ENEE698P: Multi-Modal Semantic Communication Through Transformer-Aided Compression

Yoonkyo Jung | ECE Ph.D. Student

April 9, 2025

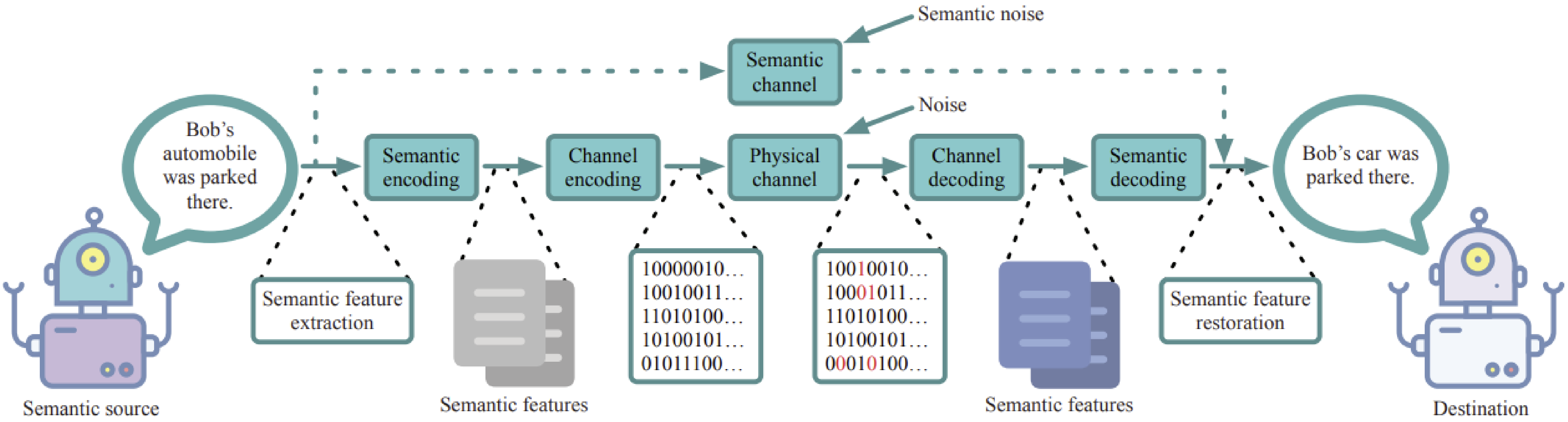
Contents

Yoonkyo Jung | Department of Electrical and Computer Engineering

Introduction

Introduction

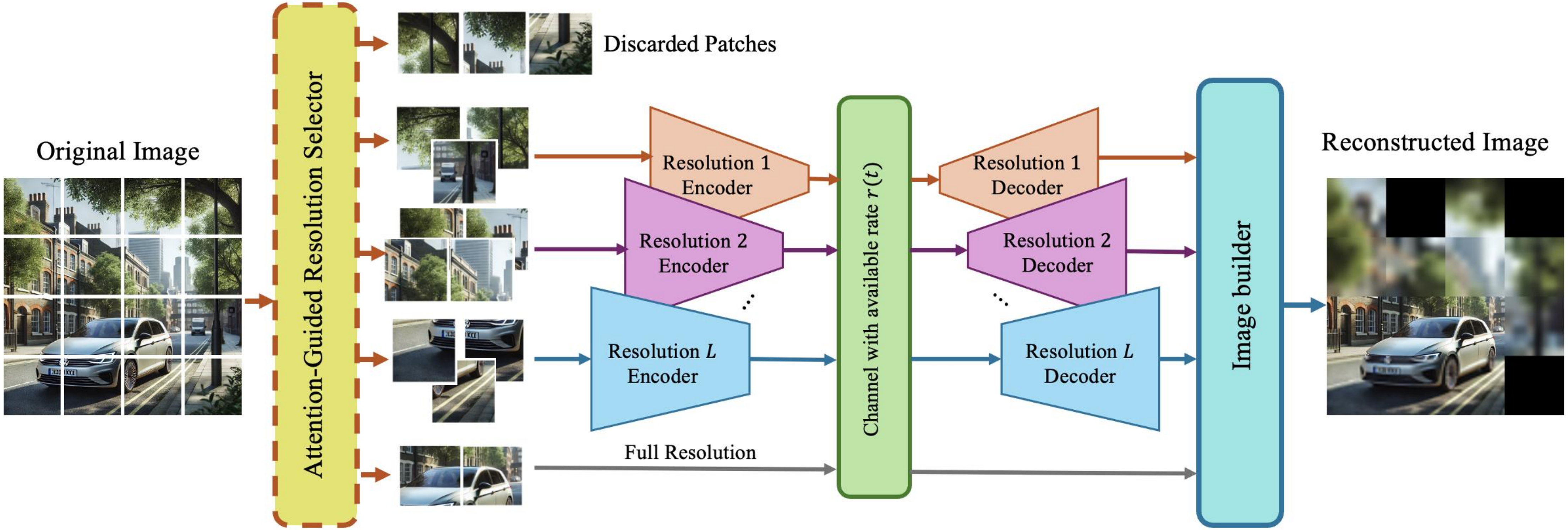
- **Semantic communications** is to extract the "meanings" or "features" of sent information from a source and "interpret" the semantic information at a destination.



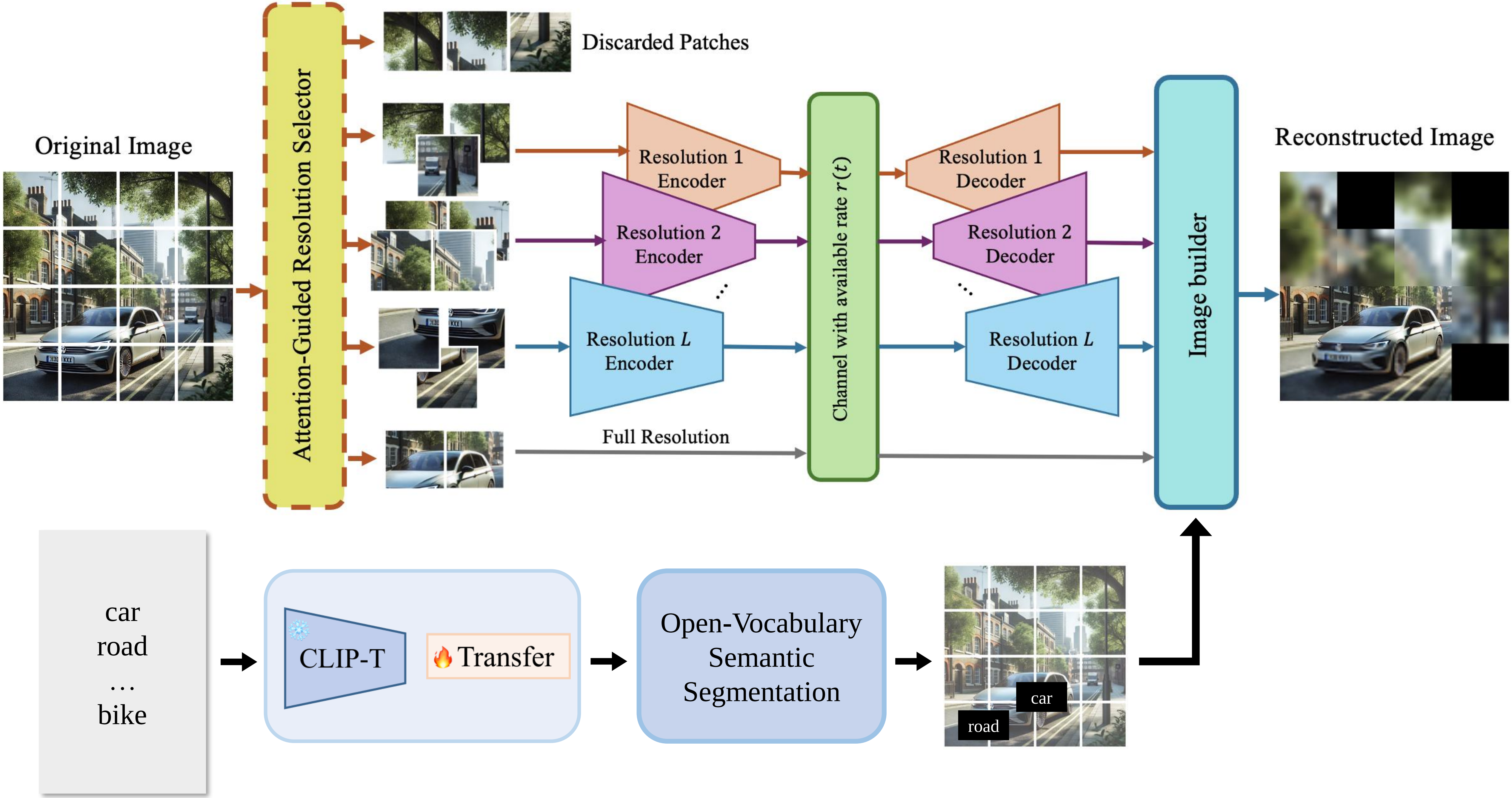
Yoonkyo Jung | Department of Electrical and Computer Engineering

Introduction

- Goal: Transmit multi-resolution data in limited bandwidth conditions.
- Developed transformer-based framework for channel-adaptive communication.



Yoonkyo Jung | Department of Electrical and Computer Engineering

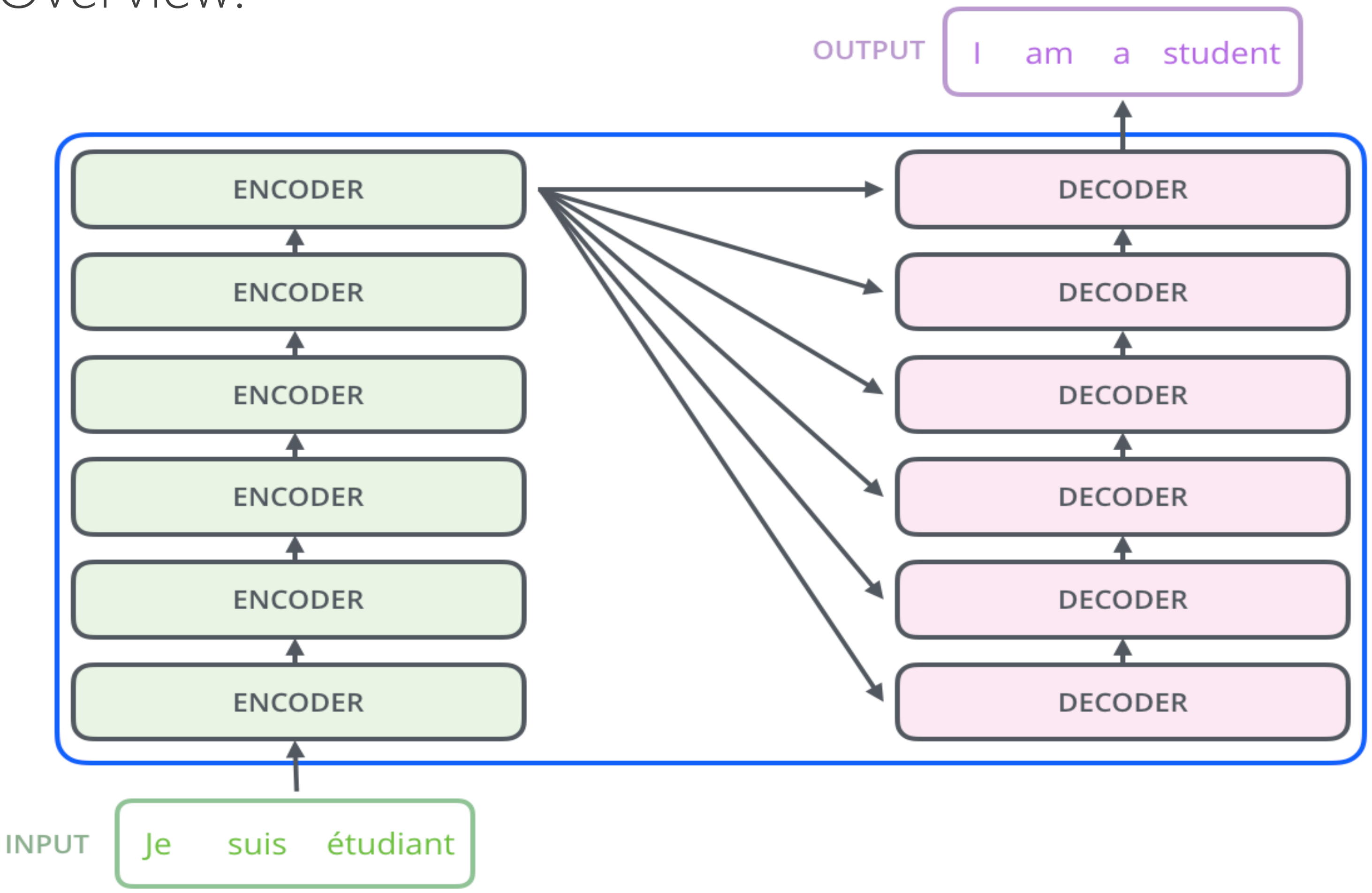


Yoonkyo Jung | Department of Electrical and Computer Engineering

Background

Background (Transformer)

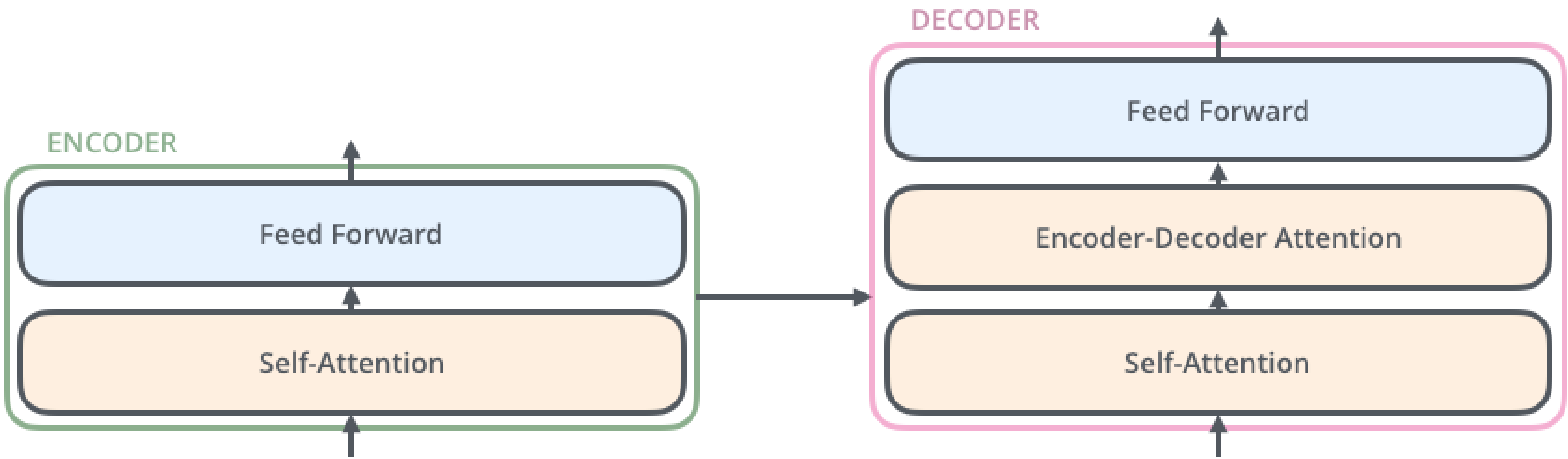
- Transformer Overview:



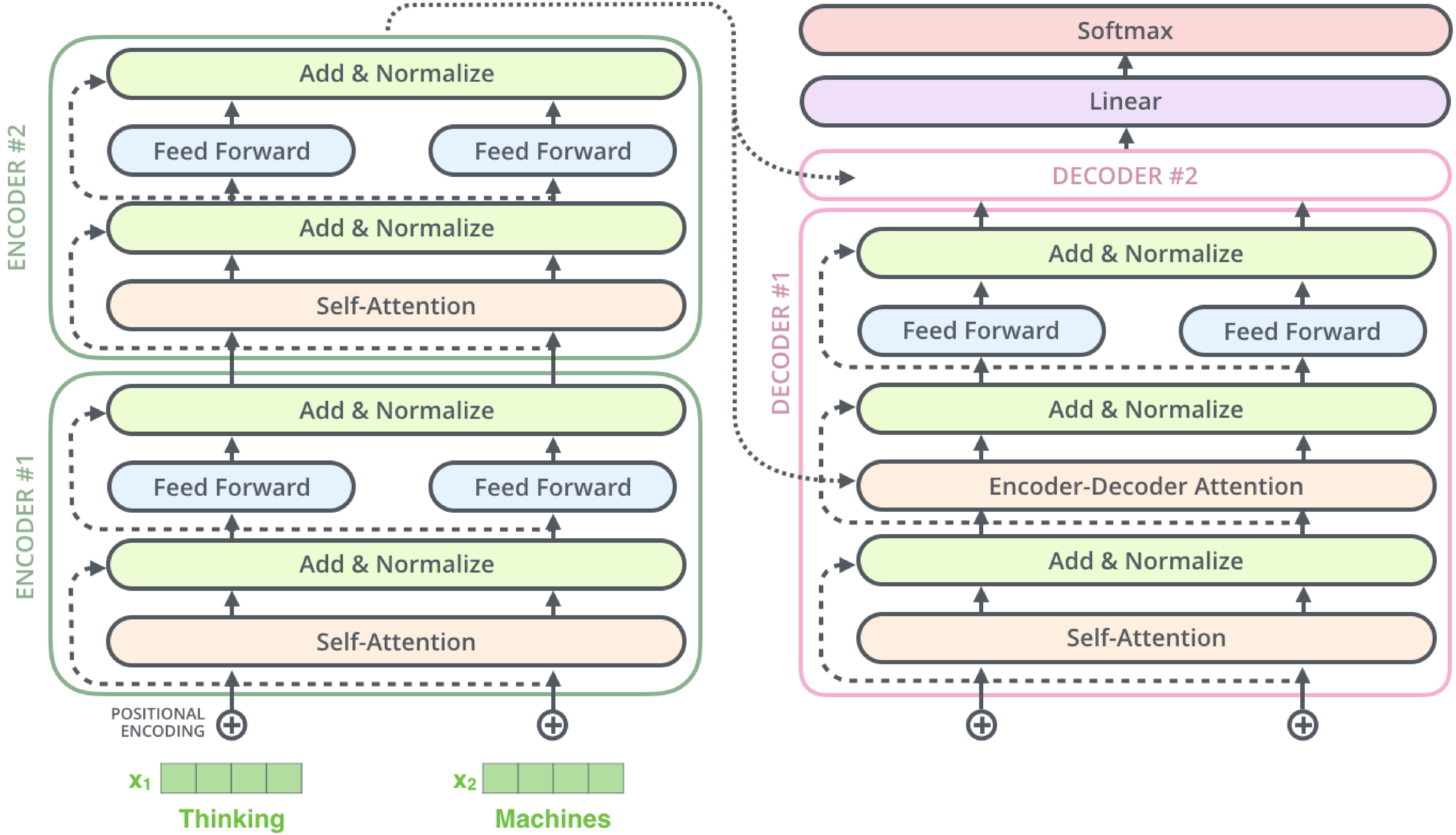
Yoonkyo Jung | Department of Electrical and Computer Engineering

Background (Transformer)

- Transformer Overview:
  - A deep learning architecture with self-attention mechanisms.
  - Enables focus on key elements in complex data.



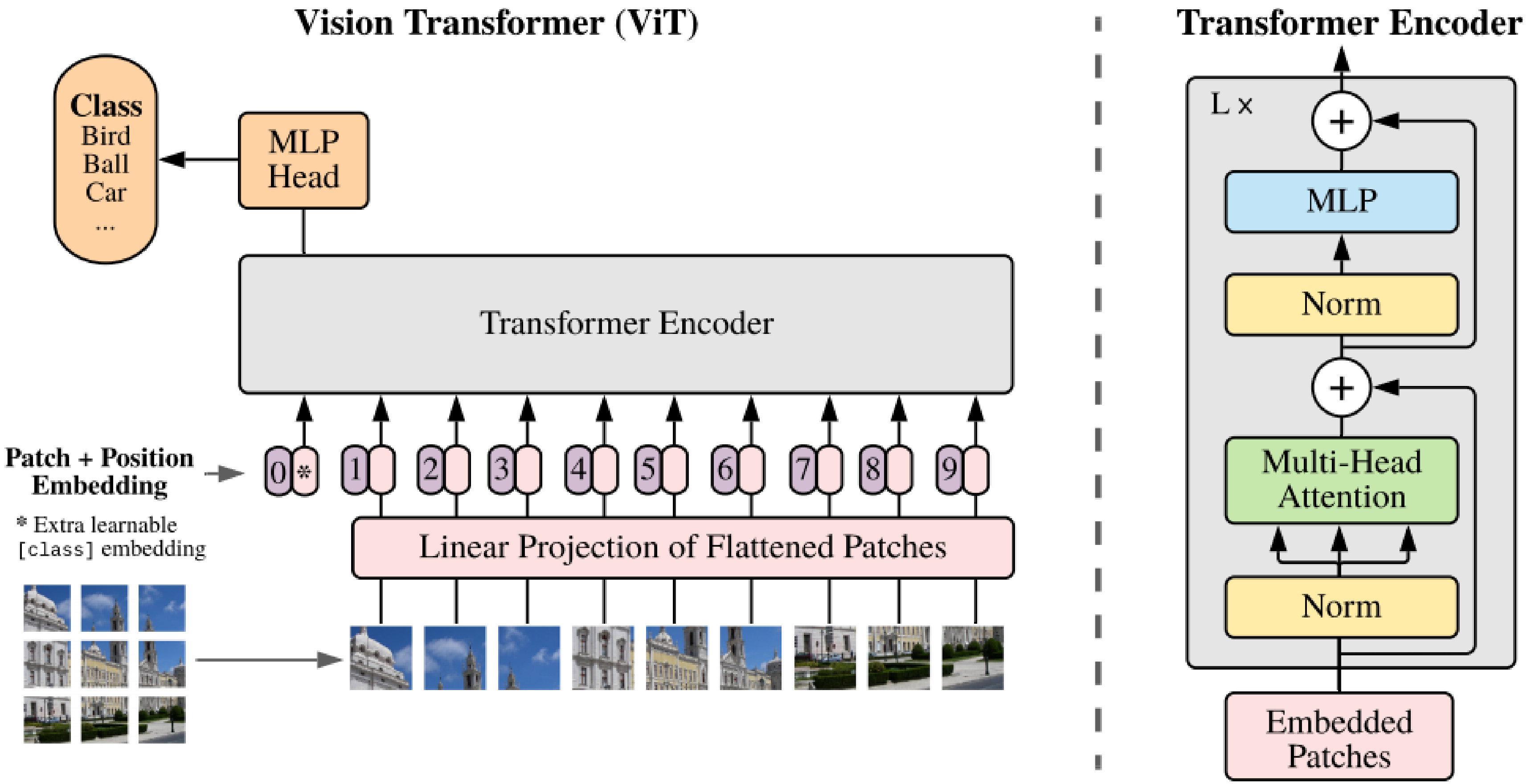
Yoonkyo Jung | Department of Electrical and Computer Engineering



Yoonkyo Jung | Department of Electrical and Computer Engineering

Background (Vision Transformer)

- Vision Transformer (ViT)



Yoonkyo Jung | Department of Electrical and Computer Engineering

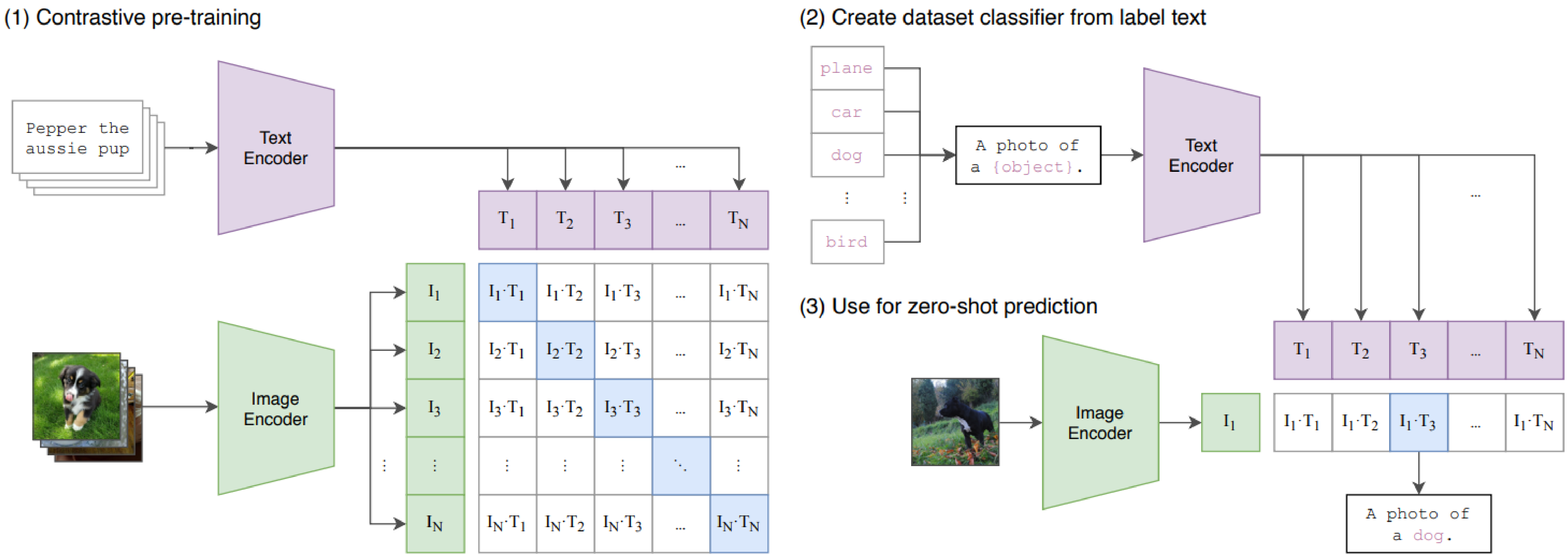
Background (Transformer & ViT)

- Key Features:
  - Multi-head attention
  - Positional encodings
  - Encoder-decoder structure
- Application for the paper:
  - Image patches encoded and compressed based on semantic content.

Yoonkyo Jung | Department of Electrical and Computer Engineering

Background (CLIP)

- Contrastive Language–Image Pretraining (CLIP)
  - A model that predicts image-text similarity

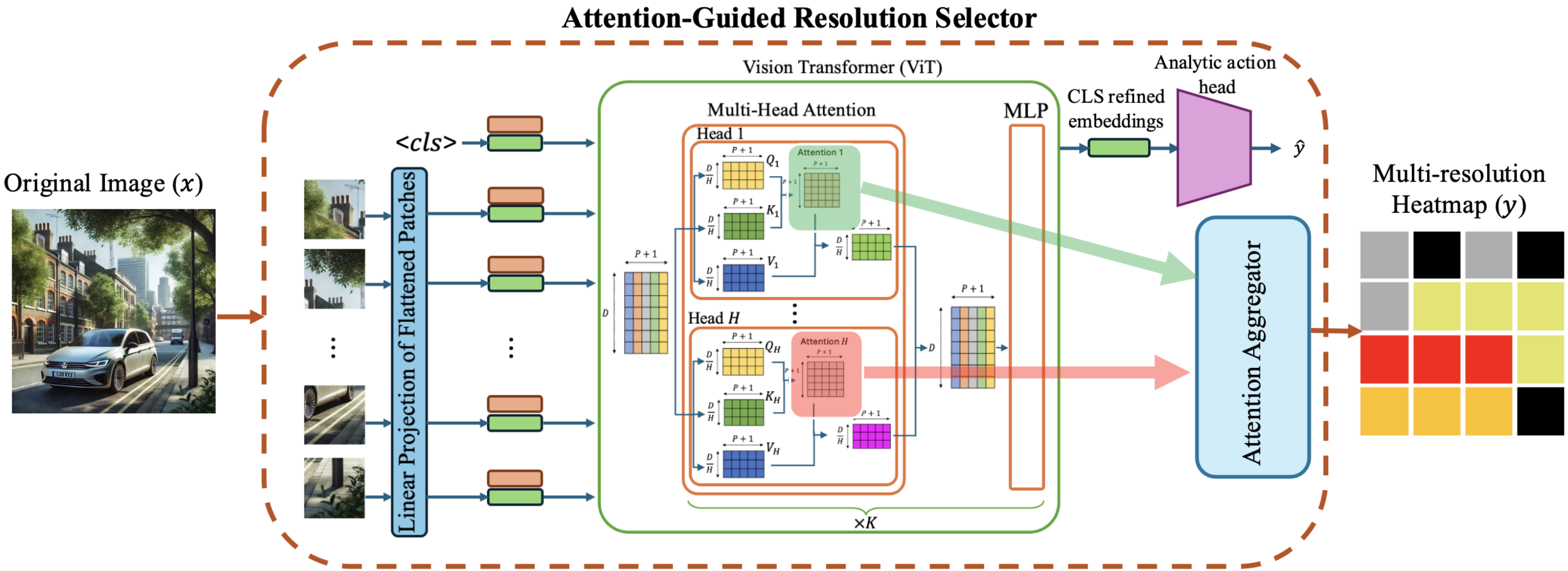


Yoonkyo Jung | Department of Electrical and Computer Engineering

System Model

System Model

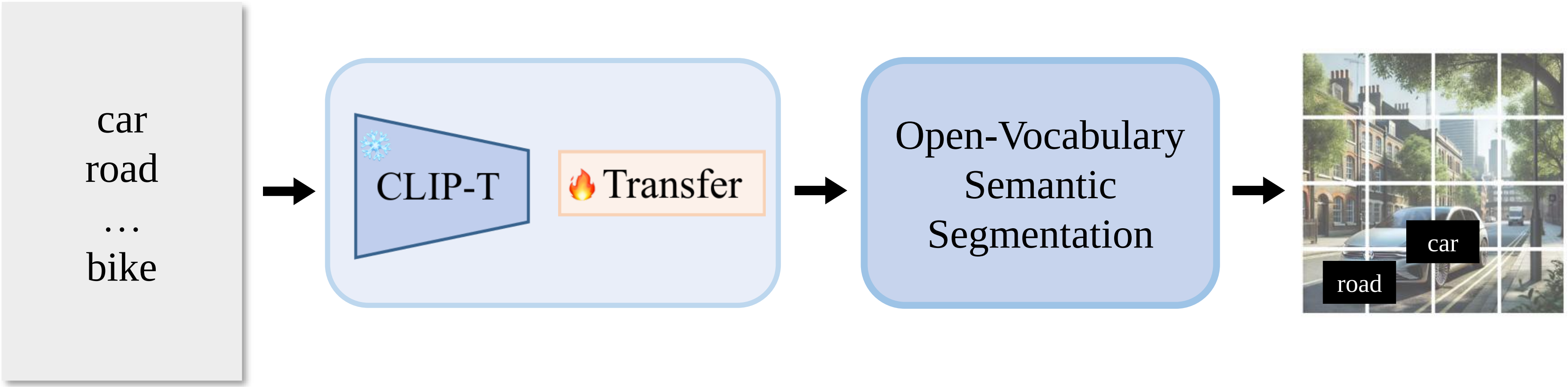
- Attention-Guided Resolution Selector determines the encoding resolution for each patch based on its semantic importance and available channel rate.



Yoonkyo Jung | Department of Electrical and Computer Engineering

System Model

- Open-vocabulary segmentation framework find an attention score of images based on the text input

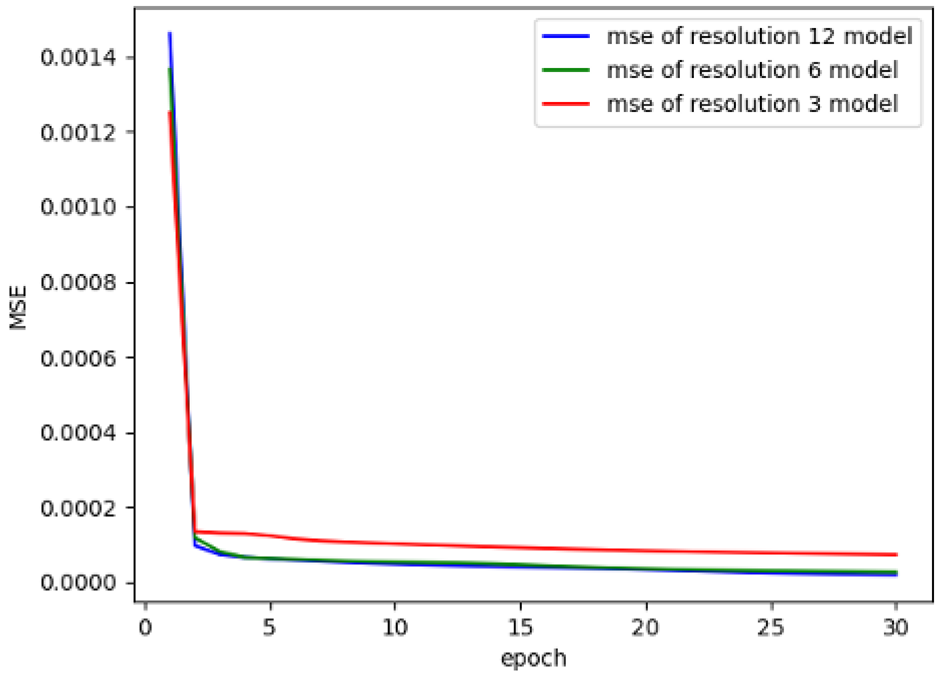


Yoonkyo Jung | Department of Electrical and Computer Engineering

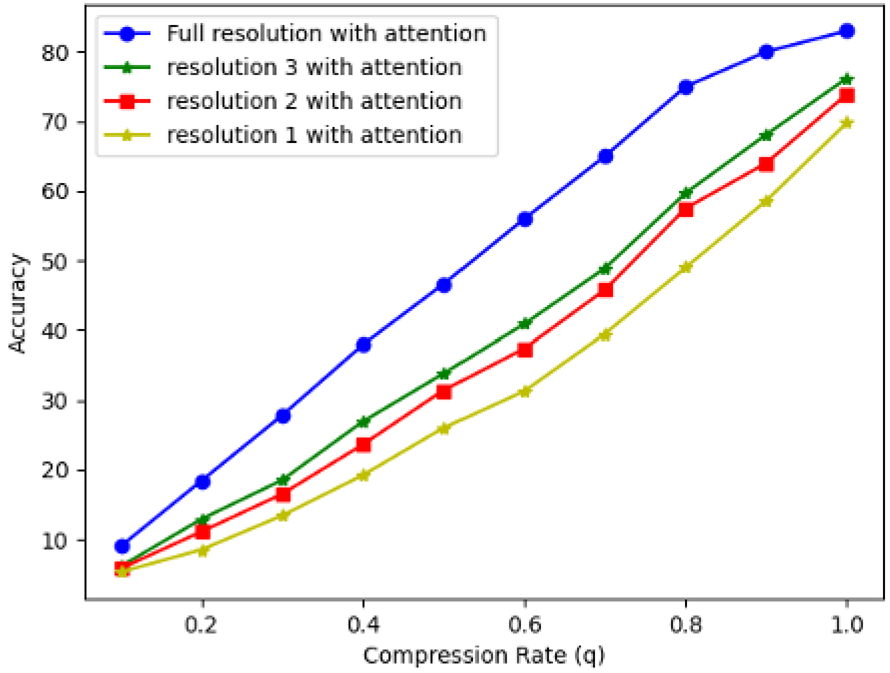
Results

Yoonkyo Jung | Department of Electrical and Computer Engineering

Results



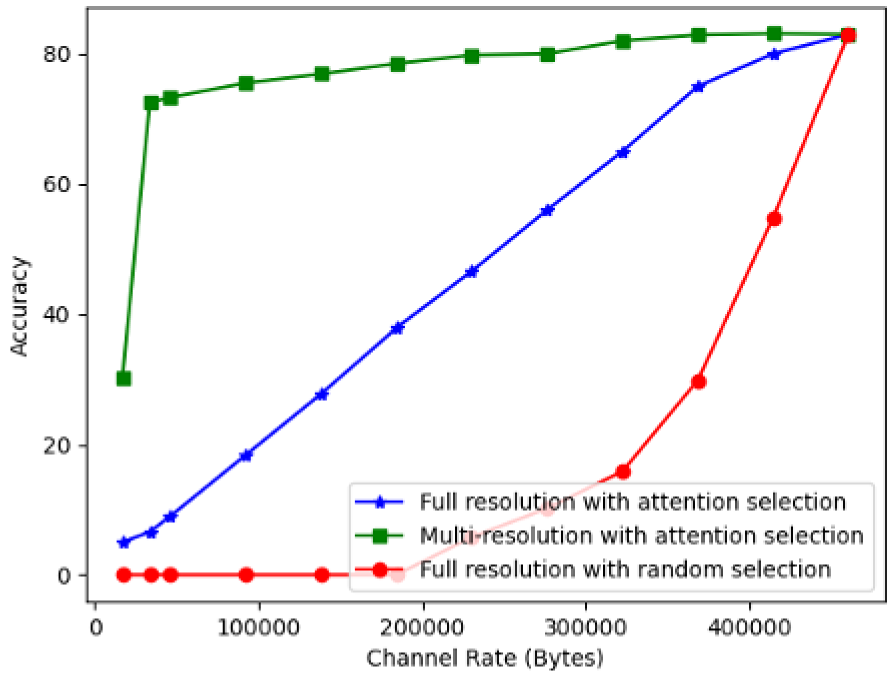
<Reconstruction result for three medium resolutions.>



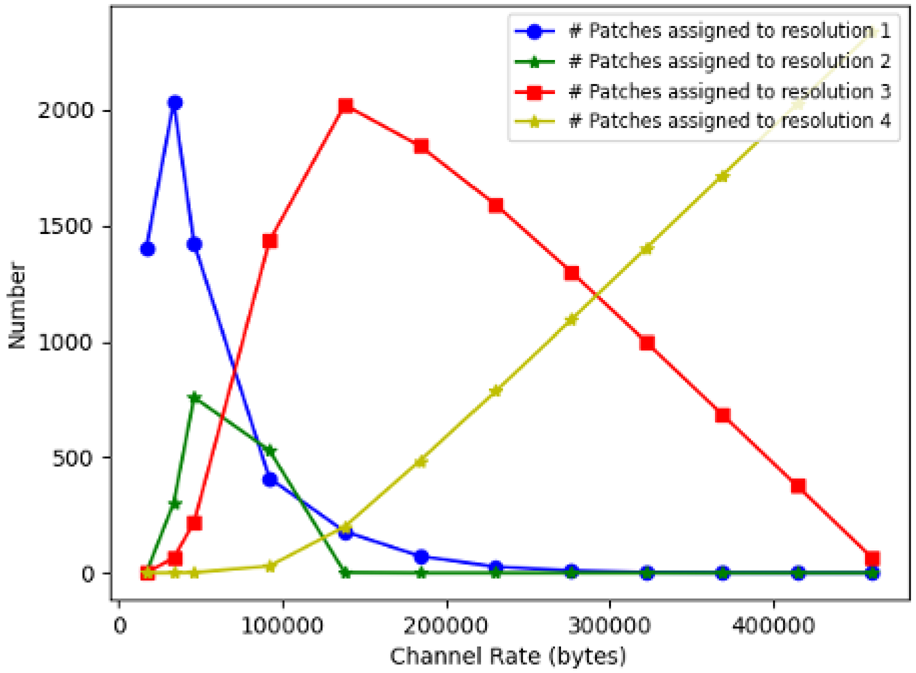
<Accuracy result for three medium resolutions.>

Yoonkyo Jung | Department of Electrical and Computer Engineering

Results



<Accuracy result for adaptive multi-resolution semantic communication framework in various channel rates.>



<Resolution assignment to patches in different channel rate.>

Yoonkyo Jung | Department of Electrical and Computer Engineering

Conclusion

Conclusion

- This work proposes a novel semantic communication framework that fuses open-vocabulary vision–language segmentation with transformer-based compression.
- We build on recent advances in open-vocabulary segmentation, which leverage large pre-trained vision–language models to break away from fixed label sets and segment arbitrary categories described by text prompts.
- But merging these models with the same scale can be a challenge.

Conclusion

- Potential Directions:
  - Extend to other multimodal data (e.g., eye-tracking and images together).
  - Extend to video data with recent techniques in computer vision fields
  - Develop task-specific optimization techniques for diverse domains (e.g., image anomaly detection).
- Key Challenges:
  - Efficient training of adaptive encoders.
  - Addressing latency in dynamic channel conditions.