Self-Attention Agreement Among Capsules

ICCV 2021 – October 2021

Rita Pucci, Christian Micheloni, Niki Martinel

Department of Mathematics Computer Science and Physics

University of Udine

Machine Learning and Perception @

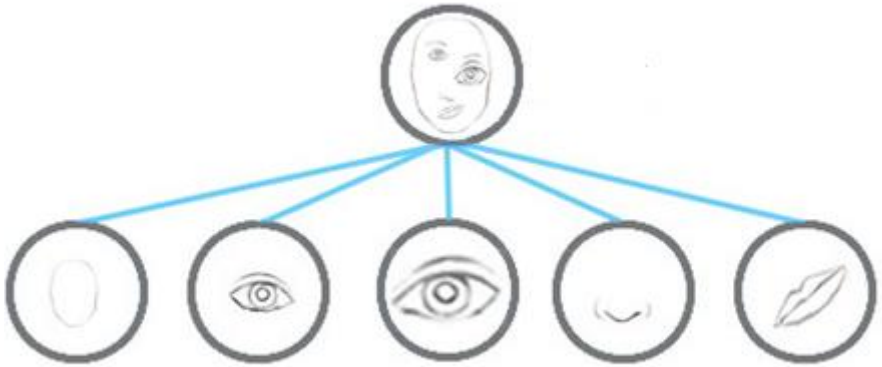
UNIVERSITÀ DEGLI STUDI DI UDINE

hic sunt futura

Classification with entities

- **Entity:** an entity is defined as an object or an object part that can be "seen" in the input image.
- **From entity to classification:** The network computes the probability of presence of an entity in the image. The prediction represents the classification of the image.

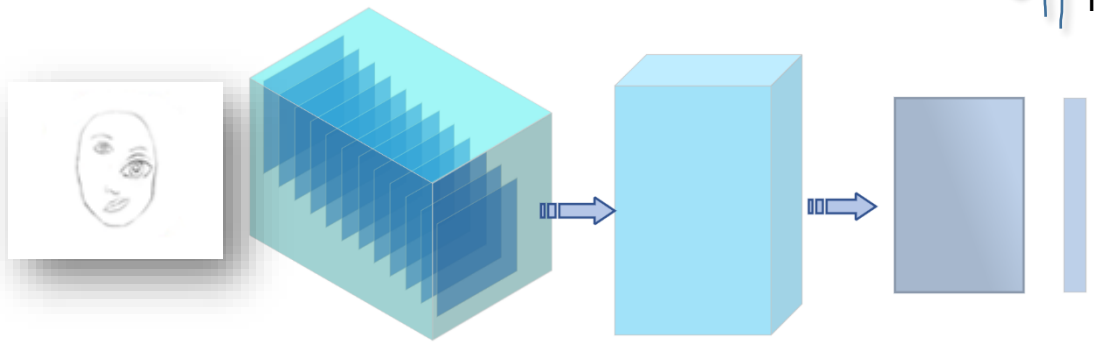
- **Entities in a classification task:** we identify the entity as the class of the image. I.e. the model predicts the presence of entities part of a face with 90% accuracy, that means that the image is classified with the class face.



How can we identify the presence of an entity? With …Capsules

- A **capsule** consists of a **group of neurons** that depicts the properties (position, size, texture) of various entities present in an image.
- While training capsules, the model **automatically selects properties** that are more representative for the recognition of the entities.

Baseline CapsNet(1) and Routing by agreement



Feature extraction *Capsules*

The output of a capsule is the **activity vector (votes)**.

The **activity vector** consists of the activation value of each neuron that composes the capsule.

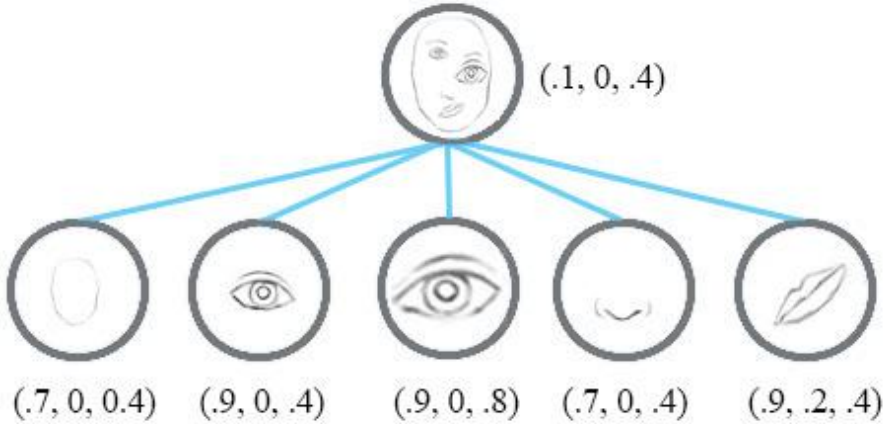
The **activity vectors** extracted by capsules are the different point of view or representation of the image.

Iterative

Increased computational costs

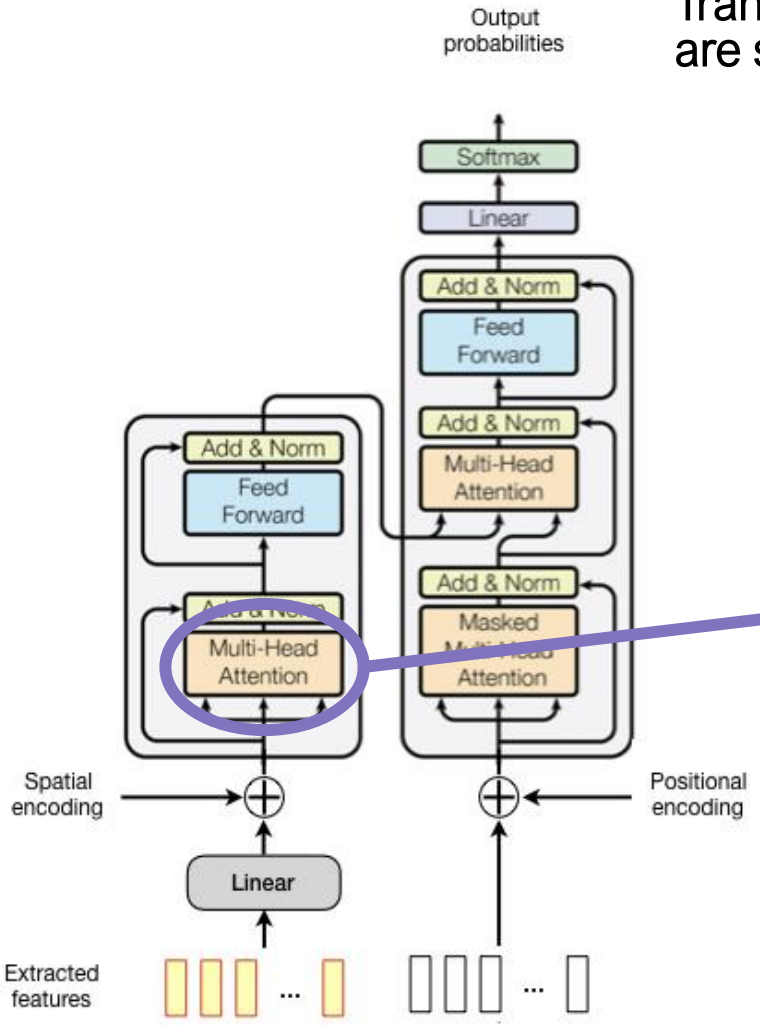
Training → Routing by agreement (aggregate and interpret entity parts)

The Activity vectors (votes) provided by the capsules are used to **compute the agreement** among the capsules



(1) Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. "Dynamic routing between capsules." (2017)

Transformers and Self Attention Layer



Transformers(2) are deep neural network architectures, where **self-attention layers are stacked on top of each other** in an encoder-decoder structure.

- Each layer computes a **representation per input** attending to the **representations of all parts from the previous layer**.

Multi Head Attention runs through an attention mechanism several times in parallel.

Each attention head gets **a different projection of the representation**.

Vectors with **larger probabilities receive additional focus** from the following layers.

(2)Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Similarities between the two mechanisms

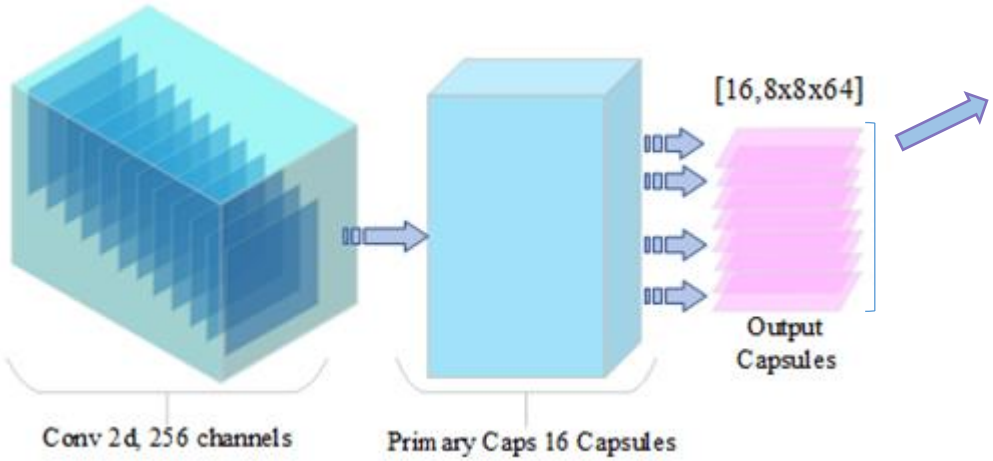
IDEA – Extracting a probability distribution that describes the relation among the lower layer and the upper layer.

	Routing by agreement	Attention Layer
	u Capsules output at lower layer	$u \Rightarrow \mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$
Votes transformation	$\hat{u}_{i\|j} = \mathbf{W}_{ij} u_i$	$\mathbf{S} = \mathbf{Q}\mathbf{K}^T \qquad \mathbf{S}_N = \frac{\mathbf{S}}{\sqrt{d_m}}$
Coefficient between lower and higher layers	$c_{ij} = \frac{exp(b_{ij})}{\sum_k exp(b_{ik})}$	
Probability of an object of been present	$s_j = \sum_i c_{ij} \hat{u}_{j\|i}$	$\mathbf{P} = softmax(\mathbf{S}_N)$
	$v_j = \frac{\|\|s_j\|\|^2}{1 + \|\|s_j\|\|^2} \frac{s_j}{\|\|s_j\|\|}$	$\mathbf{Z} = \mathbf{PV}$

Iterative mechanism

Single forward pass

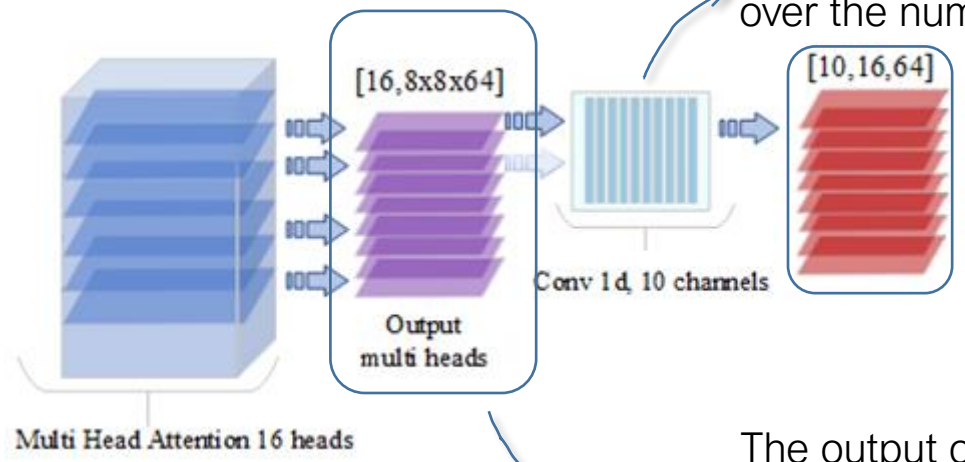
Attention Agreement-Caps Idea



Different capsules types provides different point of view over the capsules at the previous layer

u

Each attention head receives in input the output of a single capsule



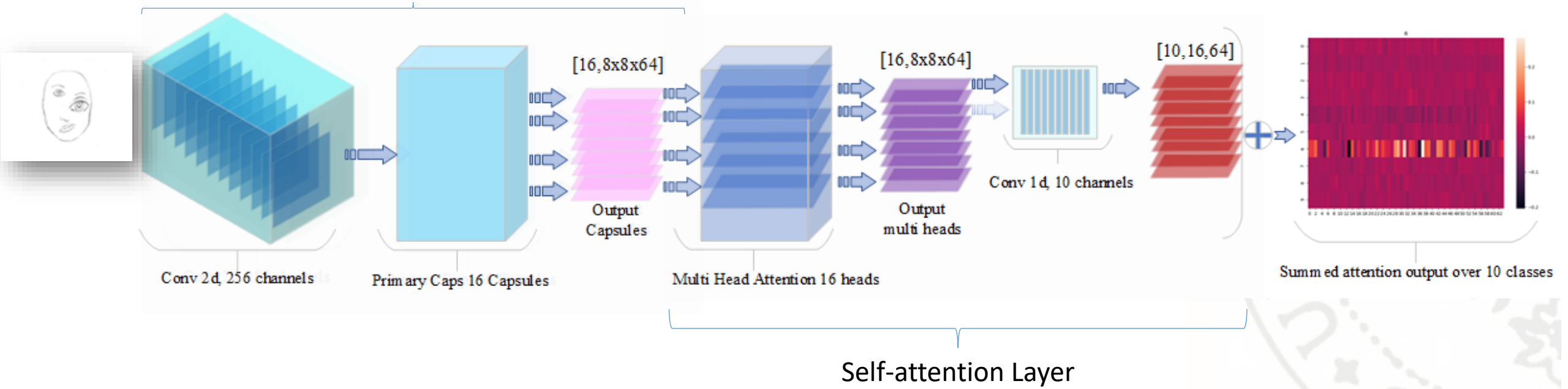
Map the distributed attention over the number of classes

Feature map obtained for each class

The output of the multi-headed attention block consists of independent attention maps

AA-Caps Architecture

Feature extraction with Capsules



Self-attention Layer

Classification results

Model	MNIST	SVHN	CIFAR10	SmallNORB	AwA2
Baseline CapsNet	99.67% (100E)	93.23% (100E)	68.70%	89.56% (50E)	12.1% (100E)
AA-Caps (Ours)	99.34% (100E)	92.13% (100E)	71.60%	89.72% (50E)	23.97% (100E)

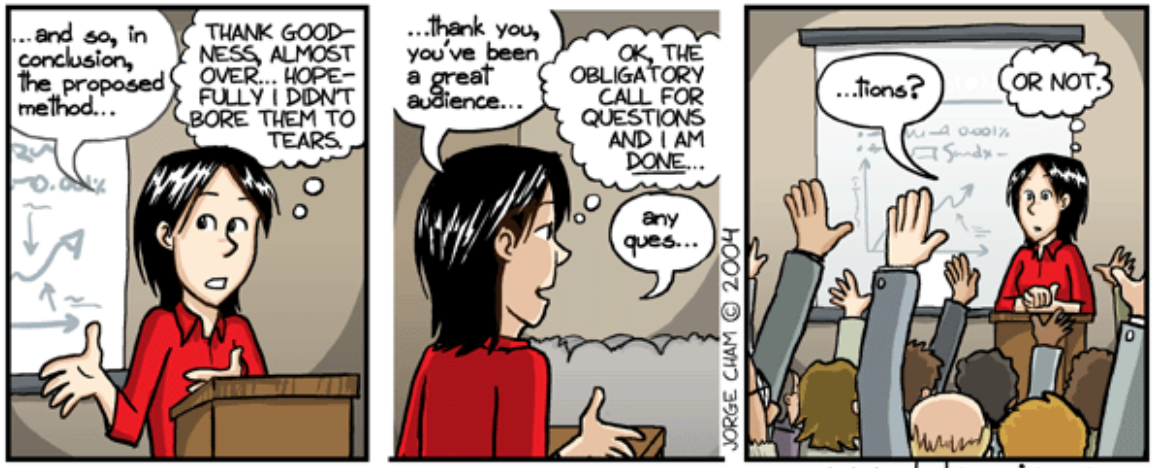
Can we do a compromise between accuracy and number of parameters?

Model	Description	Parameters	Test Acc.
Baseline CapsNet	Conv - Primary Capsules - Final Capsules	8.2M	99.67%
AA-Caps (Ours)	Conv - Primary Capsules - Self-Attention - Conv	6.6M	99.34%

We propose a competitive base line for capsules models with non-iterative aggregation mechanism.

Questions?

Thank you for listening to this presentation!



Self-Attention Agreement Among Capsules

Rita Pucci, Christian Micheloni, Niki Martinel



Contact: rita.pucci@uniud.it