# CV / VLMs

## Unit 5: State-of-the-Art Object Detection Techniques

TIL-AI
TODAY I LEARNED-AI

# 5.1.2

# Diving Deeper into Neural Networks

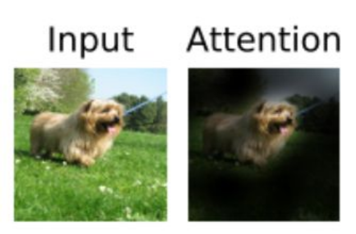
Attention mechanisms in CV

# Attention mechanisms in CV

- Given an image, one will naturally focus on the most interesting parts of a scene. Maybe it’s a colorful flower in a garden, a person’s face in a crowd, or a stunning sunset over the ocean.

- We humans can instantly identify the most relevant parts of an image effortlessly, as our brains automatically highlight them for us.

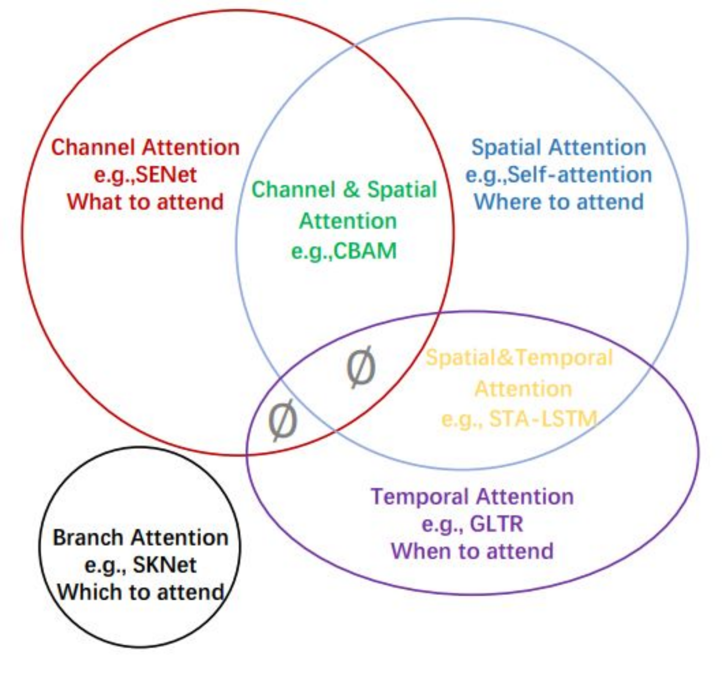
- Researchers have developed similar mechanisms - attention mechanisms with the aim of imitating this aspect of the human visual system. Such an attention mechanism can be regarded as a dynamic weight adjustment process based on features of the input image.



(left) Representative examples of attention from the output token to the input space via ViT.

- [2111.07624] Attention Mechanisms in Computer Vision: A Survey (arxiv.org)
- MenghaoGuo/Awesome-Vision-Attentions: Summary of related papers on visual attention. Related code will be released based on Jittor gradually. (github.com)

# Attention mechanisms in CV



(top) Categories of Attention mechanisms. ∅ means such combinations do not (yet) exist.

Attention mechanisms can be categorised according to data domains.

These include four fundamental categories of

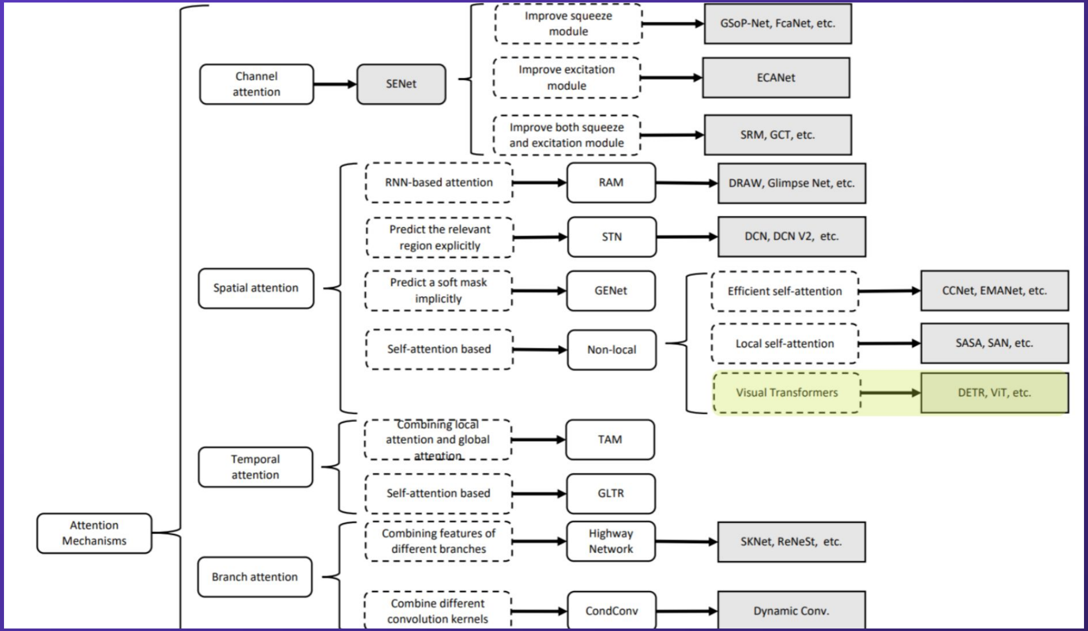
- channel attention
- spatial attention
- temporal attention
- branch attention

and two hybrid categories, combining channel & spatial attention and spatial & temporal attention.

# Attention Mechanisms in CV

- The development of attention mechanisms is branched across different networks/research objectives.

- We will proceed with more depth for Vision Transformers (DETR, ViT, etc) in the sections that follow.



Attention Mechanisms

Channel attention → SENet
- Improve squeeze module → GSoP-Net, FcaNet, etc.
- Improve excitation module → ECANet
- Improve both squeeze and excitation module → SRM, GCT, etc.

Spatial attention
- RNN-based attention → RAM → DRAW, Glimpse Net, etc.
- Predict the relevant region explicitly → STN → DCN, DCN V2, etc.
- Predict a soft mask implicitly → GENet
- Self-attention based → Non-local
  - Efficient self-attention → CCNet, EMANet, etc.
  - Local self-attention → SASA, SAN, etc.
  - Visual Transformers → DETR, ViT, etc.

Temporal attention
- Combining local attention and global attention → TAM
- Self-attention based → GLTR

Branch attention
- Combining features of different branches → Highway Network → SKNet, ReNeSt, etc.
- Combine different convolution kernels → CondConv → Dynamic Conv.