College of Health Solutions
Arizona State University

EMBEDDED MACHINE INTELLIGENCE LAB

Multi-view classification with convolutional neural networks

Presenter: Nooshin Taheri

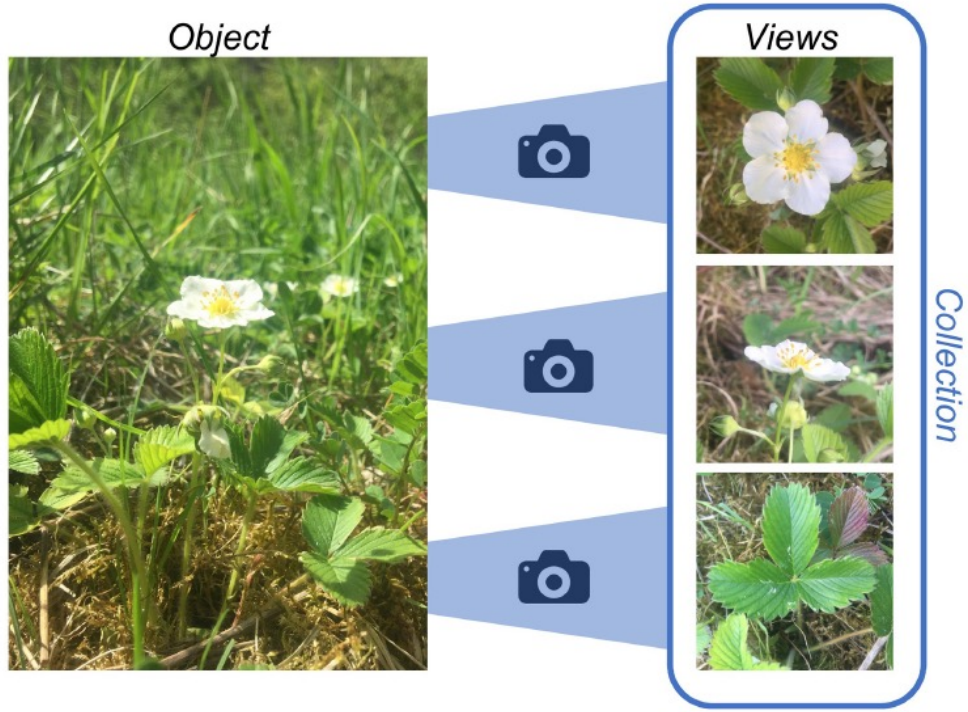
1/15/2025

Published at: PLOS one, 2021

Cited: 130

Introduction

- Humans' decision-making process often relies on utilizing visual information from different views or perspectives.
- Traditional image classification relies on single images, which may lack sufficient information for challenging tasks.
- **Solution:** Multi-view classification incorporates multiple images (views) of the same object from different perspectives to improve accuracy.



Introduction

- How can we effectively integrate multiple views of the same object into a CNN to boost classification performance?

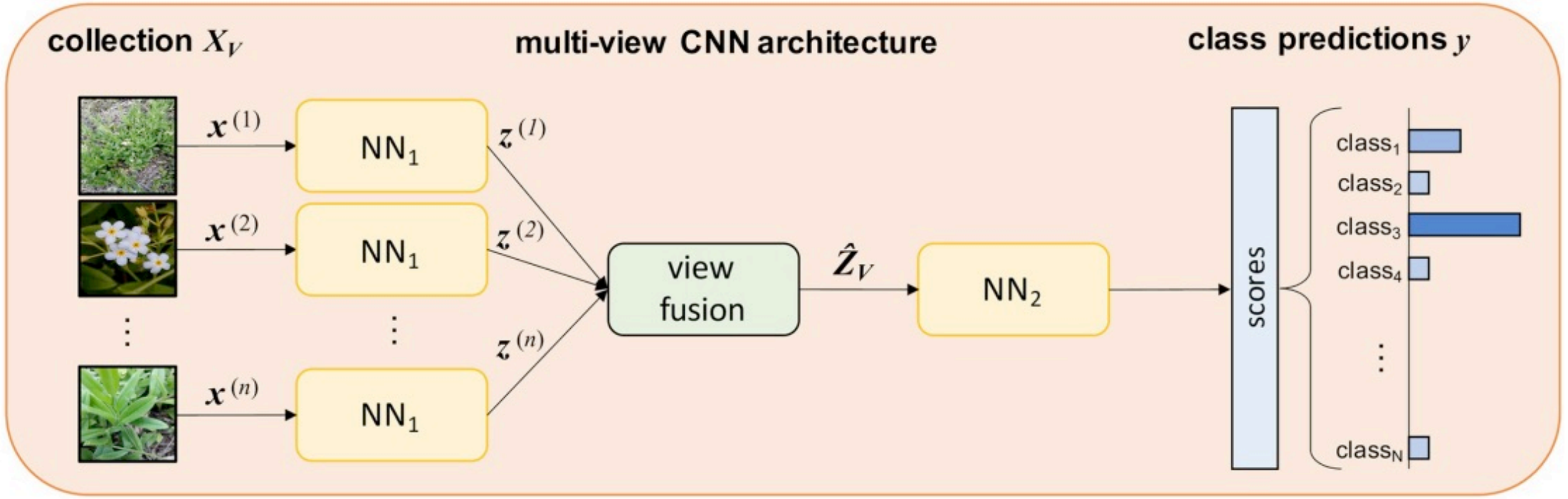
- Approach:
  - Use CNNs to extract features from each view.
  - Explore three fusion strategies:
    - Early fusion (feature map-level fusion).
    - Late fusion (latent representation fusion).
    - Score fusion (aggregation of classification scores).

Contributions

- Systematic comparison of three fusion strategies (early, late, and score).
- Evaluation of three diverse datasets: cars, plants, ants
- Demonstration of significant accuracy gains over single-view baselines.

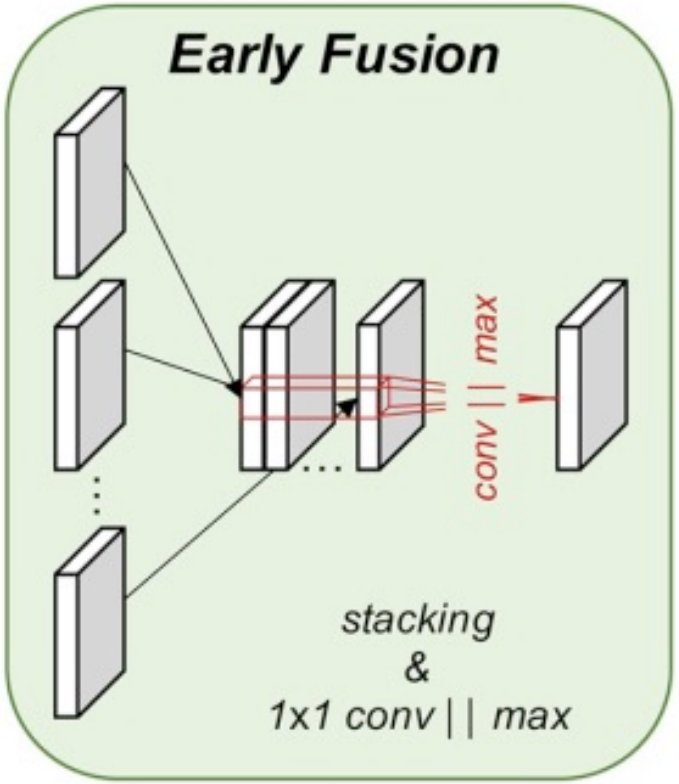
General architecture of a deep multi-view CNN

- The original CNN is split into two components:
  - **NN1:** Processes each view ($x^{(v)}$) independently, extracting intermediate representations.
  - **View Fusion Layer:** Combines the outputs ($z^{(v)}$) from all branches into an aggregated representation ($\hat{Z}$).
  - **NN2:** Processes the aggregated representation ($\hat{Z}$) to generate predictions.



Early Fusion

- Convolutional feature maps from the different CNN branches are stacked and subsequently processed together.



- **Fusion Methods**
  1. **Max-pooling (Early Fusion Max):**
     - Takes the maximum value across feature maps for each spatial position.
     - Reduces dimensionality but loses correspondence between views.
  2. **1×1 Convolution (Early Fusion Conv):**
     - Uses trainable kernels to combine feature maps across views.
     - Preserves inter-view correspondence but increases trainable parameters.

Late Fusion

- **Late fusion** relies on aggregating the output of the last layer before the classification layer, or, in the case of multiple fully connected layers at the top, the classification block, as latent representation.

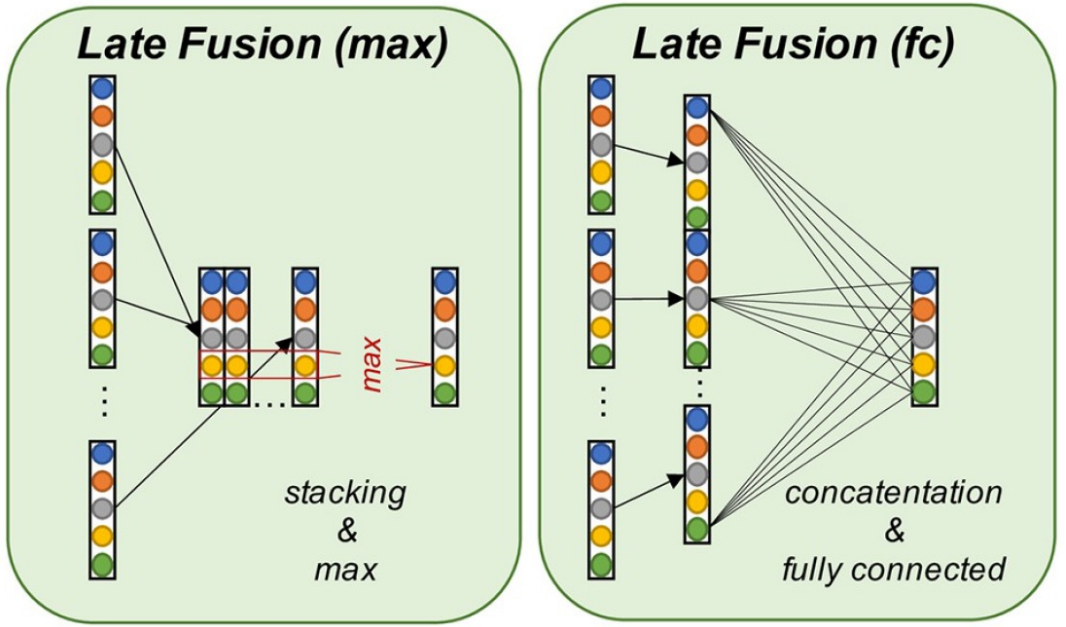
- **Fusion Methods**

  1. Max-pooling (Late Fusion Max):

  - Selects the maximum value for each dimension across feature vectors.
  - Simple, but may lose inter-view relationships.

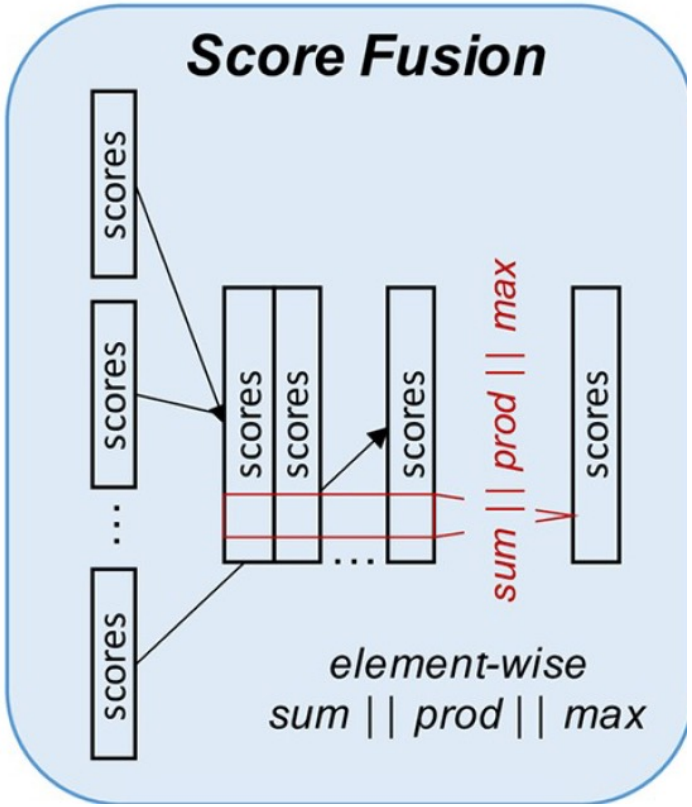
  2. Feature Concatenation (Late Fusion FC):

  - Concatenates feature vectors from all views.
  - A fully connected layer learns optimal combinations, preserving view relationships.



Score Fusion

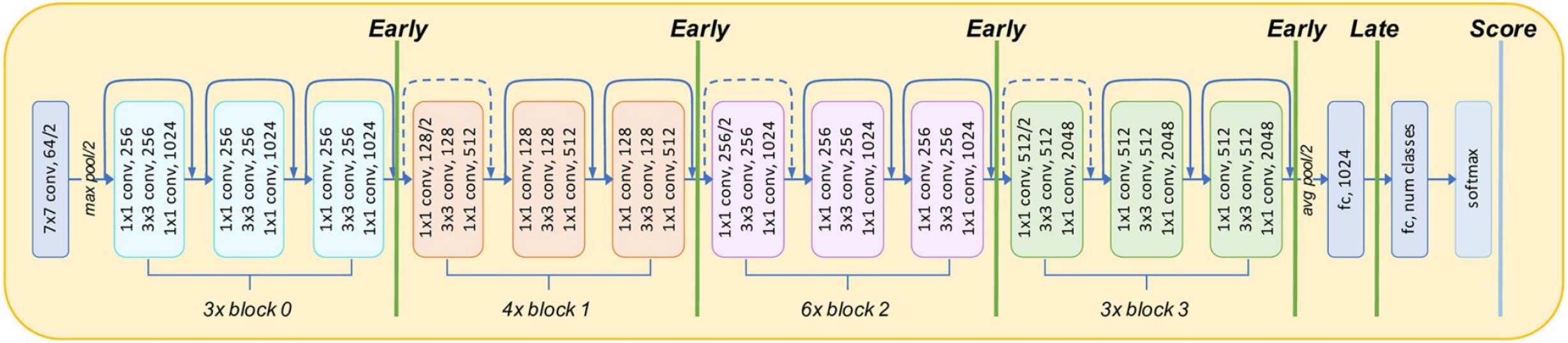
- **Score fusion** is based on the element-wise aggregation of the softmax classification scores per branch.
- Combines classification scores from multiple views after each view is processed independently.



- **Fusion Methods**
  1. **Sum Fusion:** Adds scores from all views.
  2. **Product Fusion:** Multiplies scores across views.
  3. **Max Fusion:** Selects the maximum score from all views.

Backbone Architecture

- **ResNet-50** is used as the CNN for feature extraction (NN1) and classification (NN2).
- Chosen for its modular design and superior **feature extraction capabilities**.
- Fusion layers are inserted at **various points** in the ResNet-50 architecture to test Early, Late, and Score fusion strategies.
  - A total of **14 experiments per dataset**:
    - **8 Early Fusion** experiments, **2 Late Fusion** experiments, **3 Score Fusion** experiments.
    - **1 Single-view Baseline.**



Experimental Procedure

1. **Single-view Baseline Training:**
   - Train a general-purpose CNN on single images.
   - Acts as the baseline for comparison.

2. **Multi-view Network Initialization:**
   - Duplicate NN1 branches for each view.
   - Add the respective fusion layer (Early, Late, or Score fusion) and NN2.

3. **Weight Freezing:**
   - Freeze NN1 branches to ensure consistent feature extraction.
   - Performance differences arise solely from the fusion strategy.

4. **Training with Multi-view Collections:**
   - Optimize weights of the fusion layer and NN2.
   - Use a constant number of images per batch for fair comparison.

Dataset

a)

b)

c)

Dataset	Images	Classes	Views per Collection	Description
CompCars	40,915	601	Front, front-side, rear, rear-side, side (5 views)	Fine-grained car model classification.
PlantCLEF	3,678	53	Flower and leaf (2 views)	Plant species identification using organ-based views.
AntWeb	116,742	82	Dorsal, head, profile (3 views)	Ant genera classification using structured collections.

Results

Table 3. Multi-view classification results across the three datasets.

Method	Layer	top-1 [%] & δ_{BL} [%]					
		PlantCLEF		CompCars		AntWeb	
Worst single view		65.80		65.29		79.34	
Best single view		81.32		82.33		87.65	
Avg. across single views		73.56		76.61		84.58	
Early (max)	block 0	65.80	-19.1	53.91	-34.5	77.79	-11.2
	block 1	70.98	-12.7	73.95	-10.2	85.65	-2.3
	block 2	83.62	2.8	93.64	13.7	89.73	2.4
	block 3	85.34	4.9	95.11	15.5	92.28	5.3
Early (conv)	block 0	42.82	-47.3	54.68	-33.6	87.31	-0.4
	block 1	57.76	-29.0	74.44	-9.6	90.51	3.3
	block 2	79.60	-2.1	94.13	14.3	92.35	5.4
	block 3	89.37	9.9	**95.11**	**15.5**	**95.16**	**8.6**
Late (max)	fc	**90.23**	**11.0**	92.82	12.7	**93.93**	**7.2**
Late (fc)	fc	**94.25**	**15.9**	**96.72**	**17.5**	**94.54**	**7.9**
Score ($\otimes$)	softmax	**89.66**	**10.3**	**95.74**	**16.3**	91.43	4.3
Score ($\oplus$)	softmax	86.78	6.7	94.69	15.0	89.70	2.3
Score (max)	softmax	85.34	4.9	92.88	12.8	89.48	2.1

Best methods are highlighted in 1st—green, 2nd—light green, 3rd—gray green font color. Red values indicate results worse compared to baseline results. Single view accuracy results refer to the worst performing single view, the best performing single view, as well as to the average across all available views. δ_{BL} is calculated as relative difference to the best single view result.

Conclusion

- Multi-view classification using CNNs significantly improves accuracy by leveraging information from multiple perspectives of the same object.
- **Late Fusion:**
  - Achieved the highest accuracy across all datasets.
  - Efficiently combines high-level latent representations with moderate computational cost.
  - Late fusion strategies can be seamlessly integrated into existing architectures, demonstrating flexibility and scalability.

College of Health Solutions
Arizona State University

EMBEDDED MACHINE INTELLIGENCE LAB

Thank You!

ntaheric@asu.edu