ReviewGen: Comparison of Model Generated Reviews

Eric Hansen, Cynthia Chen, Rohan Doshi

Paper Title: M-Walk: Learning to Walk over Graphs using Monte Carlo Tree Search

Paper ID: 6299

Abstract: Learning to walk over a graph towards a target node for a given query and a source node is an important problem in applications such as knowledge base completion (KBC). It can be formulated as a reinforcement learning (RL) problem with a known state transition model. To overcome the challenge of sparse rewards, we develop a graph-walking agent called M-Walk, which consists of a deep recurrent neural network (RNN) and Monte Carlo Tree Search (MCTS). The RNN encodes the state (i.e., history of the walked path) and maps it separately to a policy and Q-values. In order to effectively train the agent from sparse rewards, we combine MCTS with the neural policy to generate trajectories yielding more positive rewards. From these trajectories, the network is improved in an off-policy manner using Q-learning, which modifies the RNN policy via parameter sharing. Our proposed RL algorithm repeatedly applies this policy-improvement step to learn the model. At test time, MCTS is combined with the neural policy to predict the target node. Experimental results on several graph-walking benchmarks show that M-Walk is able to learn better policies than other RL-based methods, which are mainly based on policy gradients. M-Walk also outperforms traditional KBC baselines.

Real Reviews: This work tackles the problem of knowledge base completion (KBC) by walking through a knowledge graph. The graph walk is treated as an MDP with a known state transition model. MCTS is used to plan a trajectory through the graph, and these trajectories are used to train a value function with Q-learning. The Q-function is used to form a policy prior for the MCTS. A KBC-specific RNN structure is also proposed. Results on a range of KBC benchmarks show small but consistent gains over several baselines, including on-policy RL approaches and non-MCTS ablations. Quality MCTS is a strong planning algorithm for delayed-reward problems and seems a good fit for KBC, so this is an at the least an interesting application paper. The authors also extend previous work on combining MCTS with learned value functions and policy priors by leveraging the MCTS trajectories to learn the optimal value function of the KBC graph-traversal policy with Q-learning. This seems like a reasonable approach to improve the sample efficiency of learning the prior, compared to just updating the policy prior to match the MCTS visitation frequencies at the root node. The experiments seem to present competitive baselines and show some promise on a number of tasks. However, I would appreciate some more discussion or clarity of the exact roles of pi, Q, and V. Firstly, the background section 2 does not make the import distinction between an arbitrary policy and the optimal policy (and their corresponding value functions). In particular, note that Q-learning does not estimate Q^pi but Q* for the MDP. In M-Walk, Q is trained off-policy, and so learns about the optimal state-action value function Q* for graph-traversal. However, V is trained on-policy (wrt the MCTS tree policy, not pi_theta), but only on terminal MCTS states (not terminal MDP states, as I understand it). Could the authors elaborate on this choice (as opposed to for example using V = max_a (Q))? Meanwhile, pi is simply a soft transformation of Q into a probability distribution (the

authors note that simply using a softmax works similarly to their final implementation). Traditionally, Q-learning with Boltzmann exploration as the behaviour policy would anneal its entropy (ie temperature) to converge to a good solution. I am concerned that a fixed temperature would restrict the ability of the algorithm to develop strong confidence in its prior policy, even if it were justified by the data. In this particular setting with strongly bounded rewards and sufficient MCTS samples to correct for even an uncertain prior, it seems to work fine, but the exact mechanism by which the Q-learning transfers to the MCTS tree policy deserves some further discussion. Clarity The paper is quite clearly written on the whole. Some of the details of the architecture and intermediate state representation are a little hard to follow, and some design decisions regarding the neural architecture (s3.1) are a bit opaque. It should also be emphasized that the policy pi_theta is neither used to traverse the MDP (except indirectly as a prior) nor is its value learned by any function, which is a departure from the expectations set by section 2. The combination of MCTS with learned policy and value priors has been explored in a number of domains, but this application to KBC appears novel and includes a number of algorithmic and architectural innovations. Significance This work addresses a significant and well-studied problem, as well as a significant class of solutions (MCTS with learned value/policy priors) with methods that may transfer to other problem settings. The experimental results are promising. However, more care with the formulation and explanation of the approach wrt the exact roles of the learned value functions and consequent policy prior would be very useful for solidifying the theoretical foundations of the approach.

In this paper, authors develop a Reinforcement Learning agent - M-Walk - that, given a query, learns to walk over a graph towards a desired target (answer) node. The proposed architecture maps the state into a vector representation, and maps it into a Q-network (for the policy) and a value network. The proposed RL algorithm alternates between MCTS trajectory generation steps, and policy improvement steps, for iteratively refining the policy. Question - other than WN18RR, FB15k-237 is also a great link prediction benchmark. Why didn't you consider it in your experiments? Is it because the number of relation types in FB15k-237 is much larger compared to WN18RR, causing a significant increase in the size of the action space? It would be great to have a brief discussion on this aspect in the paper, rather than omitting it completely.

Generated Review (Reviews Only):

This paper proposes a graph-walking agent that combines a deep neural network with Monte Carlo Tree Search (MCTS) and a neural policy. The proposed agent is able to learn a policy that maximizes the terminal rewards by using MCTS and the RNN. The paper is well-written and well-motivated. The proposed method is interesting and the experimental results are convincing. However, I have two major concerns. First, the paper is not well-suited for ICLR. The main contribution of this paper is the use of Monte Carlo tree search to train the M-Walk agent. This is an important problem in graph learning and the proposed method does not address the problem well enough. Second, the proposed algorithm is not sufficiently novel. It is not clear to me what is the difference between this paper and the previous papers on this topic. In particular, I do not understand why the authors do not compare their method against the state-of-the-art methods on the three-glass-joint task.

Generated Review (Reviews Only + NER):

This paper proposes a graph-walking agent for the knowledge base completion (KBC) problem. To overcome the challenge of sparse rewards, the authors use a deep recurrent neural network (RNN) and Monte Carlo Tree Search (MCTS) to generate trajectories that yield positive rewards. From these trajectories, the network is improved in an off-policy manner using Q-learning, which modifies the RNN policy via parameter sharing. Experimental results show that M-Walk is able to learn better policies than other RL-based methods, which are mainly based on policy gradients. The paper is well-written and easy to follow. However, I have a few concerns about the paper :1. The authors do not provide a clear explanation of how the MDP transition model is known and deterministic, which makes it difficult to understand the motivation behind the proposed method.2. The proposed method is based on a deep neural network and a Monte Carlo tree search algorithm, which is not well-known in the literature.3. The experimental results on several graph- walking benchmarks are presented in Section 4.4. It is not clear to me why the authors did not compare their method to other methods that are not based on RL. For example, in the case of WN18RR, it seems that the authors don't compare the method to the original WN, which was created from the source and source-based WN. This makes it hard to judge whether the method is better than the original method.5. The results are not very convincing. It would be interesting to see how the method compares to other RL methods in terms of performance.6. In particular, it would be nice to see the performance on the three-glass-puzzle benchmark.

Generated Review (Reviews + NER + Citations + Metadata):

This paper proposes a method for learning to walk over a graph with a known state transition model from sparse rewards. The proposed method is based on a combination of a deep recurrent neural network (RNN) and Monte Carlo Tree Tree Search (MCTS). The RNN encodes the state (i.e.the state of the graph) and a policy and Q-values, which are then combined with the neural policy to generate trajectories that obtain significantly more positive rewards than using the RNN policy alone. The objective is to learn a policy that maximizes the terminal rewards of the target node with high probability. The paper is well-written and the paper is easy to follow. I would recommend the paper to all of the authors in the literature. I think the paper would have been better if the paper had been written a bit more clearly. For example, in Section 2.2, the authors proposed to combine MCTs and MCTRNN to generate a policy-improvement step to train the agent from sparse reward. In Section 3.3, they also proposed to use the M-Walk agent to learn better policies than other RL-based methods like A↞-search [11]. In section 3.4, they propose to use M-walk as a model for learning a policy for learning the policy from off-policy data. In the section 4.1, they show that M- walk outperforms the KBC baselines in terms of performance on several graph-walking benchmarks. However, the paper does not address the problem of how to learn the policy gradients from off policy data. For instance, it is not clear to me that the authors are using off policy to learn policies. For this reason, I would like to see a more thorough evaluation of the performance of the paper.2. I am not convinced that the paper can be improved by adding a policy gradient to the model. For the paper, I think it would be better to use a policy gradient instead of an off policy.3.4. I do not think that the proposed method would be much better than the proposed policy gradient.5. I also don't think that this is a good idea. It would be nice to see the results of the experiments on the paper in Section 4.6. I have a few questions :1) The paper doesn't seem to address the issue of sparse rewards, which is a problem that needs to be addressed by the authors.2) It is unclear to me why the authors didn't address this issue in the first place. It seems like the paper should be addressed in the second and third and fourth sections.3) The authors should also address the question of how the authors

learned the policy gradient in Section 5.4) I would have liked to see some of the results on the first and second sections. It is interesting to see how well the authors addressed the problem in Section 6.1. The first section is well written, but the second section is a bit confusing. The second section seems to be a bit short on details, and the third section is not very well written. It does not seem to be clear what the main contribution of this paper is, and it is hard to understand what the authors were trying to achieve. The third section does not provide any explanation of the motivation for this paper. The authors do not mention any details.

1. Summary
    a. This paper proposes a novel approach to the problem of knowledge base completion (KBC) by developing a graph-walking agent called M-Walk that consists of a deep recurrent neural network (RNN) and Monte Carlo Tree Search (MCTS). The RNN encodes the history of the walked path and maps it separately to a policy and Q-value, which is used to generate trajectories that are used to learn a policy for the target node. In order to overcome the challenge of sparse rewards, the authors propose to use a Q-learning approach to improve the policy of the RNN.
    b.
2. Clarity
    a. The paper is well written and easy to follow. The problem setting of the paper is not clear to me. It would be better if the authors could clarify this distinction in the writing. I would also suggest to the authors to clarify the gap between theory and empirical results. I am not familiar with the literature enough to tell whether the proposed method is the cleanest way to characterize the current state-of-the-art in the graph-walking problem setting, and whether it can be scaled up or scaled back to a more practical level.The paper is generally easy to read and understand.
3. Soundness
    a. The paper is well written and contains a thorough experimental design. The experiments are convincing and show that M-Walk is able to learn faster and better policies than other graph-walking methods.The paper is clearly written and the results are impressive. However, the paper does not address the problem of learning to walk on a graph with sparse rewards, and does not provide any theoretical results to support this claim. It would be better if the authors could provide some insights on why this is a problem that needs to be addressed.
4. Substance
    a. The experimental results on graph-walking benchmarks show that M-Walk outperforms the state-of-the-art in terms of performance and robustness to perturbation.The paper is well written and easy to follow. It would be better if the authors could provide more details and results on more complex problems such as hyper-parameters and hyperparameter sharing to better understand what is really important for the performance of the method. The paper is also well written, and the experiments are well-designed, but could be better.
5. Substance V2: The paper is well written and the authors have also shown that the paper can be improved with the help of a more thorough analysis of the proposed method.The paper also shows that the proposed approach would be better if the paper could be used to show that it is not

as well written as it seems, and that it would be nice to see that the method can be used in the future.The proposed method would be interesting to see if it is possible to use this method in a more general way.

Paper Title: Spatiotemporal Residual Networks for Video Action Recognition

Paper ID: 837

Abstract: Two-stream Convolutional Networks (ConvNets) have shown strong performance for human action recognition in videos. Recently, Residual Networks (ResNets) have arisen as a new technique to train extremely deep architectures. In this paper, we introduce spatiotemporal ResNets as a combination of these two approaches. Our novel architecture generalizes ResNets for the spatiotemporal domain by introducing residual connections in two ways. First, we inject residual connections between the appearance and motion pathways of a two-stream architecture to allow spatiotemporal interaction between the two streams. Second, we transform pretrained image ConvNets into spatiotemporal networks by equipping them with learnable convolutional filters that are initialized as temporal residual connections and operate on adjacent feature maps in time. This approach slowly increases the spatiotemporal receptive field as the depth of the model increases and naturally integrates image ConvNet design principles. The whole model is trained end-to-end to allow hierarchical learning of complex spatiotemporal features. We evaluate our novel spatiotemporal ResNet using two widely used action recognition benchmarks where it exceeds the previous state-of-the-art.

Real Review: In this paper, authors use ResNets architecture to learn human actions in a two-stream ConvNet architecture. They changes two-stream network in various ways: 1) Using ResNets architecture in the two-stream base-line method 2) Introducing residual connections between motion and appearance streams and also increasing the effective temporal receptive field by adding learnable residual connections over time. The authors also realized by practice that the performance of the method could be boosted using sub-batch normalization instead of batch normalization, frame rate jittering, and applying temporal max-pooling just after pool5 layer.ResNets and two-stream networks are two main ingredients of this paper. While, authors successfully have identified and used the strong aspects of these methods and applied it to their application, the paper lacks strong novelty. Since there are multiple practical ideas to boost the performance of the method, it requires more experiments to clarify the effect of each part separately. For instance, it is not clear how much of the final performance is due to the sub-mini batch method and if the method could still beat the state-of- the-art without using it. Beside this, without any intuition, using sub-mini batch method is not reliable. I am worried that it only shows some improvements in the few datasets that are tried in the paper but not be a good idea in general. It needs intuition to show it effectiveness, and also more experiments should be done to see whether or not it is a dataset specific, application specific. or a general result. While adding skip connection between two streams seems more arbitrary, adding residual connections across time method seems more interesting to me. Unfortunately, the details provided in the later method is not explored enough and explained in details. For example, is the skip connections removed in the batch boundaries in forward and backward passes? If yes how it might effect the predictions near the boundaries of the batches? After rebuttal: I believe the authors' responses in terms of novelty are convincing hence updating my scores to 3. However, I still believe it

would be good to show the specific ablation study for the different design components in the paper (sub-batch normalization, skip connection and temporal residual).

The paper extends the two-stream convolutional network architecture of [19] by introducing residual connections between the appearance and motion streams. In addition it also shows how pre-trained image ConvNets can be transformed into spatio-temporal networks by transforming the spatial dimensionality mapping filters in the residual paths to temporal filters. The approach is evaluated on UCF101 and HMDB datasets and comparisons show the proposed approach obtains state-of-art results on both datasets.Positives - The paper shows how to effectively use residual connections in the spatio-termporal setting needed for action recognition. - Results on two important action recognition datasets validates the proposed approach. Negatives - The overall novelty is only incremental. - Section 3.2 is probably one of the key contributions of the paper and could be expanded and explained better.

The paper presents a novel architecture that 1) combines residual networks with two-stream convolutional networks, and 2) injects connections from the motion stream to the appearance stream, to be able to capture spatio-temporal features. The paper shows experiments in both of the main action recognition datasets, achieving state-of-the-art accuracy in both. Overall I think the paper is great: good idea, careful experimentation, great results and clearly written. Although the basic components of the architecture are pre-existing, I think that the high performance and careful experimentation and description make it a very useful contribution. I only miss some experiments to visualize what kinds of spatio-temporal features are being learned. Since large temporal windows are important (278-280) I would add a relevant reference: Long-term Temporal Convolutions for Action Recognition, Gul Varol, Ivan Laptev, Cordelia Schmid Small typos: (L 187) "resp." (L 255) "UCF0101"

The paper proposes an extension of the two-stream approach to action recognition. It builds on top of a number of recent advances in designing Convolutional networks, most notably 1x1 dimensionality reduction and residual connections. Authors propose two modifications of the two-stream network - residual connection from motion to appearance stream and residual temporal filtering. They use these techniques to advance the state-of-the-art on two datasets, most notably on HMDB51.The paper is well written and structured. It is mostly easy to read and follow the argument. Its novelty does not lie within inventing new techniques, but rather in creatively using the known ones. I'm not sure about the architecture's potential impact - one needs to access it on a bigger data, but it does look promising. A few points to improve: First, Table 3 shows results of the baseline ResNet50 and the proposed model with both modifications enabled. It would be good to see the contribution of motion-appearance connections and temporal filtering separately. Second, the section about temporal filtering is somewhat unclear. I've reread it multiple times and still am not sure whether I understand it correctly. It would be really helpful to expand it a bit. Third, it would be really nice to see the result of this architecture on the Sports-1m dataset. However, I do understand that this dataset is huge and very computationally expensive to handle. Future work could address this, maybe?

This paper proposed a deep CNN architecture for video action recognition by extending residual networks. The proposed architecture combines both spatial stream and temporal stream into a single network, and extends to model longer period of temporal information. This work achieves state-of-the-art performance on two public action recognition datasets. This paper presents a framework that improves

two stream networks for video action recognition by extending residual network to combine information from two streams into one single network. It significantly improves over previous state-of-the-art on two popular video action recognition benchmark. The downside of this paper is the limited novelty. There are previous work tried to combine two streams into a single network [1,2], and the temporal convolution is not new either [3]. Although the way to combine two streams is slightly different from previous work, the proposed approach is still pretty straightforward. Since this paper mainly consists of incremental ideas on previous work, more details could be given on how those ideas help in solving the tasks to gain more insights. For example, separate ablation studies would be useful to understand how the techniques in section 3.1 and 3.2 help. The temporal convolution described in section 3.2 is just replacing spatial convolution with temporal convolution, which is not really related to residual connections. The term âœresidualâ could be dropped in the title of section 3.2 and line 166. Some details in the paper are unclear: - Are the HMDB51 networks initialized from UCF101 networks as in previous work? - The proposed network sees 705 frames (line 193). How do you handle videos with less than 705 frames? - Line 233: âœBatch-normalization uses a smaller batch size of 128/32.â Does that mean the batch size is 128/32 = 4 as discussed above, or the batch size is 128 or 32? This is a bit confusing. - Is the loss of ST-ResNet defined as the sum of loss of both streams? Typo: Line 255: UCF0101 -> UCF101 Overall, this paper is clearly written, and produces very good results on video action recognition datasets. [1] C. Feichtenhofer, A. Pinz, A. Zisserman. Convolutional Two-Stream Network Fusion for Video Action Recognition. https://arxiv.org/abs/1604.06573 [2] E. Park, X. Han, T. L. Berg, A. C. Berg. Combining Multiple Sources of Knowledge in Deep CNNs for Action Recognition. WACV 2016 [3] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, J. Dambre, Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. http://arxiv.org/abs/1506.01911

Generated Review (By Sections):

1. Summary: This paper presents a method for training a spatiotemporal ResNet based on two-stream convolutional networks. The authors propose to add temporal residual connections between the appearance and motion pathways of a two stream architecture to allow spatio-temporal interaction between the two streams. The proposed method is evaluated on two tasks : (i) action recognition and (ii) object-detection.The paper is well written and the experiments show promising results.The main contribution of this paper is to propose a method that combines two existing methods : (a) Residual networks (ResNets), and (b) temporal residual networks (TNNs). The ResNet is a combination of these two approaches.
2. Clarity: The paper is well-written and easy to follow, apart from a couple of places in the derivations (see my questions). I would also suggest the authors to reduce such abuse of notations (e.g., \theta^ {... }, \timescale_t,... } in the final work, because these terms are not used widely in the literature. Also, it would be nice if the authors could provide more details about the training and evaluation of the model.
3. Substance: The experiments are conducted on two standard action recognition benchmarks and the results seem to align with our intuition. However, the authors do not provide any analysis of the sensitivity of their approach to temporal residual connections, which could potentially cause confusion and extra experimentation for practitioners. The authors should provide more analysis of their method to better understand what is really important for the performance of the method.

The paper is well written and easy to follow, and the experiments are well-designed, but the results are not convincing enough to support the claims.

4. Soundness: The paper is well written and the results are interesting and mathematically sound. However, the main weakness of the paper is that it spends a lot of time on the big picture, rather than on showing experimental results in relation to the spatiotemporal domain, which is the main contribution of this paper. It is therefore difficult to reproduce the results solely based on the description. It would be better if the authors could provide some intuition on why this is a good idea and how we might want to approach this problem in the future.

Generated Review By Section (V2):

5. **Soundness:** the the, is to of. and.- a in I not The are this that for paper it) : be results on (with method well
6. **Substance:** The paper is well written and the authors have also shown that the paper can be improved with the help of a more thorough analysis of the proposed method.The paper also shows that the proposed approach would be better if the paper could be used to show that it is not as well written as it seems, and that it would be nice to see that the method can be used in the future.The proposed method would be interesting to see if it is possible to use this method in a more general way.
7. **Clarity:** the, is the the the is the, the is the is is is, is, the, is,,, the is, to to the to is to, to.. the. is.,. and and the and is and, and

Generated Reviews By Review Only:

1. **Review Only:** This paper proposes to generalize the two-stream ConvNets architecture to the spatiotemporal domain by introducing residual connections between the appearance and motion pathways of the two streams. The paper is well-written and well-motivated. The proposed method is evaluated on two standard action recognition benchmarks, where it greatly outperforms the state-of-the-art. I have the following comments: 1. In the experiments, it is not clear to me what is the difference between the performance of the proposed method and the original ResNet architecture. 2. In Figure 1, the results are not particularly strong. 3. In Fig. 2, the authors do not compare with the state of the art. 4. In Table 1, I do not understand why the Spatio-ResNet performs so much better than the baseline. 5. The experiments are not convincing.
2. **All:** This paper introduces a new method for training in the spatiotemporal domain for two-stream Convolutional Networks (Residual Networks) by introducing residual connections in two ways. First, the authors introduce residual connections between the appearance and motion pathways of a two- stream architecture to allow spatioteporal interaction between the two streams. Second, they use residual connections to allow the network to train in the spatial residual connections of the two stream architecture. Finally, they train the network in the temporal residual connections before the second residual unit at each spatial resolution of the network. The whole model is trained end-to-end to allow hierarchical learning of complex features. The paper is well-written and the paper is easy to follow. I would recommend the paper to all of my readers. I think it would be nice to see the results of the paper in a paper, and I would like to see how well the proposed method works in practice. It would be interesting to me to see if the proposed approach can be applied to other tasks. For example, the paper would be good to compare it to the current state-of-the-art in the context of human action recognition, but I would also like to know if it could be used in other domains. I am not sure if this is the case, but the paper seems to be well written

and the authors seem to have a good idea of the potential benefits of this approach. I have a few comments about the paper. First of all, it is not clear that the paper does not only apply residual connections, but not the residual connections. This is not the case with respect to the previous work, which used residual connections as a basis for training.Second, I would have liked to see a more thoroughness in the application of residual connections.The paper is very well written, but it would have been nice to be able to show the results in a more comprehensive way. For instance, I think the paper could have been better if the authors could have shown that the results were better than the existing work. Third, I don't think that the proposed model is well written. The authors should also consider using residual connections with residual connections for other tasks, as well. The proposed model would be better to show that the model is able to learn from the current work of the authors. I also think that this is a very interesting idea. However, I am still not convinced that the authors have done a thorough analysis of the results, and it is hard to see why this is important. The main contribution of this paper is that it is a novel approach to training in a spatial residual connection. It is also interesting to see what the authors are doing in terms of the performance of the model, and how it is compared to other approaches. For the most part, this paper doesn't seem to be a very novel idea, but there is a lot of work to be done in this area.

3. R**eview+NER:** This paper proposes a new Residual ResNet architecture for two-stream Convolutional Networks (ResNets). The idea is to combine the two existing ResNet architectures, which are based on the two-layer convolutional networks (ResNet and ResNet 2), by adding temporal residual connections between the appearance and motion pathways of the two streams to allow spatiotemporal interaction between the two networks. The proposed ResNet is evaluated on two standard action recognition benchmarks where it greatly boosts the state-of-the-art.The paper is well written and easy to follow. It is well-written and well-motivated. The paper is clearly written and well organized. The results are interesting. The novelty of the paper is that it is a combination of the ResNet approach and the ResNet architecture. However, there are a few things that I do not like about the paper :1. The authors do not provide a clear explanation of how the proposed Resnet is different from the existing Resnet architecture.2. They do not explain how the Resnet network is trained end-to-end to allow hierarchical learning of complex features. 3. They also do not discuss how the model is trained to learn spatially aggregating filters. 4. They did not explain why they do not use hidden fc, dropout, or max-pooling. 5. I don't understand why they didn't compare to the previous ResNet model. I would like to see the performance of the new ResNet to the state of the art in terms of accuracy and accuracy.

Paper Title: **Large Scale Multi-Domain Multi-Task Learning with MultiModel**

ID: 13063

Abstract: Deep learning yields great results across many fields, from speech recognition, image classification, to translation. But for each problem, getting a deep model to work well involves research into the architecture and a long period of tuning. We present a single model that yields good results on a number of problems spanning multiple domains. In particular, this single model is trained concurrently on ImageNet, multiple translation tasks, image captioning (COCO dataset), a speech recognition corpus, and an English parsing task. Our model architecture incorporates building blocks from multiple domains. It contains convolutional layers, an attention mechanism, and sparsely-gated layers. Each of these computational blocks is crucial for a subset of the tasks we train on. Interestingly, even if a block is not crucial for a task, we observe that adding it never hurts performance and in most cases improves it on all

tasks. We also show that tasks with less data benefit largely from joint training with other tasks, while performance on large tasks degrades only slightly if at all.

Review: The paper presents a multi-task, multi-domain model based on deep neural networks. The proposed model is able to take inputs from various domains (image, text, speech) and solves multiple tasks, such as image captioning, machine translation or speech recognition. The proposed model is composed of several features learning blocks (one for each input type) and of an encoder and an auto-regressive decoder, which are domain-agnostic. The model is evaluated on 8 different tasks and is compared with a model trained separately on each task, showing improvements on each task.

The paper is well written and easy to follow.

The contributions of the paper are novel and significant. The approach of having one model able to perform well on completely different tasks and type of input is very interesting and inspiring. The experiments clearly show the viability of the approach and give interesting insights. This is surely an important step towards more general deep learning models.

Comments:

* In the introduction where the 8 databases are presented, the tasks should also be explained clearly, as several domains are involved and the reader might not be familiar with the task linked to each database. Moreover, some databases could be used for different tasks, such as WSJ or ImageNet.

* The training procedure of the model is not explained in the paper. What is the cost function and what is the strategy to train on multiple tasks ? The paper should at least outline the strategy.

* The experiments are sufficient to demonstrate the viability of the approach, but the experimental setup is not clear. Specifically, there is an issue about the speech recognition part of the experiment. It is not clear what the task exactly is: continuous speech recognition, isolated word recognition ? The metrics used in Table 1 are also not clear, they should be explained in the text. Also, if the task is continuous speech recognition, the WER (word error rate) metric should be used. Information about the detailed setup is also lacking, specifically which test and development sets are used (the WSJ corpus has several sets).

* Using raw waveforms as audio modality is very interesting, but this approach is not standard for speech recognition, some references should be provided, such as:
P. Golik, Z. Tuske, R. Schluter, H. Ney, Convolutional Neural Networks for Acoustic Modeling of Raw Time Signal in LVCSR, in: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), 2015, pp. 26–30.
D. Palaz, M. Magimai Doss and R. Collobert, (2015, April). Convolutional neural networks-based continuous speech recognition using raw speech signal. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on (pp. 4295-4299). IEEE.
T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals. Learning the Speech Front-end With Raw Waveform CLDNNs. Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), 2015.

Revised Review:
The main idea of the paper is very interesting and the work presented is impressive. However, I tend to agree with Reviewer2, as a more comprehensive analysis should be presented to show that the network is not simply multiplexing tasks. The experiments are interesting, except for the WSJ speech task, which

is almost meaningless. Indeed, it is not clear what the network has learned given the metrics presented, as the WER on WSJ should be around 5% for speech recognition.
I thus suggest to either drop the speech experiment, or the modify the network to do continuous speech recognition. A simpler speech task such as Keyword Spotting could also be investigated.

Generated Reviews:

1. Review Only:

This paper proposes a multi-task multi-modal model for multi-language translation tasks. The model consists of a convolutional layer, an attention mechanism, and a sparsely-gated layers. It is trained on ImageNet, multiple translation tasks, a speech recognition corpus, a COCO dataset, and an English parsing task.

The paper is well-written and well-motivated. The proposed model achieves good results on all the tasks, and the proposed architecture is simple and effective. The paper is clearly written and easy to follow. The experimental results are convincing.

However, I have a few questions:

1. Why did you choose to use the spectral domain for the multi-ilingual tasks? It is not clear to me what is the advantage of using it compared to the fixed-size representation of the input tensor?

2. What is the performance of the model on large-scale image classification with Oord den Oord et al. and WaveNet? It seems to me that the main advantage of the proposed model is that it does not suffer from the limitations of RNNs. It would be interesting to see how this model performs on large scale image classification tasks.

2. Review+NER:

This paper presents a model architecture for multi-task multi-modal language processing. The model is trained on a number of tasks, including translation, image captioning, speech recognition, and parsing tasks. The architecture is based on a single model, which is trained concurrently on ImageNet, multiple translation tasks, and a speech recognition task. The proposed model architecture is a combination of several different building blocks from multiple domains. It contains convolutional layers, an attention mechanism, and sparsely-gated layers. Each of these computational blocks is crucial for a subset of the tasks we train on.The paper is well-written and easy to follow. The paper is easy to read. The writing is clear and easy-to-follow. However, I have a few questions about the paper :1. The main contribution of the paper is the introduction of the multi-model architecture. It is not clear to me what the main contribution is.2. How is the model designed? How is it trained? What is the architecture of the model? 3. What are the contributions of the proposed model? 4. How are the proposed architecture compared to the previous work? 5. How does the model compare to the existing multi-tasking models? 6. How do you compare the performance of the new model with the existing models? 7. What do you think is the difference between your model and the other models? 8. Is there a reason why your model is better than the previous models?

3. All:

This paper presents a model architecture for multi-task multi-modal learning that works well on a number of tasks, including speech recognition, image captioning, speech recognition corpus, and an English parsing task. The model architecture incorporates building blocks from multiple domains. It contains convolutional layers, an attention mechanism, and sparsely-gated layers. It is trained on ImageNet, multiple translation tasks, and a speech-recognition corpus. The paper is well-written, and the paper is easy to follow. The main contribution of this paper is to show that the model works on a variety of tasks. The proposed model architecture is well written and easy to understand.The paper is very well written, but the main contribution is not a novel one. I would have liked to see more discussion on the topic. I am not sure what the main contributions of the paper are, and I think it would be nice to know more about them. For example, what is the contribution of the model architecture to the performance of the proposed model on the proposed tasks? How does the model compare to other models? What is the performance on large tasks? I would like to see the results on large task. I think the paper would have been better if the authors could provide a more thorough explanation of the results. For instance, the authors should also include a discussion of why the model is not quite right for the large task, and why it is not good for the small task. Also, the paper should include more discussion of how the model performs on small task tasks. It would be interesting to see how well on these tasks. In addition of the attention mechanism is crucial for the performance, but I think this is a bit of a novel idea. I also think that the paper could have been improved by adding the attention function. However, I don't think that this would have improved performance on small tasks.

Section Specific Reviews:

1. Soundness: Overall, the paper is well written and contains a thorough experimental design, but lacks of high-level discussion on why it works so well on some problems and how it could do better on others. The paper is clearly written, and the experiments seem to be well-performed. However, there is no discussion of why the proposed model is better than the ones already proposed in the paper, or how it relates to competing models in terms of memory usage and memory complexity.The paper is generally well-written and contains thorough experiments, but lacking of discussion about why it is better to combine multiple models into a single model.
2. Substance: The experiments are well-designed, and the model is well-evaluated. It would be good to see a more in-depth analysis of the performance of the model on more challenging tasks, e.g.word-level translation, image captioning, and speech-related tasks, as well as a more detailed analysis of its performance on these tasks. However, I would have liked to see an analysis of how performance is influenced by the number of layers, and how attentional efficiency is influenced.
3. Clarity: The paper is well written and easy to follow, with a good balance between mathematical notation and intuitive explanations of the main ideas. I would like to see more results on the COCO dataset, but other than that, I think it is an interesting model and it is well described and well-motivated, and I think the paper will benefit from a more thorough theoretical analysis of the proposed model and its architecture, and a more detailed explanation of the key ideas and experiments. I think that the paper would benefit from more careful writing and better organization.
4. Summary: This paper presents a multi-task multi-modal model, which is trained on multiple tasks (translation, speech recognition, COCO, multiple languages, multiple tasks on a single dataset, and a speech recognition corpus on another dataset). It contains convolutional layers, an attention mechanism, and sparsely-gated layers. Each of these computational blocks is crucial for a subset of the tasks we train on, and the authors show that adding it never hurts performance and in most cases improves it on all tasks.

Paper Title: **Discrete InfoMax Codes for Meta-Learning**

ID: 18864

Abstract: This paper analyzes how generalization works in meta-learning. Our core contribution is an information-theoretic generalization bound for meta-learning, which identifies the expressivity of the task-specific learner as the key factor that makes generalization to new datasets difficult. Taking inspiration from our bound, we present Discrete InfoMax Codes (DIMCO), a novel meta-learning model that trains a stochastic encoder to output discrete codes. Experiments show that DIMCO requires less memory and less time for similar performance to previous metric learning methods and that our method generalizes particularly well in a challenging small-data setting.

Review: This paper presents DIMCO, a meta-learner that is trained by maximizing mutual information between a discrete data representation and class labels across tasks. DIMCO is inspired by an information theoretic lower bound on the generalisation gap for meta-learning, which the authors argue identifies overfitting in the task learner as the bottleneck.

This work proposes to constrain a learner to output discrete codes that are learned to capture the mutual information with class labels. While the idea of using discrete codes is interesting, its presentation in the manuscript is not well motivated and at times hard to follow. This makes it challenging to evaluate the novelty, validity, and generality of the proposed approach. Meanwhile, the empirical evaluation is somewhat lacking. Thus, I do not believe this work is ready for publication in its current form.

Detailed comments

My main concern is with respect to the primary contribution of this paper, a generalisation bound on meta-learning. The bound appears to be on a multi-task loss without task adaptation, and thus the claims made with respect to the theorem seem somewhat over-reaching. I also believe the VC-dimensionality of the encoder is missing in Eq. 4? If so, this changes the interpretation since the length of the code and the expressivity of the encoder are interrelated. Further, I would welcome a deeper analysis of the theorem and its implications. The current interpretation states that the number of tasks is independent of the size of each task, hence given many tasks, using minimal representations is an effective approach to meta-generalisation. Yet minimal representations is a well-known idea and has features in several works that use mutual information as a regularizer, most notably works on the Information Bottleneck.

Another reservation I have is the use of mutual information between encoder representations and class labels as a loss function (Eq. 1). It lacks context and a proper motivation, especially since the analysis of [1] shows that the loss function in Eq. 1 is the cross-entropy objective. The authors make a similar analysis in Appendix A and argue that Eq. 1 differs in that cross-entropy is an approximation because it adds a parametrized linear layer on top of \tilde{X}. Thus, in the absence of that layer they collapse to the same objective. As DIMCO itself directly extract class label predictions from \tilde{X}, I fail to see a difference between the loss in Eq. 1 and a cross-entropy objective.

The main motivation behind their loss objective is that it does not require a support / query set. This does not seem to be a feature of the mutual information objective itself, but rather a choice made by the authors. I would have liked a deeper discussion of this seeing as the authors make it a central tenet of DIMCO. Prior works use a support set as a principled means of doing meta-learning: meta-training explicitly takes into account that at test time, the learner will be given a small support set from which to learn how to query points. As far as I understand, DIMCO does not take this into account during

meta-training. At meta-test time however, DIMCO does use a support set to map query points (Eqs. 10 and 11). Why should we break protocols between meta training and testing? Are there any downsides to doing so?

Empirically, I find the CUB experiment compelling but would welcome some ablations. What are the trade-offs between p and d? Can DIMCO outperform N-pair when number of bits are unconstrained?

miniImagenet is a standard benchmark in few-shot learning, but I am unable to find a results table - could the authors please provide results on the standard setup so that the method can be compared against known baselines? Further, would the results currently presented hold in a N-way-5-shot setup?

As for the constrained version of miniImagenet that the authors propose, I am not convinced this is an interesting protocol. In general, the miniImagenet task distribution is created by N-way permutations of the classes in the meta-set (e.g. meta-training tasks are combinations of the 64 classes in the meta-training set). By keeping the number of classes constant but reducing the number of images per class, this protocol is not reducing samples per task: a task is always defined as 5/20-way-1-shot (Fig. 4). Instead, the effect should be that tasks are in (greater) violation of the task i.i.d. assumption. Thus, I question whether this setup demonstrates the trade-offs the authors present in Theorem 1 and whether the results can be interpreted in light of it.

Finally, that both experiments are image-based raises questions as to the generality of the method. The paper could be considerably strengthened by evaluating DIMCO on a non-image task, or if not discuss the method's limitations.

The idea of discrete codes for few-shot classification is interesting and sufficiently novel, I am likely to increase my score if my concerns are addressed and the experimental section is strengthened.

Further questions and comments:

- I am unable to parse Eq. 11 - what does the notation \prod_i p_{\tilde{x}_i, i} mean?
- It is unnecessarily hard to follow the proof of theorem 1. It would help the reader if you restated relevant definitions, such as Eq. 1, since the difference with Eq. 23 is very subtle. It would also be helpful to explain how the summand in Eq. 25 differs from either, and Eq. 26 could be expanded or briefly explained after the derivation.
- Because DIMCO is trained with backpropagation, the interpretation of Eq. 5 as d independent events seems invalid. How does it affect the method if they are not independent?
- Overloading X and Y as both random variables and mini-batch samples creates unnecessary confusion. I believe the objective in Eq. 1 is approximated, not calculated exactly? For instance, the mutual information in Eq. 8 is with respect to a mini-batch, so should it not be \hat{I}?
- p^j_{ik} in Eq. 9 is undefined.
- Eq. 9 is interpreted as an exact entropy, however it appears to be a mini-batch approximation to the true entropy?

References

[1] Achille and Soatto. Emergence of Invariance and Disentanglement in Deep Representations. JMLR. 2018.

Generated Review:

1. Review Only

This paper studies the problem of meta-learning with task-specific learners and proposes a new generalization bound for this problem. The paper is well-written and well-motivated.

The paper is clearly written and easy to follow. The main contribution of the paper is a generalisation bound for task specific learners. This is an interesting and novel contribution.

However, I have a few questions:

1. What is the expressivity of the task specific learner?

2. How does the bound compare to other generalization bounds in terms of the number of tasks and the amount of parameters?
3. How do the authors compare the generalization of DIMCO with respect to MAML et al. (2017) and Zintgraf et al (2018)?

4. In the experiments, it is not clear to me what is the computational complexity of the proposed method. For example, in Table 1, how much memory is required for training the model?

2. Review+NER

Theorem 1 shows that restricting inner-loop expressivity reduces the generalization gap in meta-learning. The paper proposes a new method called Discrete InfoMax Codes (DIMCO) that uses a stochastic encoder to output discrete codes.The paper is well-written and easy to follow. The main contribution is an information-theoretic generalization bound for meta- learning, which identifies the expressivity of the task-specific learner as the key factor that makes generalization to new datasets difficult.Experiments show that DIMco requires less memory and less time for similar performance to previous metric learning methods and that our method generalizes particularly well in a challenging small-data setting. I think the paper has a lot of potential to be improved. However, I have a few concerns with the paper :1) The paper is written in a way that makes it difficult to understand how it works, and it is not clear to me that the results are representative of the state-of-the-art.2) It is unclear to me how the experiments are comparable to the results reported in Mishra et al. (2017). 3) The experiments are not very convincing, and I am not sure how well the results compare to the performance of MAMFinn et al. (2018). 4) It would be interesting to see the results of the experiments on a Meta-Dataset, which is a very challenging dataset. 5) The results on Meta-Data are not particularly impressive. It would have been nice to see a comparison of the performance on a classifier trained with and without the same classifier. 6). The experiments on the other side of the paper would be nice to compare with the results on MAMLinn's method. 7) The experimental results are not that impressive either. I would have liked to see some comparison on a different set of datasets. 8) I would also like to see more discussion on the impact of the experimental results on generalization.

3. All

This paper proposes an information-theoretic generalization bound for meta-learning. The paper is well written and well-written. The main contribution of the paper is that it presents a novel approach to generalize to new datasets. The experiments show that DIMCO generalizes well to novel datasets, especially in a challenging small-data setting.The paper also proposes a novel generalization limit for Meta-Theoretic Generalization bound. This bound is based on the idea that the generalization of a task-specific learner is the key factor that makes generalization to a new dataset difficult. The authors show that the proposed bound is effective in generalizing to the new dataset. They also show that their

method generalizes particularly well in a small data setting. They show that this is due to the fact that the model generalizes very well to a small set of datasets. However, the paper does not provide a clear explanation of how this generalizes to larger datasets. It is not clear to me that the authors have a clear understanding of how they achieve this generalization. For example, they do not explain how they do it in the paper, but I am not sure if they have a good idea of what is going to happen to the data. Also, the authors do not provide any explanation of why they do this. I think that the paper should be updated to include a more thorough discussion of the main contributions of this paper. I would also like to see some of the other work on generalization in this area. For instance, I would like to know more about the contributions of the proposed method.

Generated Reviews By Section

1. Soundness: The paper is clearly written and the results are interesting and mathematically sound. However, I am not convinced by the theoretical results because the experiments are not very thorough and the overall results are quite significant. The paper should also include some discussion on why the proposed method is better than simply using the discriminator to learn discrete codes and why it is better to use a stochastic encoder to generalize the model to larger datasets.Overall, the paper is well-written and the experiments seem to be well-performed.
2. Substance: The experiments are well-designed, and the results seem to align with our intuition. However, I would have liked to see a more in-depth analysis of the performance of DIMCO on larger datasets to be fully persuaded. It would be great if the authors could provide such results on more challenging datasets, e.g., ImageNet or ImageNet 2, which are very similar to the ones used in the experiments. I would like to see more analysis on this topic, as it is very important to me.
3. Summary: This paper provides a generalization bound for meta-learning that identifies the expressivity of the task-specific learner as the key factor that makes generalization to new datasets difficult. Taking inspiration from this bound, the authors propose DIMCO, a novel meta learning model that trains a stochastic encoder to output discrete codes. Experiments show that the compact compactness of the model allows it to generalize to new tasks more easily, and it generalizes well in a small-data setting.
4. Clarity: The paper is well-written and easy to follow, and the experimental results show significant improvements over state-of-the-art meta-learning methods.The paper is generally well-organized and well-structured, and I enjoyed reading it. I would suggest to the authors to reconsider the structure of the paper, as some sections are hard to follow without first reading through the introduction, and some sections require rewriting to make it easier to follow. I think the paper would benefit from a more clear structure that allows readers to delve into details at different levels, and provide additional insight as to what is going on in the experiments.

Paper Title: **Generative Imputation and Stochastic Prediction**

ID: 22112

Abstract: In many machine learning applications, we are faced with incomplete datasets. In the literature, missing data imputation techniques have been mostly concerned with filling missing values. However, the existence of missing values is synonymous with uncertainties not only over the distribution of missing values but also over target class assignments that require careful consideration. In this paper, we propose a simple and effective method for imputing missing features and estimating the distribution of target assignments given incomplete data. In order to make imputations, we train a simple and effective generator network to generate imputations that a discriminator network is tasked to distinguish. Following this, a predictor network is trained using the imputed samples from the generator network to capture the

classification uncertainties and make predictions accordingly. The proposed method is evaluated on CIFAR-10 image dataset as well as three real-world tabular classification datasets, under different missingness rates and structures. Our experimental results show the effectiveness of the proposed method in generating imputations as well as providing estimates for the class uncertainties in a classification task when faced with missing values.

Review: This paper proposes a method to impute missing features using a generative model and train a predictive model on top of imputed dataset to improve classification results. They first train a GAN model where the generator outputs an imputed representation of the input and discriminator is trained to predict if an individual features (such as a pixel) is imputed or not. Given the generator and incomplete sample, they train a predictor using the output of the generator, imputed sample, as input. Their main contribution is using a MC averaging to compute the prediction by repetitively sampling from the noise variable, z, and generating different imputations from generator. They show that the proposed model improves upon the previous SOTA on final classification performance.

Overall the paper is clearly written. But I do feel it is a bit incremental over the GAIN approach. The overall GAN architecture is very similar to GAIN's and although stochastic prediction shows clear improvements it is a bit straightforward. However, I think the uncertainty of the imputations and its effect on the final prediction is interesting. I suggest the authors to extend this part with more detailed analysis.

There are several parts that are confusing/missing in the paper:

- In GAIN, they use a hint vector as an input to the discriminator. They show that without the hint vector, there is no unique solution (this is shown without the MSE loss). The authors do not use this vector in their approach (as in Figure 1) and it is not clear to me if it causes any instabilities or if multiple experiments yield similar results or if the stochastic prediction benefits from this.
- On what type of examples GI is more accurate than other models? Since stochastic prediction is the main difference from GAIN, is this related to the multi-modality of the noisy examples?
- Can you explain the difference between the results in Figure-7 and Table-2? Results between the two mismatch.
- I think the statement in the first paragraph in Section 4.4 that "MSE loss term would act as a denoising loss smoothing noisy missing pixels" could be misleading. MSE is used with mask in GAIN, hence it only applies to the observed features during training. Its effect on smoothing noisy missing pixels is not clear.

I think the paper would benefit if the authors could explain/show:
- Increasing the missing rate would also increase the possibility that the ground truth be a more multi-modal distribution. Especially in rectangular generation part where it can remove a complete object. Does stochastic averaging benefit more in this case?

Generated Review:

1. Review Only

This paper proposes a method for imputing missing features and estimating the distribution of target assignments given incomplete data. The method is evaluated on CIFAR-10 image dataset as well as three real-world tabular classification datasets, under different missingness rates and structures.

The paper is well-written and well-motivated. The proposed method is interesting and the experimental results are convincing. However, the paper is not well-suited for ICLR due to the following reasons:

1. The paper is mainly about imputing features and it is not clear whether the proposed method can be applied to other datasets. For example, in Table 1, the method is applied only to Cifar-10 and Table 2, it does not work for other datasets, e.g., UCI Landsimage dataset or Diabetes classification dataset.

2. There is no comparison with other imputation methods in the literature. For instance, in Section 3.2, it is stated that: “The authors propose a method based on deep latent variable models and importance sampling that offers a tighter likelihood bound compared to the standard VAE bound.” This is a strong statement, but the authors do not compare with other methods in terms of performance.

2. Review + NER

This paper proposes a method for imputing missing features and estimating the distribution of target assignments given incomplete data. The proposed method is evaluated on CIFAR-10 image dataset as well as three real-world tabular classification datasets, under different missingness rates and structures.The paper is well written and easy to follow. The method is well-written and well-motivated. The experimental results show the effectiveness of the proposed method in generating imputations and providing estimates for the class uncertainties in a classification task when faced with missing values. The paper is clearly written and well written. The experiments are well-constructed. The results are interesting and the paper is easy to read. However, there are a few things that I would like to see more of. For example, I would have liked to see the impact of different imputations for different features on the final hypothesis. For instance, for a specific test sample at hand, if a certain missing feature is a strong indicator of the target class, it would be interesting to see how the imputations are affected by the missing features. I would also like the authors to provide more details on the training process of the generator and predictor networks. I am not sure how to interpret the results of the experiments.

3. All

This paper proposes a simple and effective method for imputing missing data imputations and estimating the distribution of target assignments given incomplete data. The proposed method is evaluated on CIFAR-10 image dataset as well as three real-world tabular classification datasets under different missingness rates and structures.The paper is well-written and the experimental results show the effectiveness of the proposed method in generating imputations that generate imputations. The paper is written well and the paper is easy to follow. However, there are a few issues that need to be addressed. First of all, it is not clear why the paper does not address the problem of missing data in a more general way. Second, it does not seem to be clear how the method is applied in the real world. For example, the authors do not discuss the impact of the missing data on the accuracy of their method in terms of accuracy. Third, they don't discuss the effect of the method on the quality of the data. Finally, there is no comparison between the method and the method used in the paper. It is unclear why the authors didn't compare the method to the method proposed in this paper. For instance, what is the difference between missing data and missing data? How does the method compare with the method that is used for missing data, and how does it compare to the one proposed in the literature? The paper doesn't seem to address the issue of missing values, and it is hard to understand why the results are not better than the results of the previous paper. The authors do nn't address this issue.

Generated Reviews by Section

1. Soundness: The proposed method is evaluated on CIFAR-10 image dataset as well as three real-world tabular classification datasets, under different missingness rates and structures.The paper is well written and the results are interesting and mathematically sound. However, I am not

convinced by the empirical results as they seem to be quite strong in the sense that they do not match the theoretical results and do not seem to match the experimental results in the paper. The paper is clearly written, and the experiments seem to show the effectiveness of the proposed method.

2. Substance: The experiments are well-designed and the results seem to align with our intuition. However, this is not tested in the paper and the authors do not provide any analysis about the distribution of missing values or how this distribution is affected by the imputation.The paper is well-written and well-argued, and the experiments show that the proposed method works well on CIFAR-10 and UCI-10 as well as the other datasets. The paper is also well-evaluated, and it would be great if the authors could provide some additional analysis on the distributions that these distributions may be based on.
3. Summary: This paper proposes a simple and effective method for imputing missing features and estimating the distribution of target assignments given incomplete data. In order to make imputations, a generator network is trained using the imputed samples from the generator network to capture the classification uncertainties and make predictions accordingly. The proposed method is evaluated on the CIFAR-10 dataset and three real-world tabular classification datasets under different missingness rates and structures.The paper is well written and the experiments show that the proposed method can generate imputations and provide estimates for the class uncertainties in a classification task.
4. Clarity: The paper is well written and easy to follow, apart from a couple of places in the derivations (see my questions). However, I have some reservations about the presentation of the results : - It is not clear to me what the main idea of the proposed method is, and how it can be applied to a practical application.- It seems to me that the paper is missing a lot of detail about the training of the generator network, which makes it hard to understand what exactly is going on in the experiments.