Learning With Less Labels - Literature Review

Peyman Bateni

June 6, 2019

1 Introduction

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

2 Zero-Shot Learning Through Cross-Modal Transfer [17]

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 [4]

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 [8]

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 tp 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 [12]

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 [11]

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[14]

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 [3]

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 Attend, Infer, Repeat: Fast Scene Understanding with Generative Models [2]

9.1 Summary

- They propose an iterative inferential recurrent neural network for classifying, detecting and reconstructing objects in an image. The underlying generative model consists of describing N objects in the image via a Binomial distribution, where for each, latent scene descriptors z_what and z_where are then sampled from z $p_\theta^z(.|n)$. The work uses an amortized variational approximation to the true posterior by learning a distribution $q_\phi(z, n|x)$ parameterized by ϕ that minimizes the KL-divergence. To address the issues of trans-dimensionality, the number of objects being a random variable itself, and symmetry, the interchangeable order in which the objects as assigned to their respective numbers, the inference is formulated as a recurrent neural network. To simplify its sequential reasoning, n is parameterized as variable length latent vector z_pres such that for all iterations where objects are seen, the variable is 1, and when an object is not seen, the variable is 0, thus terminating the inference process. The inference network is able to perform at unprecedented speed compared to past work.
- The model parameters θ and inference parameters ϕ are learned by maximizing the marginal likelihood of an image under the model: $log_{p_\theta(x)} \geq L(\theta, \phi) = E_{q_\phi}[log(\frac{p_\phi(x,z,n)}{q_\phi(z,n|x)})]$ with L named the negative free energy and p and q describing the joint model probability and the marginal latent/count probability with respect to given example. It must be noted that as a result, the model is trained in a completely supervised manner. The gradient with respect to L is computed using Monte Carle estimates. In the case of $\frac{\delta}{\delta\phi}L$, they parameterize the dependence of each object iteration over all previously identified objects, formalized as $z^i|z^{1:i-1}$, using a recurrent function $R_\phi(.)$ implemented as a neural network such that $(w^i, h^i) = R_\phi(x, h^{i-1})$ with hidden variables h. The gradient with respect to ϕ can thus be

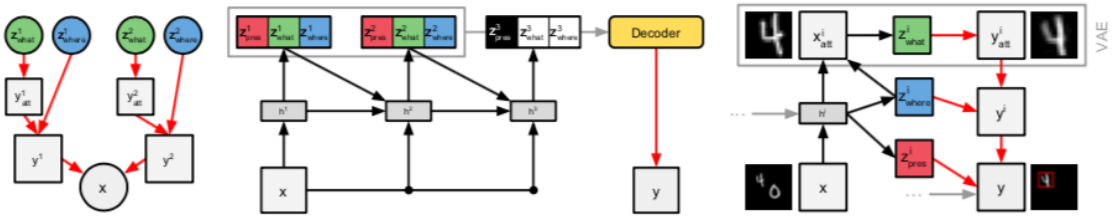

Figure 2: **AIR in practice:** *Left:* The assumed generative model. *Middle:* AIR inference for this model. The contents of the grey box are input to the decoder. *Right:* Interaction between the inference and generation networks at every time-step. In our experiments the relationship between $\mathbf{x}^i_{\text{att}}$ and $\mathbf{y}^i_{\text{att}}$ is modeled by a VAE, however any generative model of patches could be used (even, e.g., DRAW).

computed using the chain rule with $\frac{\delta}{\delta w^i} L$ where depending on whether z_i is continuous or discrete, the gradient is either backpropagated using path-wise estimator or obtained by a Monte Carlo estimate using the likelihood ratio estimator. The inference and generative networks are trained end-to-end.

- As shown in the attached figure, the inferred pose at each iteration is used to produce the attention mask x_{att} which is then used to produce the inferred code z^i_{code} and the reconstruction of the attention window y^i_{att}. The rest of the details in the figure are self-explanatory. Additionally, a modified version of AIR, namely Difference-AIR is also used on the MNIST dataset where at every time step i a partial reconstruction x^i of the data x is reconstructed which is set as the mean of the distribution $p^x_\theta(x|z^1, .., z^(i-1))$ with an error canvas defined as $\Delta x^i = x^i - x$ and the inference equation modified to be $(w^i, h^i) = R_\phi(\Delta x^i, h^i - 1)$.
- As for experiments, on the MNIST dataset, the model identified the number of digits correctly due to the opposing pressures of wanting to explain the scene and the cost that arises from instantiating an object under the prior. Also, it lactates the digits accurately and the recurrent network learns a suitable scanning policy to assure different time-steps account for different digits. Finally, the networks learns to terminate after the correct number of objects present.
- AIR was shown to be unable to effectively extrapolate to more number of objects in test images than previously trained on, although with respect to interpolating to number of object between number of objects seen at train time, the results "somewhat" improved. DAIR, by contrast, generalizes well on both cases. AIR also achieves (exluding DAIR on subsequent experiments) high accuracy on higher-level summations/ordering tasks even with scarce data, indicating the power of its disentangled, structured representation.
- On 3D scenes and the "vision as inverse graphics" task, itknow that posterior inference is either extremely expensive or prone to getting stuck in local minima and the probabilistic renderers are not capable of providing gradients. These issues are address by using finite-differencing to obtain gradients through the renderer, using the score function estimator to get gradients with respect to discrete variables and using AIR inference to handle correlated posteriors and variable-length representations. Experimental results show accuracy and reliability in inferring the identify and pose of the object. Amortization of the loss is shown to not only reduce the cost of inference, but also overcome the pitfalls of indepedent gradient optimization.

9.2 Strengths

- The inference network is very quick and the entire architecture is trained end-to-end in an unsupervised manner. Additionally, the iterative recurrent inference architecture allows for effective identification of the objects, resolving problems of symmetry and trans-dimensionality very effectively.

- The generative network is able to scale to multiple tasks whether it's classification or detection with acceptable accuracy in bounding boxes.

9.3 Weaknesses

- Model is very heavily dependent on having seen examples before for identifying the object. Demos show that the model would fail to generalize to larger object cases or those that heavily deviate from the examples seen before.
- Paper doesn't provide a study of DAIR on other experiments (more study might be needed as to why and whether there was/is potential in doing so).

9.4 Potential

- Extending the model to potentially identify zero/few-shot cases would be very useful in multi-object zero//few-shot classification/detection. Exploring DAIR for unbounded number of object detection maybe useful.
- Depending on the problem definition for learning with less labels, the unsupervised training methodology can prove helpful.

10 Generating Classification Weights with GNN Denoising Autoencoders for Few-Shot Learning [7]

10.1 Summary

- The paper proposes a meta-model consisting of a Denoising Autoencoder Network (DAE) that takes as input classification weights corrupted with Gaussian noise and learns to reconstruct the target-discriminative classification weights. The model, when evaluating, consists of two parts, a visual feature extractor producing visual feature vector z for each image and a classifier with parameters w that are used to generate scores, $[S-1,..,S_N] = [Z^T W_1,..,Z^N W_N$. The model uses cosine similarity in feature classification with the features and the classification weights both being $L2$ regularized.
- When faced with the case of few shot learning, the proposed model uses a wDAE-GNN to use the few training examples as well as the classification weights for seen classes, to produce an optimally learned set of classification weights for the newly observed class. The architecture works by first initializing the class weights same as before for seen classes and for the novel class, it averages the visual features from each of the few-shot examples as the initial vector for the classification weights of the new class. Once noise has been applied, the DAE can be trained to estimate the gradient of the energy function of the density of $p(w)$ such that $\frac{\delta log p(w)}{\delta w} \approx \frac{1}{\sigma^2}.(r(w)-w)$ with r defining the autoencoder and σ the amount of Gaussian noise.
- Given data and trained $r(w)$, the architecture performs gradient ascent: $ww + \epsilon.(r(w)-w)$ with ascent size ϵ to reach a mode of the conditional distribution $p(w|D_t r)$. The model then uses a GNN as the DAE, where the nodes are initialized using $w_i nitials$s and the edge weights are set by applying the softmax operation over the consine similarity scores of neighboring nodes, such that the outgoing edge weights sum to one.
- The aggregate function, $h^{(l)}_{N(i)} = \sum_{j \in N(i)} a_{ij}.q^{(l)}(h^{(l)}_i, h^{(l)}_i)$ such that q is a non-linear parametric function that given two vectors, it forward-passes them through the same network before adding the

output together, applying BatchNorm + Dropout + LeakyReLU, thus making the vector input positions interchangeable. The update function performs $h_i^(l-1) = [h_i^{(l)}; u^{(l)}([h_i^{(l)} l h_{N(i)}^{(l)}])$ with u being a non-linear parametric function, forwarding the input vector into an FC + BatchNorm + Dropout + LeakyReLU + L2-normalization. The update function on the last prediction GNN layer changes to $\delta w_i, , o_i = u^{L-1}([h_i^{(L-1)}; h_{N(i)}^{(L-1)}])$ such that u^{L-1} consists of FC + L2-normalization for w/Sigmoid unit for o. The update, therefore, is performed as $\hat{w}_i = w_i + o_i \odot \delta w_i$ with o_i acting as a d-dimensional getting vector.

- The model is trained with episodic training using a squared reconstruction loss of the target weights and a classification loss of the classification examples: $\frac{1}{N}\sum_{i=1}^{N} ||\hat{w} - w_i^*||^2 + \frac{1}{M}\sum_{m=1}^{M} loss(X_m, y_m|\hat{w})$ where loss is the cross entropy loss.
- During experiments, only one refinement step is applied with step size determined through cross-validation at each K-shot setting with respect to each of ImageNet-FS, MiniImageNet and tiered-MiniImageNet. The depth of the GNN was set to two and an alternative wDAE-MLP architecture was also experimented that operates without the aggregate function in the graph. Both models surpass previous work in performance on all but one K on ImageNet-FS and all Ks on MiniImageNet with wDAE-GNN being the superior of the two.
- Additionally, it's shown that ablations to the model defined as removing the Gaussian noise, traing with noisy tragets as input, and dropping either one of the losses, all reduce performance although without the reconsturction loss, the model comes very close to the best performance. It's also noted that the GNN architecture that takes into account inter-class dependencies results in a 0.4% consistent improvement.

10.2 Strengths

- The model outperforms matching and prototypical approaches to few-shot learning with reasonable margin while maintaining its ability to classify non-few-shot cases as well. In fact, this is one of the few works where few-shot accuracy reaches near overall accuracy.
- The use of DAE is regularizing weights to avoid overfitting is very novel and a potential addition that could benefit other models. Also, the weight initialization approach may be of interest in using cosine similarity such that visual features with few examples could act as initial weights for the few-shot class.
- Without having to map to a semantic space, it's arguable that in a deployed setting, this model would perform much quicker than space-mapping/KNN approaches.

10.3 Weaknesses

- The model is solely focused on few-shot learning with no or at least not easy way of extending to zero-shot learning.

10.4 Potential

- The updating strategy can be implemented or extended as part of our project as it seems to have proven effective. Additionally, there are areas to consider image context when initializing weights or defining inter-class edges.

11 Few-Shot Learning with Graph Neural Networks [15]

11.1 Summary

- The paper proposes the use of GNNs with inspired message passing across the nodes through the use of generalized learned edge features using $MLP_{\bar{\theta}}(abs(x_i^{(k)} - x_j^{(k)}))$ as a distance metric with the symmetry property fulfilled by construction and the identity property $f_{\bar{\theta}}(a, a) = 0$ learned through training. The vertices in the graph are initialized by concatenating the visual feature vectors from each example to the one-hot encoded class vector, with the unseen example substituting a uniform vector across the K-classes as oppose to the one-hot. It must be noted that in this context qK examples from K-class with q-shot examples each are used to predict ONE unlabeled example.
- To classify the unseen example, a softmax mapping is used from the node features of the unlabelled vertex to the K-simplex trained using the cross-entropy loss. Semi-supervised learning is also considered, where in addition to the few-shot setups, unlabeled "not to be classified" examples are present. Training setting is the same here, with only initialization for added unlabeled vertices following the same uniform structure as the unlabeled example in the few-shot case.
- Additionally, active learning is also considered where the model has the option to request one label from the unlabeled sub-group used in semi-supervised learning. In this case, an attention mask is used to request the label after the first layer of the GNN, where a function $f(x_i^{(1)}$ is parameterized by a two layer neural network, mapping each vertex to a scalar. Then softmax is applied to the scalars where at test time the maximum parameter is only kept. During training, however, the one value is randomly sampled on its multinomial probability. This sampled attention is multiplied by the label vector which is obtained and is the added to the current representation for that node, exactly where the uniform lable distribution was concatenated.
- The model was evaluated on Omniglot and Mini-ImageNet. On the former, the performs in the same statisticla confidence interval as SOTA on 1-shot, 5-shot 5-Way classification and 1-shot 20-Way classification while narrowly being worse than TCML on 5-shot 20-Way. This is despite the significantly less number of parameters at 300K compared to TCML's 5M. On Mini-ImageNET, it's shown than the edge metric itself combined with KNN but an increase of 2% can be achieved by leveraging the GNN. TCML outperforms the model on this dataset, although the 11M parameters compared to 400k in the GNN may be a cause of the case.
- For semi-supervised learning, it's shown that on Onmniglot, semi-supervised learning with 20% of labels matches supervised learning with 40% of labels, although accuracies are already in high nineties region. On mini-ImageNet, there is also a boost but not as noticeable.
- On active learning, random selection of an unlabelled node is shown to not perform any better than the semi-supervised case while the attention masked suggestion is shown to improve 3.4% on the Mini-ImageNet.

11.2 Strengths

- The model is able to match SOTA performance while using over an order of magnitude fewer parameters. It's also shown to be effective in leveraging semi-supervised learning and active learning in improving performance.
- The model is able to effectively use a graphical approach in performing few-shot learning.

11.3 Weaknesses

- Model doesn't seem to be extendable to zero-shot learning although easy bi-modal mappings should assist in doing so. Additionally, model needs to re-iterate for one new example, making it more time-consuming than say meta-learning models.
- Also, the proposed architecture doesn't leverage context or multi-label classification at test time. Another paper that leverage added edge labelling has shown that the model could stand to gain from considering more text examples at test time without assuming independence. I must admit that the semi-supervised case seems to come close to this idea, although without assigning labels to semi-supervised unlabeled sub-group.
- Also, one-shot encoding of the class seems rather rudimentary in the case of where it's concatenated to the visual feature vectors.

11.4 Potential

- Better class embeddings can improve performance. Additionally, use of context elements or otherwise, bi-model class definitions may aid in the task at hand.
- There is also potential in generalizing the scope of active learning to include more adoptive questions, especially with respect to visual context should we pursue it.

12 LSDA: Large Scale Detection through Domain Adaptation [9]

12.1 Summary

- The work proposes a domain adaptation approach in transforming classification networks to perform object detection, extending it to domains where annotated detection data isn't available. The model cosists of a pre-trained AlexNet where layers 1-7 are modified and fine-tuned for classification on the new categories. Afterwards, for category specific adaption, transformations are learned to map the classifier model parameters into the detector model parameters.
- Denoting set A to consist of classification examples, and set B to house classes with detection annotations, the category specific output layer (f_c8) comprises of f_cA, f_cB, δB and $f_c - BG$. For categories with annotated boundaries, this is performed by fixing f_{cB}, learning a new layer, zero initialized, δB with loss equivalent to f_{cB}, and adding together the outputs of f_{cB} and δB such that $W_i^d = W_i^c + \delta B_i$. For categories without annotated detection boxes, the KNN average of δBs are used such that $W_i^d = W_i^c + \frac{1}{k}\sum_{i=1}^{k} \delta B_{N_B(j,i)}$. Nearest neighbour is defined as the mnimal Euclidean distance between l2-normalized f_c8 parameters of the categories. Note that background is an added category since region proposal algorithms such as selective search were used to obtain background proposals as well.
- At test time, K+1 scores are extracted per region proposal, with per category score defined as $score_i - score_background$.
- The approach is evaluated on the LSVRC2013 dataset, where all woo categories are used for classification with only the first 100 having detection annotations. It's shown that LSDA results in a 50% improvement on mAP over the baseline of using the classification network with region proposals without any adaptation. The ablation study showed the most important step of the model to be adapting the feature representation, while the least important was adapting the category specific parameter.

- In terms of sources of error, while the baseline struggles with background and localization errors, the majory of false positive in the LSDA come from confusion with other categories, marking its ability localize objects in the regions and omit backgrounds.
- A version employing 7.6K categories with annotations from only 200 with non-max suppression that performs reasonably well through qualitative examples. Another version using faster region proposal and on a spatial pyramid pooling network is also provided that reduces detection time to half a second from four seconds.

12.2 Strengths

- Clearly, the model shows superior adaptation performance to that of using the simple baseline which allows good adaptation to categories where the detection domain doesn't include annotations. Additionally, the model is robust to localization errors and can be boosted to perform faster in a deployed setting.

12.3 Weaknesses

- Performance is still somewhat skewed towards already seen categories, often mistaking categories for each other. Additionally, the model relies on KNN for adapting to non-annotated cases, which comes with downfalls (best achieved at K = 10).

12.4 Potential

- Use of better metric in estimating δB may improve performance. As paper notes, single labels for images where multiple labels are present are weak label, therefore, potential in using context is present. Additionally, LSDA could be used to extend our zero/few-shot classification work to potentially also perform object detection.

13 Prototypical Networks for Few-Shot Learning [16]

13.1 Summary

- The proposed prototypical networks use an embedding function to create embeddings of the classes in a space, where the "prototype" of a class is the mean of the support examples belonging to that class as defined by a Bregman distance, notably Euclidean distance used by the authors. When given a query example, softmax probabilities are produced based on the distance between the query example and the prototypes of each class. The few-shot model is trained with episodic training, using an adopted version of the cross entropy loss, modified to minimize distance between the example and the respective class prototype while maximizing distance from other prototypes.
- Interpretive analysis shows prototypical networks to be equivalent to performing mixture density estimation on the support set with an exponential family density as long as a Bregman distance is used since then, the cluster assignment inference would be equivalent to query class prediction. Additionally, it's proven mathematical that when the model uses Euclidean distance, the model is equivalent to a linear model, where it's hypothesized that the needed non-linearity comes from the embedding function. It's also shown that the model matches matching networks in behaviour when only one

support example is provided per example, due to the respective mean of the prototypical network and the attention on the matching network being the same.

- The paper found that euclidean distance provides better performance that cosine similarity used by other works, which they argue is due to the mixture density estimation hypothesis not holding. Additionally, they found that for prototypical network training for more query classes ("way") results in better performance while maximum accuracy is achieved when the number of support examples per class ("shot") is the same in both settings.
- In the case of zero-shot embedding the model relies on class meta-data for each class using a separate meta-data embedding function such that $c_k = g_\alpha(v_k)$. Given that the query example and meta-data descriptions come from different domains, it was found helpful to constraint the prototype embedding to have unit length, while the query embedding remains unconstrained.
- Compared to matching networks and neural statistician, the model achieves new SOTA with a 1-3% margin at 98.8, 99.7, 96.0, 98.9 accuracies at 1/5shot 5-way classification and 1/5-shot 20-way classification on the Omniglot dataset. Similarly, compared to matching networks and the meta-learner LSTM, the model achieves new SOTA with a 6-8% margin at 49.4% and 68.20% accuracies at 1/5-shot 5 way classification on mini-ImageNET.
- On CUB-200 zero-shot classification, the model performs better than other embedding approaches at 54.6% accuracy using 312-dimensional continuous attribute vectors from the dataset.

13.2 Strengths

- The work showcases the benefits of using Bregman distances over say cosine similarity wildly adopted by others. Additionally, model is much simpler than matching network in methodology, although given other approaches involving GNNs, it's somewhat out of date.
- Model defines episodic training as a good approach to zero/few-shot training, and also demonstrated bi-domain mapping to the same space and its potential.

13.3 Weaknesses

- Model relies on attribute vectors, a luxury that is often not present with most datasets and is therefore, limited in its use depending on the dataset at hand.
- Model relies on basic distance metrics as oppose to more "learned" way of parameterization of the classification scheme.

13.4 Potential

- Use of Bregman distances as oppose to cosine similarity can provide better results although they must be experimented with as there isn't a strong fundamental argument for one being better than the other, although this paper provides an empirical one.

14 Transductive Multi-Label Zero-Shot Learning [6]

14.1 Summary

- Paper proposes a first-time solution to multi-label zero-shot learning. A similar structure involving a semantic space as defined by word2vec embeddings is used via a 9-layer (5-layer CNN, rest FC) multi-output deep network (named Mul-DR) with least square regressors to map raw pixel information into the semantic space. Pre-training on ImageNet is used for the convolutional layers.
- In addition to converting the multi-label task to independent binary classification, what is regarded as an extension to "Direct Attribute Prediction (DAP)", the paper proposes Direct Multi-Label Zero-Shot Prediction (DMP) and Transduction Multi-Label Zerho-Shot Prediction (TraMP).
- DMP consists of leveraging the compositionality property of the semantic space to synthesize the representations for every possible multi-label annotation in the space, resulting in 2^{m_T}. The problem is then solved through nearest neighbour classification based on the cosine distance to each of the synthesized prototypes.
- TraMP goes further by exploiting the manifold structure of the all Mul-DR outputs for the test examples. The architecture uses the known prototypes from DMP to perform transductive label propagation to the inferred semantic representations. This involved creating a power set L_p of all train/test labels totalling $2^{m_S+m_T}$ prototypes. Once there, a graphical kNN representation is formed using normalized edges weights as defined by $\frac{1}{Z_i}exp(-\frac{||distance||^2}{2\sigma^2})$ if edge is present by the kNN neighbours and zero otherwise. The learned edge weigths are then used to define $A = I - w$, partitioning the matrix into four blocks of $A_L L$, $A_L u$, $A_u L$ and $A_u u$ (left to right, top to bottom) where the label set of test instances is inferred by the following closed form solution: $\hat{L}_T = -A_{uu}^{-1} A_{uL} L_p$.
- To make Mul-DR generalize better to target domain, a simple semi-supervised learning strategy is used to perform one step of self-training to refine each prototype using the kNN predicted semantic representations $\hat{Y}_T$
- Work is evaluated on Natural Scence and IAPRTC-12 datasets. Mul-DR is shown to outperfrom Support Vector Regression and Devise for mapping into the space while DMP is superior to exDAP with TraMP showing better overall performance. The self-training step is also effective in making making Mul-DR generalize better to the target data.

14.2 Strengths

- Work extends zero-shot learning to multi-label classification and also, exploits context data through combining label representations in the semantic space.
- Additionally, it leverages SSL to use the text set as whole to adopt labelling more effectively while also taking measures to generalise the mapping function better to the semantic space.

14.3 Weaknesses

- Work attempts to combine the visual features for multiple labels using the same mapping function. This means that varying number of classes present are still expected to map the same 100-dim representation which seems a bit naive of an approach.

14.4 Potential

- As the paper motivates, many combinations of labels in the power set are unlikely to happen together, thus pruning unlikely prototypes could be helpful in improving performance.
- Exploiting more explicitly representations of each object proposal in the image as oppose to combining all together may prove more effective. Also, the NLP semantic space is not updated at all, which although more reading is needed to fully confirm this, the word2vec embeddings and context mapping may not be as useful in the visual case.

15 Multi-Label Zero-Shot Learning with Structured Knowledge Graphs [10]

15.1 Summary

- The paper proposes use of pre-learned knowledge graphs with structured graph propagation to improve multi-label classification both for ML-classification and ML-ZSL. The model leverages a pretrained ResNet-152 to obtain visual feature vectors of the input images. A graphical representation of each possible label is formed where nodes are defined using the class labels and the edges are one of super-subordinate, positive or negative correlation between the labels. The relations are obtained using WordNet with super-subordinate relations defined by ISA relations present as part of the corpus and the other two, calculated using WUP similarity followed by thresholding the soft similarities.
- The node values are first initialized using learned input function $F_I(x, w_v)$ where x defines the feature vector from the aforementioned pre-trained ResNet and w_v describes the label representation for the node. The propagation mechanism through subsequent steps from each label node u to a connecting node v is governed by propagation weights a_{vu} which are produced from the relation function F_R^k with k denoting the relation in play. Using learned propagation weigth matrix A, the belief states at each node are updated as follows: $h_v^{(0)} = F_I(x, w_v)$, $u_v^(t) = tanh(A_v^T[h_1^{(t-1)T}...h_{|S|}^{(t-1)T}])$, $h_v^{(t)} = GRUCell(u_v^{(t)}, h_v^{(t-1)})$. At time step t, the actual class probabilities on each node is obtained using learned output function F_O through $p_v^{(t)} = F_O(h_v^{(t)})$.
- Matrix A consists of non-zero weights for adjacent nodes and zero otherwise with $a_{vu} = F_R^K(w_v, w_u)$ with w_v and w_u representing word vectors for each of the class labels. Note that since word embeddings are pre-set and F_R^k is learned, the matrix weights is always the same. Also, the same architecture is used for both ML-classification and ML-ZSL with the latter extending the graph to unseen classes and constraining propagation only from seen to unseen labels, and between unseen labels. Network is trained using BCE (binary cross-entropy) loss.
- The model is evaluated for ML-classification on both Microsoft Coco and NUS-WIDE datasets, and for ML-ZSL, it's evaluated on NUS-WIDE with 1000 noisy labels used as seen labels and the 81 dedicated ground-truth concepts as the unseen labels. The model is compared to WSABIE, WARP, and logistic regression baselines while also comparing to Fast0Tag on both ZSL/non-ZSL tasks. The model achieves comparably results on MS-COCO and NUS-WIDE with reasonable margins over baselines. Fast0Tag is shown to have greater recall due to simply always selecting top K labels where as the model proposed here is more flexible with respect to number of output labels.
- The model is shown to have 1-2% performance gain over baselines on ML-ZSL and generalized ML-ZSL (where model is evaluated on both seen/unseen classes with F1 scores of 30.6 and 24.2 respectively. The ablation study shows the importance of the propagation mechanism where significant gains in F1 is

achieved through even fewer iterations with the subsequent ones have less of an impact in ML-non-ZSL, ML-ZSL and even generalized ML-ZSL. classification.

15.2 Strengths

- Model is able to exploit context and pre-known knowledge structures in allowing better classification. It's also able to show good improvements over baselines using a graphical representation approach.
- Model is able to use inter-label information gathered more explicitly through edges which makes it more interpretable.

15.3 Weaknesses

- It's hard to see how the model could extend to detection, as with some other work as well, the model uses a more image-global ResNet embeddings as oppose to specific region proposals.
- Inter-label relations are only limited to three cases.

15.4 Potential

- Better knowledge structures could improve the performance through more explicit definition of co-occuring classes although it seems that this work does a pretty good job of extrapolating that from word2vec embeddings.
- Extention to region proposals and potential detection may prove insightful.

16 Multi-Label Zero-Shot Learning with Transfer-Aware Label Embedding Projection [19]

16.1 Summary

- The paper proposes a transfer-ware mapping of both word embeddings of the class labels and the visual feature embeddings from a VGG16 neural network to a low dimension space where the similarity matching score between the two can be obtained through the inner product of their project representations. The architecture in essence learns two mapping matrices W and U where the loss is defined as a calibrated separation of the max-margin ranking with distance to the 0 vector in the mapping space used as a baseline where unrelated classes are aimed to be less similar and related classes are trained to be more similar than.
- The loss is further regularized through the Frobenius norm of the weights. Additionally, transfer-aware objective $H(u)$ is added as a regulization term to maximize similarity between seen and unseen classes while minimizing the similarities between the unseen classes. $H(u)$ helps produce better inter-class relationship structure for cross-class knowledge transfer. Furthermore, auxiliary information is incorporated using a manifold regularization term to exploit separately two sources of auxilary information, one involving inter-label connection based on ISA hops on the WordNet embeddings and the other consisting of the co-occurrence statistics obtained by the hit rate of the visual labels on Flickr. This latter regularization term is directly added inside $H(u)$. For more mathematical details, visit the paper.

- The loss is reconstructed using the dual formulation of the loss function using the Standard Lagrangian of the max-margin learning problem allowing for a closed form solution to U while recovering W from the vudal variables Ψ. An iterative optimization algorithm is used, initializing weights as zero and then iteratively performing: 1. coordinate optimizing Ψ while keeping other rows fixed through a simple quadratic minimization problem and 2. fixing Ψ and minimizing U using the closed form solution.
- The model is evaluated on PASCAL VOC2007 and VOC2012 with four related baselines approaches involving conex combination of sematic embeddings (ConSE), latent embedding method (LatEm), DMP and Fast0Tag. Three versions of this work's model (TAEP) is also experimented with consisting of the original, TAEP-H using WordNet auxiliary hierarchical information and TAEP-C using Flickr Image Hit-counts. Overall, TAEP outperforms baselines on zero-shot multi-label tagging with MiAP of 57.42 with both extension performing better than the plain model and TAEP-C performing the best at 57.62. Similar trends are seen on the generalized multi-label zero-sho learning task. An ablation study reducing hyperparameters that regulate the impact of the regularizers show performance drops both with respect to the transfer-aware regularization unit and the additional auxiliary relations.

16.2 Strengths

- This paper actually comes closest to the idea I had in mind, although it certainly relies much more on direct closed form optimization while I was thinking more of multi-layer deep networks used for mapping to the same space. The performance is clearly superior and the model is able to classify zero-shot cases effectively while exploiting transfer-aware regularization.
- The use of the calibrated max-margin ranking loss with the twist of the zero-vector comaprison is an interesting proposal which I wonder whether it's much responsible for the gain in performance.
- The model also exploits auxiliary information, which is novel and also, kind of very close to my idea especially in the case of visual hit counts.

16.3 Weaknesses

- Model relies on single layer weight matrices to map into the embedding space without leveraging non-linearities or deeper architectures which may prove useful.
- Model only leverages inter-class relations in form of a regularizer which I think may be limited in use.

16.4 Potential

- TAEP-C motivates and in fact heavily motivates use of class embeddings not defined by an NLP corpus but rather by contextual visual hit counts. This could even motivate an idea I had regarding forming a visual GloVe space.
- Deeper architectures could prove useful in increasing performance. Also, this work too relies on single global visual representations of multi-label images. Use of region proposals can be helpful especially in extending the work to zero-shot detection.

17 Multi-Label Zero-Shot Learning via Concept Embedding [13]

17.1 Summary

-

17.2 Strengths

-

17.3 Weaknesses

-

17.4 Potential

-

18 LaSO: Label-Set Operations networks for multi-label few-shot learning [1]

18.1 Summary

- The paper proposes LaSO, an architecture consisting of three feature-transform networks of intersection, union and subtraction which use visual multi-label feature embeddings as input to form the corresponding embeddings in the feature space that match the set of labels at the intersection, union or difference or the label sets of the input images. Note that this model uses one-hot encoding of labels combined together to form label sets. The classifier is trained using BCE multi-label classification loss.
- The classifier is trained using only the loss from direct image-based label set prediction where as LaSO operation networks loss is defined using BCE on the expected label set outputs with respect to each of the operations. Additionally, two sets of MSE losses are applied, one for enforcing symmetry across LaSO units and the other performing reconstruction loss involving LaSO operations to rebuild the original visual feature embeddings on the input images. LaSO networks are implemented using 3-4 block MLPs containing an FC layer followed by batch normalization, leaky RELU and dropout.
- The model is evaluated on MS-COCO for multi-object recognition and on CelebA for multi-attribute few-shot learning of facial attributes. Feature vector synthesis on MS-COCO on seen classes is shown to perform comparably in the cases of intersection and union in terms of mAP compared to that of the original non-manipulated feature vectors, with subtraction performing a bit over half as well. On unseen classes, performance is considerably worse but still much better than guessing thus demonstrating the capabilities of LaSO. A similar trend is seen with respect to kNN retrieval with k = 1, 3 and 5, although on unseen classes, intersection and union performance come much closer to performance on original non-manipulated cases.
- Analytic approximations to set operations vs. learned are performed matching union, intersection and subtraction with max, min and RELU operations on the input feature embeddings respectively. On

both datasets, learned parameters achieve higher mAP. Furthermore, data augmentation in few-shot learning is evaluated where episodic training using learned (and analytic for slightly worse performance) union augmentation or the few-shot training set, outperforming benchmarks including no augmentation, basic, mixUP and lastly learned/analytic intersection augmentation which performed second best after union on both 1/5-shot.

- The CelebA experiment interestingly showed all operations to perform with worse mAP tha original feature vector case. However, this time union and subtraction came closest with intersection performing far worse. This observation is attributed to the fact that the intersection network is training in an unbalanced and biased manner towards negative attributes while mAP is more affected by accurately predicting the positive labels.

18.2 Strengths

- The work provides an interesting approach to data augmentation for the task of few-shot learning. The approach of manipulating feature embeddings depending on the associated label sets shows to be effective that zero augmentation.
- The use of the reconstruction loss is an interesting added touch.

18.3 Weaknesses

- Overall, it's relatively unclear how the LaSO operations attempt to exploit multi-label context outside of shared embeddings. Also, the architecture opts to remove bounding boxes and uses global features, a trend that seems to be happening with other papers too. I wonder if this is actually grounded in the presumption that global visual embeddings perform better.

18.4 Potential

- Use of operational label-set transformation can prove to be interesting add-ons to the work on few-shot learning. Although, this seems to be one of those cases where extension to zero-shot would require major changes and an entirely different architecture.

19 Graph R-CNN for Scene Graph Generation [18]

19.1 Summary

-

19.2 Strengths

-

19.3 Weaknesses

-

19.4 Potential

-

20 Recent Advances in Zero-Shot Recognition [5]

20.1 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] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogério Schmidt Feris, Raja Giryes, and Alexander M. Bronstein. Laso: Label-set operations networks for multi-label few-shot learning. *CoRR*, abs/1902.09811, 2019.

[2] S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, infer, repeat: Fast scene understanding with generative models. *CoRR*, abs/1603.08575, 2016.

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

[4] 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.

[5] 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.

[6] Yanwei Fu, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-label zero-shot learning. *CoRR*, abs/1503.07790, 2015.

[7] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning, 05 2019.

[8] 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.

[9] Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Jeff Donahue, Ross B. Girshick, Trevor Darrell, and Kate Saenko. LSDA: large scale detection through adaptation. *CoRR*, abs/1407.5035, 2014.

[10] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. *CoRR*, abs/1711.06526, 2017.

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

[12] 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.

[13] Ubai Sandouk and Ke Chen. Multi-label zero-shot learning via concept embedding. *CoRR*, abs/1606.00282, 2016.

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

[15] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[16] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4077–4087. Curran Associates, Inc., 2017.

[17] 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.

[18] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. *CoRR*, abs/1808.00191, 2018.

[19] Meng Ye and Yuhong Guo. Multi-label zero-shot learning with transfer-aware label embedding projection. *CoRR*, abs/1808.02474, 2018.