Learning With Less Lables - Literature Review

Peyman Bateni

May 10, 2019

1 Introduction

The following outlines the literature review completed as part of this study.

2 Zero-Shot Learning Through Cross-Modal Transfer [8]

2.1 Summary

- Using Huang et al.'s context-based word-prediction dataset of word vectors and visual features from images by Coates et al.'s unsupervised algorithm, they formulate a transfer two-layer transfer network that maps visual feature vectors from the input image to the semantic space of word vectors. The network is trained using a sum-of-squared loss objective function based on word vector representations of the labels of images and the resulting mapping of the input visual feature vector.
- Introduce a binary novelty random variable V that indicates novelty of the potential label for the input image. The prediction would be $p(y|x, X_s, F_s, W, \theta) = \sum_{V \in s,u} P(y|V, x, X_s F_s, W, \theta) P(V|x, X_s, F_s, W, \theta)$ where x is the input image, X_s describes seen images from training, F_s consists of the seen visual features, W are the the word embeddings in the semantic space and $theta$ is the learned mapping function.
- They propose two strategies for novelty detection. The first relies on Gaussians of each class such that for $y \in Y_s$, $P(x) = P(f|F_y, w_y) = N(f|w_y, \Sigma_y)$ where w_y is the respective word embedding of the class and $/Sigma_y$ is estimated based on the seen classes. Here, the novelty is estimated based on $P(f|F_y, w_y) < T_y$ where T is manually assigned threshold experimentally chosen and the novelty is assigned if $yP(f|F_y, w_y) < T_y$.
- The second approach uses a modified form of Kroger's outlier detection to formulate a Gaussian Error probability bounded from below to weigh the seen and unseen classifiers given belief about the outlierness of the input test image.
- When classifying, if object is determined to have been seen, a classical Softmax classifier is used. Otherwise, isometric Gaussian distributions around each of novel class word vectors are assumed with classes being predicted based on the likelihood. An outlier probability threshold is used to decide which to use in the case of the second approach to novelty detection.
- The model is able to achieve near SOTA (at the time) performance on seen images at 82.2% while maintaining reasonable accuracy on zero-shot at 52.7%, although improved by 10% using a two layer network suggesting there are performance gains to be made using deeper models.
- The first novelty detection approach is found to be more liberal in assigning unseen labels to test images, where as the second, is very conservative especially when the visual feature representation of

the input image is within the manifold of seen images. The results also show a none-linear trade-off between seen and unseen performance, choosing thresholds accordingly.

- Database used: CIFAR-10/100 dataset of images, Wikipedia corpus for word vectors

2.2 Strengths

- The work proposes zero-shot learning that outperforms previous attribute-based models without the need for explicit class definitions for unknown classes. In fact, theoretically, the model has implicit unsupervised understanding of all concepts as learned from the language corpus. Additionally, by introduction of the novelty random variable, the model is able to perform open-set classification, performing reasonably on both seen and unseen examples, suggesting that there's value in modulating the two tasks as separate classifiers.
- The work uses seen examples to evaluate the manifold of the space more effectively that simple KNNs in evaluating whether the input belongs to a previously seen class or not.

2.3 Weaknesses

- Classification isn't generalized in the case of seen/unseen labels. The model uses the semantic space for evaluating whether the example has been seen before but afterwards opts to use essentially nearest neighbors in form of isometric Gaussians to evaluate unseen examples and a softmax classifier for seen examples.
- The model is limited to items defined by the corpus and assumes that the NLP semantic space is adequate for unseen examples given the transformation function. There is potential in using seen examples from known classes to verify whether distances between class vectors in word space are accurate with respect to visual features.
- Thresholds are assigned through experimentation as oppose to learned. Episodic training could be used here to potentially learn those parameters on training.

2.4 Potential

- Use of learned probabilistic distributions on the novelty variable may help to generalize the model as a whole such that confidence levels are obtained on how confident the model would be to classify the input as an unseen example vs seen as oppose to using the novelty variable to decide which to pursue.
- The word vector space is extremely powerful. Exploration of how the model could potentially add to them if entirely new examples are seen would be interesting. Additionally, considering visual context and extending the work to multi-object zero-shot classification, and/or zero-shot detection may be of potential.

3 DeViSE: A Deep Visual-Semantic Embedding Model [2]

3.1 Summary

- The work uses a pre-trained space of word embedddings widely used in other works and also, in parallel pre-trains a softmax classifier for image classification. Afterwards, the final layer of the visual classifier is replaced with a 4096 to d ($d = 500$ and $d = 1000$ performed the best) FC-layer that intends to

predict the visual vector embedding as close to the assigned label as possible. This new model is fine-tuned using hinge rank loss which is shown to perform twice better than L2 (used by other works) as intuitively, it's believe to resemble KNN more closely. At test time, the computed vector for the image is assigned the closest label to it in the semantic space, obtained by KNN (could be more quickly obtained through a tree/hashing approach).

- The core model was kept constant during fine-tuning. Later experiments extending the fine-tuning to the core visual model in later stages showed to improve accuracy by 1-3%. Additionally, the it was found to be expedient to randomize the loss through: 1. restricting the set of false text terms to possible images in the ranking, and 2. truncating the sum after the first margin-violating false term was encountered. It must be noted that all embeddings were unit norm.
- With precision@k as the secondary metric based on the label hierarchy provided by ImageNet where label accuracy is extended by also considering how close the incorrect label was to the actual one semantically, the model performed close to the visual softmax classification bench mark on labeling accuracy. However, on precision@k, the model showed to beat the softmax classifier, demonstrating its ability to make semantically sound mistakes as oppose to completely irrelevant ones. This evaluation was based on the ImageNeT ILSRC 2012 1K-label on seen examples.
- With respect to zero shot learning, 2-hop and 3-hop label sets based on label hops away from the 1K seen examples were used with 1,589/7,860+1K labels respectively. The model showed strong generalization capabilities achieving 18.1%/5.3% top5 accuracy on zero-shot only and 7.9%/3.4% accuracy on zero-shot only and open-set classification. On full ImageNet 2011 21K, with 20,841 unseen labels, performance was 2.5%/1.9% in the same order.
- On 800 seen to 200 unseen classification, model outperforms Mensink et al. 2012 and Rohrbach et al. 2011 on open-set classification at 9% relative to 2%. However, on zero-shot only model performs worse than than Mensink et al. 2012 on zero-shot only.

3.2 Strengths

- Model matches SOTA performance on seen-class classification of images while also demonstrating better understanding of class semantics as demonstrated by the more semantically sounds mistakes and better performance on precision@k hierarchical accuracy.
- Model effectively leverages the underlying corpus of vocabularies motivating extension to larger vocabularies to be used to extend the work to more classes. Additionally, despite SOTA performance on seen classes, model is able to generalize well to other classes, especially if they are close to the seen example semantically.
- Work establishes hinge rank loss as twice more effective at training mapping networks than the previously widely used L2.

3.3 Weaknesses

- Model shows bias towards seen classes heavily when classifying. This bias halves accuracy on zero-shot classification as oppose to when seen classes are removed. Better generalization could fix this problem.
- Model still relies on simple distance methods and KNN to identify new classes, not leveraging number of existing examples to create better decision boundaries.

3.4 Potential

- Closest distance may not be fully accurate as in a 500-dimensional semantic space, decision boundaries with different classes may be different to that of just simple Euclidean. More investigation could be fruitful.
- Model doesn't use contextual information which could improve performance on multi-label if that's the goal. Additionally, model relies on label semantic embeddings based on a language corpora without fine-tuning using visual information. This is worth exploring as it may further optimize the space.

4 Neural Graph Matching Networks for Fewshot 3D Action Recognition [4]

4.1 Summary

- The work proposes nneural graph matching (NGM) for few-shot learning to exploit 3D action data when classifying newly seen actions. The networks consist of two stages trained end-to-end. The first stage consist of generating the graph where annotated object/poses (or if not available using a pre-trained classifier) are used to generate the nodes with associated features extracted from raw pixels of the image. As for edge, they use a differentiable method due satisfy end-to-end training using an MLP for updating edges with node-specific feature tranforming functions that are used in step, such that both the nodes and edges are updated iteratively to take into account the surroundings of both the nodes and the edges.
- The second stage consists of inexact graph matching where they use graph tensors as representation consisting of three dimensional tensors whose size is based on node types and the note feature dimension where each $cell_ij$ corresponds to number of nodes if $i = j$ and the sum of feature edges otherwise. The squared distance between the two graph tensors is used as the matching metric. Afterwards, the model follows the prototypical approach, defining graph tensor prototypes as the average the respective examples, using the graph matching metric as the distance when performing (nearest based) softmax classification.
- Model is trained through "episodic" training often used in few-shot learning with negative log softmax of the true label stochastic gradient descent based on the seen data.
- The model is evaluated on CAD-120 sub-activities and the PiGraphs capturing common activities (comes with voxel annotations). It's compared to 3 baselines: PointNet which takes in the point coordinates cloud as input and achieves "current" SOTA (+ feature representations concatenated for fairness), Part-Aware LSTM, and NGM without edges (reducing graph generation and matching to only nodes without messages being passed between them. NGM outperforms baselines significantly at 78.5%, 91.1% on 1/5-shot classification on CAD-120 and 80.2%/88.3% on PiGraphs. Without edges, NGM performs over 10% lower, more comparable to baselines.
- An ablation study shows that while heuristic proximity or human-object measures on the edges can improve performance, NGM's true performance comes through the explicitly use of the graph generation network in obtaining more subtle relations. Additionally, it's shown that both adjacency information and feature information with respect to each node is needed to maximize performance which the graph tensor is able to combine both.

4.2 Strengths

- Work sets new state-of-the-art performance on the work at hand. The end-to-end training combined with the episodic optimization allows for better capturing of few-shot learning.
- Additionally, while this particular paper focuses on 3D action, it demonstrates the importance of multi-object context and to extend learning end-to-end to capture the graph generation wanted. The proposed differentiable iterative edge and node updates are effective.
- The graph tensor way of generating matching scores otherwise used as the distance metric in the graph space is effective, and can be used in our work, should we pursue the graphical bi-modal representation approach.

4.3 Weaknesses

- The model uses the graphs as metrics only, and there is an argument that similarities between graph structures outside of just the metric can allow the model to learn generalizable insights about part-actions. The use of simple closest distance classification seems a bit naive given the complexity of the model.
- It's unclear whether use of more expressive visual features or potentially adding a CNN for producing such features to the end-to-end training could improve performance.

4.4 Potential

- Using the graphical representation could be very beneficial in modelling contextual relation in multi-object zero/few-shot classification. While this is subject to more exploration, the suggest graph tensor and distance metric can prove useful in measuring distances between graphical representations of support/query examples.
- The graph generation algorithm, especially since it allows end-to-end training, can be leveraged to generate the graphs in the initially proposed graphical context-based deep learning approach.

5 One-Shot Learning with Hierarchical Nonparametric Bayesian Model [6]

5.1 Summary

- The proposed model uses three hierarchical levels (albeit more like "two" with respect to complexity), where at the lower level (level-1) the distribution of each category c is assumed to be Guassian with category specific means and precision matrices, the aggregation of which form θ_1, the level-1 category parameters. The second level models super-categories where every category c is assigned to super-category k where a Normal-Gamma conjugate prior is placed over level-1 means and precision matrices where level-2 parameters consist of $\theta_2 = \{\mu_k, \tau^{-1}, \alpha_k\}_k = 1^K$ for all K super categories with Normal, Exponential and Inverse Gamma conjugate priors respectively. They further diffuse a Gamma prior over $\theta_3 = \{\alpha_0, \tau^0\}$, the set of more global random variable used in assigning the level-2 conjugate priors.
- The model generalizes to nonparametric unbounded number of super/basic categories using a Nested CRP, which extends CRP to nested sequence of partitions, one for each of the two main levels of the

tree. Observation is first assinged to the super-category where recursively, the basic category is also drawn. The probabilities for assigning to each category at each level follows classical CRP definition. Additionally, a Gamma prior is placed over the concentration parameter of the CRP.

- For sampling level 1 and 2 parameters, the means and precision matrices for the basic categories are completed using conditional distributions taking Normal, Gamma, Normal and Inverse-Gamma forms respectively. The complications arise in sampling α_k where the none-closed form solution is approximated using a Gamma distribution, with values being accepted based on the Metropolis-Hastings rule.
- As for sampling assignments z, the posterior is computed through combining the likelihood term of model parameters with nCRP prior, which formulate a Bayesian proportionality equation. The work further exploits the conjugacy in the hierarchical model, using the fact that Normal-Gamma prior is conjugate prior of a normal distribution, to calculate the marginal likelihood on an example, integrating out basic-level parameters, thus making sampling more efficient.
- In the case of one-shot learning, the nCRP is used to decide to whether add the new example to an existing super-category or form one of its, the same follows with the basic-level. Of course, at basic level, the knowledge regarding mean and precision parameters of the super-category are used to transfer super-category learned characteristic to the new cluster. When presented with a text example, the probability of belong to the new category is measured by the normalized Bayesian of the conditional probability of the category given the example where the prior is given by the nCRP. The log-likelihood is given using the feature-focused categorical distributions, where of course, more "precise" features are given more weigtht.
- With AUROC of 0.81 on the MNIST dataset, withholding 100 examples belonging to 9, the model outperformns HB-FLAT (HB with one shared super category), HB-VAR (using covariances, ignoring means of super-categories), Euclidean and MLE (completely excluding super-categories). On MSR Cambridge Data, similar results are achieved at AUROC of 0.77. It's important to note that the model was able perform comparably after seeing one example to that of HB-FLAT and MLE after seeing four examples.
- In terms of one-shot learning, the model shows reasonable performance when tasked to face three unsupervised unseen classes. However, the performance gets much better with only a little more data. At 18 unlabelled images, after running Gibbs sampler for 100 steps, the model is able both classify familiar examples and also form new clusters for new unseen examples with appropriate super-categories as verified through a qualitative study.

5.2 Strengths

- Well, model is able to leverage Bayesian nonparametrics in from of the nCRP to potentially learn an unbounded number of labels, and discover new clusters or attach to previously familiar ones with reasonable accuracy.
- Use of the hierarchy allows for better transfer of priors to new examples, for instance helping to distinguish between cows and chimneys using the super-category, a luxury that others fail to capture.
- Model requires astonishingly low amounts of data to train to high-levels of accuracy, and does well with even forming new clusters despite having no prior knowledge embeddings of any sort describing the class definition; this is a major leg over deep learning approaches that need attributes or class descriptions to perform one-shot or zero-shot learning.

5.3 Weaknesses

- This is a personal bias, but I have a tendency to believe that real-life data distributions are too complex to be modeled using Gaussains or Gamma distributions (unless getting to ridiculous number of mixtures) and thus, have a deep learning bias. Maybe using "deep" visual representations and training end-to-end would be of interesting performance gains.
- The model doesn't consider contextual information. Additionally, and I have a feeling that depending on the task at hand, if this were to be used for zero-shot learning as oppose to one-shot, while you may be able to recognize the difference of the new example, assigning just a vectorized label to the zero-shot learned example may not be of interest especially if the model is to be used by a individuals who otherwise want comprehensible labels.

5.4 Potential

- Factoring in context, using better deep learned features, introducing multi-modal learning potentially combining the NLP concepts with visual examples to essentially create a Hierarchical Bayes model on the space of examples from multiple modes and a lot more :)

6 Human-level concept learning through probabilistic program induction [5]

6.1 Summary

- The paper proposes BPL, where simple probabilistic generative models as structural procedures are used as concept abstraction in the language. The framework brings together three ideas: compositionality, causality and learning to learn. The model breaks down alphabetic examples to more primitive structures that resembling pen movement on the paper. The generative model therefore samples the appropriate primitive pieces and combines the associated parts reflecting the causal structure of writing a character.
- More specifically, the model defined the joint distribution on types ψ given M tokens of that type $\theta^(1), ..., \theta^(M)$ and the corresponding binary images $I^(1), ..., I^(M)$ as: $P(\psi, \theta^(1), ..., \theta^(M), I^(1), ..., I^(M)) = P(\psi)\Pi_{m=1}^{M} P(I^(m)|\theta^(m))P(\theta^(m)|\psi)$. $P(\psi)$ is obtained using the generate type function that samples the number of parts and there after, the sub-parts. Then it proceeds to sampling the sub-part sequence and sampling the relation. The obtained parts, subparts and relation is passed as θ to the generate token function, which obtains $P(\theta^(m)|\psi)$ through adding motor variance, sampling part's start location and composing trajectory before sampling the afine transform and finally sampling the image itself.
- The model learns to learn by fitting each conditional distribution to a background set of characters from 30 alphabets, using both and image and the stroke data. The model is also able to perform well with respect to generating new examples on one-shot basis thanks to placing a nonparametric prior on the type level.
- The model was evaluated against two deep learning model (a normal convolutional network and a Siamese network intended for one-shot learning), a hierarchical deep model along with two variations of additional variations of the model dropping learning to learn through disrupting learned hyper-parameters of the generative model and reducing compositionality by allowing just one spline-based stroke. The model is shown to perform extremely well, 3.3% on one-shot classification against human

one-shot classification of 4.5% vastly over the Siamese model at 8% with similar performance across seen examples (multiple-shot).

- Additionally, on generative tasks were the model was evaluated on a Turing test against human-generated copies of the characters, the model achieved a 52% identification level with 50% being equal to human beings, although on generating new concepts from type and unconstrained, it went below 50% ID value. It must be noted that model lesions without learning to learn and compositionality resulted in 5-10% worse performance on classification and over 30% lower Turing ratio.

6.2 Strengths

- The model is able to perform human-level classification while outperforming recent deep learning approaches. The model is able to generalize well as showcase by the near 50% Turing scores in ratios. Overall, the performance of the model is exemplary for the task at hand.

6.3 Weaknesses

- The model relies heavily on domain assumptions with respect to alphabet characters consisting of sub-parts formed as by-product of resembling writing with a pen. I'm not sure how generalizable this assumption is with respect to visual task involving more complex objects.
- The model also solely focuses on written characters as oppose to other visualizations and while the compositionality principle is very powerful, extending the model to composition of more complex and even hierarchical entities may prove to require model extensions relative to the line and dot parts used here.
- Paper lacks certain specifications as to what distributions were used, which motivated looking deeper into their code which is thankfully available online.

6.4 Potential

- Extending this model to more visually complex examples such as objects can be interesting and fruitful. This of course heavily relies on finding a generalizable sub-compositional scheme.
- Additionally, it seems like model avoids grouping or clustering entities together or potentially creating relations outside of the scope of the compositionality of parts which makes sense in this setting, but with respect to more visual images such as object identification, it may be interesting to look into ways to extend the model to consider object relations and even potentially, context for that matter.

7 Edge-Labeling Graph Neural Network for Few-shot Learning[7]

7.1 Summary

- The proposes EGNN a framework where all support and query examples are formulated as a full-connected graph where node's represent examples using their feature representations from the final layer of a CNN as the embedding. The edges hold two value of similarity and dissimilarity are initialized based on the labels for the seen examples where if two nodes belong to the same cluster, edge similarity is 1, otherwise zero and if one of the nodes belongs to an unseen example, edge similarity is set to 0.5 with edge dissimilarity being equal to 1 - dissimilarity.

- The node features and edge values are updated iteratively. The node feature update is firstly conducted by a neighbor aggregation of features where separate sums of the neighboring nodes weighted by sim/dissimilarity values of the edges respectively is used as input to the feature transformation network. The edge feautre update is then performed using newly updated nodes by combining the previous edge and updated similarities/dissimilarities using a metric network for computing dis/similarities, and dividing by the L1-norm of the resulting value at the end. Note that the edge updates don't just consider the relationship between the two nodes, but also relationships between other nodes. For specific formulas visit the paper.
- The final result for a node to be classified is a softmax classifier using the summation of the recalculated edge weights (1 if same, zero otherwise) based on the classification label of the neighboring nodes as per each potential classification.
- Model is evaluated on miniImageNet (100 classes, 60,000 images) and the tieredImageNet (700k image, 608 classes), both subsets of ILSVRC-12. The model is also trained with episodic training, using episodic query edge-label binary cross entropy loss. EGNN shows best performance in 5-way 5-shot setting, on both transductive (outperforming TPN) and non-transductive (outperforming Prototypical Networks) settings, with transductive settings consisting of processing all queries all together as oppose to independently.
- Additionally, EGNN outperforms node-labelling GNN, showcasing the effectiveness of edge-labelling. It's also shown that in a semi-supervised setting, the model outperforms node-only model especially when extending the training data to unlabelled examples as it's able to exploit semi-supervised relations between examples more effectively. There are notable increases from 68% accuracy to 73.2'% accuracy when the number of layers increases to 2, with marginal improvements at three layers with 76.37% accuracy, showcasing better performance with more layers.
- Use of separate inter and intra-cluster aggregation on the edges improves performance with greater generalizability to cases where few-shot settings are different to that of training.

7.2 Strengths

- Work showcases the performance gains that can be earned by using a graphical representation of the support/query example space without summarizing clusters into just their respective means. This can be leveraged elsewhere where more explicit use of examples and their inter-connections can improve performance.
- Additionally, the graphical network proposed uses edges to explicitly formulate inter-example relations and as a result, by considering all query examples at once, stands to perform better where as previous models usually didn't have this luxury.

7.3 Weaknesses

- The optimization takes substantially longer and its unclear how many passes are needed at test time to fully optimize the graphical space.
- Based on my understanding, the network requires re-calibration every time new query samples are introduced, where as one might argue that using heuristic you can avoid recomputing edge/node aggregated values on samples are potentially irrelevant/or very loosely relevant.
- The evaluation mainly focuses on few-shot learning of query cases, often missing out on how the performance sits on known classes.

7.4 Potential

- There is clear potential in extending the edge relationship using contextual information in terms of appearance of object together, potentially even further extend using some information padded from an NLP corpus.
- I wonder if we could leverage some nonparametric Bayesian model using the edge weights as space distances to decide on whether the sample belongs to an existing class or we need to find/create a new class for it. Lots of potential here.

8 Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks [1]

8.1 Summary

- MAML proposes an algorithm based on the intuition that underlying parameters can be learned independent of the task with maximum adoptation through few gradient descent steps. The algorithm uses a distribution over tasks with two step size hyperparameters and randomly initializes the parameter set θ. Until convergence, the model samples batch of tasks and then evaluates each task with respect to K examples, computing the adopting parameters and then aggregating the descent steps for all sampled tasks to then update the parameter set as a whole.
- The meta-gradient update is applied using an additional backward pass through the task function to compute the Hessian-vector to obtain the gradient through the task=specific gradient. MAML adapted for Few-Shot supervised learning uses a separate validation sample set from each task to calculate loss from the sampled tasks. This means that the validation loss with respect to each task in the batch is used as the meta-training loss. Similar extension are made to reinforcement learning, although since redundent for the research at hand, they were ingored for the purpose of the summary.
- A special experiment using a sine wave function with varying amplitude and phase of the sinusoid. Compared to two baselines of a transfer learned network and an oracle that uses the knowledge of the sine function, it's shown that using 1 gradient step and only 5 examples, the MAML model is able to closely mimic the underlying function and significantly improve performance with just 1-2 gradient steps, while the transfer learned pre-trained network converges much more slowly, reaching just approximately half MAML levels even after 10 steps.
- On 1/5-shot image classification 5-way accuracy on Omniglot and MiniImagenet, MAML narrowly outperforms state of the art on Omniglot while significantly outperforming memory-augmented netowrks and meta-learner LSTM, on both datasets. Additionally, major computational cost of MAML comes from calculating second derivatives when updating through gradients. A simple approximation using only first-order derivatives is shown to nearly mimic the same accuracy while saving 33% in computation costs. Experiments on reinforcement learning also shown significant improvements in fast adaptability to new tasks.

8.2 Strengths

- Clearly, MAML significant reduces the amount of fine-tuning needed to adopt to new tasks. Additionally, this is achieved in 1-2 gradient steps, minimizing the adaption needed for few-shot learning without overfitting which is key in learning from few examples.

- MAML is model agnostic, thus it could potentially be applied to models of great complexity or specific nature to a particular setting, without requirements in terms of the model structure limiting the possible range of designs.

8.3 Weaknesses

- There is unclear evidence as to how effective this would be on zero-shot learning or on adopting to tasks it hasn't seen for which can completely different in nature to the previous observed scenes.
- Furthermore, there doesn't seem to be much evidence with respect to open-set or both seen/unseen classification. While the task at hand is adapted effectively, I wonder if the resulting model significantly loses its previous knowledge on all tasks.

8.4 Potential

- We could potentially use MAML in training larger models that leverage priors, NLP corpus class knowledge or graphical representations.
- More exploration is needed by potentially generalizing the overall model such that its potentially able to recognize tasks that are completely different from previously seen ones before and use different sets of initializations.
- NOTE TO SELF: Explore this more, look at other papers as well to gain a better understanding!

9 DeViSE: A Deep Visual-Semantic Embedding Model [2]

9.1 Summary

-

9.2 Strengths

-

9.3 Weaknesses

-

9.4 Potential

-

10 Recent Advances in Zero-Shot Recognition [3]

10.1 Summary

- Provides definition for zero-shot learning (unseen visual, unseen class definition released at test time), few-shot learning (average of 1-5 examples seen during training, 1 is known as one-shot) and open-set classification where model must classify with respect to both seen and unseen labels. At training you are given the source/auxiliary dataset providing the support examples while at test time, you get the target/test dataset with query examples.
-

References

[1] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400, 2017.

[2] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc., 2013.

[3] 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.

[4] Michelle Guo, Edward Chou, De-An Huang, Shuran Song, Serena Yeung, and Li Fei-Fei. Neural graph matching networks for fewshot 3d action recognition. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[5] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350:1332–1338, 12 2015.

[6] Ruslan Salakhutdinov, Joshua Tenenbaum, and Antonio Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver, editors, *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pages 195–206, Bellevue, Washington, USA, 02 Jul 2012. PMLR.

[7] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018.

[8] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, pages 935–943, USA, 2013. Curran Associates Inc.