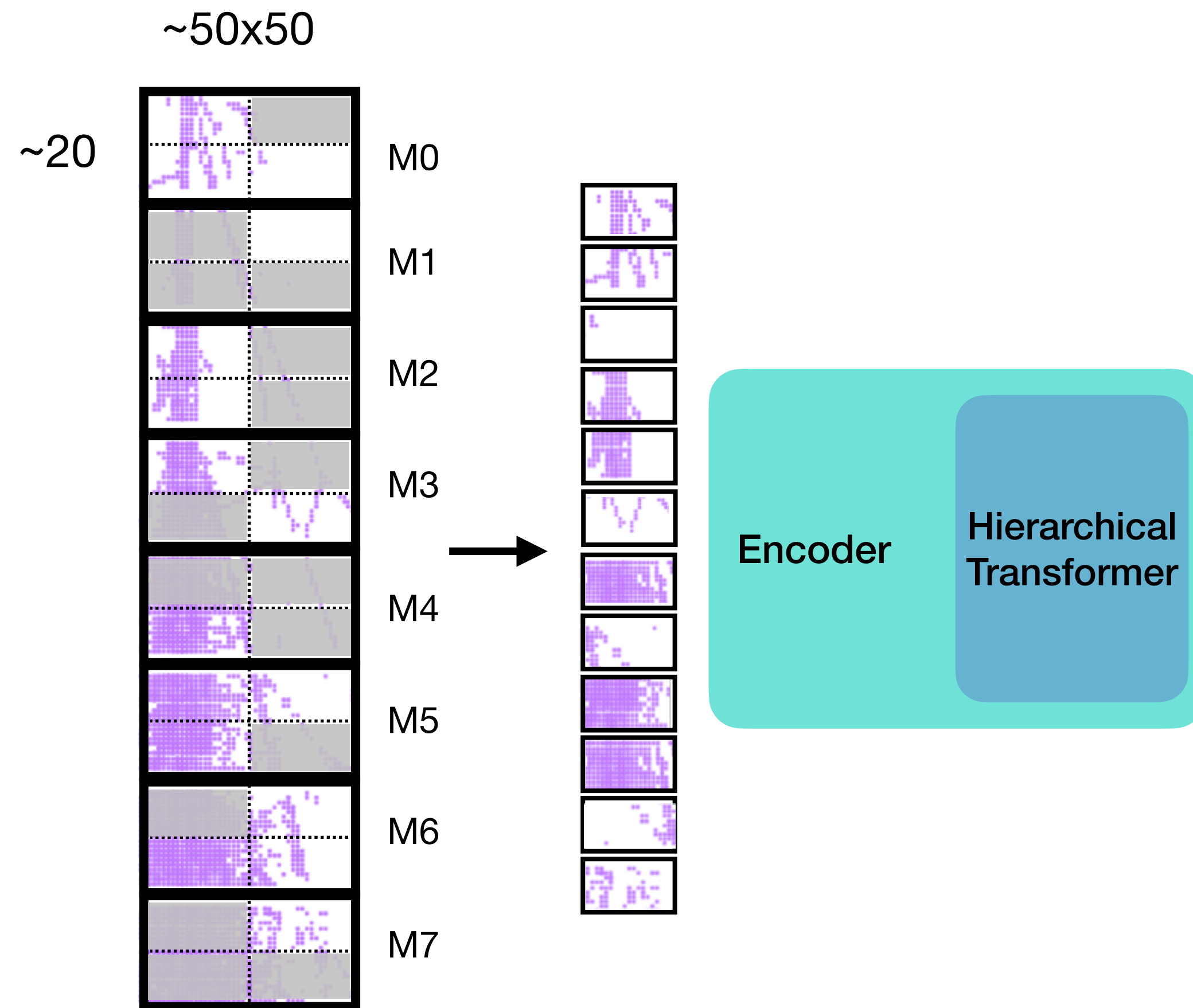
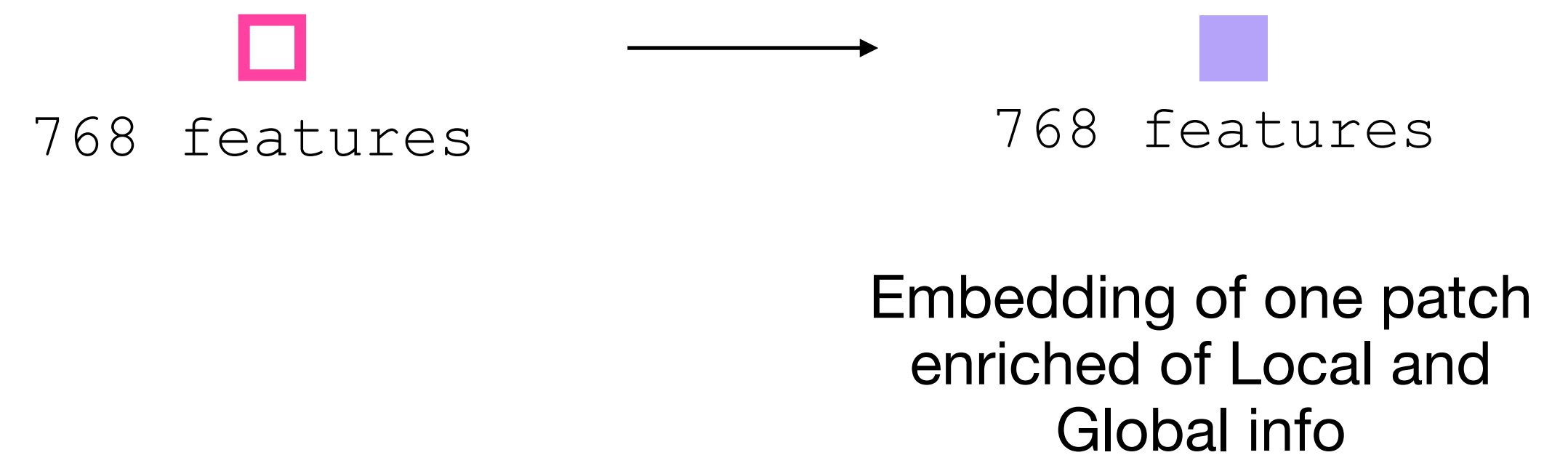


# Pre-Training

## Hierarchical Transformer



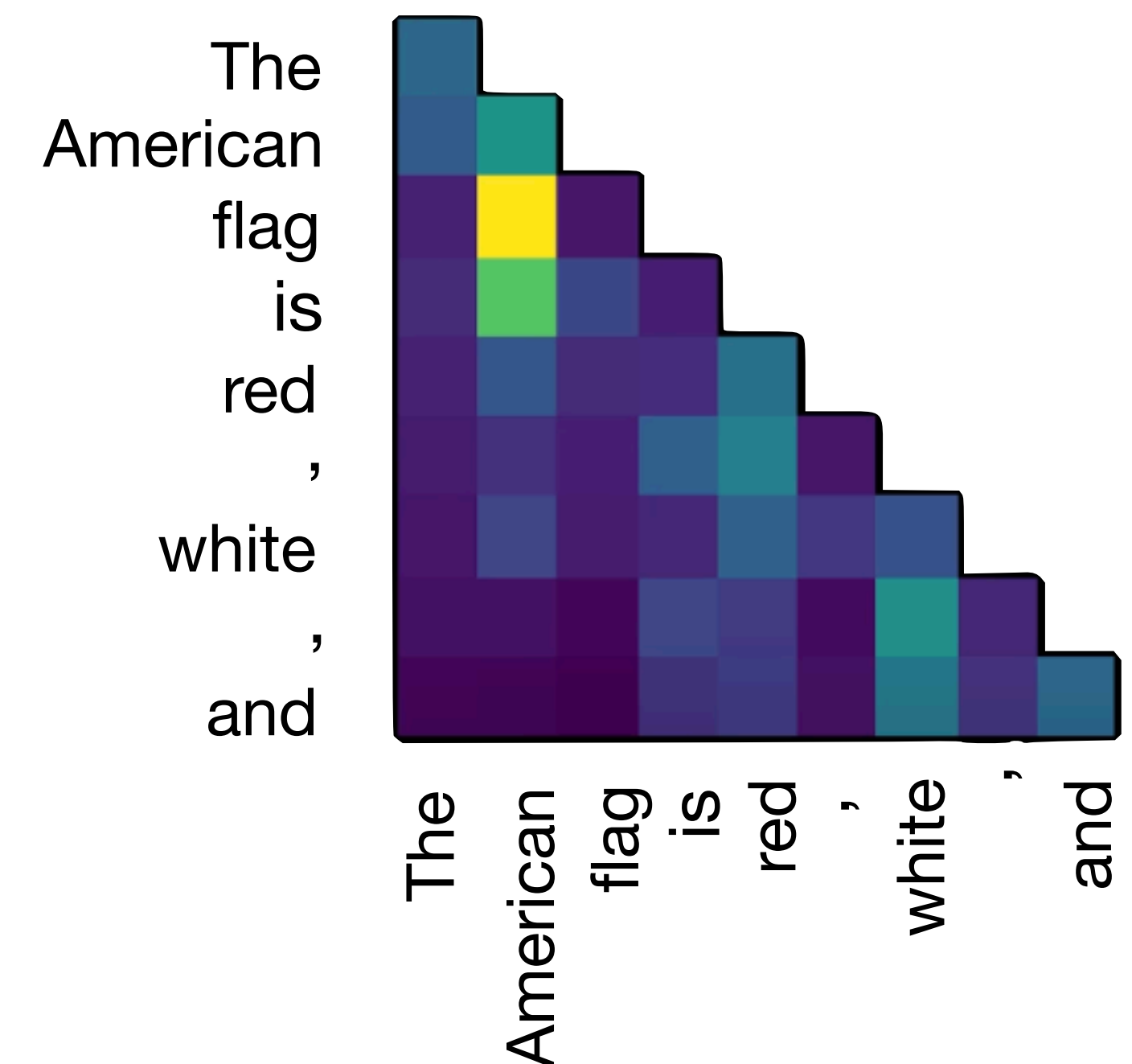
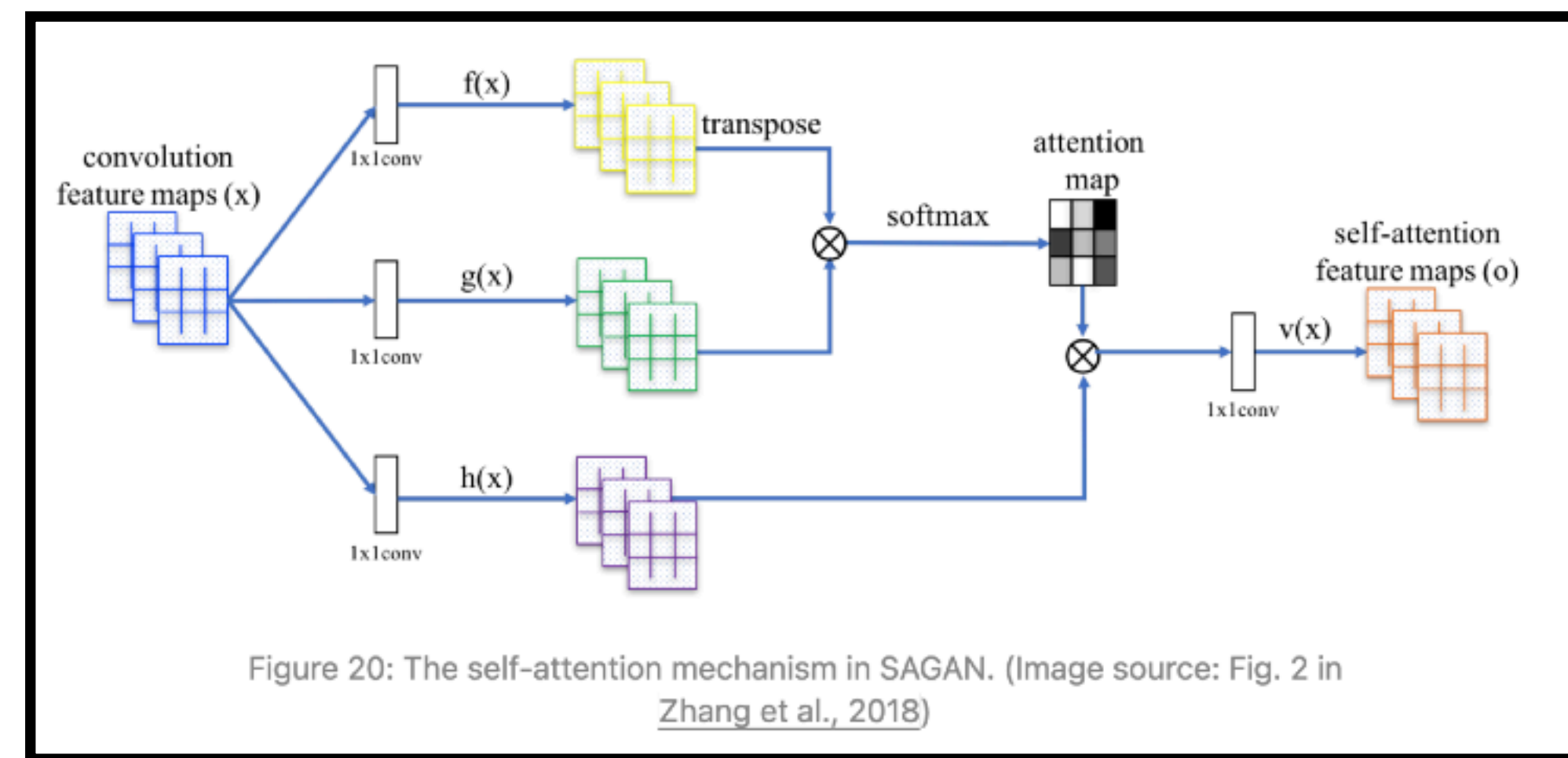
- **Intra-Module Attention (The Local View):**  
Summarize hits *within* each detector module independently
- **Inter-Module Attention (The Global View):**  
Combine module summaries → learn long-range relations & event topology. Also uses global information from other sub detectors (RealCal, HCal, MuTag)



# Attention

## Transformer

- All the sentence processed at the same time
- What is an Attention Map?
  - a grid that tells how much **focus** a model places on different parts of the input text when processing a specific token
- Example: "flag" → "American"
  - The brightest square (highlighted) shows it's paying extremely high attention to the token "**American**"
  - **Attention allows a model to dynamically create relationships between tokens**



**Bright squares** = Strong connection.  
**Dark squares** = Weak connection.