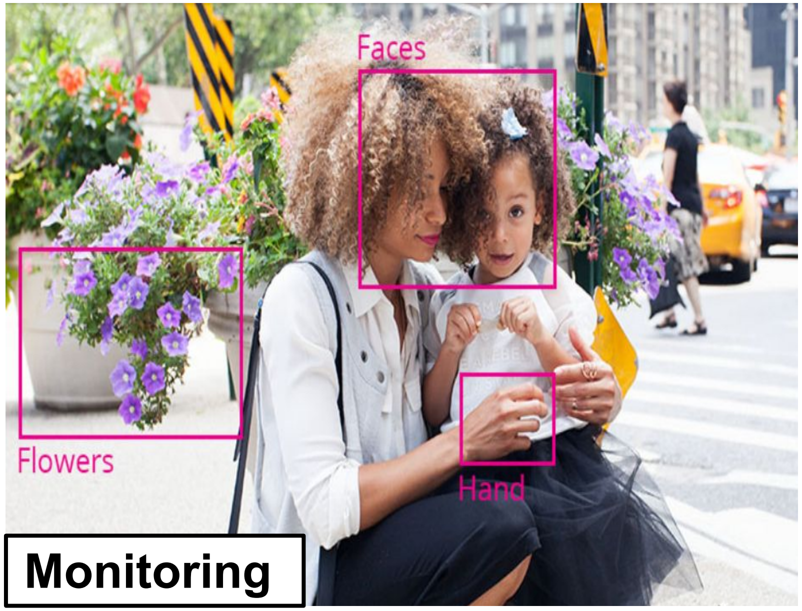
# On The Universality of Visual and Multimodal Representations

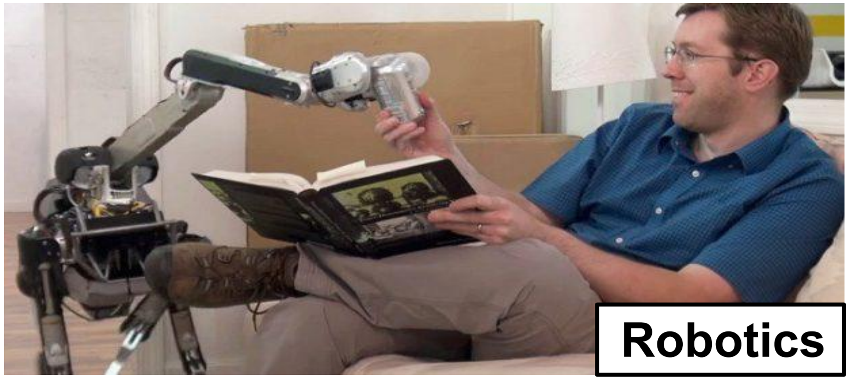
**Jury**
Mathieu Cord
Céline Hudelot
Hervé Le Borgne
Pablo Piantanida
Philippe-Henri Gosselin
Iasonas Kokkinos
Florent Perronnin

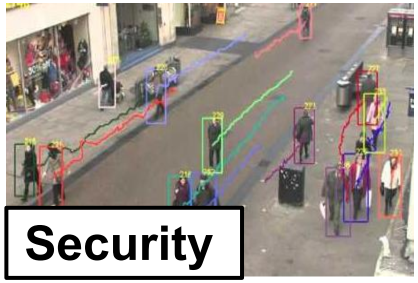
**Youssef Tamaazousti | Ph.D. Defense**

June 1st, 2018

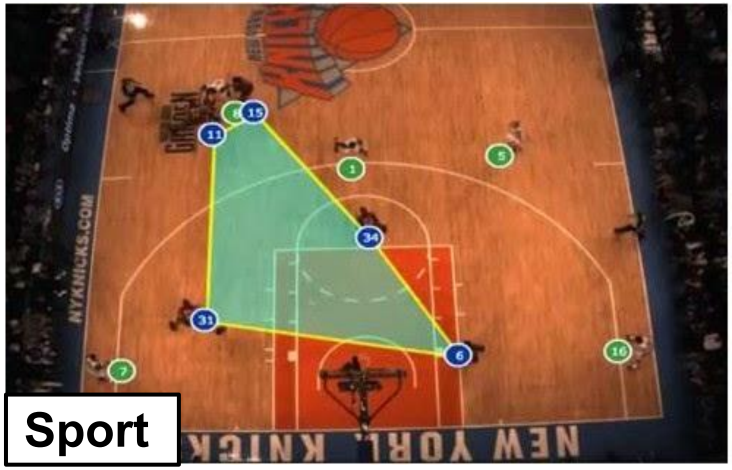
# AI Today: performing systems in many tasks and domains

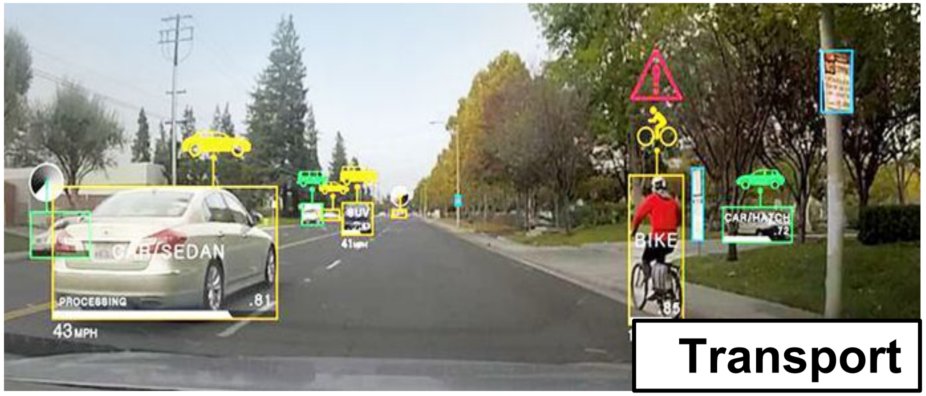


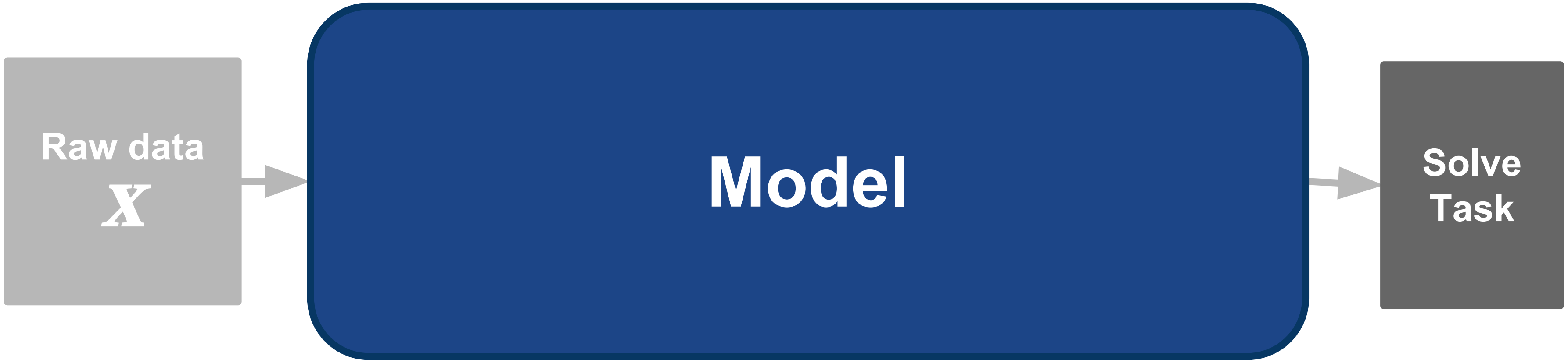







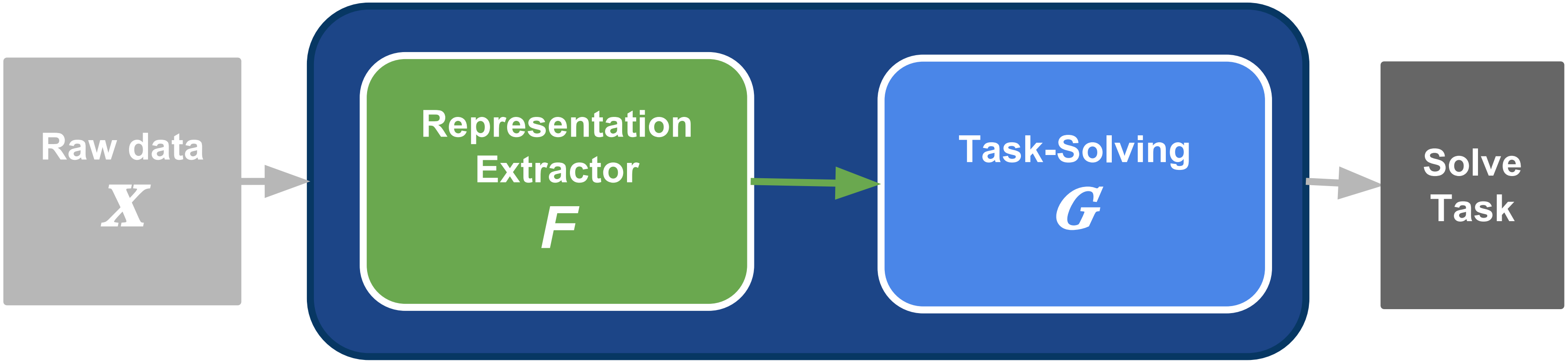



# Learning-based AI



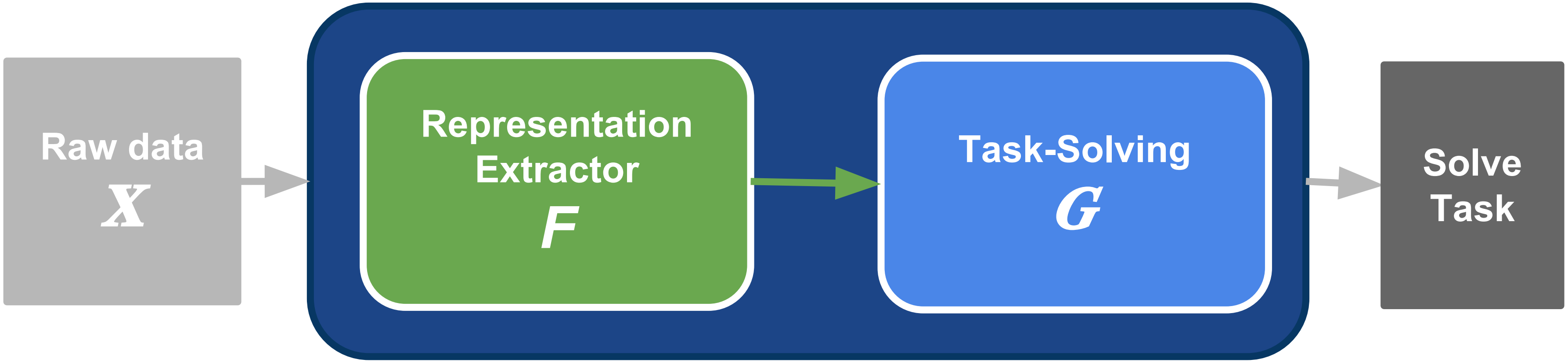

- **Learning-based AI**
  - **Aims at** performing tasks from raw data

# Learning-based AI




- **Learning-based AI**
  - **Aims at** performing tasks from raw data
  - **Consists in** a Representation-extractor (***F***) and a Task-solving (***G***)

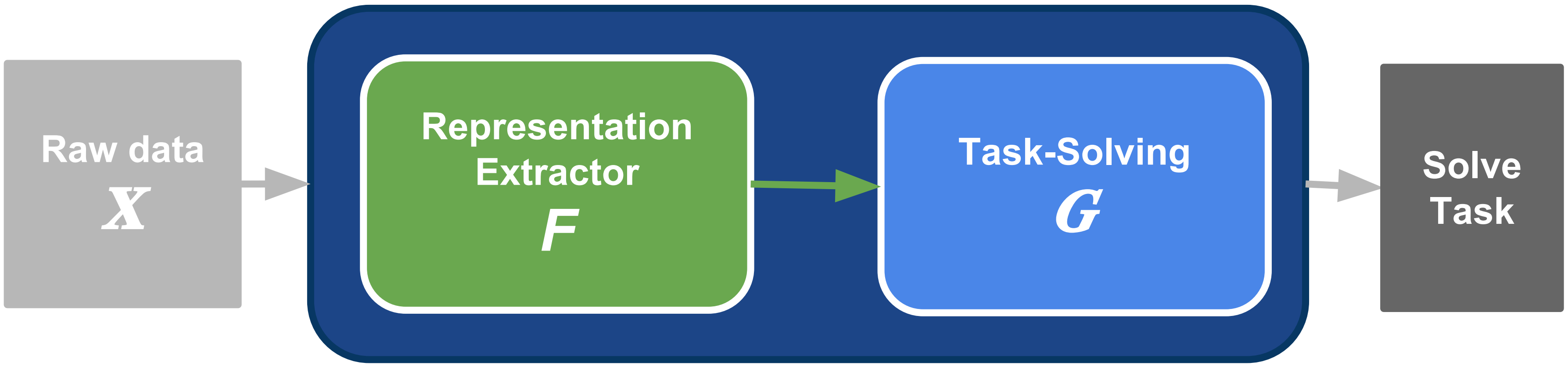
# Learning-based AI




- **Learning-based AI**
  - **Aims at** performing tasks from raw data
  - **Consists in** a Representation-extractor (***F***) and a Task-solving (***G***)

- **Main Characteristics:**
  - ***F*** learned from data
  - ***F*** and ***G*** learned jointly
  - ***G*** could be omitted, ***F*** used with another ***G*** to solve another task: ``Transferability''

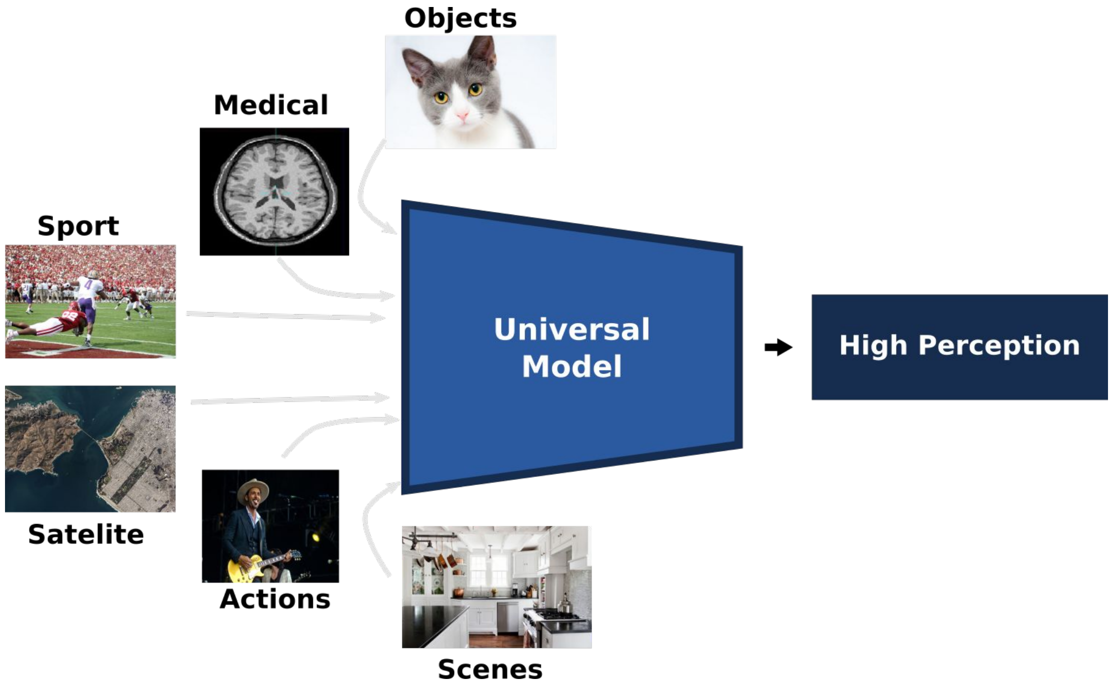
# Learning-based AI




- **Goal in the literature:**
  - Learning a model (***F*** and ***G***) in order to excel at a given task

# Challenge

- **Learning a universal model:**
  - Model that provides **high-level representation** of raw data from different nature (modalities, visual domains and semantic domains)
  - **high task-solving abilities** for different tasks (recognition, detection, segmentation, etc.).



# Motivation

- **Humans:**
    - able to perform an enormous variety of different tasks.

- **Machines:**
    - able to perform one task at time (``expert model'')


# Motivation

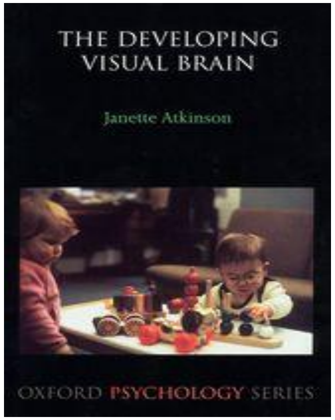
- **Humans:**
  - able to perform an enormous variety of different tasks.

- **Machines:**
  - able to perform one task at time (``expert model'')

I am good at painting, counting, talking, and so many other things !!

I am an expert in cat recognition !

**Humans develop powerful internal representation in their infancy and re-use it later in life to solve many problems** [Atkinson, OPP'00]



# Motivation

- **Universality: recent growing interest in AI community**

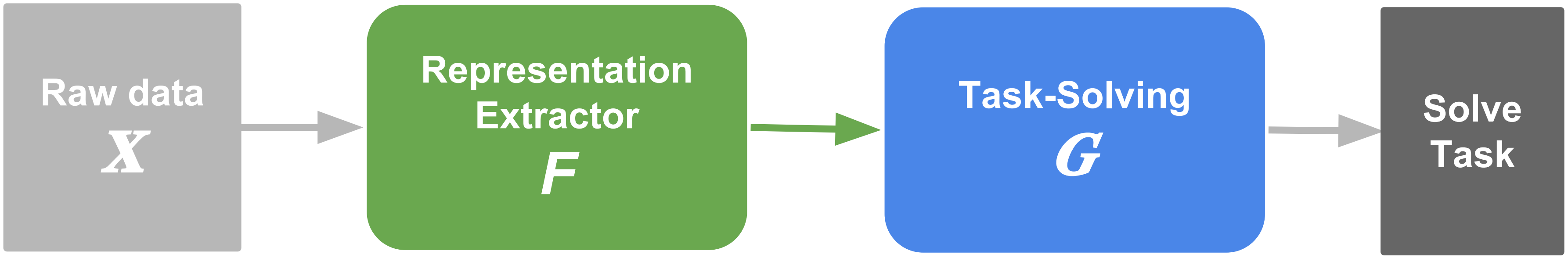
- **Motivations of other works**

  - Same motivation than us: **``mimic'' humans**
    - [Bilen & Vedaldi, ArXiv'17]; [Rebuffi *et al.*, NIPS'17]; [Nie *et al.*, ArXiv'17]; [Rebuffi *et al.*, CVPR'18]

  - **Practical motivation**: even if we want to build an expert AI, it is always beneficial to have a good starting point (universal model)
    - [Conneau *et al.*, EACL'17]; [Conneau *et al.*, EMNLP'17]; [Cer et al., ArXiv'18]; [Subramanian & Bengio, ICLR'18];

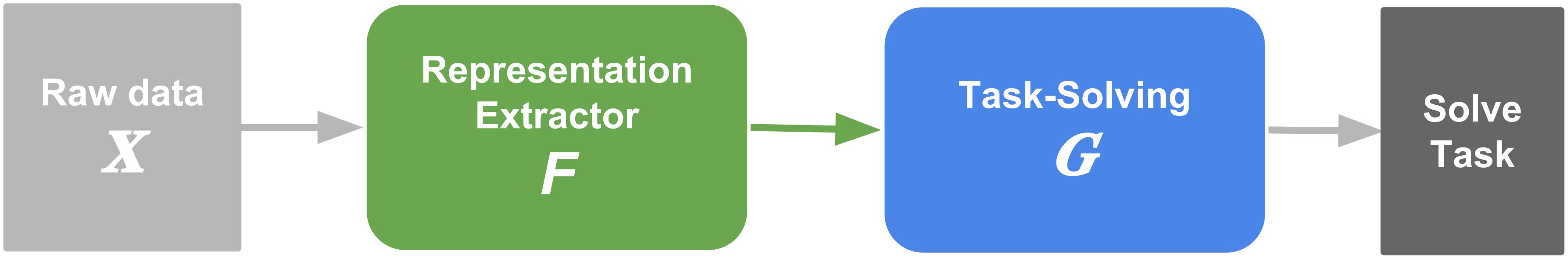
  - Build a ``swiss-knife'' that may be useful for **general AI**
    - [Kokkinos, CVPR'17]; [Wang et al., WACV'18]

# General Problem Formulation



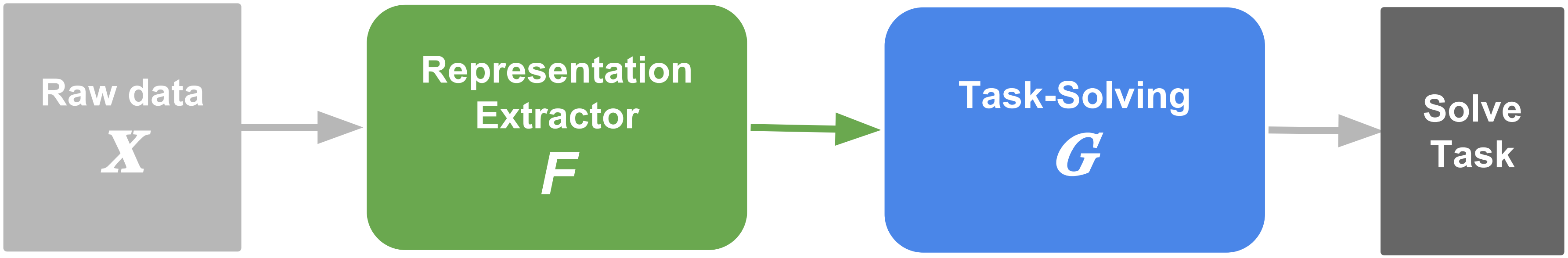

- At least, two different aspects to address the problem

# General Problem Formulation




- **At least, two different aspects to address the problem**
  - Universal Task-Solving: make ***G*** able to handle the largest set of tasks **GENERAL AI**
    [Kokkinos, CVPR'17]; [Wang *et al.*, WACV'18]

# General Problem Formulation




- **At least, two different aspects to address the problem**
  - Universal Task-Solving: make ***G*** able to handle the largest set of tasks **GENERAL AI**
    [Kokkinos, CVPR'17]; [Wang *et al.*, WACV'18]
  - Universal Representation-Extractor: make ***F*** able to handle the largest set of modalities, visual & semantic domains
    **UNIVERSAL REPRESENTATIONS**
    [Bilen & Vedaldi, ArXiv'17]; [Rebuffi *et al.*, NIPS'17]; [Nie *et al.*, ArXiv'17]; [Rebuffi *et al.*, CVPR'18]; [Conneau *et al.*, EACL'17]; [Conneau *et al.*, EMNLP'17]; [Cer et al., ArXiv'18]; [Subramanian & Bengio, ICLR'18]

# Problem Formulation (1/4)

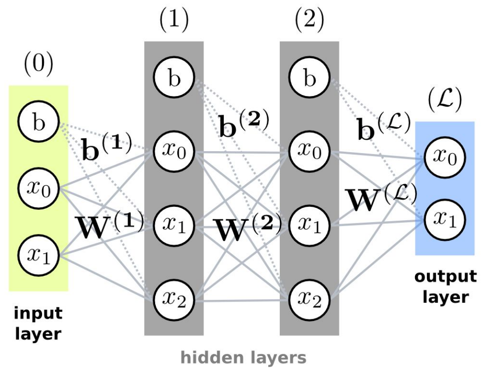
- **A priori, no representation is completely universal**

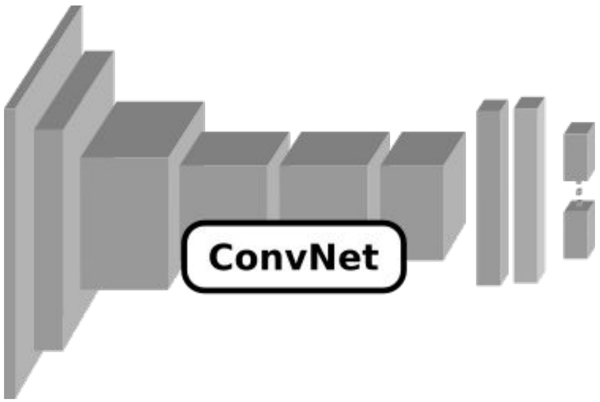
- **Learned representations contain some level of universality**

- **Our goal:**
    - Increase the universality of the representation

# Problem Formulation (2/4)

- **Learning algorithm:**
  - (Deep) neural-networks




- **Data:**
  - Visual or Multimodal (visual & textual)

# Problem Formulation (3/4)

- **Learning strategy:**
  - According to a **supervised** approach
    - better than semi-supervised and unsupervised approaches
  - With **many annotated data**

# Problem Formulation (4/4)

- **Evaluation scenario of universality:**

**Close to [Atkinson, OPP'00]**: Humans learn a visual representation of the world in their infancy and use it (as-is) later in life to solve different problems

a. **In Transfer-Learning scheme,**
   **Infancy**: source-task; **later**: target-task
b. **As-is: w/o modifying the learned representation**
c. **Different problems: Large set of Undetermined Target-Tasks (UTT)**

**Close to the real-world**: most tasks (in academy & industry) contain few annotated data because hard to collect & annotate

d. **UTT with few annotated data**

e. **Aggregated performance on set of UTT**

# Outline

- **State-Of-The-Art (S.O.T.A)**
- **Contributions**
  - Evaluation of Universality
  - Universality in Features Learned with Explicit Supervision
  - Universality in Features Learned with Implicit Supervision
  - Universality via Multimodal Representations
- **Conclusions**
- **Perspectives**

# S.O.T.A: Positioning

| Works | Univ. Aspect |
|---|---|
| [Conneau et al., EACL'17] [Conneau et al., EMNLP'17] | Repres-entation |
| [Cer et al., ArXiv'17] | |
| [Subramanian & Bengio, ICLR'18] | |
| [Kokkinos, CVPR'17] [Wang et al., WACV'18] | Task Solving |
| [Bilen & Vedaldi, ArXiv'17] [Rebuffi et al., NIPS'17] | Repres-entation |
| [Rebuffi et al., CVPR'18] | |
| This Thesis | |

# S.O.T.A: Positioning

| Works | Univ. Aspect | Mod. |
| --- | --- | --- |
| [Conneau et al., EACL'17] [Conneau et al., EMNLP'17] | Repres-entation | Textual |
| [Cer et al., ArXiv'17] | | |
| [Subramanian & Bengio, ICLR'18] | | |
| [Kokkinos, CVPR'17] [Wang et al., WACV'18] | Task Solving | Visual |
| [Bilen & Vedaldi, ArXiv'17] [Rebuffi et al., NIPS'17] | Repres-entation | |
| [Rebuffi et al., CVPR'18] | | |
| This Thesis | | Visual & Multimodal |

# S.O.T.A: Positioning

| Works | Univ. Aspect | Mod. | Eval. Scenario |
|---|---|---|---|
| [Conneau et al., EACL’17] [Conneau et al., EMNLP’17] | Repres-entation | Textual | Transfer Learning |
| [Cer et al., ArXiv’17] | | | |
| [Subramanian & Bengio, ICLR’18] | | | |
| [Kokkinos, CVPR’17] [Wang et al., WACV’18] | Task Solving | Visual | End2End |
| [Bilen & Vedaldi, ArXiv’17] [Rebuffi et al., NIPS’17] | Repres-entation | | |
| [Rebuffi et al., CVPR’18] | | | Fine Tuning |
| **This Thesis** | | Visual & Multimodal | Transfer Learning |

# S.O.T.A: Positioning

| Works | Univ. Aspect | Mod. | Eval. Scenario | SP Domain-Task |
|---|---|---|---|---|
| [Conneau et al., EACL'17] [Conneau et al., EMNLP'17] | Repres-entation | Textual | Transfer Learning | 1 domain - 1 task |
| [Cer et al., ArXiv'17] | Repres-entation | Textual | Transfer Learning | 1 domain - No annotation |
| [Subramanian & Bengio, ICLR'18] | Repres-entation | Textual | Transfer Learning | Multi-task |
| [Kokkinos, CVPR'17] [Wang et al., WACV'18] | Task Solving | Visual | End2End | Multi-task |
| [Bilen & Vedaldi, ArXiv'17] [Rebuffi et al., NIPS'17] | Repres-entation | Visual | End2End | Multi-domain - 1 task |
| [Rebuffi et al., CVPR'18] | Repres-entation | Visual | Fine Tuning | Multi-domain - 1 task |
| This Thesis | Repres-entation | Visual & Multimodal | Transfer Learning | 1 domain - 1 task |

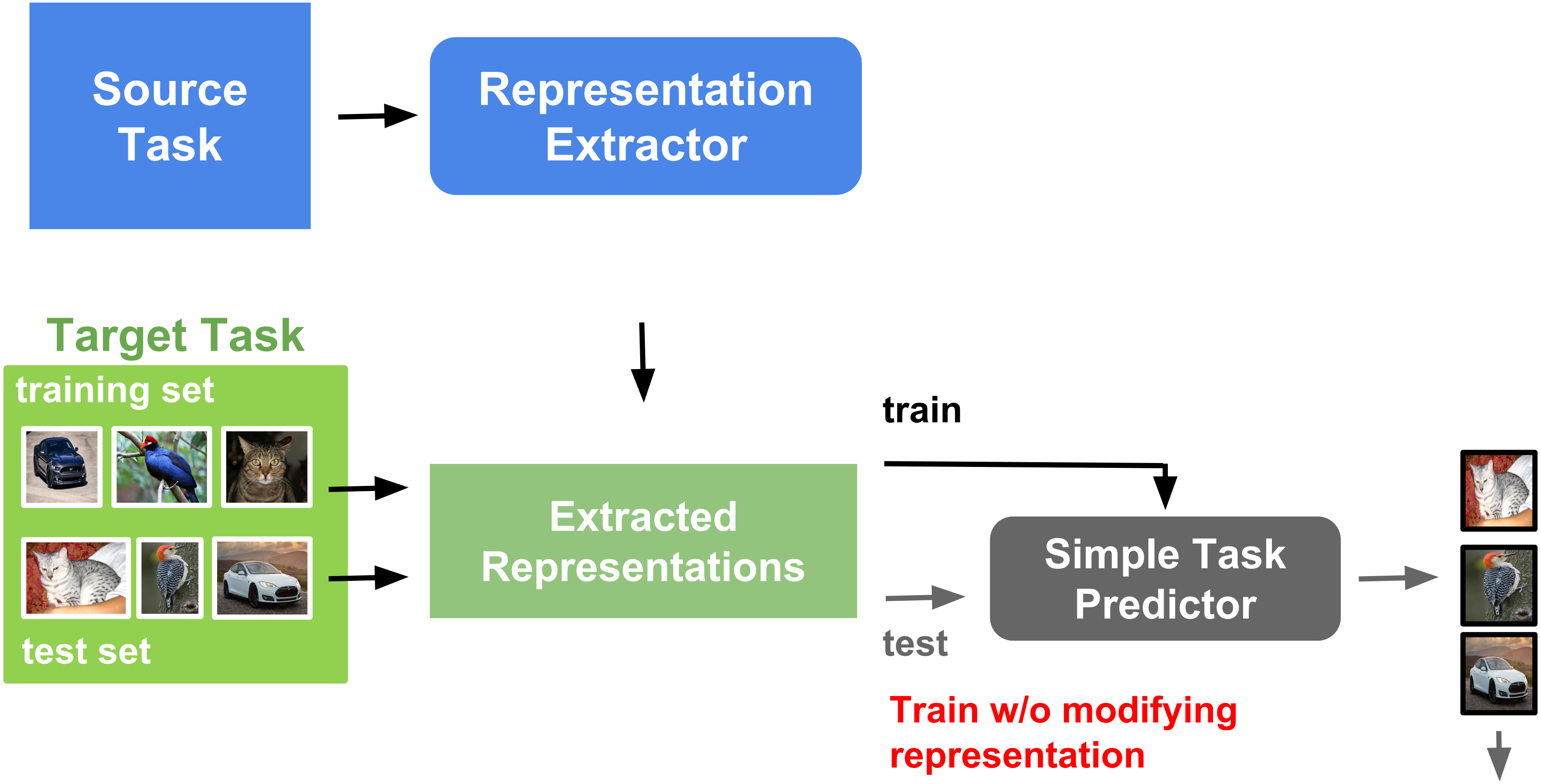
# S.O.T.A: Positioning

| Works | Univ. Aspect | Mod. | Eval. Scenario | SP Domain-Task | Approach |
|---|---|---|---|---|---|
| [Conneau et al., EACL'17] [Conneau et al., EMNLP'17] | Repres-entation | Textual | Transfer Learning | 1 domain - 1 task | Best task & algorithm |
| [Cer et al., ArXiv'17] | | | | 1 domain - No annotation | Tricks to auto. get annotations |
| [Subramanian & Bengio, ICLR'18] | | | | Multi-task | Best tasks & algorithm |
| [Kokkinos, CVPR'17] [Wang et al., WACV'18] | Task Solving | Visual | End2End | Multi-task | Get better learning algorithm |
| [Bilen & Vedaldi, ArXiv'17] [Rebuffi et al., NIPS'17] | Repres-entation | | | Multi-domain - 1 task | Domain-Specific Scaling parameters |
| [Rebuffi et al., CVPR'18] | | | Fine Tuning | Multi-domain - 1 task | |
| **This Thesis** | | Visual & Multimodal | Transfer Learning | 1 domain - 1 task | Automatically get **more** annotations |

# Outline

- State-Of-The-Art (S.O.T.A)
- **Contributions**
  - Evaluation of Universality
  - Universality in Image Representations Learned w/ Explicit Supervision
  - **Universality in Image Representations Learned w/ Implicit Supervision**
  - **Universality In Multimodal Representations Learned w/ Implicit Supervision**
- Conclusions
- Perspectives

# Evaluation of Universality




[Donahue et al., ICML'14], [Zeiler & Fergus, ECCV'14], [Agrawal et al., ECCV'14], [Oquab et al., CVPR'14], [Razavian et al., CVPRW'14], etc.

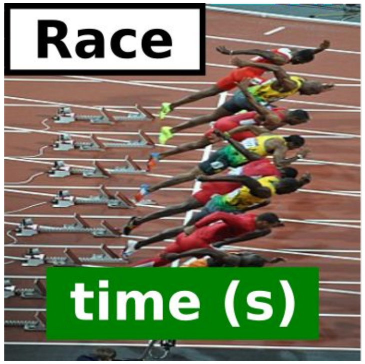
# Evaluation of Universality





| | | | |
|---|---|---|---|
| 24s | 1.7m | 3/10 | 8/10 |

# Evaluation of Universality





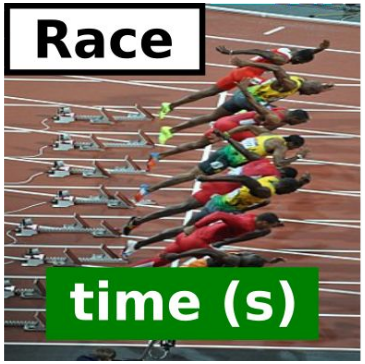


| | | | |
|---|---|---|---|
| 24s | 1.7m | 3/10 | 8/10 |
| 17s | 1.5m | 4/10 | 9/10 |

**What is desirable for the evaluation:**

- Coherent aggregation

# Evaluation of Universality





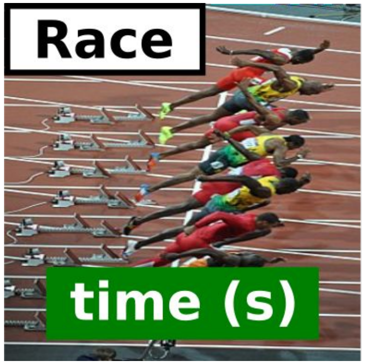


| | | | |
|---|---|---|---|
| 24s | 1.7m | 3/10 | 8/10 |
| 17s | 1.5m | 4/10 (+1) | 9/10 (+1) |

**What is desirable for the evaluation:**

- Coherent aggregation, Merit bonus

# Evaluation of Universality

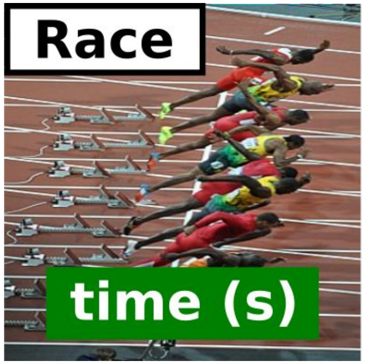



| | | | |
|---|---|---|---|
| 24s | 1.7m | 3/10 | 8/10 |
| 17s | 1.5m | 4/10 | 9/10 |

**What is desirable for the evaluation:**

- Coherent aggregation, Merit bonus, **Penalty for damage**

# Evaluation of Universality





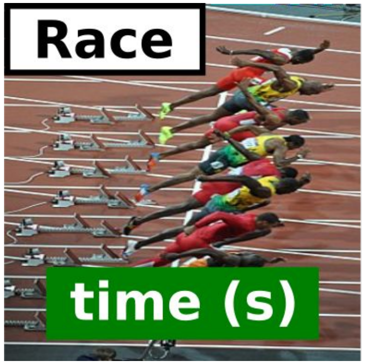


| Race | High jump | Penalty shootout | Dribble |
|---|---|---|---|
| 24s | 1.7m | 3/10 | 8/10 |
| 50s | 4.5m | 1/10 | 1/10 |

**What is desirable for the evaluation:**

- Coherent aggregation, Merit bonus, Penalty for damage, Independent to outliers

# Evaluation of Universality







| | | | |
|---|---|---|---|
| 24s | 1.7m | 3/10 | 8/10 |
| 17s | 1.5m | 4/10 | 9/10 |

**What is desirable for the evaluation:**

- Coherent aggregation, Merit bonus, Penalty for damage, Independent to outliers, consistence with time

# Evaluation of Universality

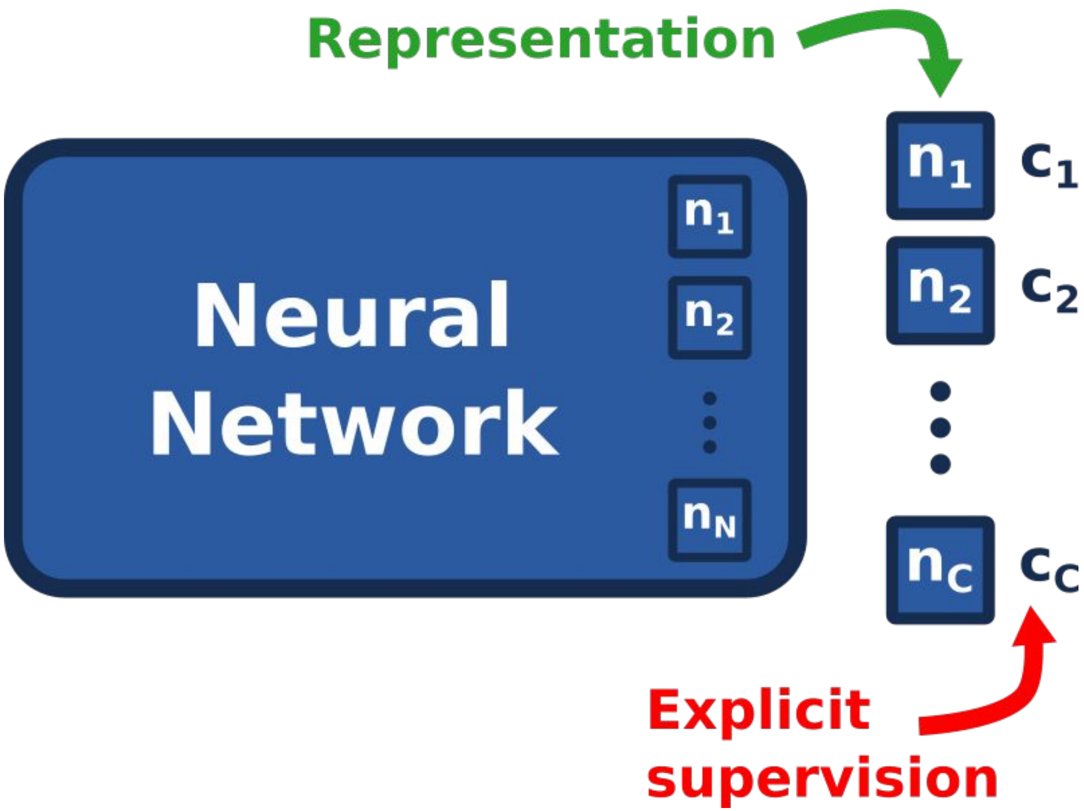
- **Average raw scores (Avg) [Baseline]**
- **Visual Decathlon Challenge (VDC) [Rebuffi *et al.*, NIPS'17]**
  - Average error classification gain over baseline
- **Borda Count (BC) [ours]**
  - Based on order statistics
- **Average/Median Relative Gain (aRG / mRG) [ours]**
  - Based on relative gain compared to reference

| Property | Avg | VDC | BC | aRG | mRG |
|---|---|---|---|---|---|
| Coherent aggregation | | ✓ | ✓ | ✓ | ✓ |
| Merit bonus | | ✓ | | ✓ | ✓ |
| Penalty for damage | ✓ | | ✓ | ✓ | ✓ |
| Indep. to ref. method | ✓ | | ✓ | | |
| Indep. to outliers | | | ✓ | | ✓ |
| Consistent with time | ✓ | ✓ | | ✓ | ✓ |

# Outline

- **State-Of-The-Art (S.O.T.A)**
- **Contributions**
  - Evaluation of Universality
  - Universality in Image Representations Learned w/ Explicit Supervision
  - Universality in Image Representations Learned w/ Implicit Supervision
  - Universality in Multimodal Representations Learned w/ Implicit Supervision
- **Conclusions**
- **Perspectives**

# Starting Point: Semantic Features



# Starting Point: Semantic Features

- **Implementation of** **[Ginsca *et al.*, MM'15]**
    - Independent classifiers (on top of internal layer of CNN)
    - Generic and specific classifiers

# Starting Point: Semantic Features

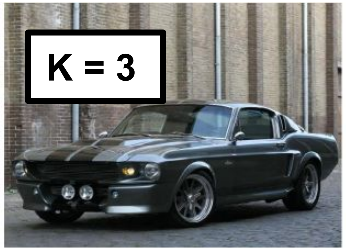
- **Advantages to increase of universality:**
    - Adding classes w/o retraining all CNN
    - No limit of capacity (cover large range of data)

# Starting Point: Semantic Features

- Increase universality by increasing capacity

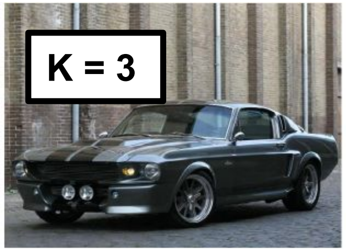
# Starting Point: Semantic Features

- **Increase universality by increasing capacity**
- **Problems**
  - When N large, statistical redundancy between neurons
    - Sparsity adapted to each sample image

# Starting Point: Semantic Features

- **Increase universality by increasing capacity**
- **Problems**
  - When N large, statistical redundancy between neurons
    - Sparsity adapted to each sample image

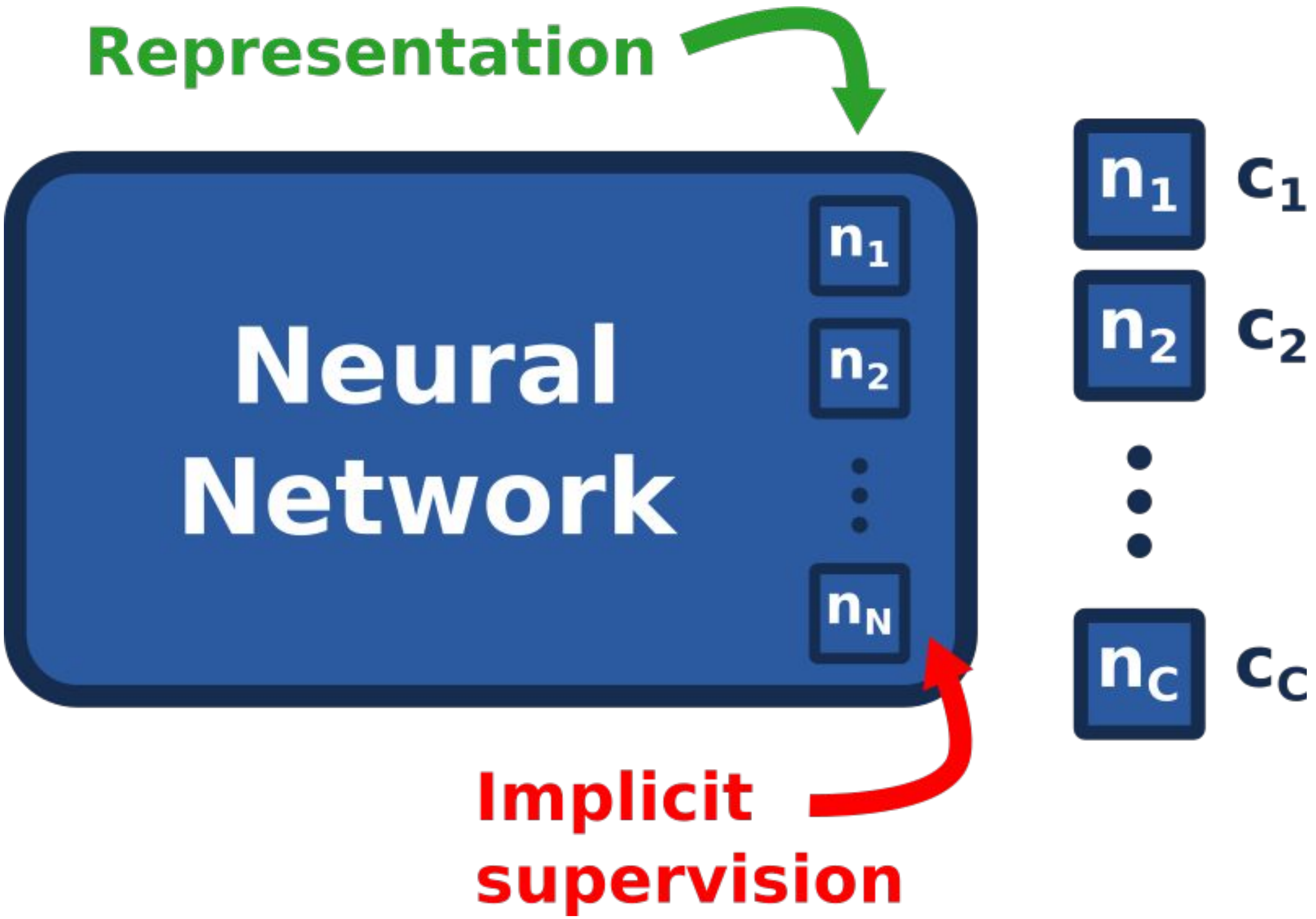




  - Do not benefit from generic classifiers (because low intra-class variance ⇒ low output scores)
    - Boosting outputs of generic classifiers with scores of their child nodes

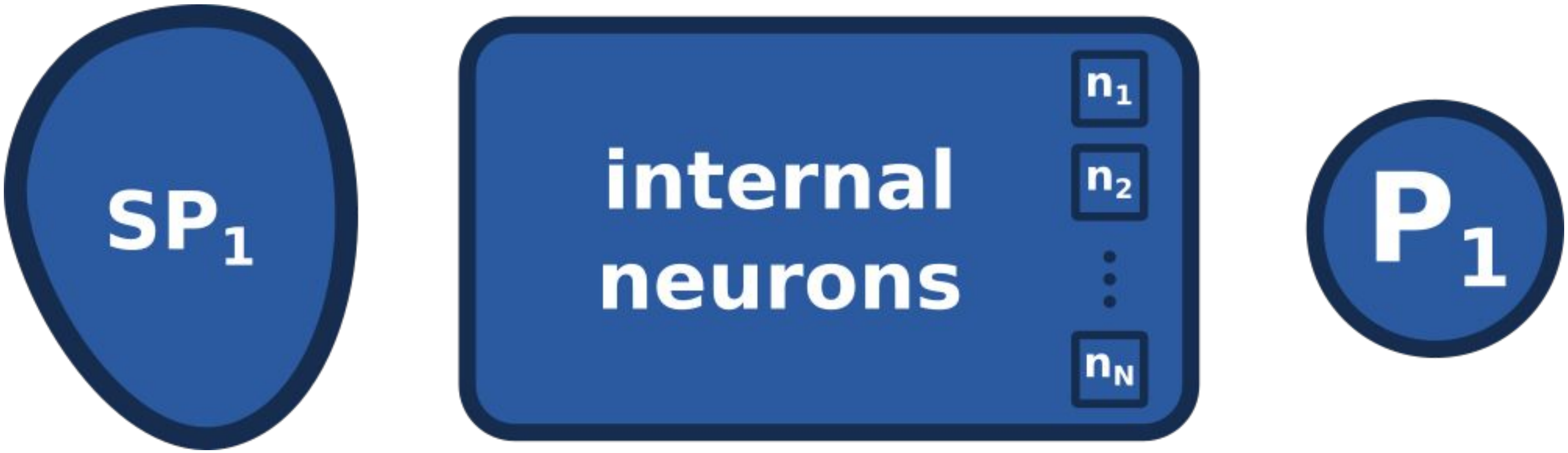
# Outline

- **State-Of-The-Art (S.O.T.A)**
- **Contributions**
  - Evaluation of Universality
  - Universality in Image Representations Learned w/ Explicit Supervision
  - Universality in Image Representations Learned w/ Implicit Supervision
  - Universality in Multimodal Representations Learned w/ Implicit Supervision
- **Conclusions**
- **Perspectives**

# Starting Point: Internal layers of CNN



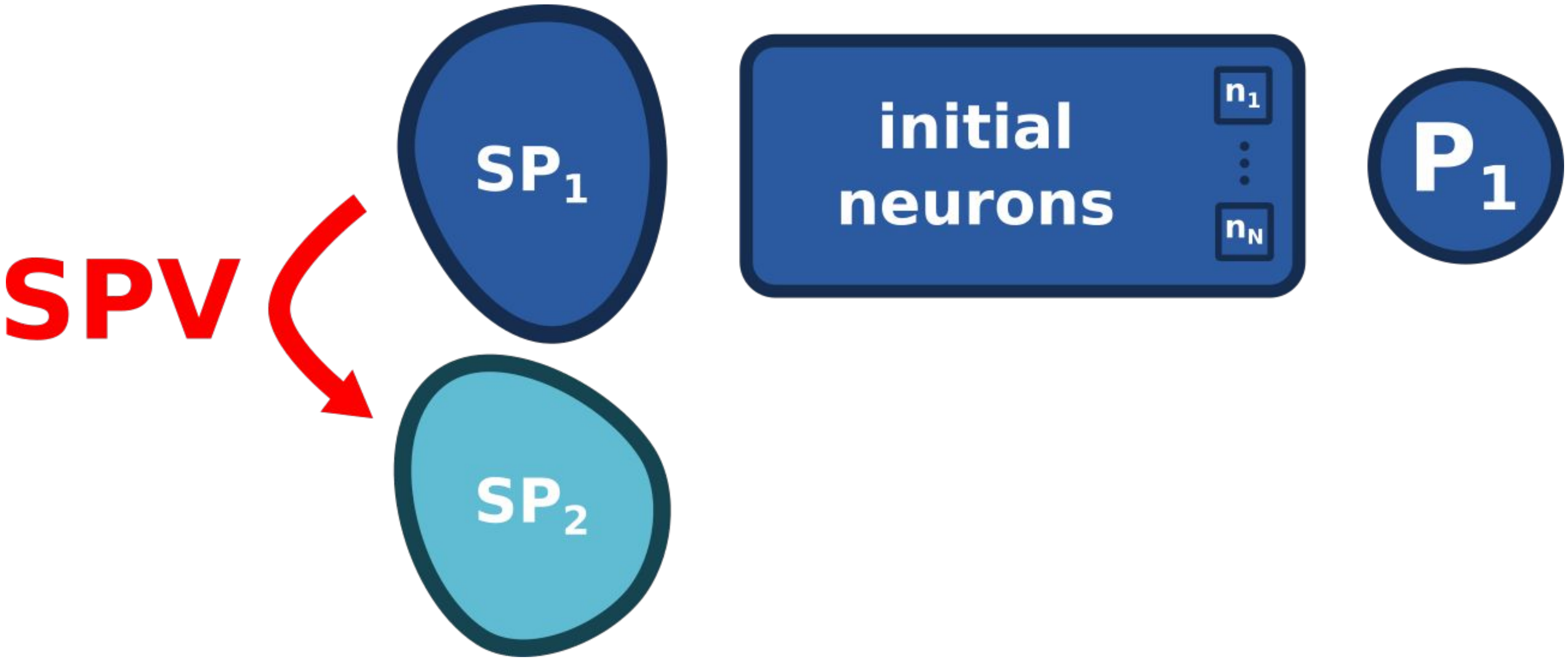
# Starting Point: Internal layers of CNN




- **Source-problem (SP);**
- **Network trained on SP**
  - According learning-strategy + architecture

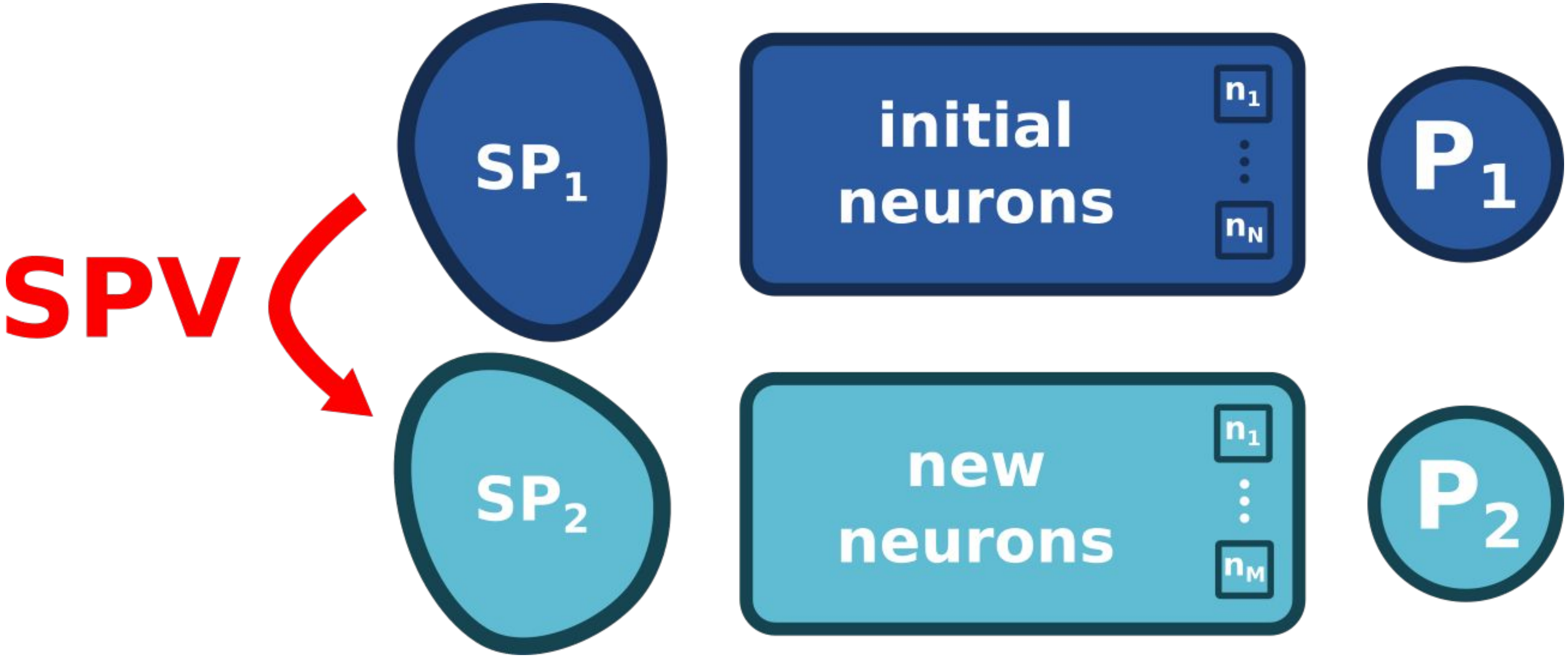
**⇒ Set of learned neurons**

# Proposed Approach: Step 1/4



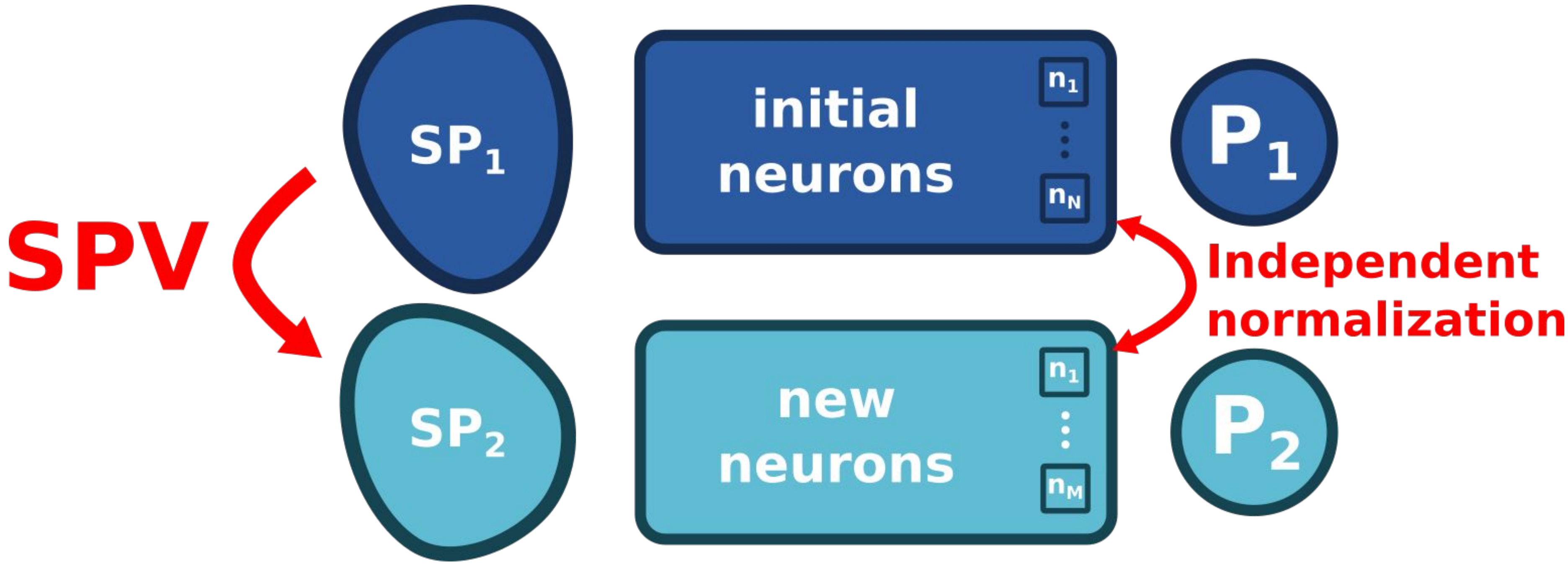

1. **Source Problem Variation (SPV)**
   - Automatic variation of raw data (pixels) and/or labels

# Proposed Approach: Step 2/4



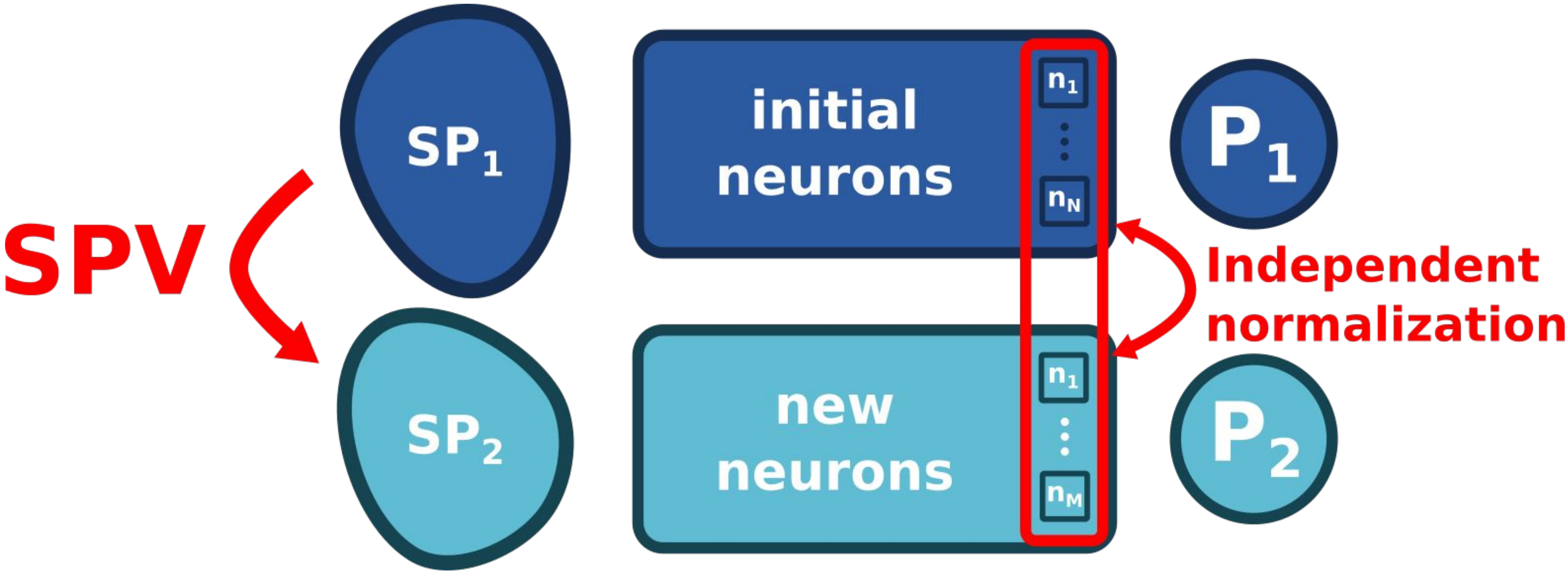

1. **Source Problem Variation (SPV)**
2. **Train new neurons**
    - 1 network on each new SP w.r.t same strategy & archi.

# Proposed Approach: Step 3/4



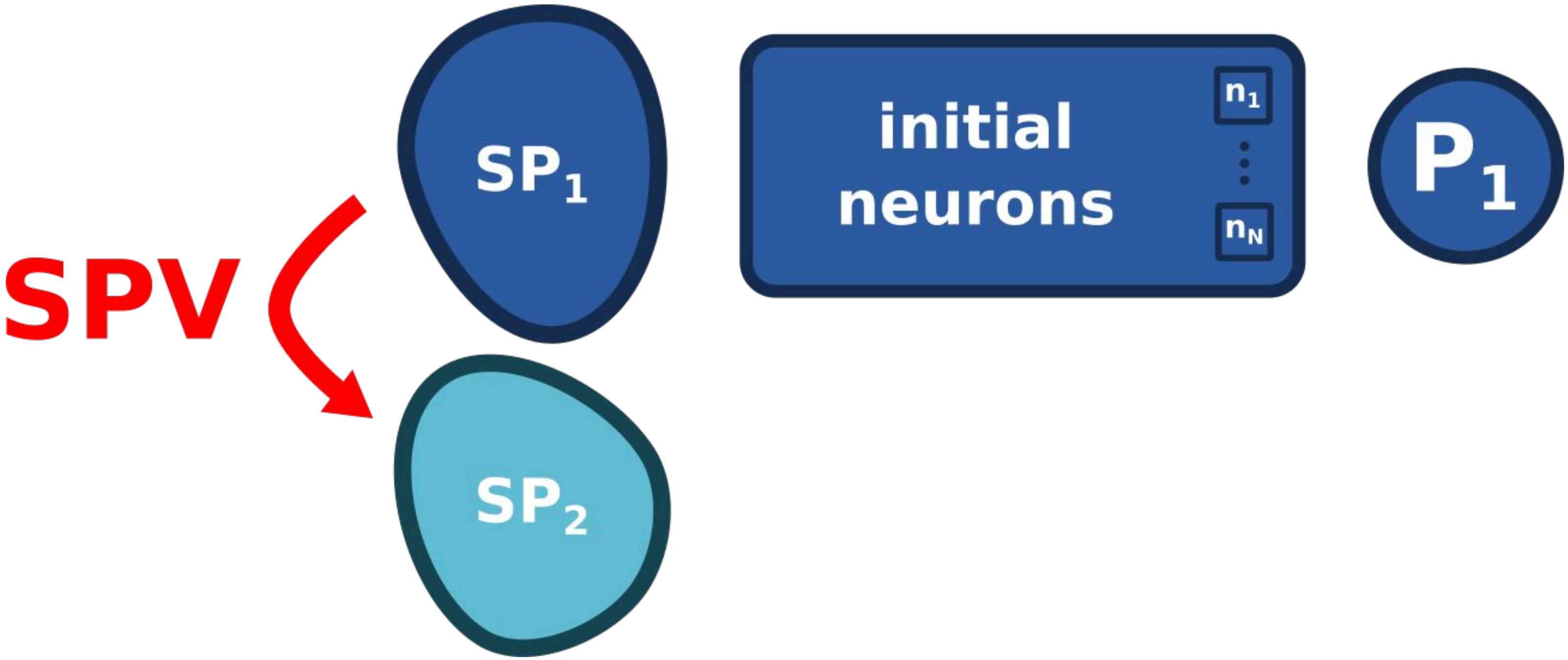

1. **Source Problem Variation (SPV)**
2. **Train new neurons**
3. **Representation**
   - Independent normalization

# Proposed Approach: Step 4/4




1. **Source Problem Variation (SPV)**
2. **Train new neurons**
3. **Representation**
   - Independent normalization
   - Combination (concatenation) + Dim. Reduction (FSFT)

# Proposed Approach: Step 1/4




1. **Source Problem Variation (SPV)**
    - Automatic variation of raw data (pixels) and/or labels

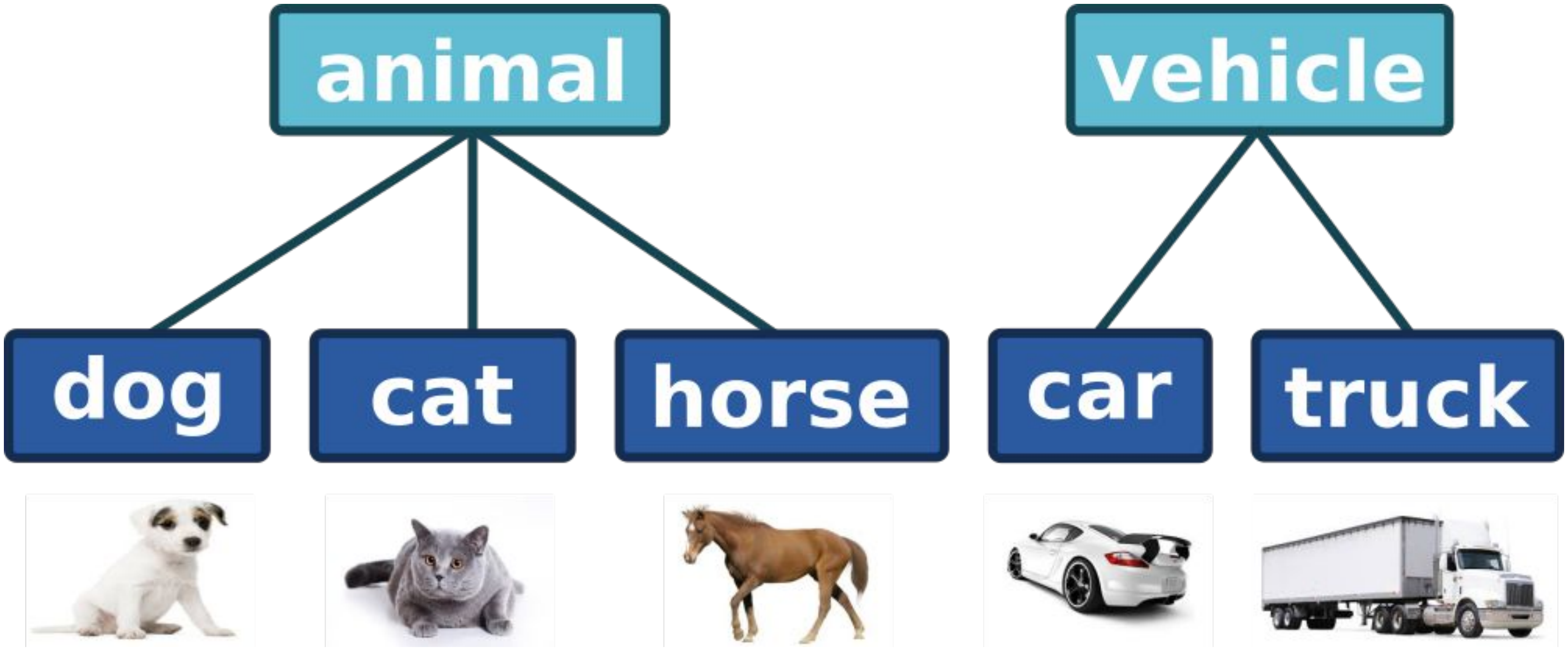
# How to get new SPs?

dog cat horse car truck

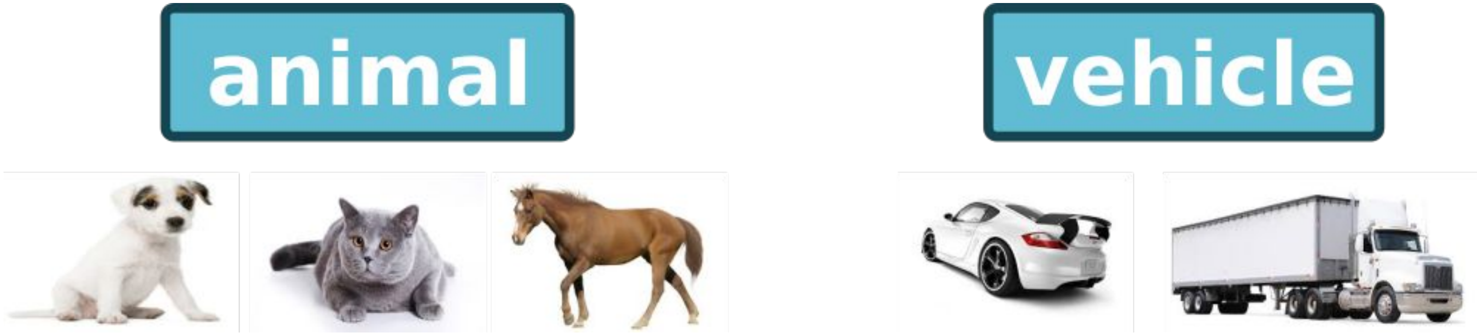
- **Starting point**
  - images associated to labels

# New SPs by Grouping-SPV




- **Getting generic labels**
  - by random grouping
  - using clustering
  - using an external ontology (*e.g.*, ImageNet, WordNet)

# New SPs by Grouping-SPV




- **Re-annotation of images**
  - according obtained generic labels
- **Generic classes contain:**
  - more images per class
  - more diverse set of images

# Getting Generic Labels according Categorical-Levels

- **Human Categorization according three levels** [Rosch, 1978] [Jolicoeur, 1984]
    - Concepts mostly known and used by Humans:
        - Superordinate (*vehicle*)
        - Basic-level (*car*)
        - Subordinate (*ford mustang*)

**⇒ Getting generic labels according categorical-levels**

# Experimental Settings

- Source-task: ILSVRC-half (subset of ImageNet)
- 10 Target-datasets (classification, many domains, few data)

| Datasets | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| **ILSVRC* [13]** | objects | 483 | 1,2K | ✗ | 569,000 | 48,299 | Acc. |
| **ILSVRC [13]** | objects | 1K | 1,2K | ✗ | 1.2M | 50,000 | Acc. |
| **VOC07 [5]** | objects | 20 | 250 | ✓ | 5,011 | 4,952 | mAP |
| **NWO [3]** | objects | 31 | 700 | ✓ | 21,709 | 14,546 | mAP |
| **CA101 [6]** | objects | 102 | 30 | ✗ | 3,060 | 3,022 | Acc. |
| **CA256 [7]** | objects | 257 | 60 | ✗ | 15,420 | 15,187 | Acc. |
| **MIT67 [10]** | scenes | 67 | 80 | ✗ | 5,360 | 1,340 | Acc. |
| **stACT [16]** | actions | 40 | 100 | ✗ | 4,000 | 5,532 | Acc. |
| **CUB [15]** | birds | 200 | 30 | ✗ | 5,994 | 5,794 | Acc. |
| **FLO [9]** | plants | 102 | 10 | ✗ | 1,020 | 6,149 | Acc. |
| **FOOD [2]** | food | 101 | 50 | ✗ | 5050 | 5050 | Acc. |
| **AIRC [8]** | airplanes | 100 | 66 | ✗ | 6,667 | 3,333 | Acc. |

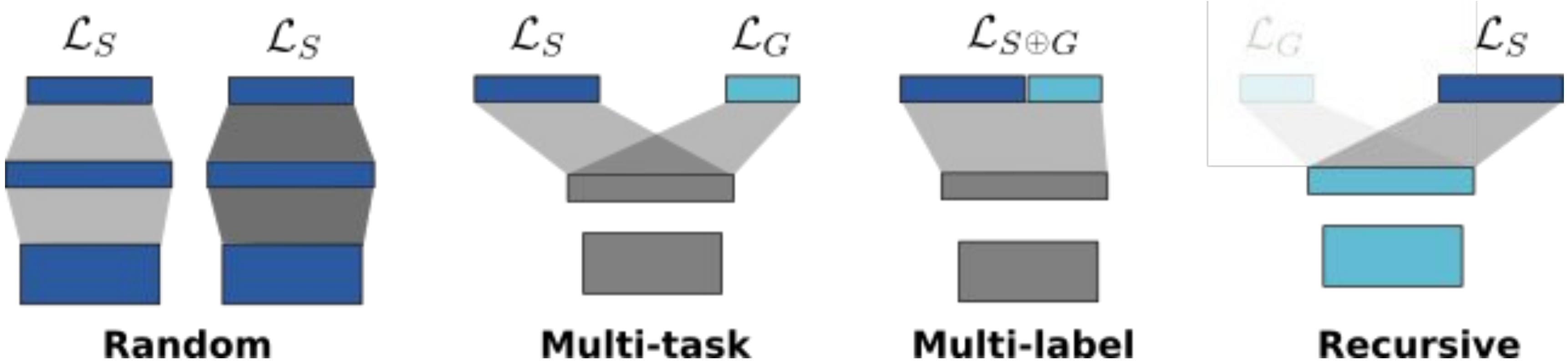
- For each class: one-vs-rest SVM classifier

# Comparison to S.O.T.A

| Method | VOC07 mAP | VOC12 mAP | CA101 Acc. | CA256 Acc. | NWO mAP | MIT67 Acc. | stACT Acc. | CUB Acc. | stCA Acc. | FLO Acc. | mRG |
|---|---|---|---|---|---|---|---|---|---|---|---|
| REFERENCE | 66.8 | 67.3 | 71.1 | 53.2 | 52.5 | 36.0 | 44.3 | 36.1 | 14.4 | 50.5 | n/a |
| Azizpour *et al.*, PAMI'15 | 66.6 | 67.5 | 74.7 | 54.7 | 53.2 | 37.4 | 45.1 | 36.0 | 13.7 | 51.9 | +1.5 |
| Mettes *et al.*, ICMR'16 | 67.7 | 68.1 | 73.0 | 54.3 | 50.5 | 37.1 | 44.9 | 36.8 | 14.6 | 50.3 | +1.4 |
| Chami *et al.*, ICMR'17 | 61.1 | 62.1 | 58.7 | 40.6 | 45.8 | 24.3 | 32.7 | 26.1 | 13.1 | 36.4 | -17.7 |
| Huh et al., NIPS-W'16 | 64.0 | 62.7 | 69.4 | 50.1 | 45.6 | 33.7 | 41.9 | 15.0 | 12.5 | 42.8 | -7.5 |
| Wu *et al.*, ACM'16 | 62.5 | 65.4 | 68.8 | 50.7 | 28.5 | 37.9 | 42.6 | 34.0 | 13.3 | 50.0 | -4.3 |
| Wang *et al.*v1, CVPR'17 | 68.4 | 68.3 | 73.1 | 54.7 | 49.3 | 38.4 | 46.5 | 37.5 | 14.7 | 54.8 | +3.5 |
| Wang *et al.*v2, CVPR'17 | 69.1 | 69.0 | 74.8 | 55.9 | 50.4 | 40.0 | 48.4 | **38.6** | 14.8 | 56.1 | +6.0 |
| MulDiP-Net (Ours) | 69.5 | 69.8 | 76.0 | 56.8 | **54.7** | 41.3 | 48.5 | 35.6 | 15.7 | 54.8 | +7.7 |
| MulDiP+FSFT (Ours) | **69.8** | **70.0** | **77.5** | **58.3** | 47.9 | **43.7** | **50.2** | 37.4 | **16.1** | **59.7** | **+9.8** |

# Some insights…

- **Comparison to "baseline" universalizing methods**
  - Reference: *specific*
  - Ours: *specific* + *generic*
  - Random: 2 *specific* with ≠ Initialization
  - Multi-task
  - Multi-label
  - Recursive: Fine-tune *generic* on *specific*

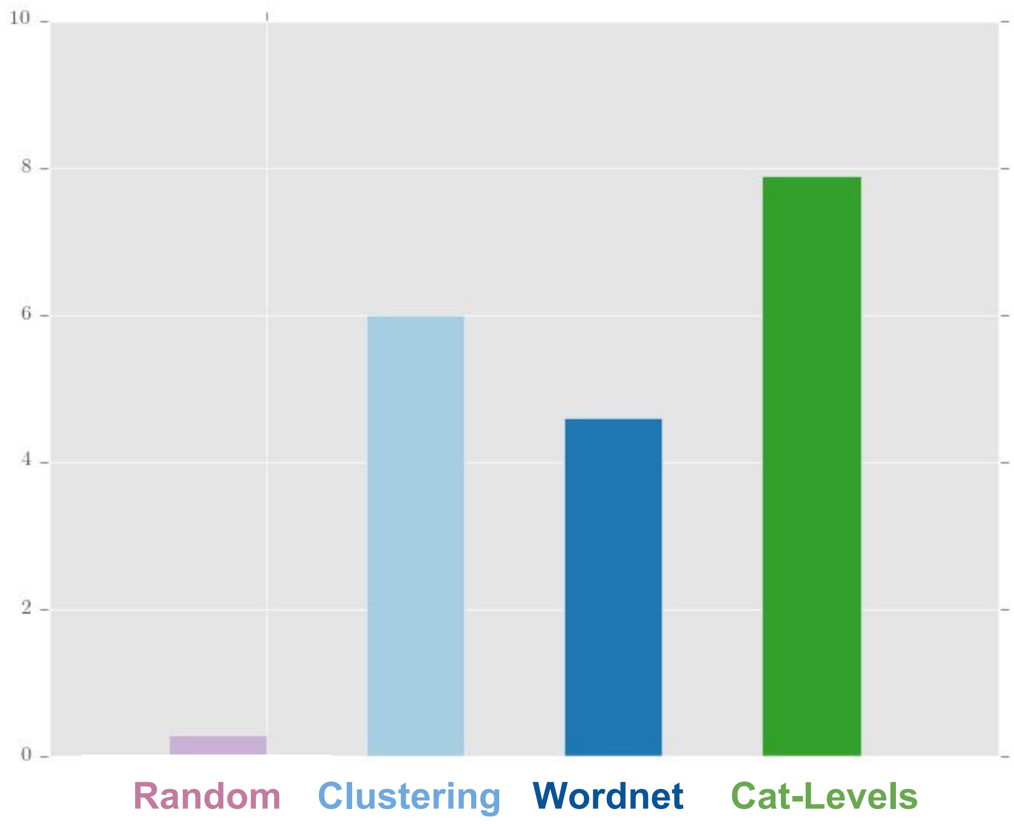


# Some insights…

| Method | VOC07 mAP | CA101 Acc. | CA256 mAP | NWO Acc. | MIT67 Acc. | stACT Acc. | CUB Acc. | FLO Acc. | mRG |
|---|---|---|---|---|---|---|---|---|---|
| REFERENCE | 66.8 | 71.1 | 53.2 | 52.5 | 36.0 | 44.3 | 36.1 | 50.5 | n/a |
| Random | 67.8 | 72.2 | 54.5 | 52.0 | 37.2 | 45.0 | 34.7 | 51.8 | +2.3 |
| Multi-Task | 61.5 | 61.8 | 45.4 | 49.4 | 30.7 | 36.4 | 25.6 | 38.7 | -16.2 |
| Multi-Label | 44.7 | 46.8 | 26.4 | 25.1 | 27.2 | 28.0 | 15.2 | 38.1 | -45.0 |
| Recursive | 65.3 | 68.6 | 50.8 | 52.4 | 33.4 | **50.8** | 29.4 | 45.5 | -4.8 |
| Ours | 69.5 | 76.0 | 56.8 | **54.7** | 41.3 | 48.5 | 35.6 | 54.8 | +7.9 |
| Ours+FSFT | **69.8** | **77.5** | **58.3** | 47.9 | **43.7** | 50.2 | **37.4** | **59.7** | **+10.7** |

# Some insights...

- Comparison of “grouping methods”




  - Cognitive knowledge (Categorical-levels) useful !

# Deeper networks, More data

- **Comparison with deeper architectures**
  - AlexNet [Krizhevsky *et al.*, NIPS'12]
  - VGG-16 [Simonyan & Zisserman, ICLR'14]
  - DarkNet [Redmond *et al.*, CVPR'16]: fully convolutionnal; base of YOLO

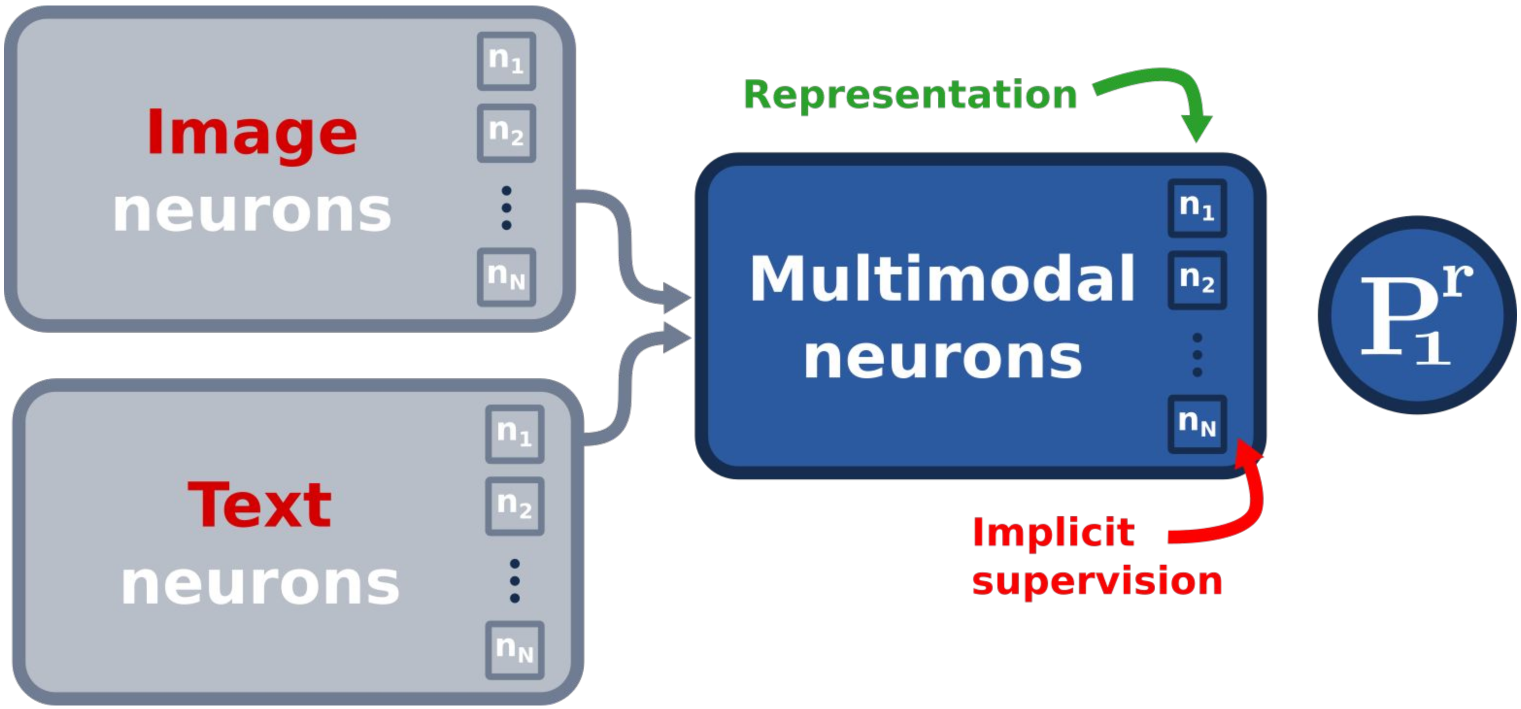
| Method | Network | VOC07 mAP | CA101 Acc. | CA256 mAP | NWO Acc. | MIT67 Acc. | stACT Acc. | CUB Acc. | FLOW Acc. | BC |
|---|---|---|---|---|---|---|---|---|---|---|
| **Reference** | AlexNet | 71.7 | 79.7 | 62.4 | 58.3 | 46.9 | 51.2 | **36.3** | 58.4 | 19 |
| **SPV$_G^{cat}$** | AlexNet | 71.5 | 77.4 | 60.4 | 57.8 | 42.8 | 49.3 | 19.5 | 52.4 | 10 |
| **MulDiP-Net** | AlexNet | **74.4** | **82.5** | **65.2** | **60.8** | **48.4** | **54.2** | 36.1 | **62.5** | 25 |
| **Reference** | VGG-16 | 86.1 | 88.8 | 78.0 | 71.8 | 66.7 | 73.5 | 69.8 | 78.9 | 52 |
| **SPV$_G^{cat}$** | VGG-16 | 85.7 | 87.6 | 76.9 | 70.3 | 65.8 | 72.2 | 67.0 | 75.0 | 43 |
| **MulDiP-Net** | VGG-16 | **87.5** | **92.0** | **80.9** | **72.6** | **68.9** | **75.0** | **71.5** | **81.9** | **68** |
| **Reference** | DarkNet-20 | 82.7 | 91.0 | 78.4 | 70.5 | 64.8 | 72.2 | 59.5 | 80.0 | 44 |
| **SPV$_G^{cat}$** | DarkNet-20 | 83.2 | 91.5 | 78.1 | 73.2 | 64.4 | 72.6 | 52.5 | 78.9 | 46 |
| **MulDiP-Net** | DarkNet-20 | **84.1** | **92.7** | **80.1** | **73.9** | **66.4** | **74.5** | **61.2** | **82.1** | 62 |

- **Deeper network are not always more universal**
- **Net-G > Net-S with Darknet**

# Outline

- State-Of-The-Art (S.O.T.A)
- Contributions
  - Evaluation of Universality
  - Universality in Image Representations Learned w/ Explicit Supervision
  - Universality in Image Representations Learned w/ Implicit Supervision
  - Universality in Multimodal Representations Learned w/ Implicit Supervision
- Conclusions
- Perspectives

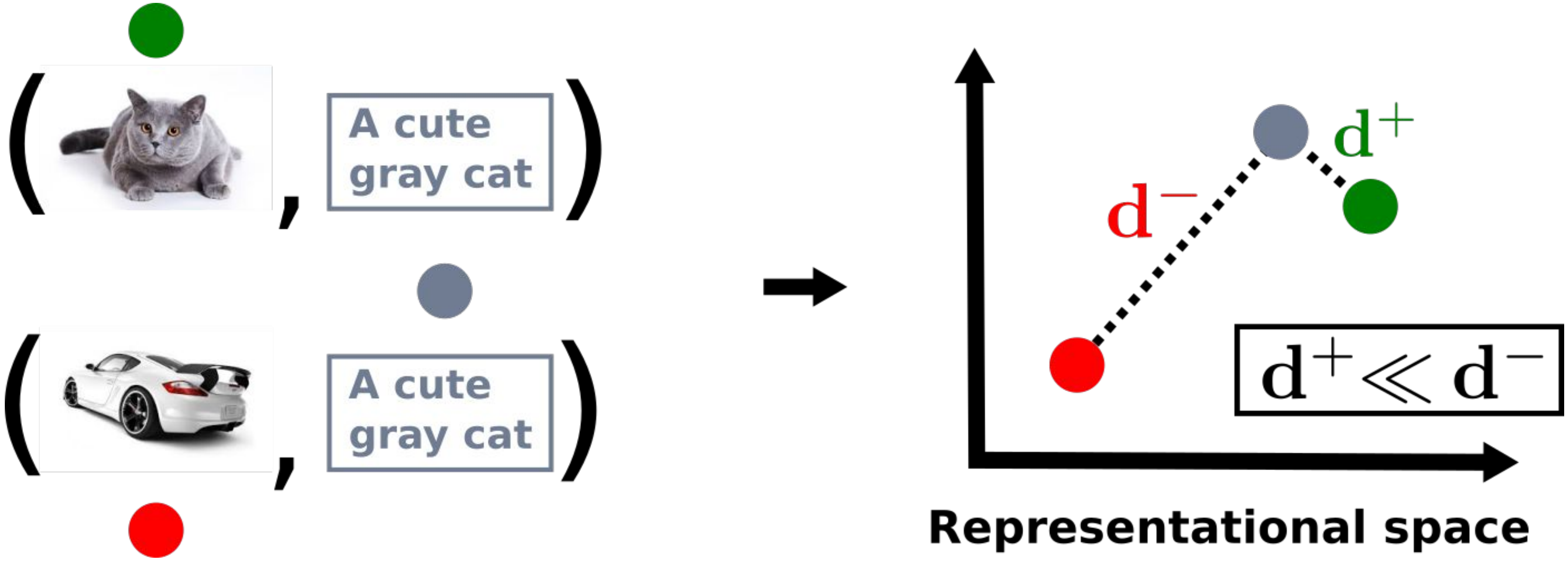
# Starting point: Multimodal Representations



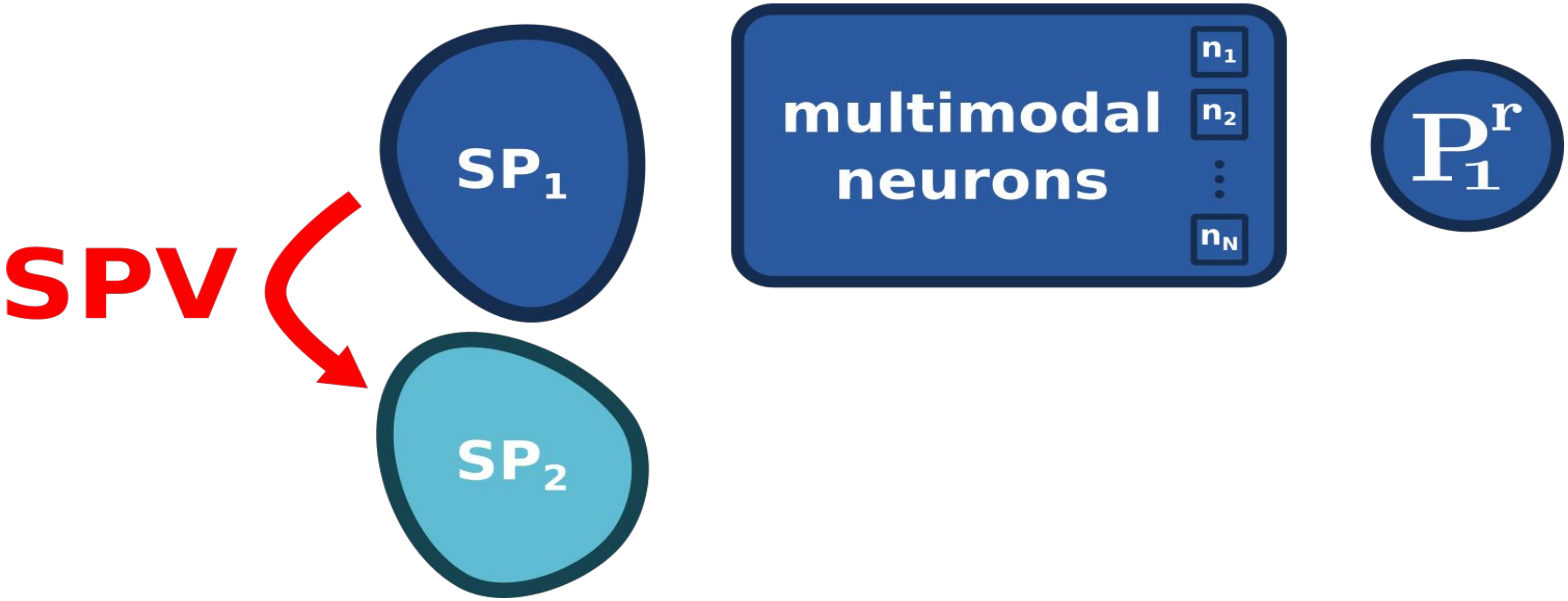

- [Wang & Lazebnik, CVPR'16] [Wang & Lazebnik, TPAMI'18] Two-branches networks
- [Salvador et al., CVPR'17]: Adding semantic loss for regularization
- [Zheng *et al.*, Arxiv 2017]: Dual-Path Convolutional Image-Text Embedding
- [Engilberge et al., CVPR 2018] : Semantic-visual embedding with localization

# Starting point: Multimodal Representations

- **Ranking Objective**
  - Make **positive** couple of data as **close** as possible
  - Make **negative** couple of data as **far** as possible

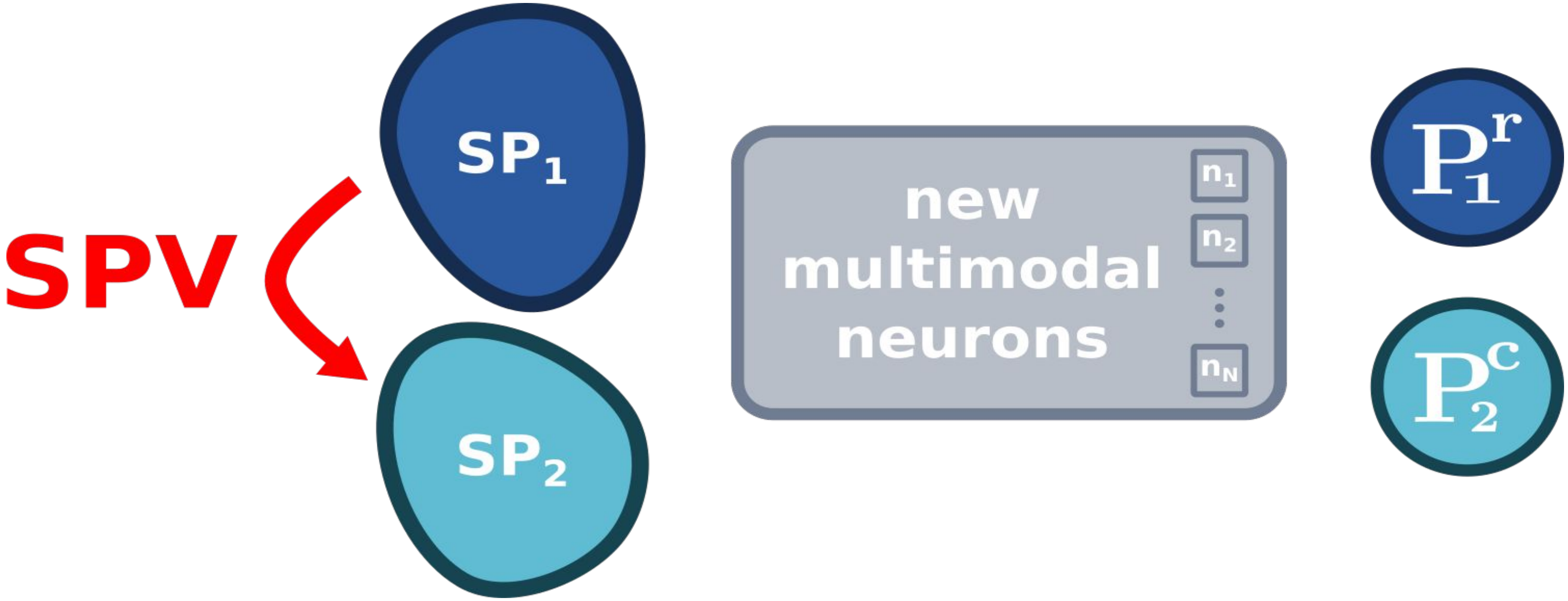


# Proposed Approach: step 1/2




1. **Source Problem Variation (SPV)**

# Proposed Approach: step 1/2




1. **Source Problem Variation (SPV)**
2. **Retrain new neurons**
   - According multi-task objective (joint training)

# New SPs by Grouping-SPV

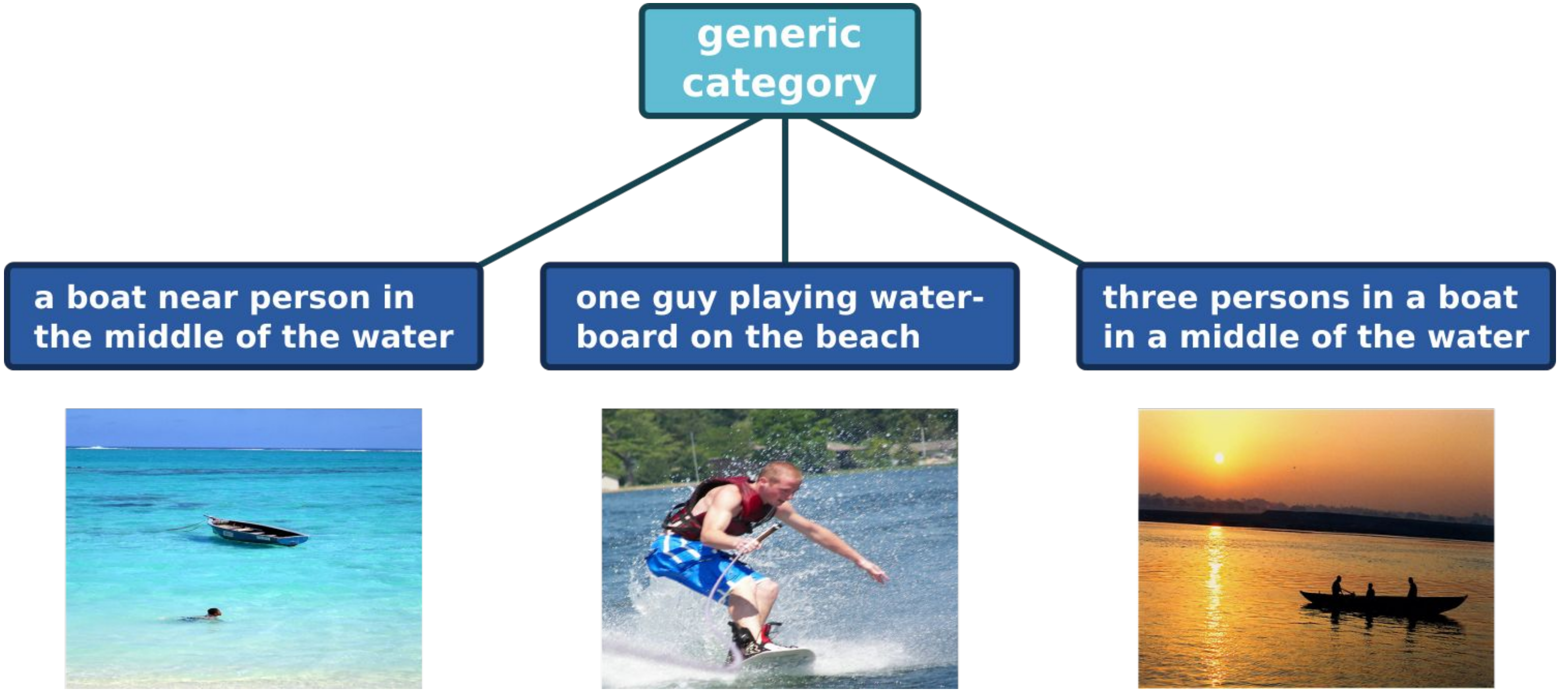
a boat near person in the middle of the water

one guy playing water-board on the beach

three persons in a boat in a middle of the water

- **Starting point**
  - Complex images and textual descriptions

# New SPs by Grouping-SPV




- **Getting generic labels**
  - Not easy to rely on an existing ontology (complex data)
  - using clustering of visual & textual representations

# Experimental Settings

- **Task: Cross-modal retrieval**
  - Image annotation
  - Text illustration
- **Metric: Recall (R@K)**
- **Flickr-30K**
- **End-to-end scheme**
- **Simple predictors (L2 normalized representations + k-NN)**

# Comparison to S.O.T.A

| Method | Image-Annotation | | | Text-Illustration | | | Avg |
|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| Wang & Lazebnik, CVPR'16 | 26.6 | 53.0 | 65.9 | 22.7 | 50.1 | 63.3 | 46.9 |
| Karpathy *et al.*, NIPS'14 | 12.6 | 32.9 | 44.0 | 10.3 | 31.4 | 44.5 | 29.3 |
| Kiros *et al.*, ArXiv'14 | 14.8 | 39.2 | 50.9 | 11.8 | 34.0 | 46.3 | 32.8 |
| Karpathy *et al.*, CVPR'15 | 22.2 | 48.2 | 61.4 | 15.2 | 37.7 | 50.5 | 38.7 |
| Dong *et al.*, ToM'18 (txt2im) | 16.8 | 40.3 | 53.0 | 0.1 | 5.6 | 10.0 | 21.0 |
| Chami *et al.*, ICMR'17 (txt2im) | 18.3 | 41.3 | 53.5 | 6.1 | 9.8 | 12.2 | 23.5 |
| Chami *et al.*, ICMR'17 (im2txt) | 9.6 | 13.5 | 19.2 | 20.0 | 49.8 | 64.2 | 29.4 |
| NAMRank (Ours) | **30.4** | **55.3** | **69.1** | **23.8** | **52.0** | **65.5** | **49.4** |

# Outline

- State-Of-The-Art (S.O.T.A)
- Contributions
  - Evaluation of Universality
  - Universality in Features Learned with Explicit Supervision
  - Universality in Features Learned with Implicit Supervision
  - Universality via Multimodal Representations
- **Conclusions**
- Perspectives

# Conclusions

- **Unified framework to tackle universality of representations**

- **A new protocol to evaluate the increase of universality**
  - Identify desirable properties
  - 3 new metrics

- **A new approach for learning more universal representations**
  - Without additive data
  - Very low annotation cost
  - Relying on cognitive knowledge about Human categorization
  - Efficient universal & dimensionality reduction method (FSFT)

- **Extend the universality question to the multimodal aspect**

# Publications

- **Journals (1 international)**
  - Tamaazousti, Le Borgne, Popescu, Gadeski, Ginsca and Hudelot, **Vision-Language Integration using Constrained Local Semantic Features**, **CVIU** 2017
  - Tamaazousti, Le Borgne, Popescu, Gadeski, Ginsca and Hudelot, **Déscripteur Sémantique Local Contraint Basé sur un RNC Diversifié,** Traitement du Signal, 2017
- **Conferences (5 international)**
  - Tamaazousti, Le Borgne and Hudelot, **MuCaLe-Net: Multi Categorical-Level Networks to Generate More Discriminating Features**, **CVPR** 2017 (poster)
  - Chami*, Tamaazousti*, Le Borgne, **AMECON: Abstract Meta Concept Features for Text-Illustration,** **ICMR** 2017 (oral)
  - Daher, Besançon, Ferret, Le Borgne, Daquo, and Tamaazousti, **Supervised Learning of Entity Disambiguation Models by Negative Sample Selection**, **CICling** 2017
  - Daher, Besançon, Ferret, Le Borgne, Daquo, and Tamaazousti, **Désambiguïsation d'entités nommées par apprentissage de modèles d'entités à large échelle**, CORIA 2017
  - Tamaazousti, Le Borgne and Hudelot, **Diverse Concept-Level Features for Multi-Object Classification**, **ICMR** 2016, (oral)
  - Tamaazousti, Le Borgne and Popescu, **Constrained Local Enhancement of Semantic Features by Content-Based Sparsity**, **ICMR** 2016 (oral)
  - Tamaazousti, Le Borgne and Hudelot, **Descripteurs à divers niveaux de concepts pour la classification d'images multi-objets,** RFIA 2016
  - Tamaazousti, Le Borgne and Popescu, **Agrégation de descripteurs sémantiques locaux contraints par parcimonie basée sur le contenu**, RFIA 2016
- **Patents**
  - Tamaazousti, Le Borgne and Hudelot. **Procédé d'obtention d'un système de labellisation d'images, programme d'ordinateur et dispositif correspondant, système de labellisation d'images**, filled INPI N° 1662013, dec 2016.

# Outline

- State-Of-The-Art (S.O.T.A)
- Contributions
  - Evaluation of Universality
  - Universality in Features Learned with Explicit Supervision
  - Universality in Features Learned with Implicit Supervision
  - Universality via Multimodal Representations
- Conclusions
- Perspectives

# Perspectives

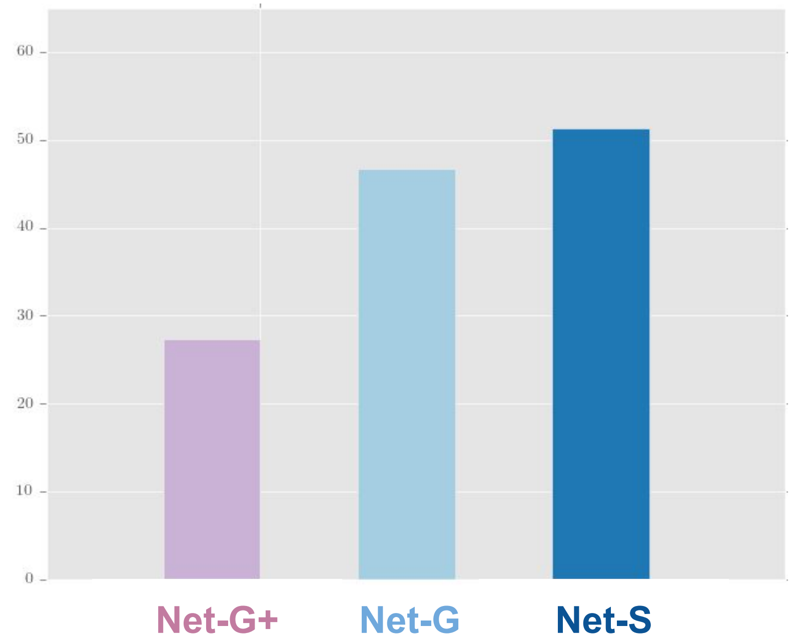
- **Independent training of networks**
  - Costly in terms of amount of parameters ⇒ Efficient parametrization
    - Decrease #parameters ?
      Pruning [Mallya & Lazebnik, CVPR'18], Knowledge distillation [Hinton, Arxiv'15], Mapping from master-net to others [in manuscript]
    - Learn efficiently ?
      Learning by growing capacity [Wang *et al*., CVPR'17]

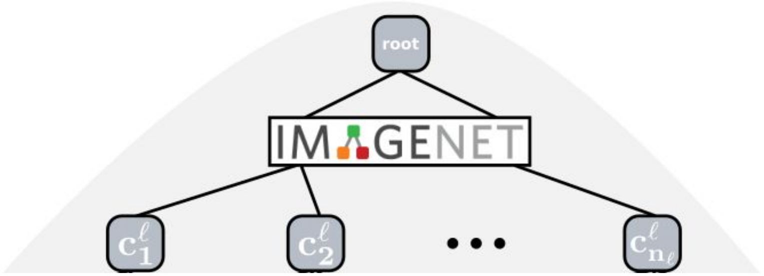
- **1 task in target-tasks** (classification or cross-modal retrieval)
  ⇒ Evaluate on other tasks (detection, segmentation, VQA, etc.)

- **Multimodal representations on top of fixed image & textual representations**
  ⇒ Learn them all together

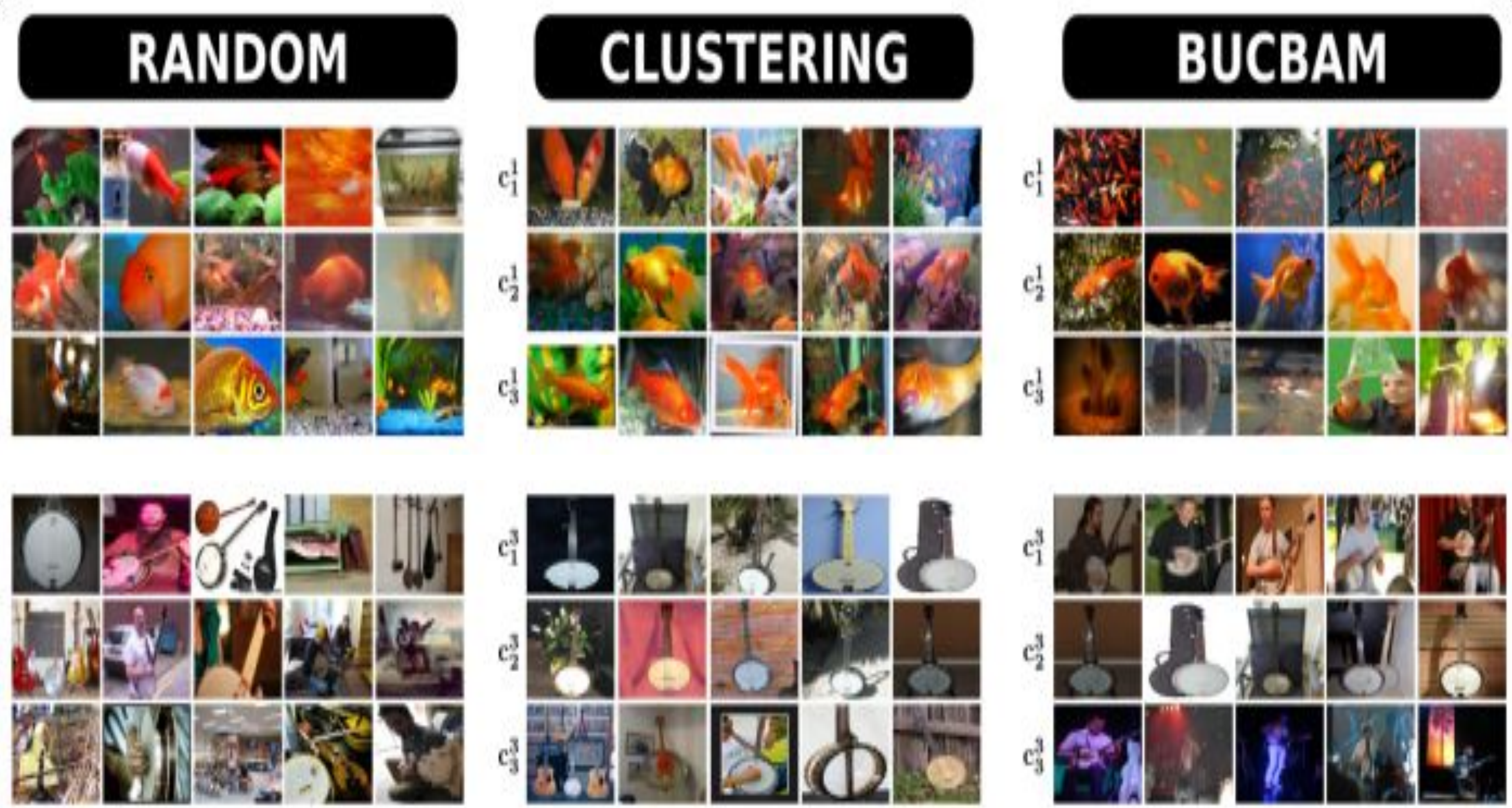
# Perspectives

- **In 2nd technical contribution**
  - Net-G+ < Net-G < Net-S
- **Learn a Net-S+, on more specific labels (poses, context, etc.)**
- **Problem: no annotations available**



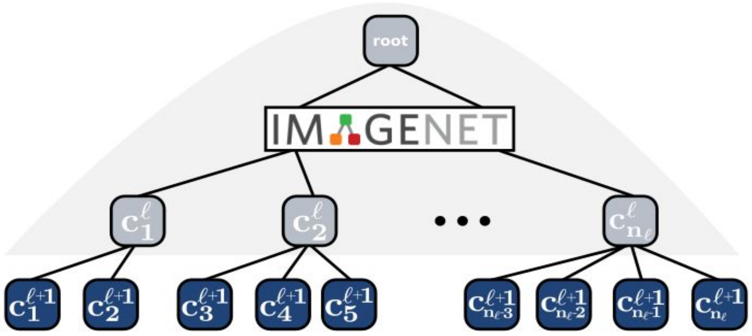
# Perspectives



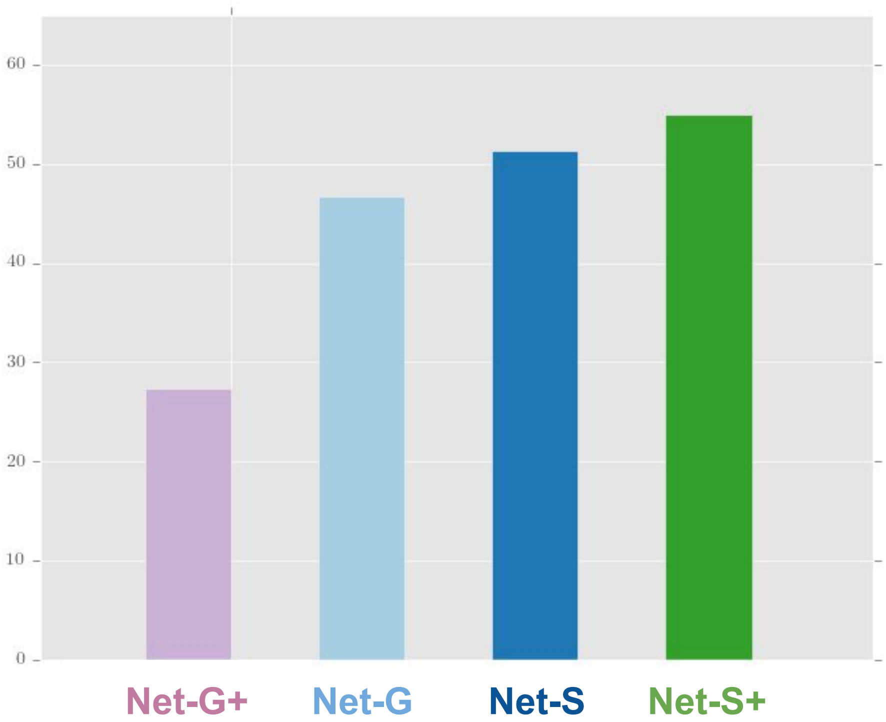
# Perspectives

- **Proposal: BUCBAM**
  - Splitting each category
  - A new level to ImageNet hierarchy

**Under review at BMVC and patent filled**




- **Results:**
  - +5 (avg) compared to Net-S
  - With ensembling: +8 (avg)




- **Above Grouping or Splitting, it seems that the most interesting aspect is SPV !**

- **How to variate the SPs ?**

# Thank you

Commissariat à l'énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr