Fine-tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally

Zongwei Zhou[a], Jae Y. Shin[a], Lei Zhang[a], Suryakanth R. Gurudu[b], Michael B. Gotway[c], and Jianming Liang[a]

[a] Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ 85259
[b] Division of Gastroenterology and Hepatology, Mayo Clinic, Scottsdale, AZ 85259
[c] Department of Radiology, Mayo Clinic, Scottsdale, AZ 85259

ARIZONA STATE UNIVERSITY

MAYO CLINIC

CVPR July 21-26 HONOLULU 2017

Motivation: Annotating biomedical images is very challenging.

It is not only tedious and time consuming, but also demanding of costly, specialty-oriented knowledge and skills, which are not easily accessible.

Key Ideas:

1. **Active selection:** consistency among the patches generated from a candidate.
2. **Handling noisy labels:** majority selection.
3. **Continuous fine-tuning:** fine-tuning the fine-tuned CNN.

Advantages:

1. Starting with a completely **empty** labeled dataset.
2. Incrementally improving the learner through **continuous fine-tuning** rather than repeatedly re-training.
3. Naturally exploiting expected consistency among the patches associated for each candidate to select samples "**worthy**" of labeling.
4. Automatically handling **noisy labels** as only a portion of the patches in each candidate participates in the selection process.
5. Computing entropy and diversity **locally** on a small number of patches within each candidate, saving computation time considerably.

References:

N. Tajbakhsh, et.al. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE TMI, 2016.

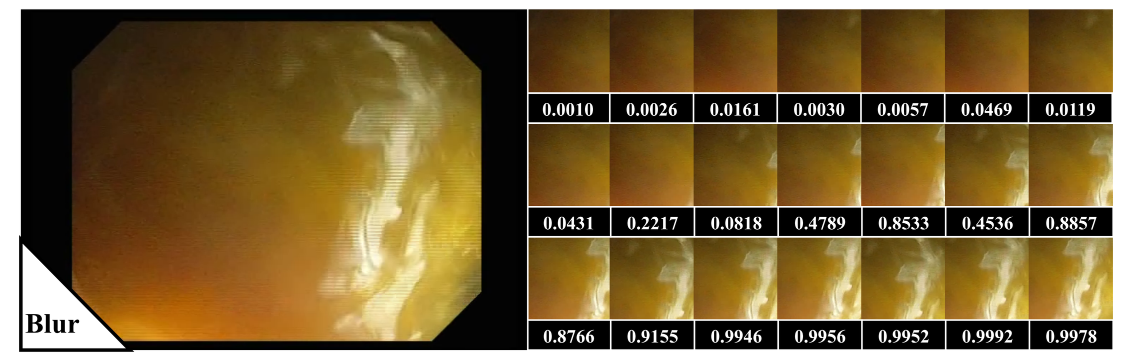
I. Guyon, et al. Active Learning Challenge. Microtome Publishing, 2011.

Method: Integrating active learning and transfer learning.

Algorithm 1: Active incremental fine-tuning method.

Input:
$\mathcal{U} = \{\mathcal{C}_i\}, i \in [1, n]$ {$\mathcal{U}$ contains n candidates}
$\mathcal{C}_i = \{x_i^j\}, j \in [1, m]$ {$\mathcal{C}_i$ has m patches}
$\mathcal{M}_0$: pre-trained CNN
b: batch size
α: patch selection ratio
Output:
$\mathcal{L}$: labeled candidates
$\mathcal{M}_t$: fine-tuned CNN model at Iteration t
Functions:
$p \leftarrow P(\mathcal{C}, \mathcal{M})$ {outputs of $\mathcal{M}$ given $\forall x \in \mathcal{C}$}
$\mathcal{M}_t \leftarrow F(\mathcal{L}, \mathcal{M}_{t-1})$ {fine-tune $\mathcal{M}_{t-1}$ with $\mathcal{L}$}
$a \leftarrow mean(p_i)$ {$a = \frac{1}{m}\sum_{j=1}^{m} p_i^j$}
Initialize:
$\mathcal{L} \leftarrow \varnothing$
repeat
  for *each* $\mathcal{C}_i \in \mathcal{U}$ **do**
    $p_i \leftarrow P(\mathcal{C}_i, \mathcal{M}_{t-1})$
    if $mean(p_i) > 0.5$ **then**
      $\mathcal{S}'_i \leftarrow$ top α percent of the patches of $\mathcal{C}_i$
    else
      $\mathcal{S}'_i \leftarrow$ bottom α percent of the patches of $\mathcal{C}_i$
    end
    Build matrix R_i using Eq. 3 for $\mathcal{S}'_i$
  end
  Sort $\mathcal{U}$ according to the numerical sum of R_i
  Query labels for top b candidates, yielding $\mathcal{Q}$
  $\mathcal{L} \leftarrow \mathcal{L} \bigcup \mathcal{Q}$; $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{Q}$
  $\mathcal{M}_t \leftarrow F(\mathcal{L}, \mathcal{M}_{t-1})$
until *classification performance is satisfactory*;

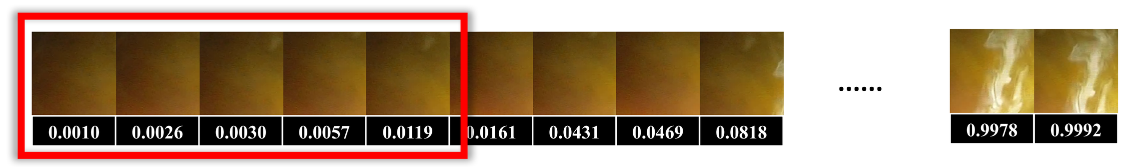
1 Get the prediction of unlabeled data.



2 Compute the average probabilistic prediction of all of its patches.

$$a_i = \frac{1}{m}\sum_{j=1}^{m} p_i^j = \frac{1}{21}\sum_{j=1}^{21} p_i^j = 0.4705 < 0.5$$

3 Select the top α percent patches when $a_i > 0.5$, otherwise bottom α.

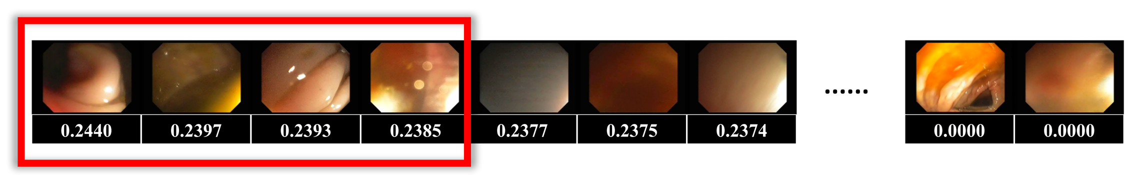


4 Construct the score matrix using either entropy or diversity quota.

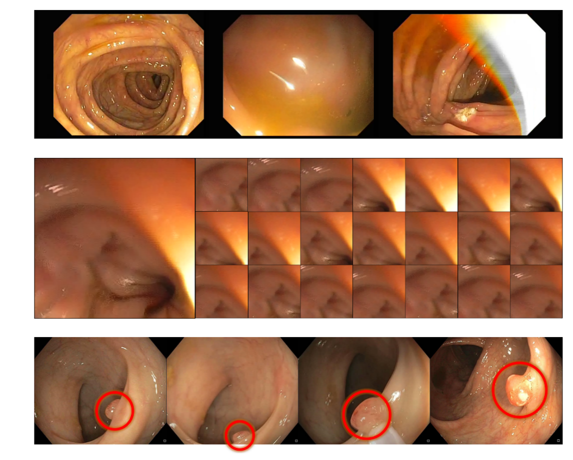
$$e_i = -\sum_{k=1}^{|Y|}\sum_{j=1}^{\alpha m} p_i^{j,k} \cdot \log p_i^{j,k} = -\sum_{k=1}^{2}\sum_{j=1}^{5} p_i^j \cdot \log p_i^j = 0.2440$$

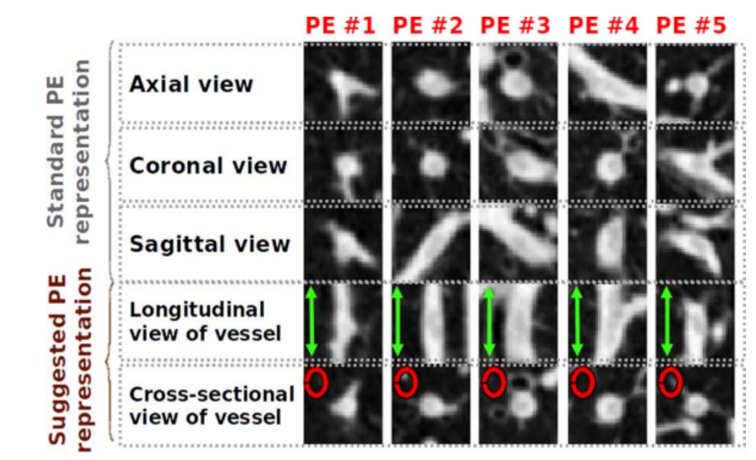
$$d_i = \sum_{k=1}^{|Y|}\sum_{j=1}^{\alpha m}\sum_{l=1}^{\alpha m} \left(p_i^{j,k} - p_i^{l,k}\right) \cdot \log \frac{p_i^{j,k}}{p_i^{l,k}} = 0.2964$$

5 Query labels for top frames according to their quota value.

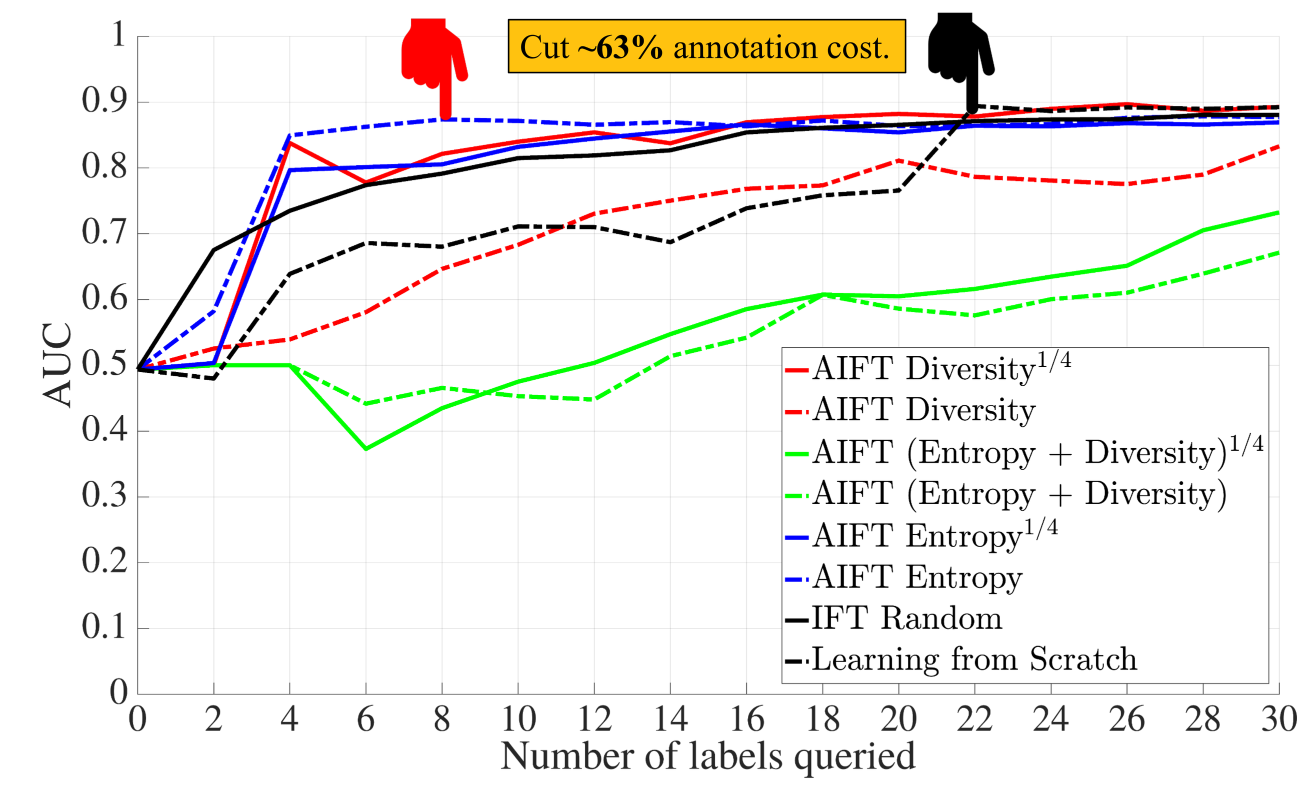


Applications: Cutting annotation cost at least in half.

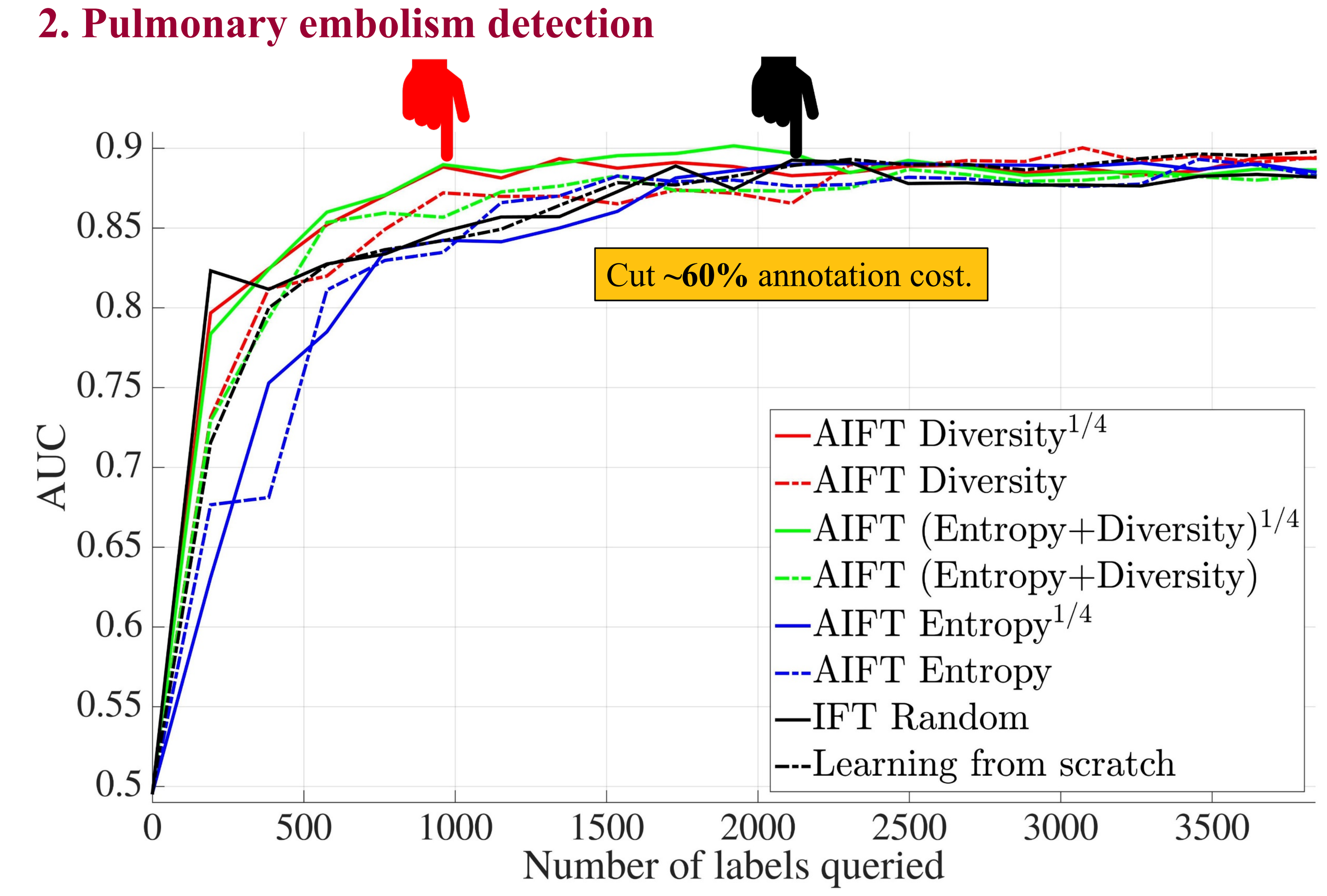




1. Colonoscopy frame classification



2. Pulmonary embolism detection



3. Polyp detection

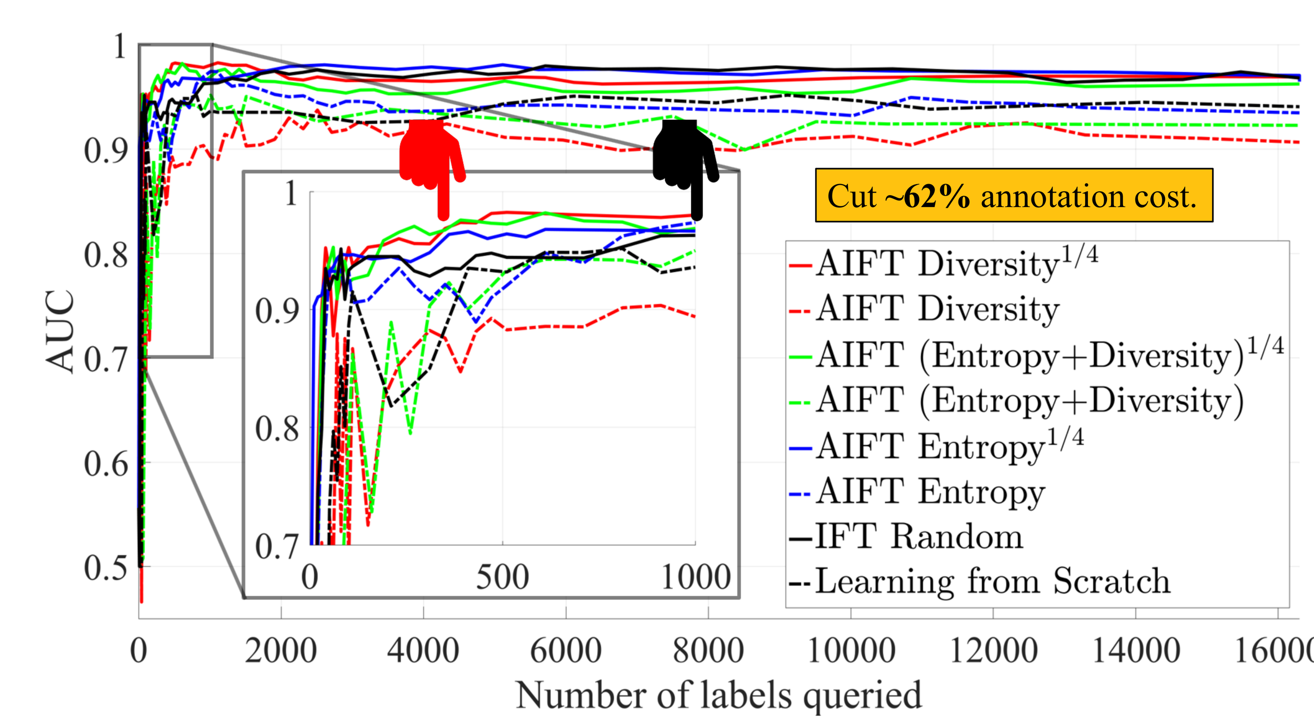
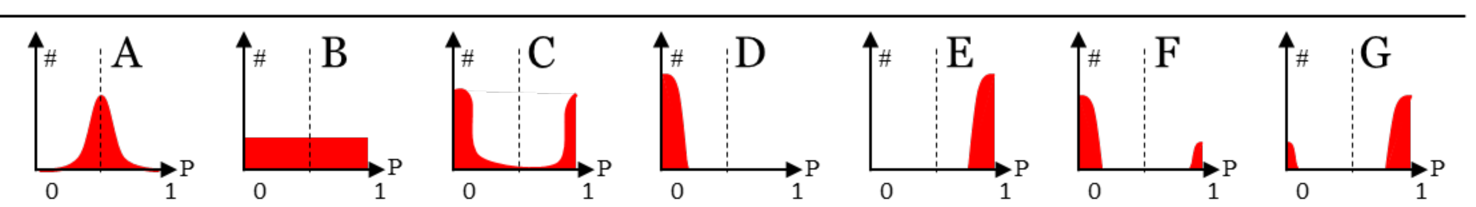
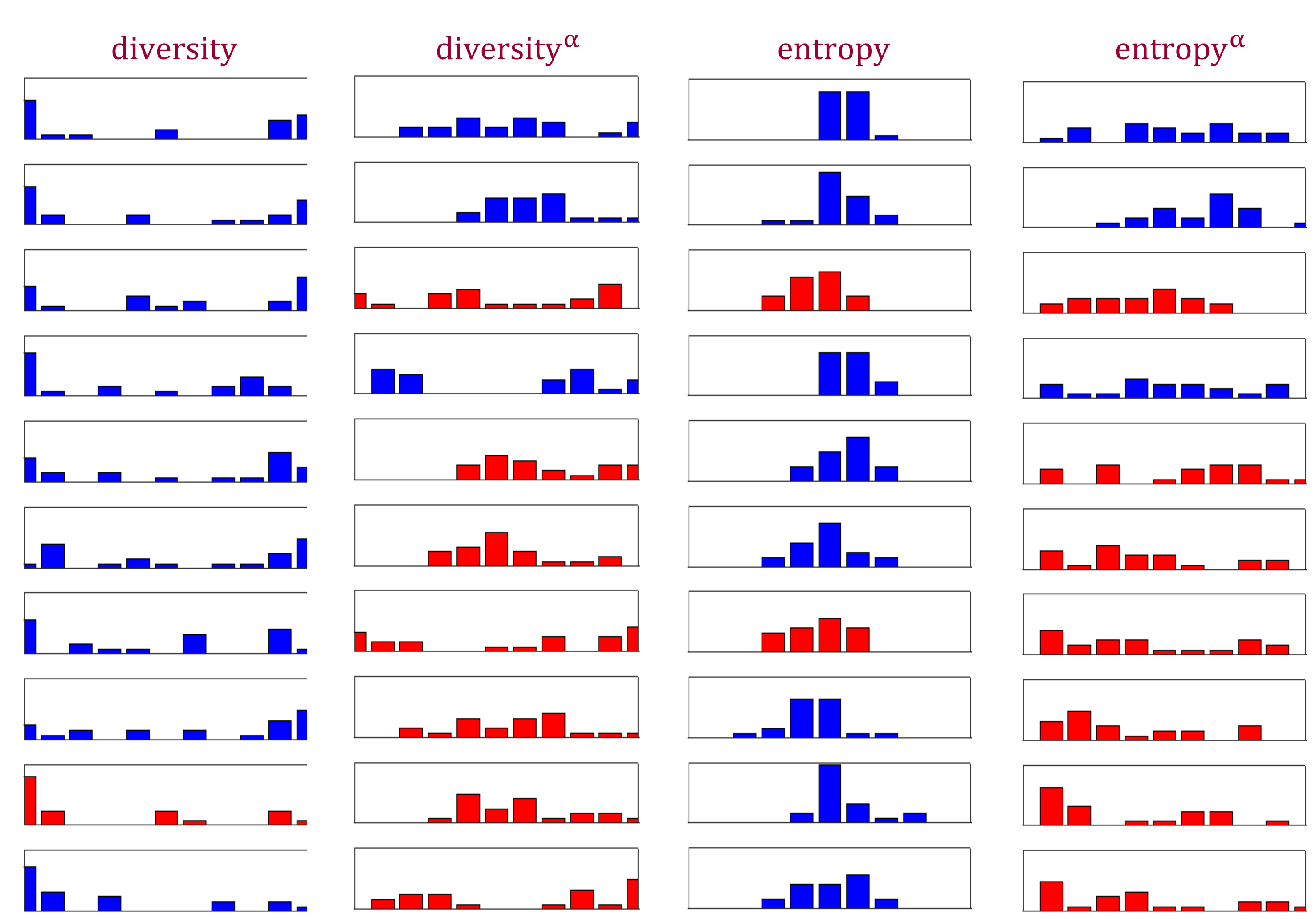


Illustration: Seven fundamental prediction patterns.



pattern	A	B	C	D	E	F	G
Example	0.4 0.5 0.4 0.5 0.4 0.5 0.5 0.6 0.5 0.6 0.6	0.0 0.6 0.1 0.7 0.2 0.8 0.3 1.0 0.4 1.0 0.4	0.0 0.9 0.0 1.0 0.0 1.0 0.1 1.0 0.1 1.0 0.9	0.0 0.0 0.0 0.1 0.0 0.1 0.0 0.1 0.0 0.1 0.0	0.9 1.0 0.9 1.0 0.9 1.0 0.9 1.0 1.0 1.0 1.0	0.0 0.2 0.0 0.2 0.0 0.3 0.1 0.9 0.1 1.0 0.1	0.0 0.9 0.1 0.9 0.7 1.0 0.8 1.0 0.8 1.0 0.9
entropy	7.52	4.57	1.30	1.30	1.30	3.24	3.24
entropy$^\alpha$	2.02	0.83	0.00	0.00	0.00	0.33	0.33
diversity	4.38	1237.21	2816.66	189.54	189.54	1076.87	1076.87
diversity$^\alpha$	0.00	20.79	0.00	0.00	0.00	13.54	13.54



Observations:

1. Patterns A and B are dominant in the earlier stages as the CNN has not been fine-tuned properly to the target domain.
2. Patterns C, D and E are dominant in the later stages of AIFT as the CNN has been largely fine-tuned on the target dataset.
3. The majority selection is effective in excluding Patterns C, D, and E.
4. prefers Pattern B while Diversity prefers Pattern C. This is why AIFT Diversity may cause sudden disturbances in the CNN's performance.
5. Patterns B, F, and G generally make good contributions to elevating the current CNN's performance.

Acknowledgements:

This research has been supported partially by the NIH NHLBI under Award Number R01HL128785, by an ASU-Mayo Seed Grant, and by an ASU-Mayo Innovation Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.