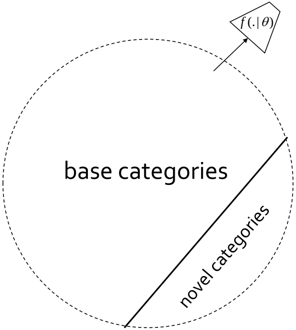
Associative Alignment for Few-Shot Image Classification

Arman Afrasiyabi, Jean-François Lalonde, Christian Gagné
Université Laval

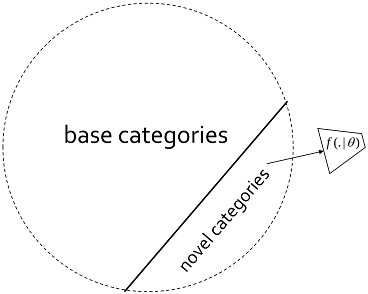
few-shot learning

few-shot learning is supervised transferring pre-trained knowledge to novel categories with only few examples



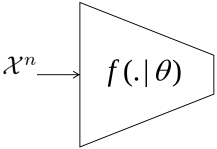
few-shot learning

few-shot learning is supervised transferring pre-trained knowledge to novel categories with only few examples

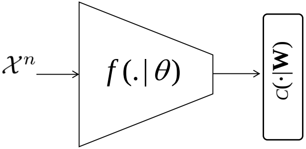


few-shot frameworks

I) meta-learning

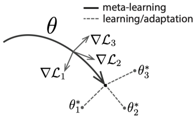


II) standard transfer learning



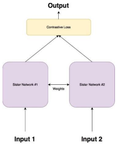
meta-learning

I) initialization based

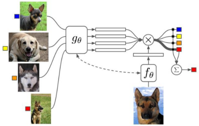


MAML
Finn et al., 2017

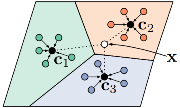
II) distance based



Siamese Net.
Koch et al. (2015)

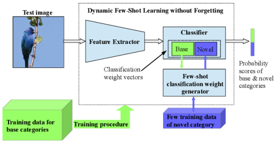


Matching Net.
Vinyals et al. (2017)

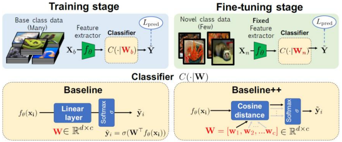


Proto. Net.
snell et al. (2017)

standard transfer learning



forgetting
Gidaris et al., 2018



baseline++
Chen et al., 2019

pros and cons

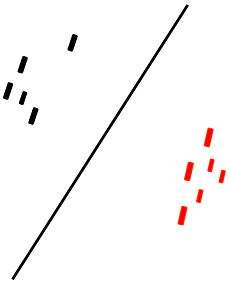
I) meta-learning

- overfitting: robust
- learning capacity: limited

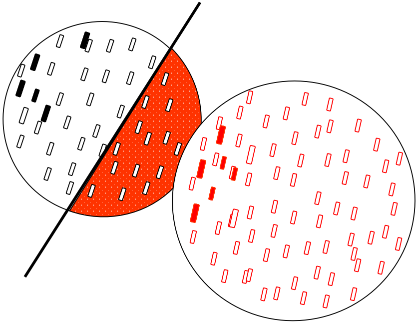
II) standard transfer learning–fine-turning

- overfitting: not robust
- learning capacity: high

sampling problem

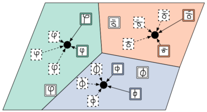


sampling problem

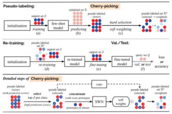


data-augmentation

I) semi-supervised based

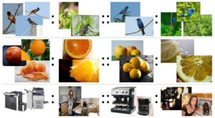


meta semi-supervised
Ren et al. 2018

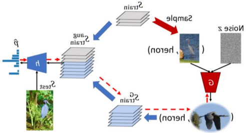


self-train semi-supervised
Li et al. 2019

II) generative or mapping based



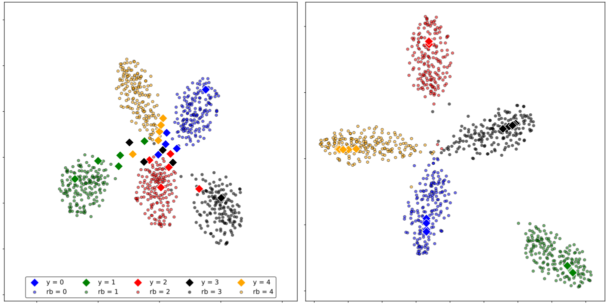
FH
Hariharan et al. 2017



Imaginary data
Wang et al. 2018

Associative Alignment for Few-Shot Image Classification

associative alignment



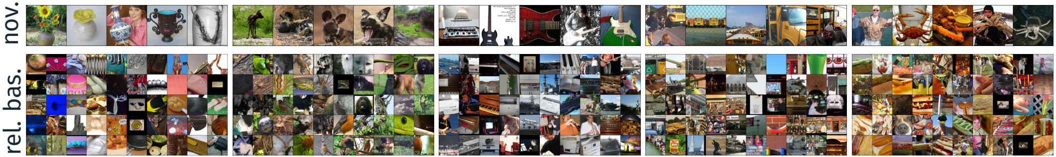
(a) before alignment (b) after alignment

related base matrix

$$m_{i,j} = \frac{1}{|\mathcal{X}_i^b|} \sum_{(\mathbf{x}_l^b,\cdot)\in\mathcal{X}_i^b} \mathbb{I}\left[j =_{k=1}^{K^n} \left(c_k(f(\mathbf{x}_l^b|\theta) \mid \mathbf{W})\right)\right]$$

$c_k(f(\mathbf{x}|\theta)|\mathbf{W})$ is the classifier output $c(\cdot|\mathbf{W})$ for class k
then, B base classes with the highest score for a given novel class are kept as the related base

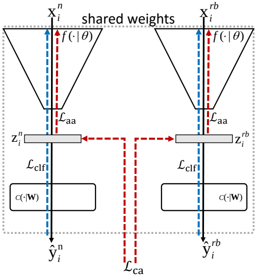
related base algorithm–results



5-way 5-shot scenario
first row: columns represents a different novel class
second row: corresponding related classes for $B = 10$

alignment strategies

centroid alignment



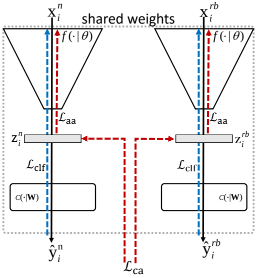
centroid alignment

- the centroid of $\mathcal{X}_i^{rb}$ is computed by

$$\mu_i = \frac{1}{|\mathcal{X}_i^{rb}|} \sum_{(\mathbf{x}_j,\cdot)\in\mathcal{X}_i^{rb}} f(\mathbf{x}_j|\theta)$$

- which allows to define the centroid alignment loss as

$$\mathcal{L}_{\mathrm{ca}}(\mathcal{X}^n) = -\frac{1}{N^n N^{rb}} \sum_{i=1}^{K^n} \sum_{(\mathbf{x}_j,\cdot)\in\mathcal{X}_i^n} \log \frac{\exp[-\|f(\mathbf{x}_j|\theta) - \mu_i\|_2^2]}{\sum_{k=1}^{K^n} \exp[-\|f(\mathbf{x}_j|\theta) - \mu_k\|_2^2]}$$



centroid alignment

I. network's encoder θ is updated by $\mathcal{L}_{\text{ca}}$

II. all parameters $(\theta; \mathbf{W})$ updates by a $\mathcal{L}_{\text{clf}}$ (classification loss)

Algorithm 1:
Centroid alignment.

Input: pre-trained model $c(f(\cdot|\theta)|\mathbf{W})$, novel class $\mathcal{X}^n$, related base set $\mathcal{X}^{rb}$.
Output: fine-tuned $c(f(\cdot|\theta)|\mathbf{W})$.
while *not done* **do**

$\widetilde{\mathcal{X}}^n \leftarrow$ sample a batch from $\mathcal{X}^n$
$\widetilde{\mathcal{X}}^{rb} \leftarrow$ sample a batch from $\mathcal{X}^{rb}$

evaluate $\mathcal{L}_{\text{ca}}(\widetilde{\mathcal{X}}^n, \widetilde{\mathcal{X}}^{rb})$, (eq. 3)
$\theta \leftarrow \theta - \eta_{\text{ca}} \nabla_\theta \mathcal{L}_{\text{ca}}(\widetilde{\mathcal{X}}^n, \widetilde{\mathcal{X}}^{rb})$

evaluate $\mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^{rb})$, (eq. 7)
$\mathbf{W} \leftarrow \mathbf{W} - \eta_{\text{clf}} \nabla_{\mathbf{W}} \mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^{rb})$

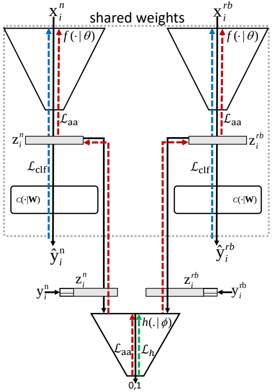
evaluate $\mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^n)$, (eq. 7)
$\mathbf{W} \leftarrow \mathbf{W} - \eta_{\text{clf}} \nabla_{\mathbf{W}} \mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^n)$

$\theta \leftarrow \theta - \eta_{\text{clf}} \nabla_\theta \mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^n)$

end

---

adversarial alignment



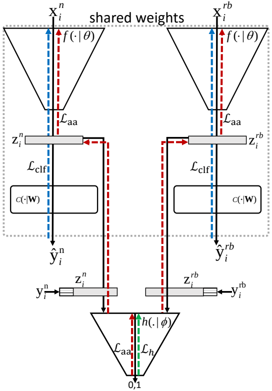
adversarial alignment

- the critic $h(\cdot|\phi)$ is trained with loss

$$\mathcal{L}_h(\mathcal{X}^n, \mathcal{X}^{rb}) = \frac{1}{N^{rb}} \sum_{(\mathbf{x}_i^{rb}, y_i^{rb}) \in \mathcal{X}^{rb}} h\left([f(\mathbf{x}_i^{rb}|\theta)\ y_i^{rb}] \,|\, \phi \right) - \frac{1}{N^n} \sum_{(\mathbf{x}_i^n, y_i^n) \in \mathcal{X}^n} h\left([f(\mathbf{x}_i^n|\theta)\ y_i^n] \,|\, \phi \right) ,$$

- while the encoder parameters θ updates by

$$\mathcal{L}_{\mathrm{aa}}(\mathcal{X}^n) = \frac{1}{K^n} \sum_{(\mathbf{x}_i^n, y_i^n) \in \mathcal{X}^n} h\left([f(\mathbf{x}_i^n|\theta)\ y_i^n] |\phi \right)$$



adversarial alignment

I. the parameter update of critic $h(\cdot|\phi)$ using $\mathcal{L}_{\mathrm{h}}$

II. update $f(\cdot|\theta)$ using $\mathcal{L}_{\mathrm{aa}}$

III. all parameters $(\theta; \mathbf{W})$ updates by a $\mathcal{L}_{\mathrm{clf}}$ (classification loss)

Algorithm 2:
Adversarial alignment

Input: pre-trained model $c(f(\cdot|\theta)|\mathbf{W})$, novel class $\mathcal{X}^n$, related base set $\mathcal{X}^{rb}$.
Output: fine-tuned $c(f(\cdot|\theta)|\mathbf{W})$.
while *not done* **do**
  $\widetilde{\mathcal{X}}^n \leftarrow$ sample a batch from $\mathcal{X}^n$
  $\widetilde{\mathcal{X}}^{rb} \leftarrow$ sample a batch from $\mathcal{X}^{rb}$

  for $i = 0, \ldots, n_{\mathrm{critic}}$ **do**
    evaluate $\mathcal{L}_h(\widetilde{\mathcal{X}}^n, \widetilde{\mathcal{X}}^{rb})$, (eq. 5)
    ▷ update critic:
    $\phi \leftarrow \phi + \eta_h \nabla_\phi \mathcal{L}_h(\widetilde{\mathcal{X}}^n, \widetilde{\mathcal{X}}^{rb})$
    $\phi \leftarrow \mathrm{clip}(\phi, -0.01, 0.01)$
  end

  evaluate $\mathcal{L}_{\mathrm{ca}}(\widetilde{\mathcal{X}}^n, \widetilde{\mathcal{X}}^{rb})$, (eq. 3)
  $\theta \leftarrow \theta - \eta_{\mathrm{ca}} \nabla_\theta \mathcal{L}_{\mathrm{ca}}(\widetilde{\mathcal{X}}^n, \widetilde{\mathcal{X}}^{rb})$

  evaluate $\mathcal{L}_{\mathrm{clf}}(\widetilde{\mathcal{X}}^{rb})$, (eq. 7)
  $\mathbf{W} \leftarrow \mathbf{W} - \eta_{\mathrm{clf}} \nabla_{\mathbf{W}} \mathcal{L}_{\mathrm{clf}}(\widetilde{\mathcal{X}}^{rb})$

  evaluate $\mathcal{L}_{\mathrm{clf}}(\widetilde{\mathcal{X}}^n)$, (eq. 7)
  $\mathbf{W} \leftarrow \mathbf{W} - \eta_{\mathrm{clf}} \nabla_{\mathbf{W}} \mathcal{L}_{\mathrm{clf}}(\widetilde{\mathcal{X}}^n)$

  $\theta \leftarrow \theta - \eta_{\mathrm{clf}} \nabla_\theta \mathcal{L}_{\mathrm{clf}}(\widetilde{\mathcal{X}}^n)$
end

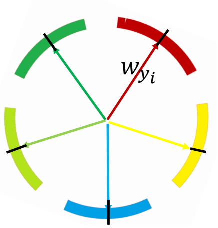
summary of the idea

summary of the idea

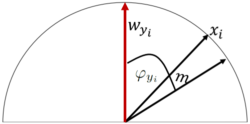
classification loss

classification loss

- let $\mathbf{z} = f(\mathbf{x}|\theta)$ be the representation of $\mathbf{x}$
- we transform the logit as $\mathbf{w}_j^\top \mathbf{z} = \|\mathbf{w}_j\| \|\mathbf{z}\| \cos \varphi_j$
  - φ_j is the angle between $\mathbf{z}$, and $\mathbf{w}_j$
  - $\mathbf{w}_j$, the j-th column in the weight matrix $\mathbf{W}$



$$\mathcal{L}_{\text{clf}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s \cos(\varphi_{y_i} + m))}{\exp(s \cos(\varphi_{y_i} + m)) + \sum_{\forall j \neq y_i} \exp(s \cos \varphi_j)}$$



results

experiments-Detests

- tieredImageNet
  351, 97, 160 categories for base, validation, test
- *mini*ImageNet
  64, 16, 20 categories for base, validation, test
- CUB–generic object recognition
  100, 50, 50 categories for base, validation, test
- FC100–derived from CIFAR-100
  60, 20, 20 categories for base, validation, test

experiments-backbones

- Conv4– four layer convolutional network (input size: 84×84)
- ResNet-18
- Wide Residual Network ("WRN-28-10")

results: Conv4

	Method	*mini*-ImageNet 1-shot	*mini*-ImageNet 5-shot	CUB 1-shot	CUB 5-shot
meta learning	Meta-LSTM [35]	43.44 ± 0.77	55.31 ± 0.71	–	–
	MatchingNet‡ [46]	43.56 ± 0.84	55.31 ± 0.73	60.52 ± 0.88	75.29 ± 0.75
	ProtoNet‡ [41]	49.42 ± 0.78	68.20 ± 0.66	51.31 ± 0.91	70.77 ± 0.69
	MAML‡ [10]	48.07 ± 1.75	63.15 ± 0.91	55.92 ± 0.95	72.09 ± 0.76
	RelationNet‡ [43]	50.44 ± 0.82	65.32 ± 0.70	62.45 ± 0.98	76.11 ± 0.69
tr. learning	softmax†	46.40 ± 0.72	64.37 ± 0.59	47.12 ± 0.74	64.16 ± 0.71
	softmax†◇	46.99 ± 0.73	65.33 ± 0.60	45.68 ± 0.86	66.94 ± 0.84
	cosmax†	50.92 ± 0.76	67.29 ± 0.59	60.53 ± 0.83	79.34 ± 0.61
	cosmax†◇	52.04 ± 0.82	68.47 ± 0.60	60.66 ± 1.04	79.79 ± 0.75
	our baseline (sec. 5)	51.90 ± 0.79	69.07 ± 0.62	60.85 ± 1.07	79.74 ± 0.64
align.	adversarial	52.13 ± 0.99	70.78 ± 0.60	**63.30** ± 0.94	**81.35** ± 0.67
	centroid	**53.14** ± 1.06	**71.45** ± 0.72	62.71 ± 0.88	80.48 ± 0.81

† our implementation ◇ with early stopping ‡ implementation from [3] for CUB

results: ResNet-18, WRN-28-10

	Method	*mini*-ImageNet 1-shot	*mini*-ImageNet 5-shot	tieredImageNet 1-shot	tieredImageNet 5-shot
ResNet-18	TADAM [32]	58.50 ± 0.30	76.70 ± 0.30	–	–
	ProtoNet‡ [41]	54.16 ± 0.82	73.68 ± 0.65	61.23 ± 0.77	80.00 ± 0.55
	SNAIL [31]	55.71 ± 0.99	68.88 ± 0.92	–	–
	IDeMe-Net [4]	59.14 ± 0.86	74.63 ± 0.74	–	–
	Activation to Param. [34]	59.60 ± 0.41	73.74 ± 0.19	–	–
	MTL [42]	61.20 ± 1.80	75.50 ± 0.80	–	–
	TapNet [53]	61.65 ± 0.15	76.36 ± 0.10	63.08 ± 0.15	80.26 ± 0.12
	VariationalFSL [56]	61.23 ± 0.26	77.69 ± 0.17	–	–
	MetaOptNet [24]	**62.64** ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
	our baseline (sec. 5)	58.07 ± 0.82	76.62 ± 0.58	65.08 ± 0.19	83.67 ± 0.51
	adversarial alignment	58.84 ± 0.77	77.92 ± 0.82	66.44 ± 0.61	85.12 ± 0.53
	centroid alignment	59.88 ± 0.67	**80.35** ± 0.73	**69.29** ± 0.56	**85.97** ± 0.49
WRN-28-10	LEO [38]	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.09	81.44 ± 0.12
	wDAE [15]	61.07 ± 0.15	76.75 ± 0.11	68.18 ± 0.16	83.09 ± 0.12
	CC+rot [13]	62.93 ± 0.45	79.87 ± 0.33	70.53 ± 0.51	84.98 ± 0.36
	Robust-dist++ [38]	63.28 ± 0.62	81.17 ± 0.43	–	–
	Transductive-ft [7]	65.73 ± 0.68	78.40 ± 0.52	73.34 ± 0.71	85.50 ± 0.50
	our baseline (sec. 5)	63.28 ±0.71	78.31 ±0.57	68.47 ±0.86	84.11 ±0.65
	adversarial alignment	64.79 ±0.93	82.02 ±0.88	73.87 ±0.76	84.95 ±0.59
	centroid alignment	**65.92** ± 0.60	**82.85** ± 0.55	**74.40** ± 0.68	**86.61** ±0.59

ProtoNet‡ results are from [3] for *mini*-ImageNet and from [24] for tieredImageNet.

results: FC100, CUB

	FC100		CUB	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Robust-20 [8]	–	–	58.67 ± 0.65	75.62 ± 0.48
GNN-LFT [44]	–	–	51.51 ± 0.80	73.11 ± 0.68
RelationNet‡ [43]	–	–	67.59 ± 1.02	82.75 ± 0.58
ProtoNet‡ [41]	40.5 ± 0.6	55.3 ± 0.6	71.88 ± 0.91	87.42 ± 0.48
TADAM [32]	40.1 ± 0.4	56.1 ± 0.4	–	–
MetaOptNet† [24]	41.1 ± 0.6	55.5 ± 0.6	–	–
MTL [42]	45.1 ± 1.8	57.6 ± 0.9	–	–
Transductive-ft [7]	43.2 ± 0.6	57.6 ± 0.6	–	–
our baseline (sec. 5)	40.84 ± 0.71	57.02 ± 0.63	71.71 ± 0.86	85.74 ± 0.49
adversarial	43.44 ± 0.71	58.69 ± 0.56	70.80 ± 1.12	88.04 ± 0.54
centroid	**45.83** ± 0.48	**59.74** ± 0.56	**74.22** ± 1.09	**88.65** ± 0.55

‡ implementation from [3] for CUB, and from [24] for FC100

results: ablation on B

B	1-shot	5-shot
1	55.76 ± 1.20	**79.34** ± 0.69
5	58.20 ± 1.14	78.65 ± 0.94
10	**58.84** ± 0.77	77.92 ± 0.82

(a) Adversarial alignment

B	1-shot	5-shot
1	58.04 ± 0.98	77.54 ± 0.73
5	58.97 ± 1.06	79.14 ± 0.91
10	**59.88** ± 0.67	**80.35** ± 0.73

(b) Centroid alignment

results: ablation on ways

	Method	**5-way**	**10-way**	**20-way**
meta-l.	MatchingNet‡ [46]	68.88 ± 0.69	52.27 ± 0.46	36.78 ± 0.25
	ProtoNet‡ [41]	73.68 ± 0.65	59.22 ± 0.44	44.96 ± 0.26
	RelationNet‡ [43]	69.83 ± 0.68	53.88 ± 0.48	39.17 ± 0.25
transfer-l.	softmax [3]	74.27 ± 0.63	55.00 ± 0.46	42.03 ± 0.25
	cosmax [3]	75.68 ± 0.63	63.40 ± 0.44	50.85 ± 0.25
	our baseline (sec. 5.1)	76.62 ± 0.58	62.95 ± 0.83	51.92 ± 1.02
	B	10	5	3
align.	adversarial	77.92 ± 0.82	64.87 ± 0.96	52.46 ± 0.99
	centroid	**80.35** ± 0.73	**68.17** ± 0.79	**54.67** ± 1.02

‡ implementation from [3]

acknowledgement

We thank to Ihsen Hedhli, Marc-André Gardner, and Annette Schwerdtfeger for proofreading of the manuscript.

This project was supported through funding from Mitacs, Prompt-Québec, NSERC-Canada, and E Machine Learning.

my special thanks go

Thank you for your attention!