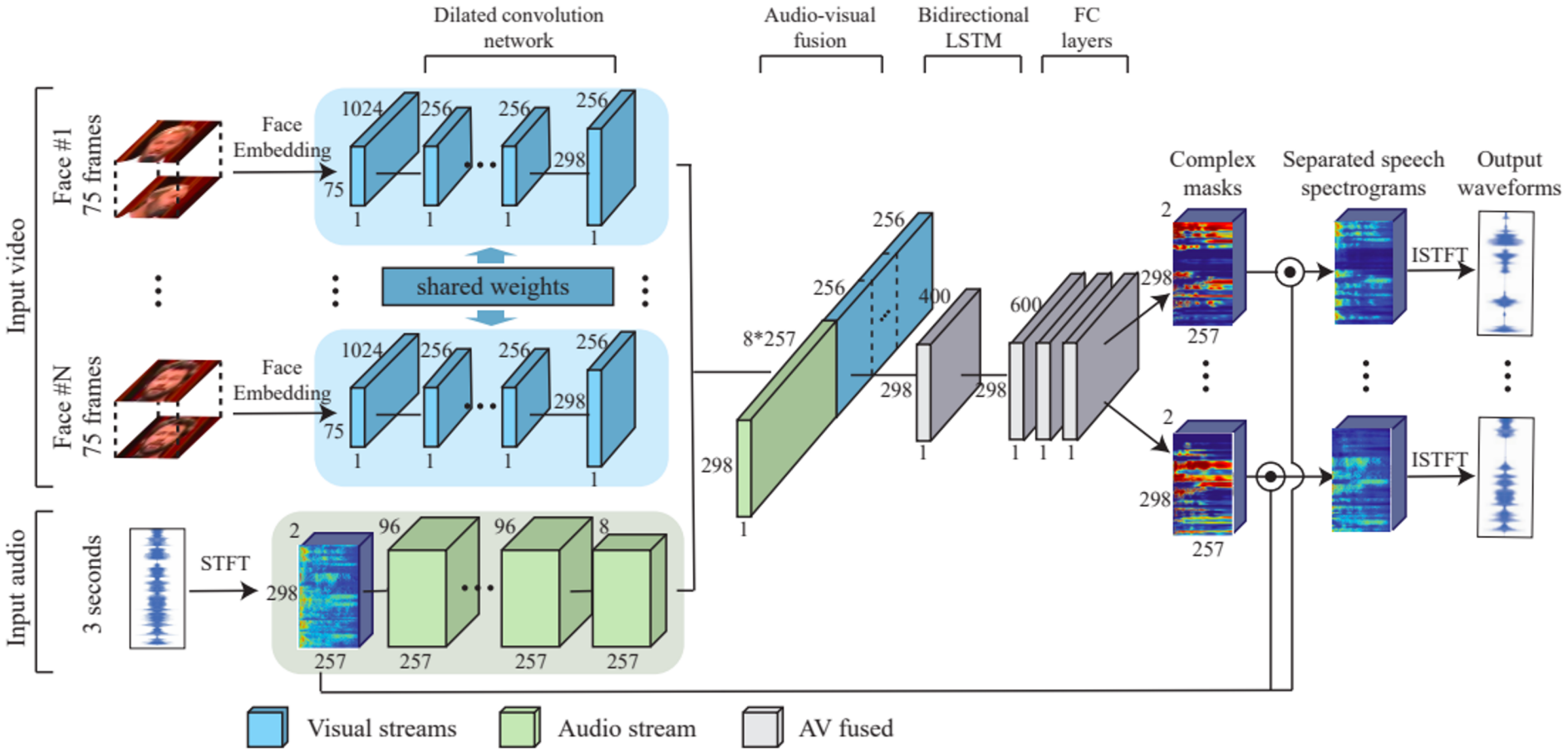
[Ephrat et al., SIGGRAPH 2018]

[https://www.youtube.com/watch?v=rVQVAPiJWKU]

3.4 Cross modal reasoning

Looking to listen at the cocktail party

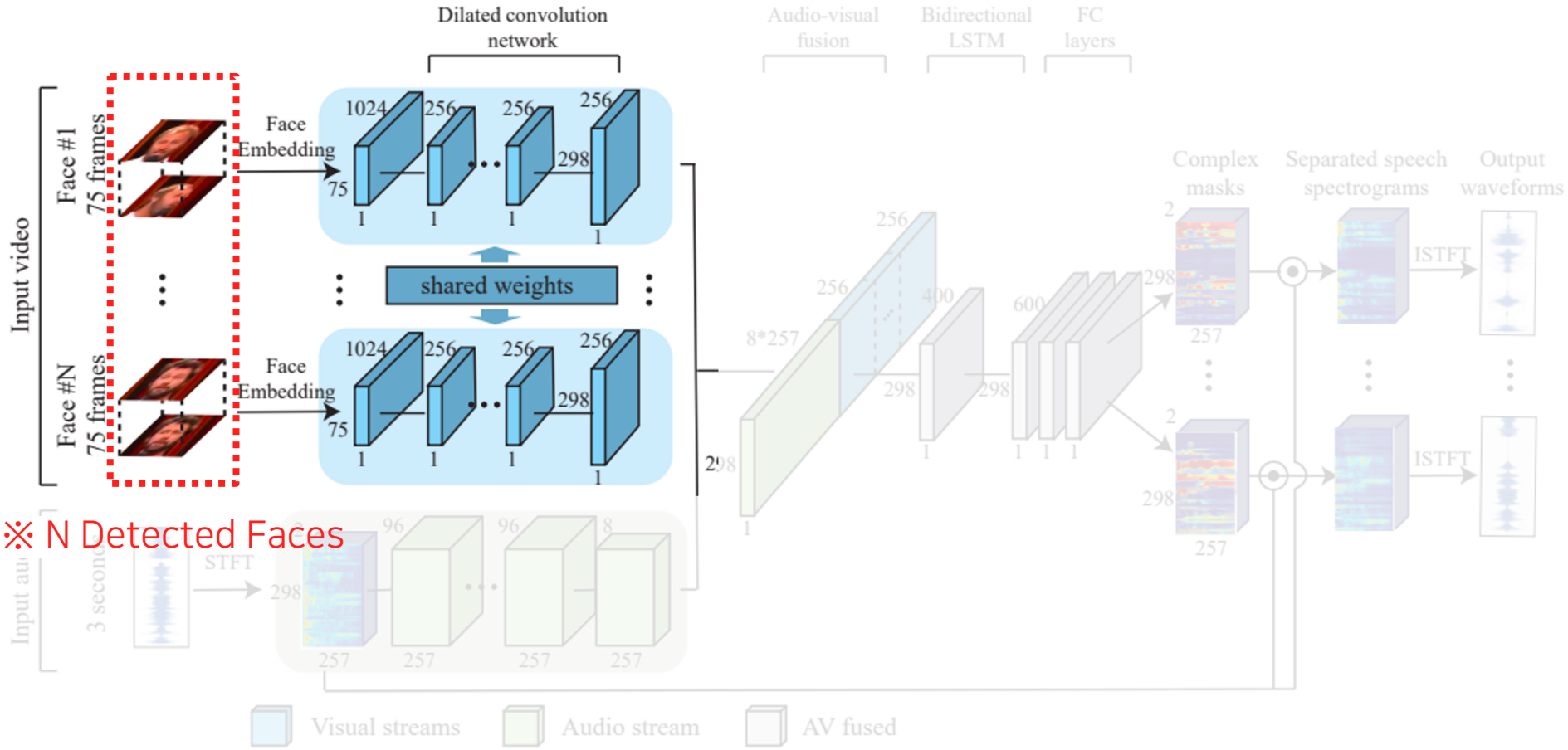
[Ephrat et al., SIGGRAPH 2018]



3.4 Cross modal reasoning

Looking to listen at the cocktail party – Visual stream

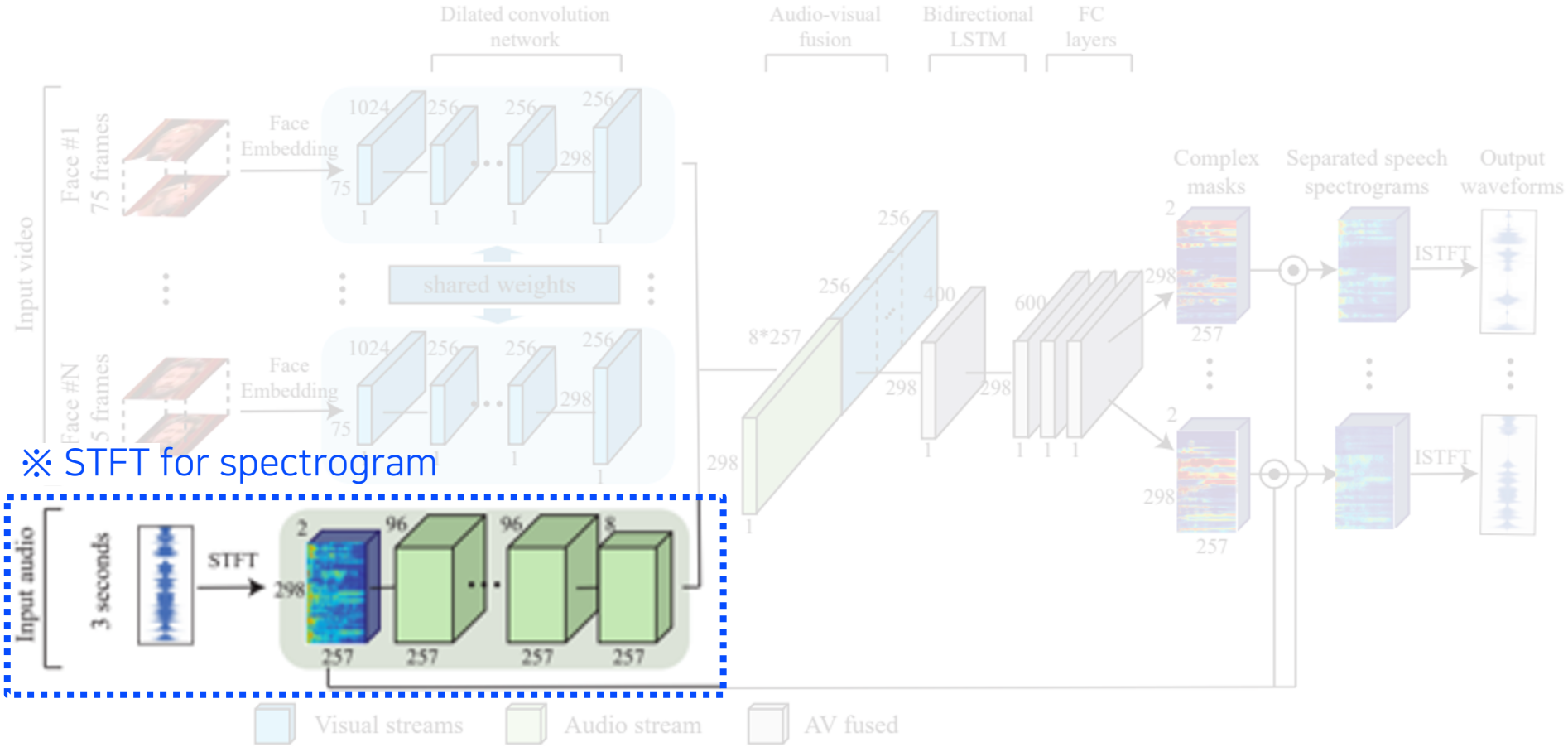
[Ephrat et al., SIGGRAPH 2018]



3.4 Cross modal reasoning

Looking to listen at the cocktail party – Audio stream

[Ephrat et al., SIGGRAPH 2018]

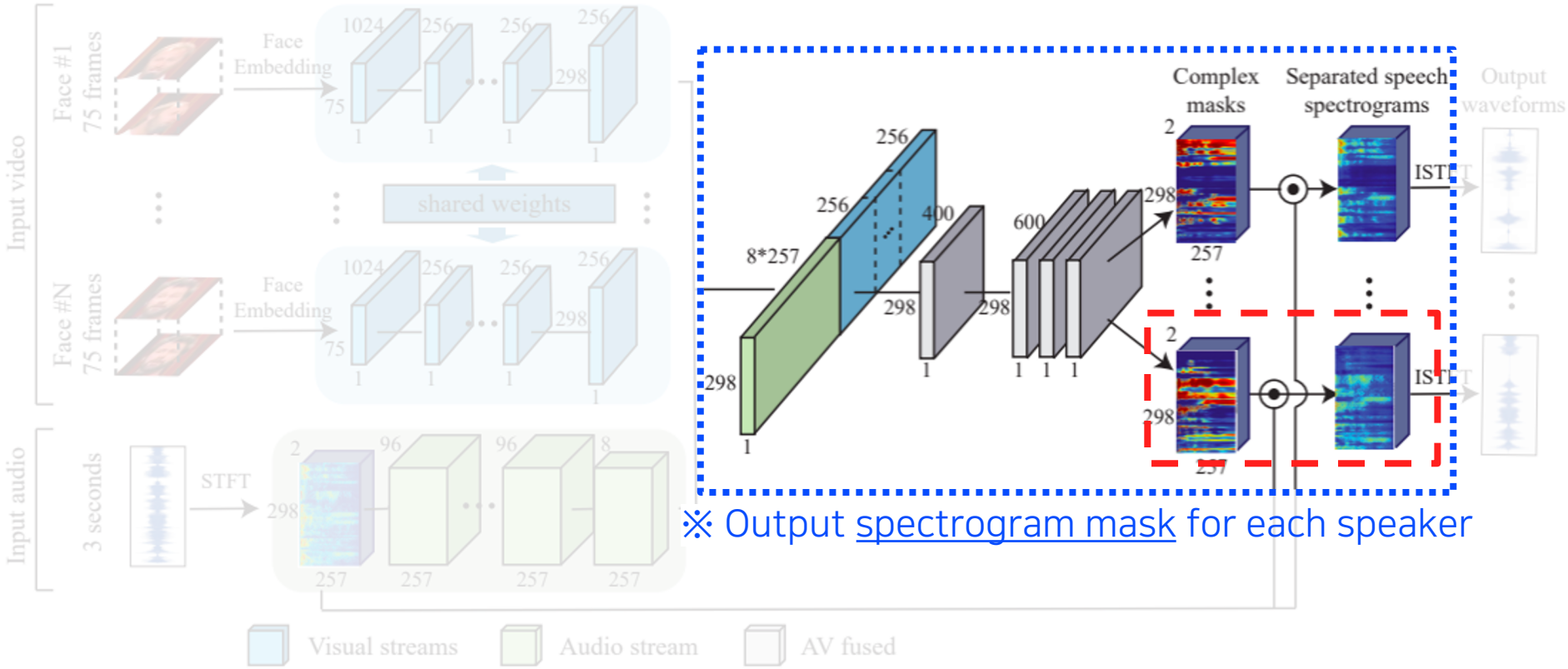


3.4 Cross modal reasoning

Looking to listen at the cocktail party – Audio-visual fusion

[Ephrat et al., SIGGRAPH 2018]

- Training data: synthetically generated by combining two clean speech videos
- Loss: L2 loss between “clean spectrogram” and “enhanced spectrogram”



※ Output spectrogram mask for each speaker

3.4 Cross modal reasoning

Lip movements generation - Synthesizing Obama example

[Suwajanakorn et al., SIGGRAPH 2017]

Speech source

Synthesized Obama

[https://youtu.be/9Yq67CjDqvw]

Conclusion. Beyond image, text and audio

Autopilot - Tesla self-driving

[https://youtu.be/tlThdr3O5Qo]

1. Multi-modal learning overview

- Wang et al., What Makes Training Multi-Modal Classification networks Hard?, CVPR 2020

2. Multi-modal (1) – Text

- Srivastava and Slakhutdinov, Multimodal Learning with Deep Boltzmann Machines, JMLR 2014
- Marin et al., Recipe 1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images, TPAMI 2019
- Vinyals et al. , Show and Tell: A Neural Image Caption Generator, CVPR 2015
- Xu et al., Show, attend and tell: Neural image caption generation with visual attention, ICML 2015
- Reed et al., Generative Adversarial Text to Image Synthesis, ICML 2016
- Antol et al., VQA: Visual Question Answering, ICCV 2015

Reference

3. Multi-modal (2) – Audio

- Aytar et al., SoundNet: Learning Sound Representations from Unlabeled Video, NIPS 2016
- Senocak et al., Learning to Localize Sound Sources in Visual Scenes: Analysis and Applications, CVPR 2018
- Oh et al., Speech2Face: Learning the Face Behind a Voice, CVPR 2019
- Hsu et al., Text-Free Image-to-Speech Synthesis Using Learned Segmental Units, arXiv 2020
- Ephrat et al., Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation, SIGGRAPH 2018
- Suwajanakorn et al., Synthesizing Obama: Learning Lip Sync from Audio, SIGGRAPH 2017
- Chen et al., Lip Movements Generation at a Glance, ECCV 2018

End of Document
Thank You.

상위 카테고리 입력란