Multi-Modal Semantic Communication Through Transformer-Aided Compression

Yoonkyo Jung | ECE Ph.D. Student

April 2, 2025

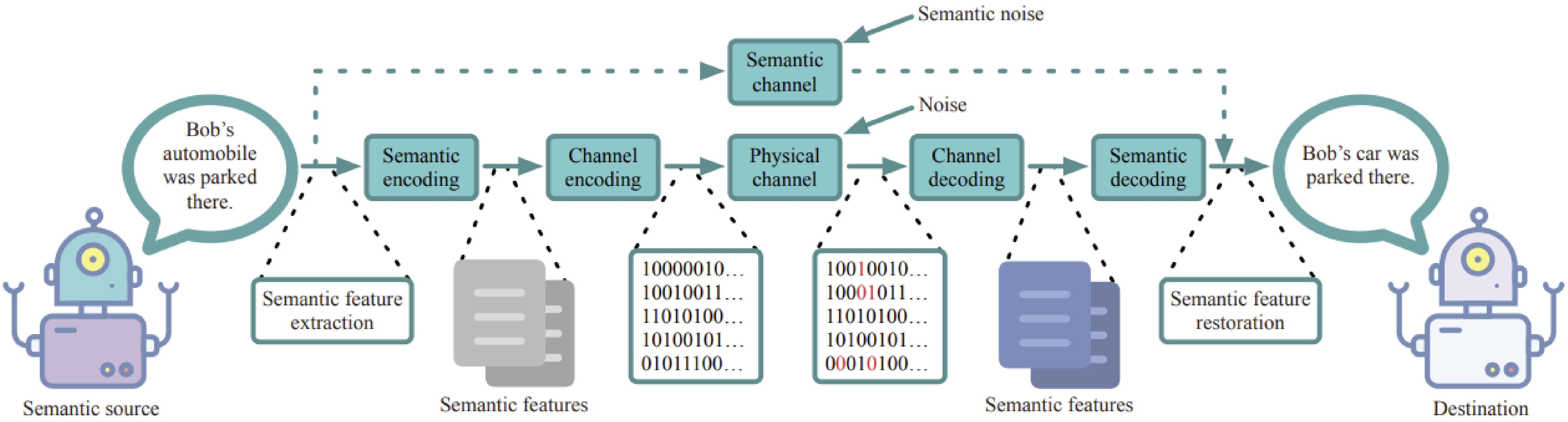
Contents

- Introduction
- Background
- System Model
- Results
- Conclusion

Yoonkyo Jung | Department of Electrical and Computer Engineering

Introduction

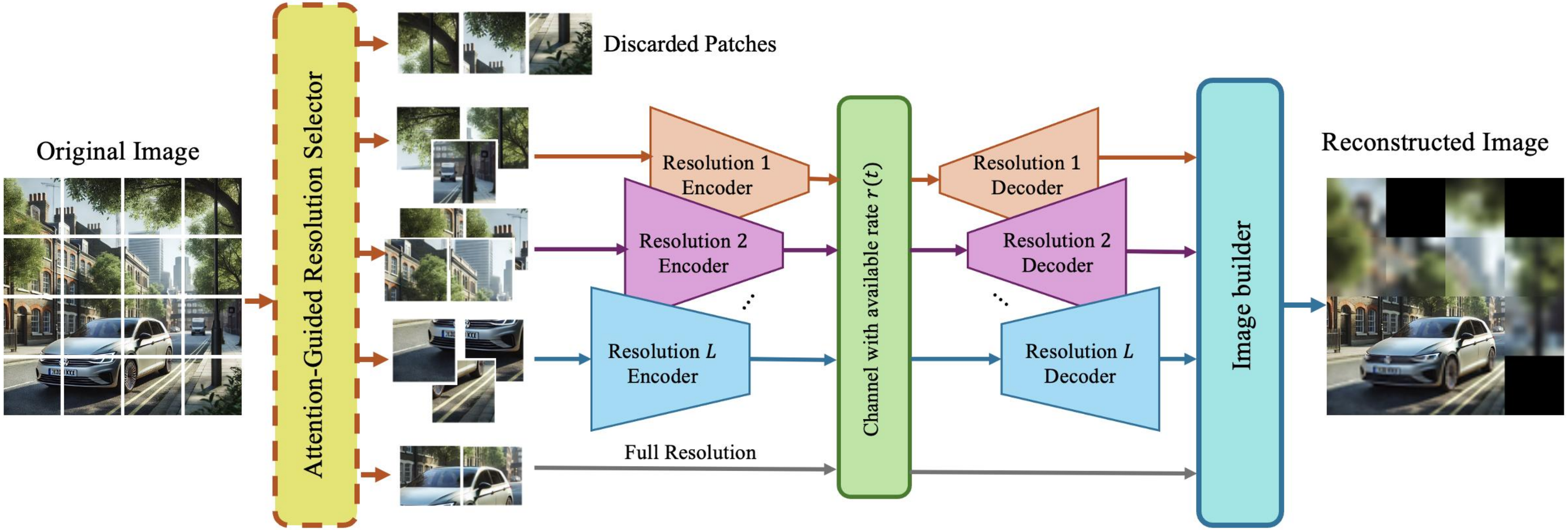
- **Semantic communications** is to extract the "meanings" or "features" of sent information from a source and "interpret" the semantic information at a destination.

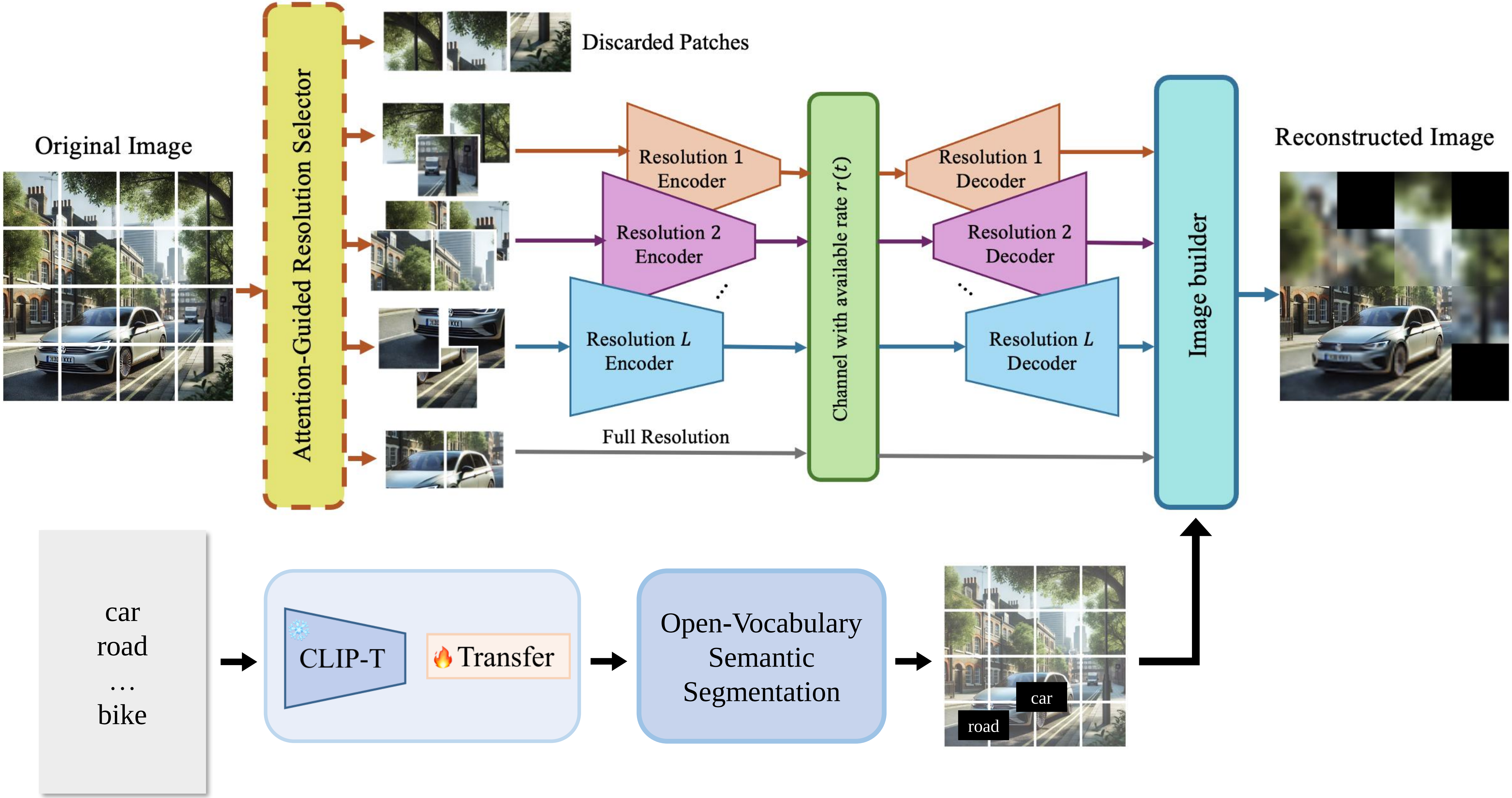


Yoonkyo Jung | Department of Electrical and Computer Engineering

Introduction

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



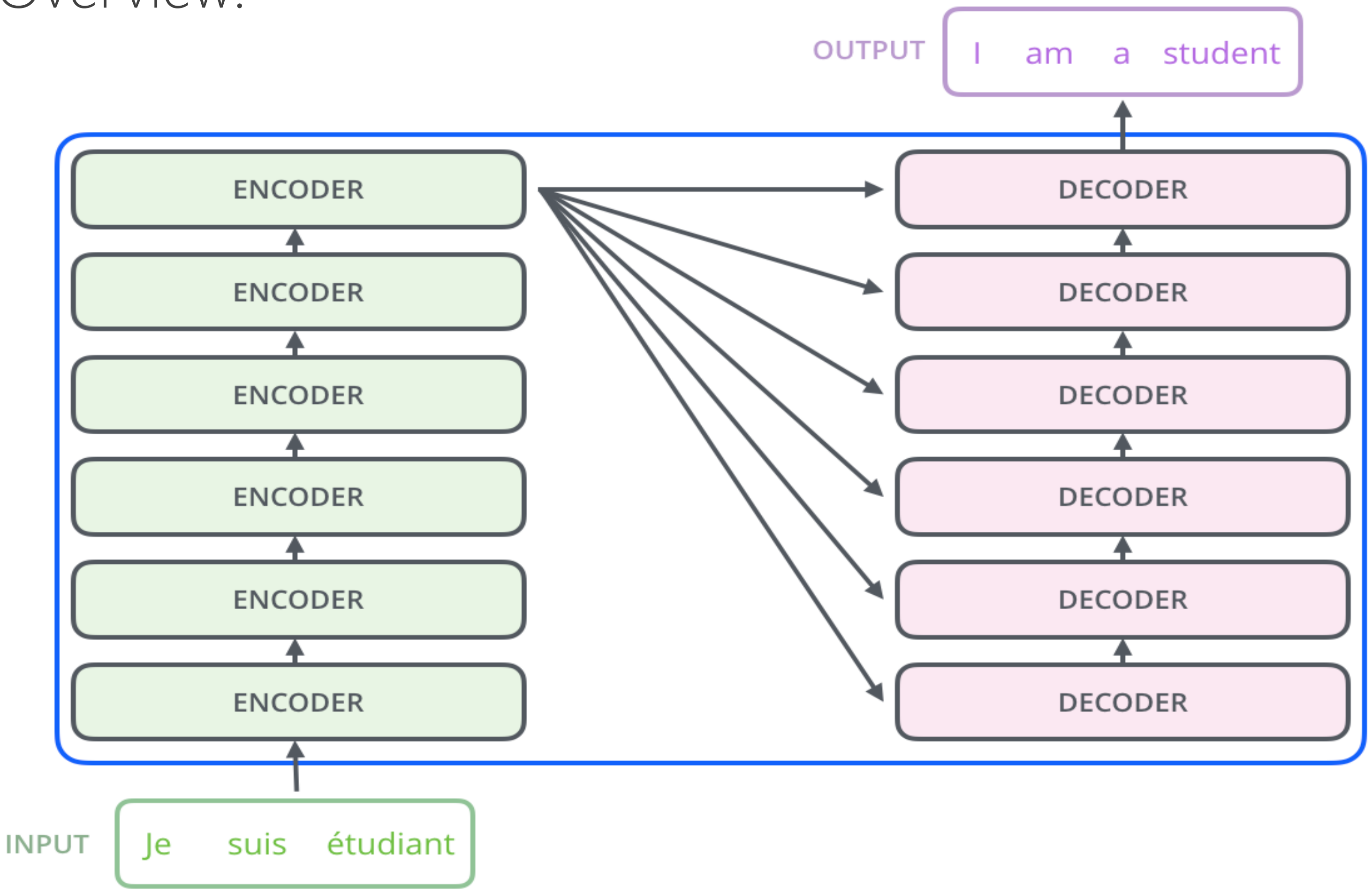


Yoonkyo Jung | Department of Electrical and Computer Engineering

Background

Background (Transformer)

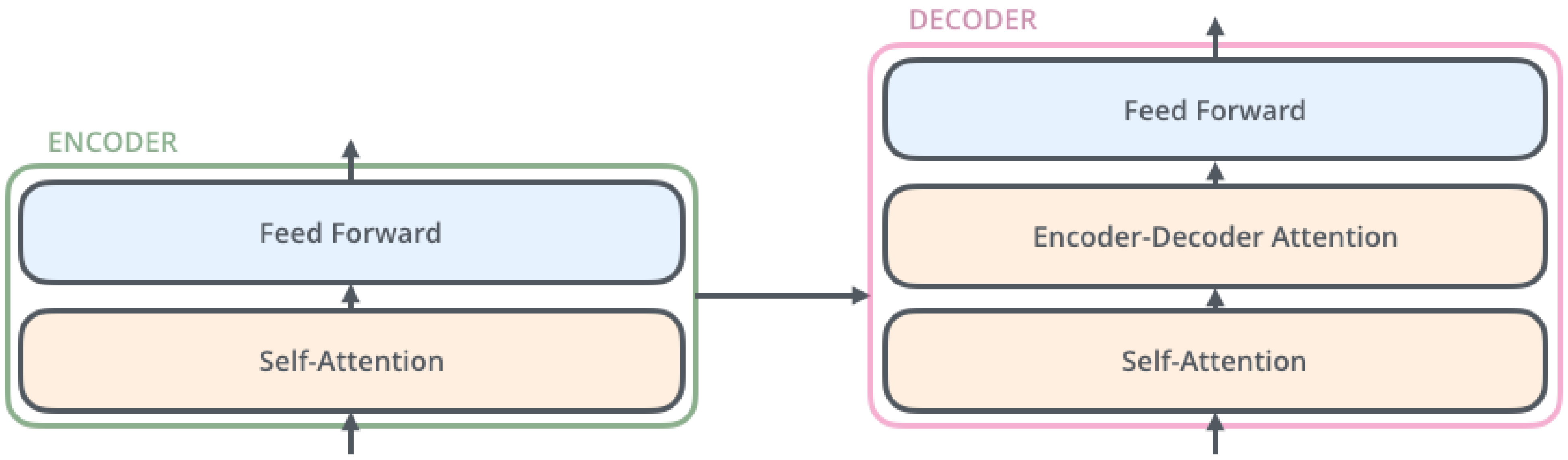
- Transformer Overview:



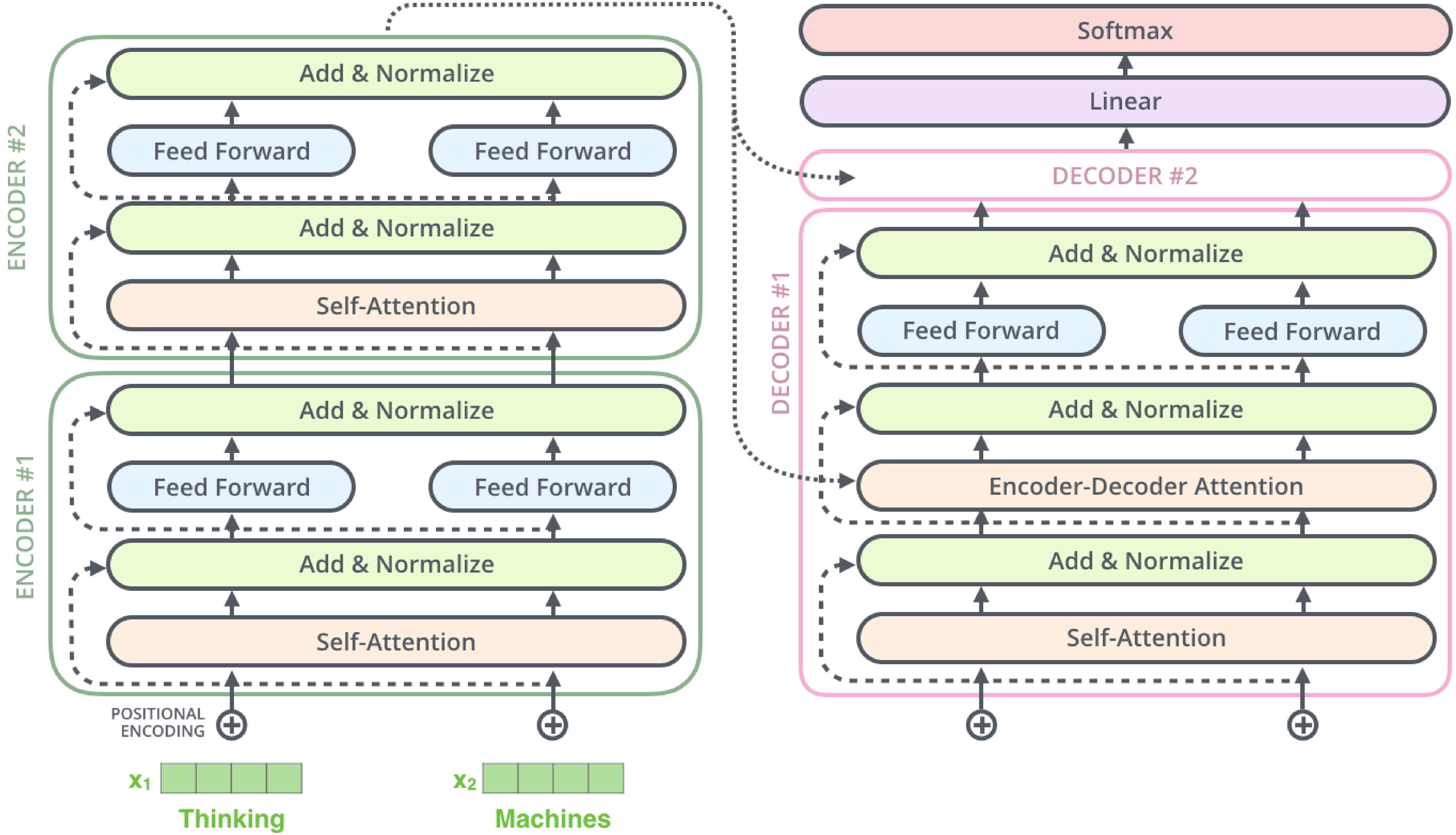
Yoonkyo Jung | Department of Electrical and Computer Engineering

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


Softmax
Linear
DECODER #2
DECODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Encoder-Decoder Attention
Add & Normalize
Self-Attention
ENCODER #2
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
POSITIONAL ENCODING
x1
x2
Thinking
Machines

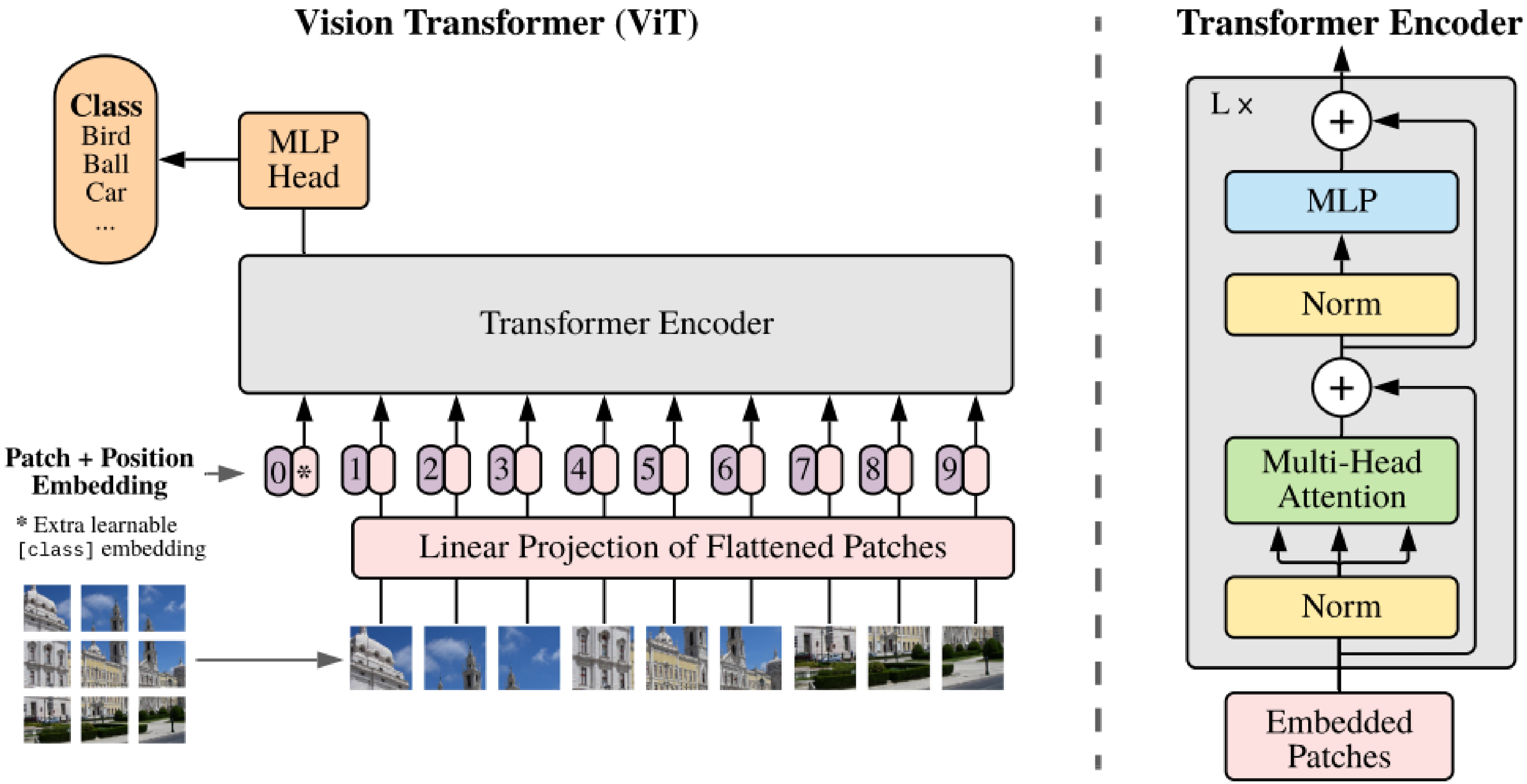
Background (Transformer)

- 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 (Vision Transformer)

- Vision Transformer (ViT)



Yoonkyo Jung | Department of Electrical and Computer Engineering

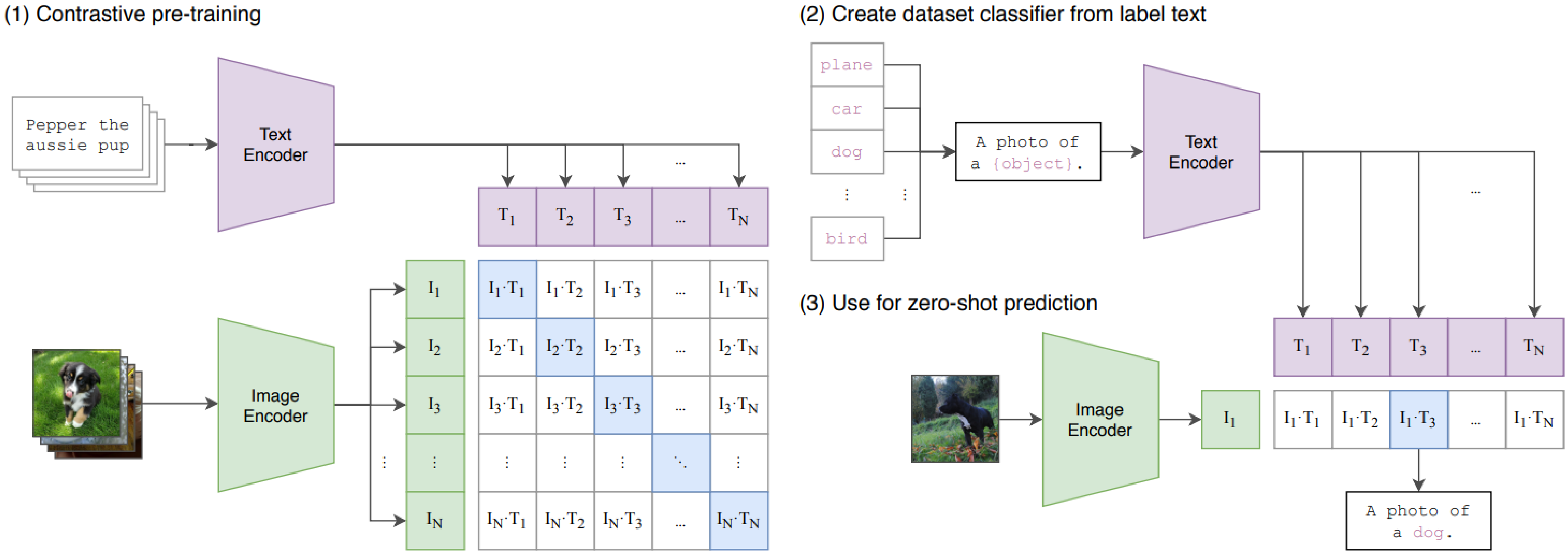
Background (Vision Transformer)

- Vision Transformer (ViT):
  - Image Patching: The input image is divided into fixed-size patches, typically 16x16 pixels.
  - Patch Embedding: Each patch is flattened and linearly embedded into a lower-dimensional vector.
  - Positional Encoding: Positional embeddings are added to retain spatial information.
  - Transformer Encoder: The sequence of patch embeddings is processed by a standard Transformer encoder.
  - Classification: A special "classification token" is added to the sequence for final prediction

Yoonkyo Jung | Department of Electrical and Computer Engineering

Background (CLIP)

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



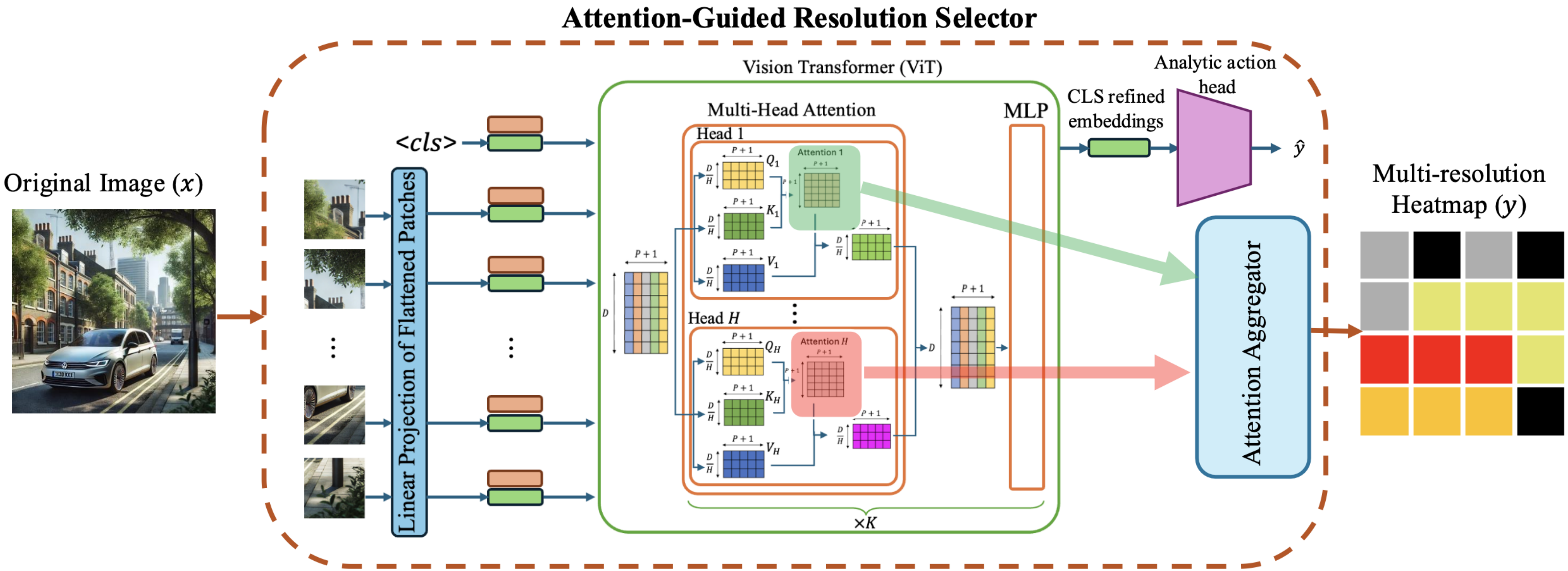
Yoonkyo Jung | Department of Electrical and Computer Engineering

System Model

Yoonkyo Jung | Department of Electrical and Computer Engineering

System Model

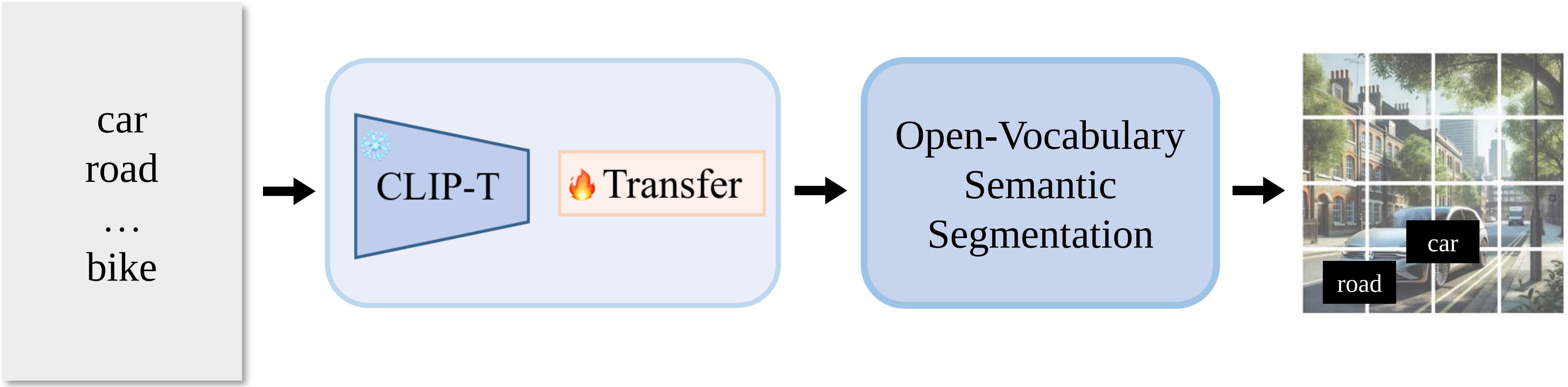
- Attention-Guided Resolution Selector determines the encoding resolution for each patch based on its semantic importance and available channel rate.



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

System Model

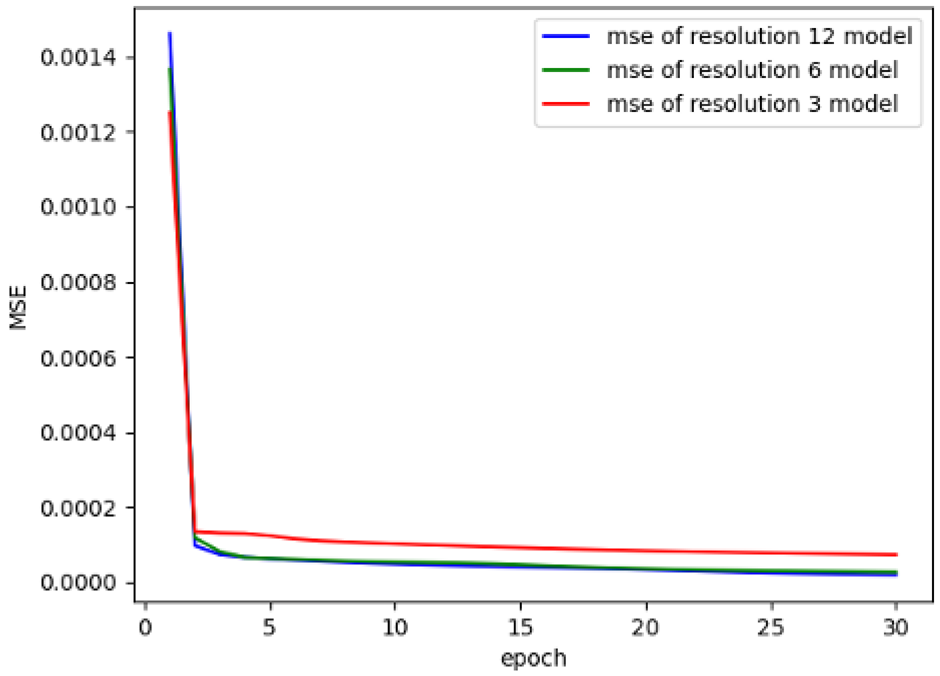
- Key contributions:
  - Multi-resolution encoding for preserving semantic fidelity.
  - Dynamic adaptation to varying channel conditions.
- Results:
  - Optimized bandwidth utilization.
  - Maintained high task accuracy and data quality.

Yoonkyo Jung | Department of Electrical and Computer Engineering

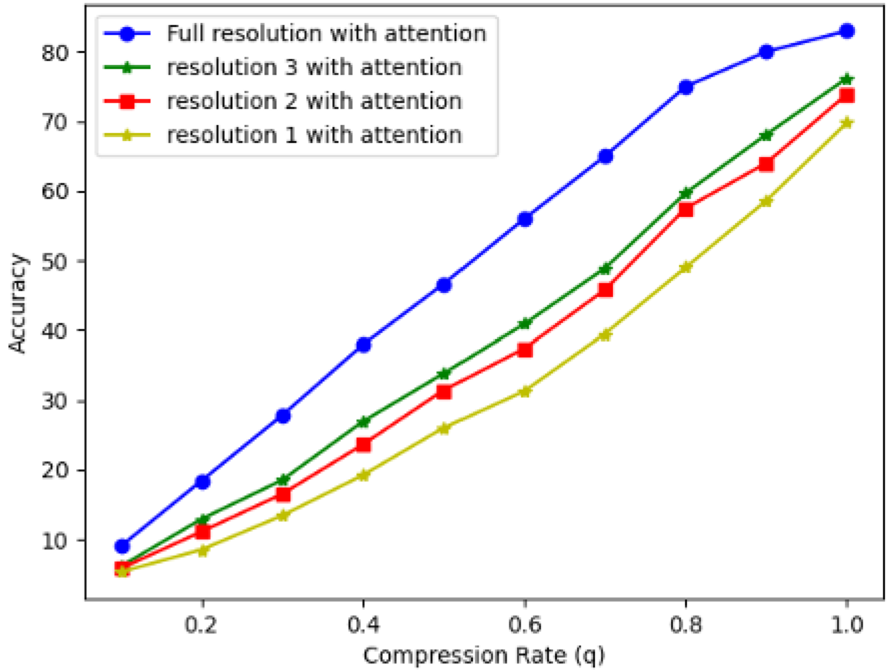
Results

Yoonkyo Jung | Department of Electrical and Computer Engineering

Results



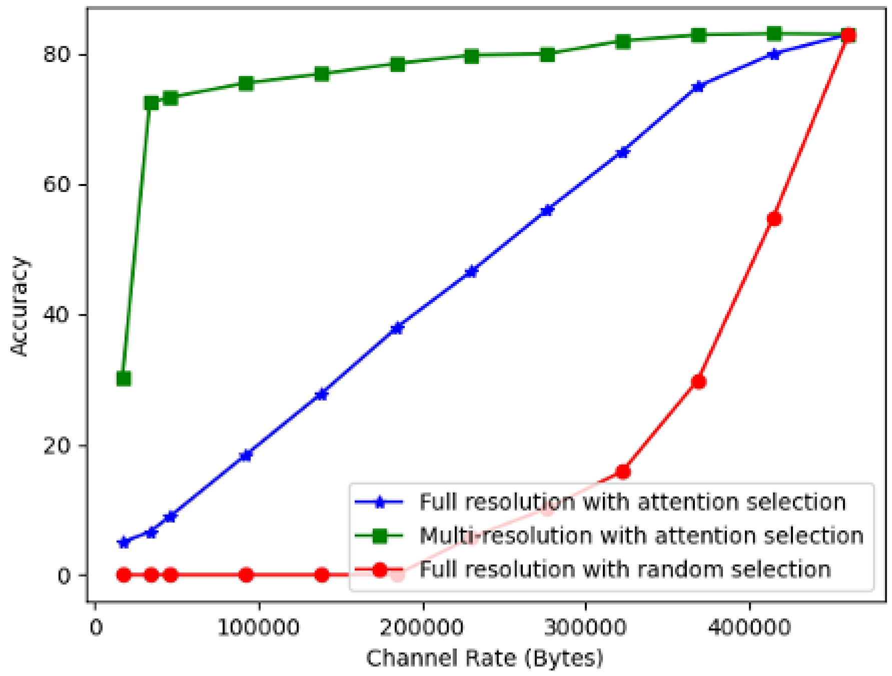
<Reconstruction result for three medium resolutions.>



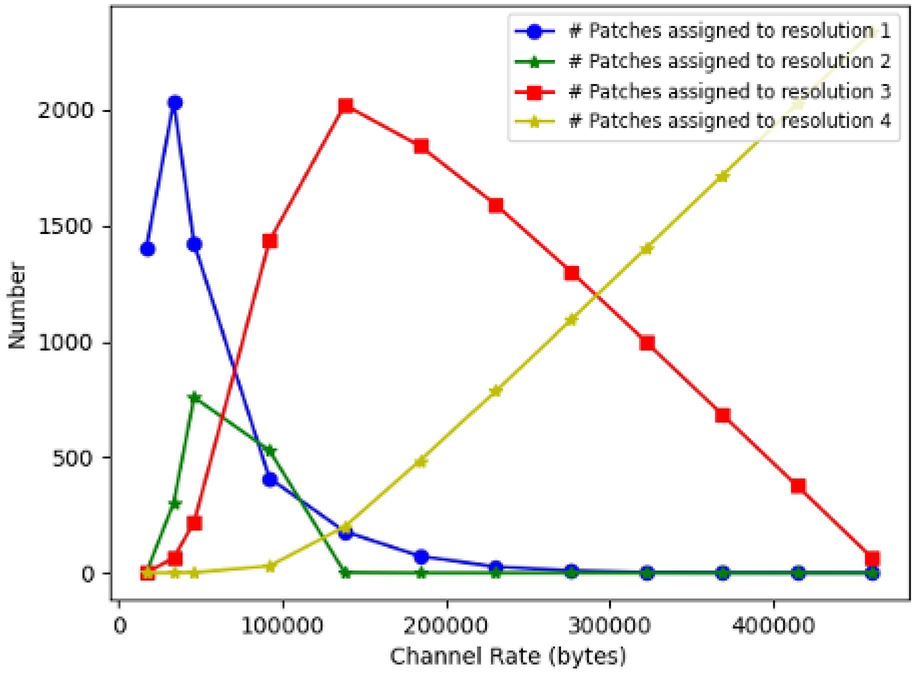
<Accuracy result for three medium resolutions.>

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

Discussion and Conclusion

Discussion

- Potential Directions:
  - Extend to multimodal data (e.g., text 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.

Yoonkyo Jung | Department of Electrical and Computer Engineering

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.

Yoonkyo Jung | Department of Electrical and Computer Engineering