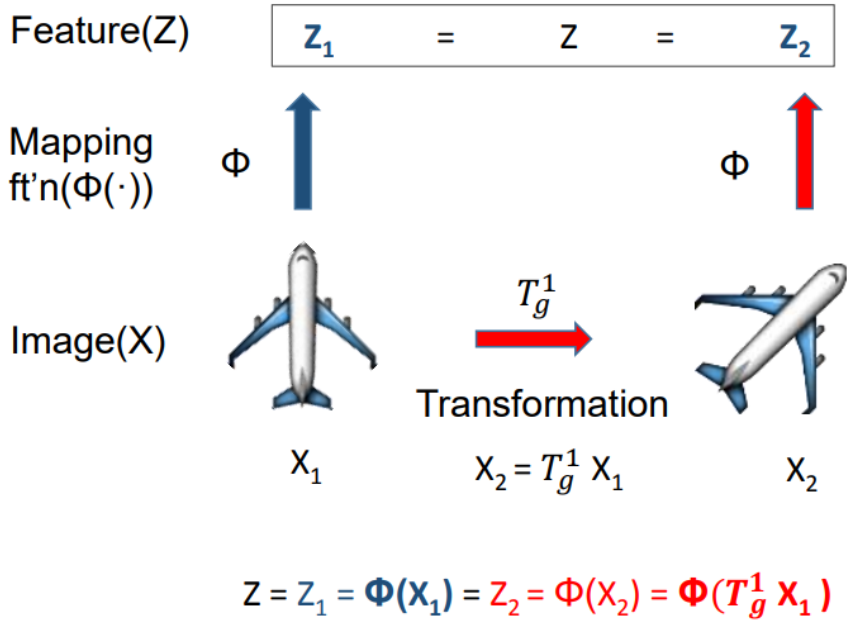
How to represent part-whole hierarchies in a neural network

Geoffrey Hinton

Minho Park

What's wrong with CNN?

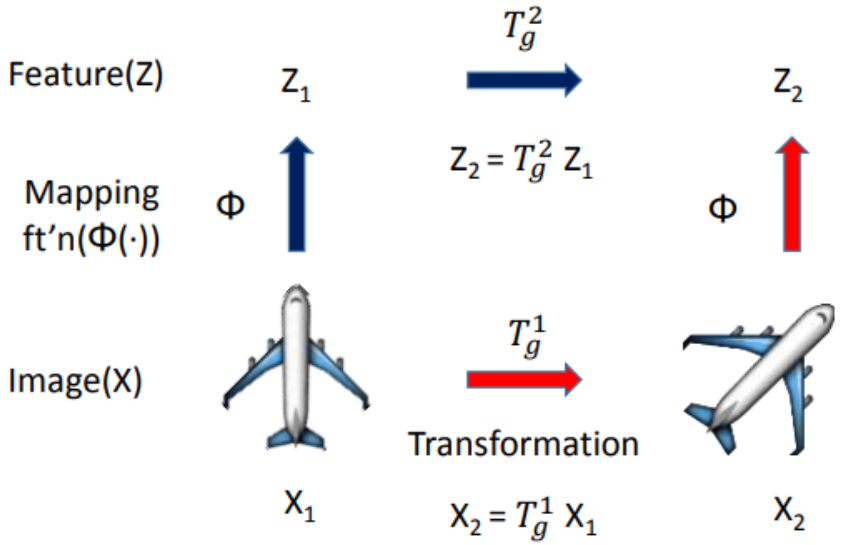
- Invariance



$Z = Z_1 = \Phi(X_1) = Z_2 = \Phi(X_2) = \Phi(T_g^1 X_1)$

What's wrong with CNN?

- Equivariance

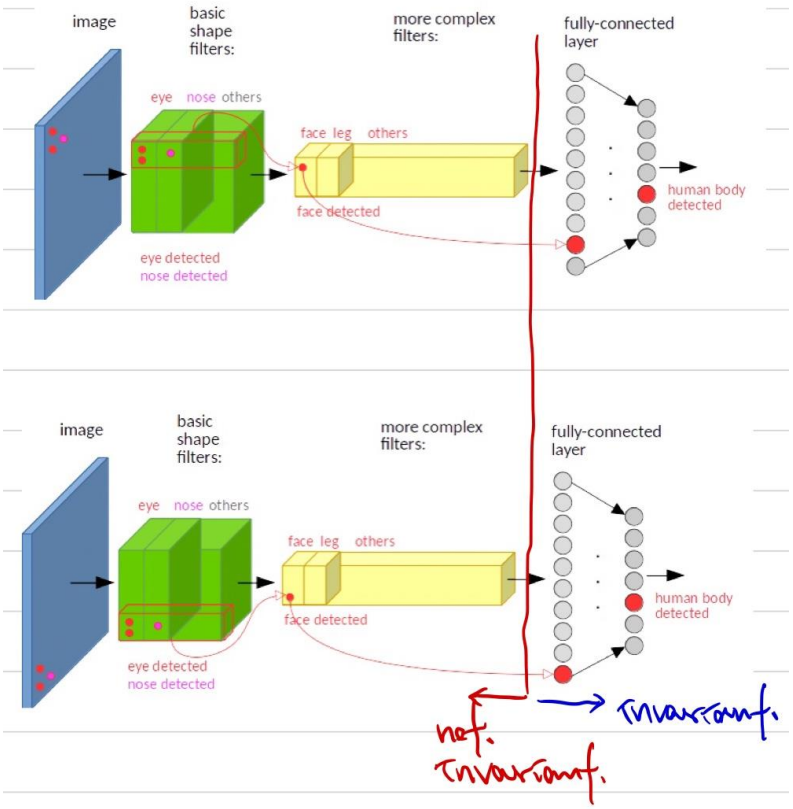


$Z_1 \neq Z_2$ but keeps the relationship $Z_2 = T_g^2 Z_1 = T_g^2 \Phi(X_1) = \Phi(T_g^1 X_1)$

: Invariance is special case of equivariance where T_g^2 is the identity.

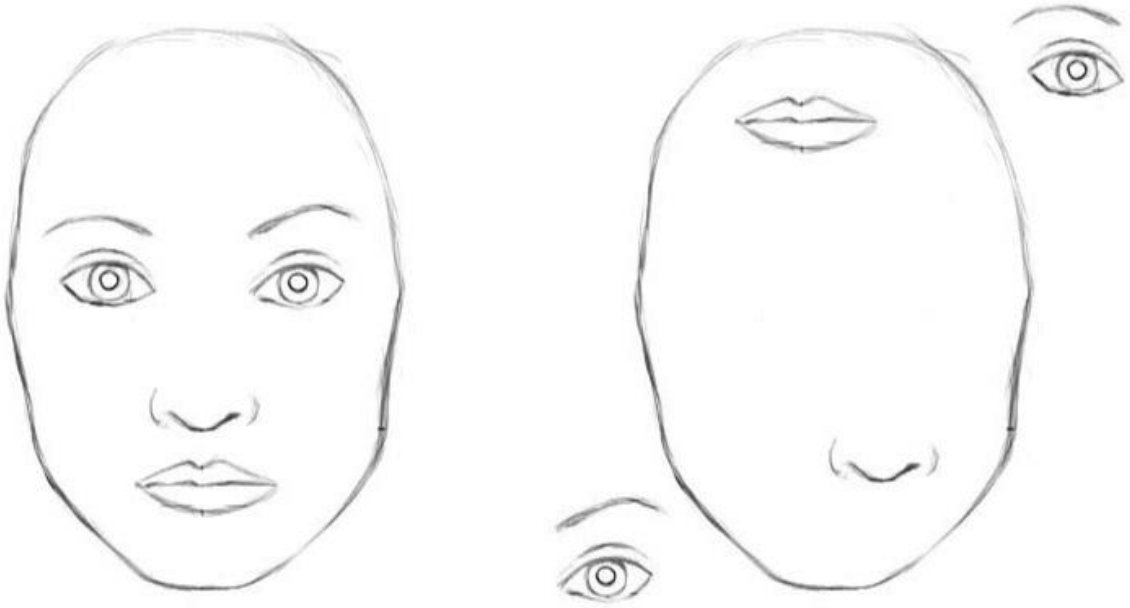
CNN cause invariance

- Because of pooling and fc-layer
- For translation invariance, we use data augmentation



CNN cause invariance

- As a result, CNN recognizes both pictures as human faces



Capsule

- A group of neurons (Vector)
  - Even the same object can have different properties
- Length: Probability of the entity
- Direction(Vector elements): Properties of the entity
- Ex. CapsNet (Dynamic Routing)

Overview of the idea

- Goal: Represent part-whole hierarchies in a neural network
- Why?: Strong psychological evidence
  - People parse visual scenes into part-whole hierarchies
  - and model the viewpoint-invariant spatial relationship between a part and a whole

Overview of the idea

GLOM architecture

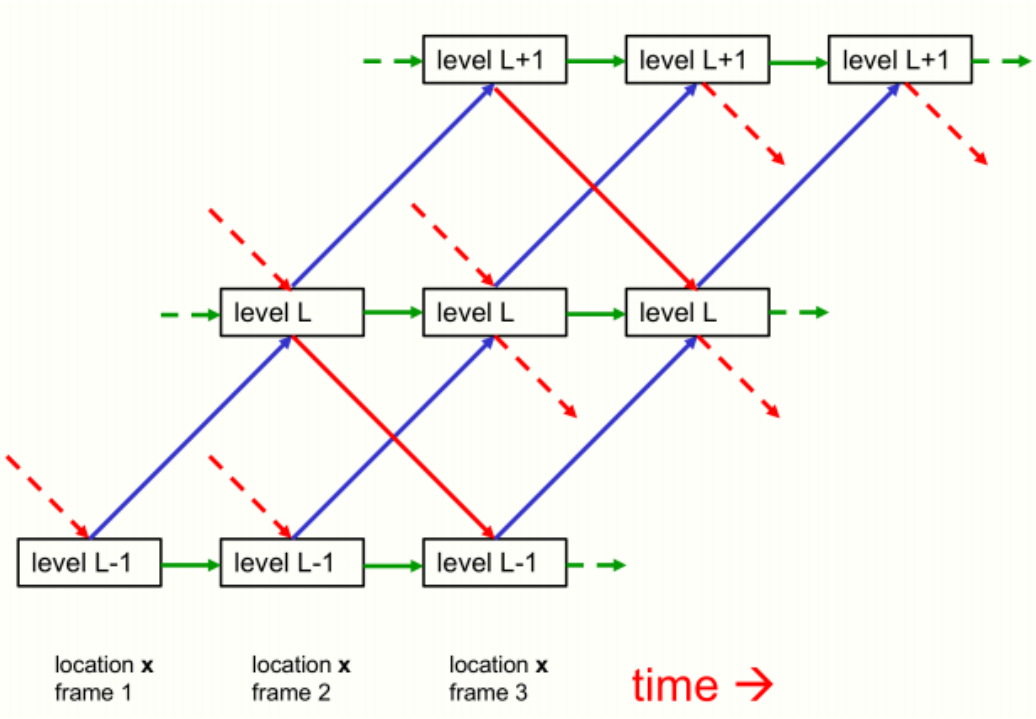
- Columns (autoencoders)

Columns share weight

- Bottom-up encoder
  - from Feature Extractor (e.g. CNN)
- Top-down decoder
  - from Neural Fields
- Previous time and Consensus attention
  - from BERT

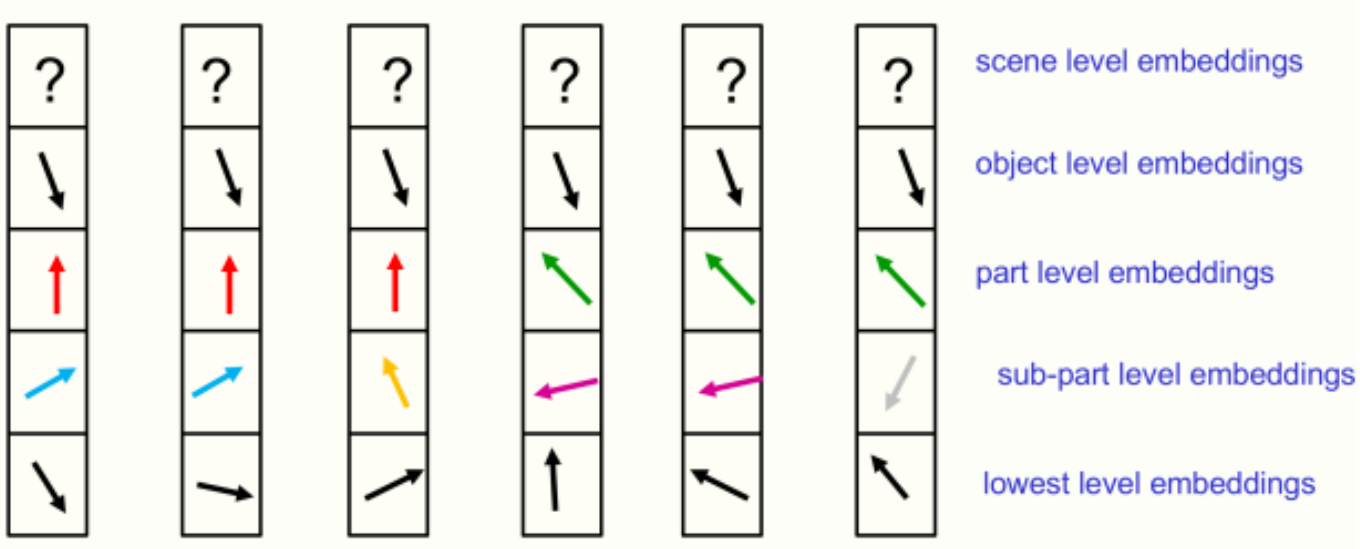
Columns share weight

- Every column share the same weight.



Attention in the same level

- $\mathrm{Softmax}(XX^T)X$ at the same level in nearby columns at the previous time



Discuss design decisions

- How many levels are there?
  - The paper's choice is 5 (biological reason)
- How fine grained are the locations?
  - Pixels or image patches
  - Grid of locations does not have to remain at all levels
- Does the bottom-up net look at nearby location?
  - It can, but it is not pure version of GLOM

Discuss design decisions

- How does the attention work?
- The visual input
  - The output of convolution net could be used the primary

Color and texture

- Imagine answer the color example
  - The color of a part is straightforward, but what color is the whole object?

Color and texture

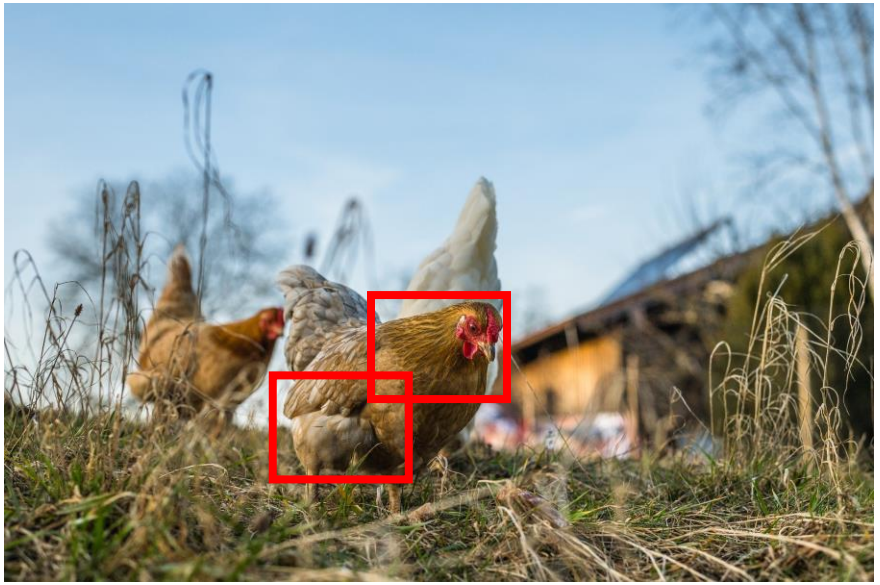
- This is one of the motivation for GLOM
  - The whole object has a compound color which might be called “red-green-or-blue”

Learning Islands

- Assume that GLOM is trained to reconstruct
  - if the regions are sufficiently large, top level will be helpful

Learning Islands

- Contrastive learning
  - An embedding at one location is free to choose which embeddings at other locations, it should resemble



Discuss

- The part-whole hierarchies seems to be a good representation
- Is the representation too large?
- Does it learn exactly the representation we want?
  - We want an interpretable representation
  - Any ideas?

Reference

- CapsNet, GLOM arxiv paper
- Invariance vs. Equivariance: https://www.slideshare.net/ssuser06e0c5/brief-intro-invariance-and-equivariance?from_action=save
- Yannic Kilcher, CapsNet: https://youtu.be/nXGHJTtFYRU
- Jaejun Yoo, CapsNet: https://youtu.be/_YT_8CT2w_Q
- Yannic Kilcher, GLOM: https://youtu.be/cllFzkvrYmE