Problem Definition

Peyman Bateni

June 27, 2019

1 Important Clarifications - Check with Frank/Leonid

- Is detection as defined by the BAA a multi-object task? If we are only focusing on placing bounding boxes on a single object in the image, then we may (with lots of emphasis on the may) need to update some assumptions.
- In the object detection case, how are number of labels defined? Is an image with 5 distinct object instances, each having separate boundary boxes one label? Are the boundary boxes each a label regardless of how many there are in a single image? Are we supposed to account for either case? Then in the active learning case, do we get to request for all object instances and their boundary boxes in a single photo?

2 Important (Unconventional) Notes from BAA

- Goal is to make 6 orders of magnitude reduction in training data and 2 orders of magnitude reduction in data needed to adapt the model. Specific goals and milestones relative to the two phases (each consisting of 18 months) are provided in Figure 1. For this work, we focus on object detection.
- It's important to note that similar to other works in few-shot learning, we are able to leverage external publicly available dataset(s) or corpoa in enhancing performance. Additionally, it's noted that "[these] algorithms can make use of as much unlabeled data as they wish and they may choose specific examples for labeling". Thus, active learning is not only an option but seems to be encouraged throughout the BAA. It's not specified where the limit is drawn on amount on unlabeled data available but it's fair to be presumed to be orders of magnitude higher than the labeled amount.

3 Problem Definition

- Assumptions for the classification case:
  - $T^{train}_{source} = \{(X_{1,1}, C_1), ..., (X_{n,1}, C_1), ..., (X_{1,m}, C_m), ..., (X_{1,m}, C_m)\}$ where $X_{i,j}$ is the ith training example belonging to source class j, where T^{train}_{source} consist of m class with n examples each with $n \geq 1000$ (NOTE: we set this last time but reviewing the LwLL description, I think this may be too high given the label limitations imposed).
  - $T^{train}_{target} = \{(X_{1,m+1}, C_m+1), ..., (X_{k,m+1}, C_m+1), ..., (X_{1,m+h}, C_m+h), ...(X_{k,m+h}, C_m+h)\}$ where we have k target classes with h examples each with $1 \leq k \leq 10$.

	Phase 1		Phase 2	
Challenges	**Train (TA1.1)**	**Adapt (TA1.2)**	**Train (TA1.1)**	**Adapt (TA1.2)**
Object detection[1] Train: LSVRC (open) Adapt: TBD Metric: mAP @ # labels	80% @ 10^5	80% @ 10^3	80% @ 10^2	80% @ 10^2
Object classification[2] Train: LSVRC (open) Adapt: TBD Metric: mAP @ # labels	97% @ 10^6	97% @ 10^3	97% @ 10^3	97% @ 10^2
Activity recognition[3] Train: TRECVid MED task Adapt: TBD Metric: mAP @ # labels			41% @ 10^3	41% @ 10^2
Machine translation[4] Train: OpenMT task Adapt: TBD Metric: BLEU @ # labels			47% @ 10^3	47% @ 10^2

Figure 2: Program Goals for TA1 (accuracy @ N labeled examples. See terms here[1,2,3,4])

[1] LSVRC = ImageNet Large Scale Visual Recognition Challenge
[2] mAP = Mean Average Precision
[3] TRECVid MED = NIST TREC Video Multimedia Event Detection
[4] BLEU = BiLingual Evaluation Understudy metric

Figure 1: DARPA goals and milestones defined on the four noted tasks.

- $T^{train}_{zero-shot} = \{C_{m+k+1}, ..., C_{m+k+d}\}$ defined the zero shot case where no examples are provided for the d zero-shot classes. While the notation allows for T^{train}_{target} and $T^{train}_{zero-shot}$ to both be part of the problem definition, they are usually dealt with separately in the few-shot learning case (with d = 0) and the zero-shot learning case (with h = 0).
- At test time separate T^{test}_{source} and T^{test}_{target} are provided although the number of examples per class in T^{test}_{source} and T^{test}_{target} may be different from that of T^{train}_{source} and T^{train}_{target} (ie. the assumption $n_{test} = n_{train}$ and $h_{test} = h_{train}$ doesn't necessarily hold). With that being said, the set distribution with respect to each class remains the same: $\forall C_i \in C, (\frac{card(T^{train}_{source,C_i})}{card(T^{train}_{source})} = \frac{card(T^{test}_{source,C_i})}{card(T^{test}_{source})}) \wedge (\frac{card(T^{train}_{target,C_i})}{card(T^{train}_{target})} = \frac{card(T^{test}_{target,C_i})}{card(T^{test}_{target})})$.
- $T^{test}_{zero-shot} = \{(X_{1,m+k+1}, C_{m+k+1}), ..., (X_{r,m+k+1}, C_{m+k+1}), ..., (X_{1,m+k+d}, C_{m+k+d}), ..., (X_{r,m+k+d}, C_{m+k+d})\}$ at test time where $1 \leq r \leq 10$. Note that the zero-shot test set actually has examples, consisting of tuples as oppose to the set of only classes that's available during training.
- Assume for each class $C_i \in \mathbb{Z}$, with $1 \leq C_i \leq n + k + d$, there exists function $D : \mathbb{Z} \rightarrow \Sigma^+$ such that (as per definition), $\Sigma = \{"a", "b", "c", "d", ..., "z"\}$.
- Assuming the presence of a secondary source of class string label embeddings such as word2vec [4] or GLoVe embeddings [5], there exists a function $J : \Sigma^+ \rightarrow \mathbb{R}^e$ where e is the dimension of the embedding (often set to 50 or 100).

- **As suggested by the BAA, we also have acccess to $T^{train}_{unlabelled} = \{X_1, X_2, ..., X_g\}$ where $g \geq 10^{10}$. The assumption here is that we have near limitless unlabelled data for training. Additionally, there exists a function such that for $X_i \in \mathbb{Z}^{height,width}$ and corresponding class $C \in \mathbb{Z}$, $M : \mathbb{Z}+^{height \times width} \rightarrow \mathbb{Z}$, providing the ability to request labels for unlabelled examples available. However, we can only use this function a total of p times with $1 \leq p \leq 100$. Note that here, height and width are defined by the dimensions of the input image.**

- Assumptions for the detection case:
  - Problem setting is somewhat similar with the addition of boundary boxes to each example/class topple in each of the sets. Changes would be as follows:
    - $T^{train}_{source} = \{(X_{1,1}, B_{1,1}, C_1), ..., (X_{n,1}, B_{n,1}, C_1), ..., (X_{1,m}, B_{1,m}, C_m), ..., (X_{n,m}, B_{n,m}, C_m)\}$
    - $T^{train}_{target} = \{(X_{1,m+1}, B_{1,m+1}, C_m+1), ..., (X_{k,m+1}, B_{k,m+1}, C_m+1), ..., (X_{1,m+h}, B_{1,m+h}, C_m+h), ..., (X_{k,m+h}, B_{k,m+h}, C_m+h)\}$
    - where $B_{i,j}$ is an 8-tuple of positive pixel indices and $(0, 0)$ is assumed to be the bottom left corner of the image.
  - The zero-shot training set remains the same with $T^{train}_{zero-shot} = \{C_{m+k+1}...C_{m+k+d}\}$. However, for the zero-shot test set, the set is update to include boundary boxes:
    - $T^{test}_{zero-shot} = \{(X_{1,m+k+1}, B_{1,m+k+1}, C_{m+k+1}), ..., (X_{r,m+k+1}, B_{r,m+k+1}, C_{m+k+1}), ..., (X_{1,m_k+d}, B_{1,m_k+d}, C_{m+k+d}), ..., (X_{r,m_k+d}, B_{r,m_k+d}, C_{m+k+d})\}$
  - Similar functions D and J are present for mapping. **Additionally, $T^{train}_{unlabelled}$ and function M are also available in the detection case. Note that here the examples are unlabelled meaning they don't have the boundary boxes either.**
  - **For unlabelled example $X_i \in \mathbb{Z}^{height,width}$, there exists function $B : \mathbb{Z}^{height,width} \rightarrow \mathbb{Z}^8$ that maps the unlabelled image to an 8-tuple describing the boundary boxes. Similar to fucntion M, B can only be used on a total of p unlabelled images with $1 \leq p \leq 100$.**

- Distinctions to make for problems at hand (unless otherwise noted, definitions were taken from [2]):
  - Few-shot learning:
    - Parameter setting:
      - $n \geq 1000$ (you are given a reasonable number of examples per source class)
      - $1 \leq k \leq 10$ (on few-shot, otherwise known as target classes, you have a maximum of 10 examples per class)
      - Often $m > n$, although this is subject to the problem setting in the paper.
      - $d = 0$ (no zero-shot or zero example classes)
    - Performance is only based on classification/detection accuracy on T^{test}_{target}.
    - Common jargon used includes N-way M-shot learning being defined as a few-shot learning task involving N target classes and M examples per target class.
  - Zero-shot learning:
    - Parameter setting:
      - $n \geq 1000$ (you are given a reasonable number of examples per source class)
      - $k = 0$ (there are no target classes that have few examples)

· $d > 0$ (there are d zero-shot classes)

* Performance is only based on classification/detection accuracy on $T^{test}_{zero-shot}$
* Similar notation is held with respect to N-way zero-shot learning specifying a zero-shot learning setting with N target classes.

– Generalized few-Shot learning:

* Same parameter setting as few-shot learning
* Performance is based on classification/detection accuracy on both T^{test}_{source} and T^{test}_{target}

– Generalized zero-Shot learning:

* Same parameter setting as few-shot learning
* Performance is defined as classification/detection accuracy on both T^{test}_{source} and $T^{test}_{zero-shot}$.

– Open-set learning [3]:

* With regards to few-shot learning and zero-shot learning respectively, openset learning defined the case where h (ie. number of few-shot cases) in the first case and d (ie. number of zero-shot cases) in the second case are not given (are unbounded at test time).
* Work here often involves label discovery as the network should be able to determine whether a presented example is of a previously seen class or new, for which in the latter case it must then learn a new class representation during test time, or effectively deal with it some other way.

– Semi-supervised learning [1]:

* In any of the aforementioned cases, problem would be defined as semi-supervised learning if we're given access to $T^{train}_{unlabelled}$ during training.

– Active learning [1]:

* Extension to the semi-supervised learning task where we are also given access to functions M (and B in the detection case) during training, thus allowing us to request labels for a limited (≤ 100) unlabelled examples.
* In the context of the DARPA project, this would be taking away from the number of starting labels (as the total number of labels used would be intended to be lower than the DARPA specified thresholds).

4 Datasets

- **The DARPA challenge with respect to object classification and detection focus on the image dataset from ImageNet Large Scale Visual Recognition Challenge (ILSVRC).** An in-depth description of the dataset has been included as part of the dataset description.
- **The dataset for the adaptation task has yet to be announced.**

References

[1] Rinu Boney and Alexander Ilin. Semi-supervised few-shot learning with prototypical networks. *CoRR*, abs/1711.10856, 2017.

[2] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. Recent advances in zero-shot recognition. *CoRR*, abs/1710.04837, 2017.

[3] Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *CoRR*, abs/1811.08581, 2018.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[5] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.