Response Letter

March 29, 2024

1 On choice of regularizer

1.1 Comparing KL and ℓ_1 regularizer

We thank you for the question regarding the use of element-wise regularizer and questions on the use of l1 regularization (given as $\|p\|_1$) and the centered mean regularizer (given as $|mean(p) - \frac{s}{d}|$). In the following, we take $\frac{s}{d} = 0.05$.

When we minimize the objective with the centered mean regularizer and monitor the value of $mean(p)$, we see that starting from $p = 0.5$ the loss can decrease to $p = 0.05$ but not more, where it becomes stationary and does not change over 10 thousands of iterations (Figure- centered mean). During this phase, this penalty has the same gradient as the ℓ_1 norm regularizer. However, after $mean(p)$ reaches 0.05, the mean p becomes stationary and the loss seems to get stuck, although the penalty might behave differently than ℓ_1. So, the overall effect of ℓ_1 and the centered mean regularizer are similar.

Now, comparing ℓ_1 regularizer to the KL regularizer, we notice across various experiments that ℓ_1 regularization is less stable to the choice of the regularization strength (which we reported in the appendix). This is because ℓ_1 regularizer encourages sparser solutions (for centered mean, (p-0.05) is sparse) than KL regularizer. This enforces a bulk of p values to collapse on the same point. Hence the relative ranking gets lost due to this effect.

For example, when the logits corresponding to the three regularizers are plotted in Figure-4, the logits in KL regularization seems to be more well spread than the ℓ_1 and centered mean regularizer. When we look at the corresponding layerwise architecture in Figure-3, we see that the middle layers are severely pruned by ℓ_1 and centered mean regularization which may lead to layer collapse. We intentionally plot the sparsity percentage on the log scale to show the severity of this effect.

So, based on this empirical observation, we think that sparser solutions may not be ideal for bringing the data misfit loss down (since loss of rank importance may lead to layer collapse). Furthermore, enforcing sparsity shrinks the search space of gradient descent, so it may be more likely to get stuck in local minima. We will add a brief discussion in the paper.

1.2 On using pointwise regularization

Reviwer ySfU raised a very reasonable argument that using pointwise regularization can dampen the values of p to be very close to p_0. This is indeed true if we use a large regularization strength.

However, for KL regularization penalty, the ranking would remain preserved for very large range of moderate values of regularization coefficient λ, especially when compared to other pointwise regularization choices like $mean|p - p_0|$. This is because for KL regularization, the regularizer takes very low values around a large window $[p_0 - \epsilon, p_0 + \epsilon]$ (see Figure). This is untrue for linear pointwise regularizers such as ℓ_1.

We want to emphasize that pointwise regularization may allow the implementation of non-uniform prior p_0 across various weights in Unet. That's why we presented a more generic implementation, so if the user has some prior knowledge on what parameters are more important, they have the flexibility to modify the corresponding prior value.

Otherwise, we do not think pointwise regularization is essential and putting the mean of p in the KL regularization term would work just as well. We observed that when we used $KL = p_{mean} \log \frac{p_{mean}}{p_0} + (1 - p_{mean}) \log \frac{1-p_{mean}}{1-p_0}$ as the regularizer, it still preserved the ranking. If we want to use non-uniform layerwise

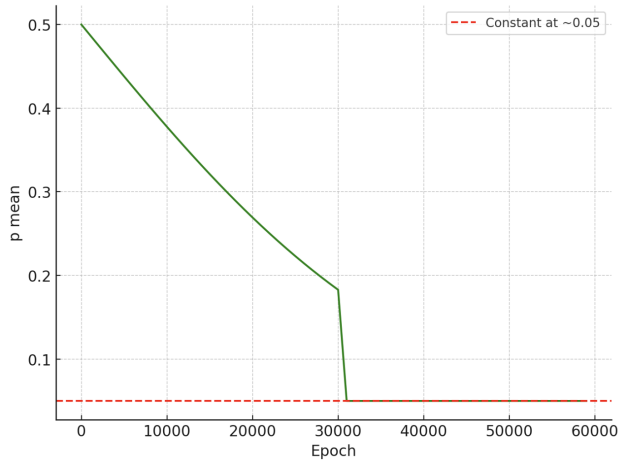


Figure 1: Mean of p across various epochs when the regularization used is $|mean(p) - \frac{s}{d}|$. In this particular experiment, $\frac{s}{d} = 0.05$.

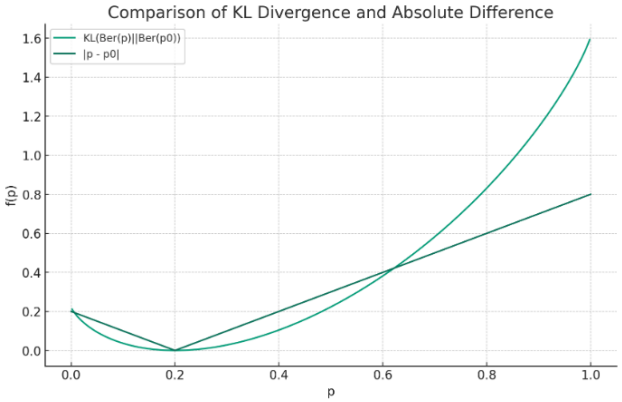


Figure 2: Comparison for KL regularization and pointwise centre ℓ_1 regularization for a scalar value. Around the prior value p_0, the KL is much smoother than L1 regularizer. $f'(p) = \log(\frac{\frac{p}{1-p}}{\frac{p_0}{1-p_0}})$ for KL, which is very close to 0 when $p \in [p_0 - \epsilon, p_0 + \epsilon]$. However, for L1 regularization, $f'(p) = 1$ or -1 for all points except $p = p_0$.

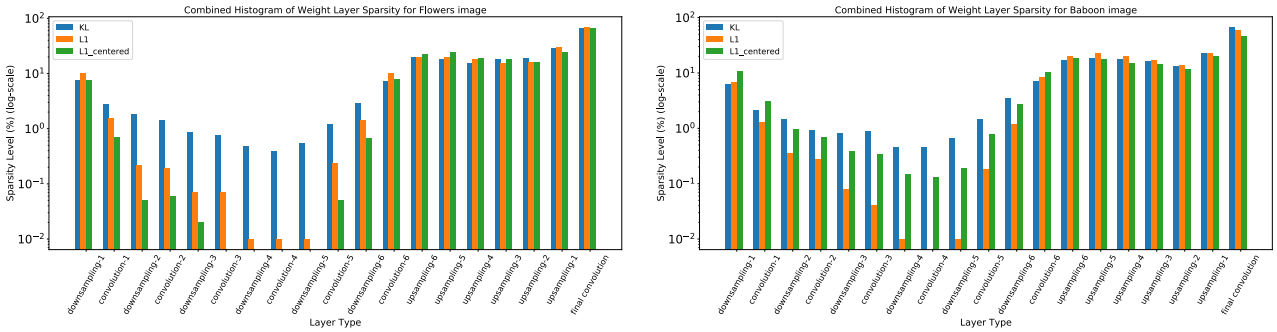
prior, we can still modify the regularizer to use use the $mean(p)$ for each layer. We will add a discussion of the alternatives in the final version.

2 OES pruning for MRI reconstruction

We extend the OES pruning and sub-network training framework to the setting of multi-coil magnetic resonance image (MRI) reconstruction from undersampled k-space measurements. In previous literature [4], dense networks based DIP was used for MRI reconstruction as folows:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{c=1}^{N_c} \|\mathbf{A}^{(c)} G(\boldsymbol{\theta}, \mathbf{z}) - \mathbf{y}^{(c)}\|_2^2 \qquad \text{(P1:Vanilla DIP)}$$

For multi-coil MRI, let there be N_c number of coil sensitivity maps denoted as $\mathbf{S}_c \in \mathbb{C}^{q \times q}$, $c = 1, 2.., N_c$. The corresponding $\mathbf{A}^c$ denotes the forward linear operator $\mathbf{A}^c(\mathbf{M}) = \mathbf{M}\mathcal{F}\mathbf{S}_c$. $\{\mathbf{M} \in \{0,1\}^{q \times q}\}$ is the sampling mask in k-space, $\mathcal{F} \in \mathbb{C}^{q \times q}$ denotes the Fourier Transform operator and $\mathbf{y}^{(c)} \in \mathbb{C}^q$ denotes the undersampled k-space measurements. $G(\boldsymbol{\theta}, \mathbf{z})$ is an overparameterized Unet with two channels that processes



(a) Flowers (b) Baboon

Figure 3: Layerwise architecure pruning (sparsity percentage in log-scale) by OES at initialziaiton using three different choices of regularization, KL, ℓ_1 and centered ℓ_1 for Baboon image and Flowers image in Set-14 dataset. Centered ℓ_1 means the centered mean regularizer.

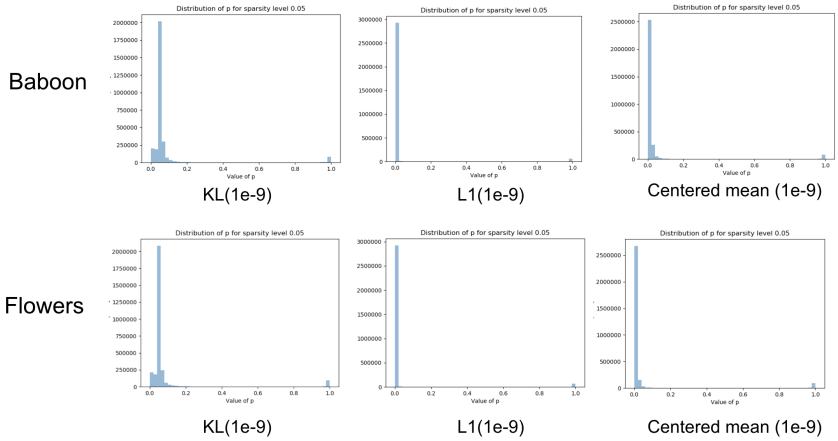


Figure 4: Histogram of logits of p when OES is ran across images with KL, ℓ_1 and centered mean regularizer. In our implementation we minimize $|\sum_i p_i - (\frac{s}{d} * numel(p))|$, to both ℓ_1 regularization and centered mean regularizer on the same scale.

the real and complex channel separately and with trainable parameters $\boldsymbol{\theta}$ and fixed input $\mathbf{z}$. For our experiments, we use multi-coil fastMRI knee and brain datasets [1,2] which are available publicly. The coil sensitivity maps were obtained using the BART toolbox [3]. When the dense network [4] is trained with generic optimizer like ADAM, the above suffers from overfitting (Figure 7a). In the OES framework, we first learn the mask for the subnetwork, denoted as $\mathbf{m}^*(\mathbf{A}, \mathbf{y})$ (not to be confused with the k-space mask $\mathbf{M}$), where $\mathbf{A}(\mathbf{M}) = [\mathbf{A}^c(\mathbf{M})]_{c=1}^{Nc}$ and $\mathbf{y} = [\mathbf{y}^c]_{c=1}^{Nc}$. For the sake of notation, we will omit the coil dependency c as the loss can be combined across coils and written in terms of one forward operator $\mathbf{A}$ and measurements $\mathbf{y}$.

$$\begin{aligned} \mathbf{m}^*(\mathbf{y}, \mathbf{A}) = C(\mathbf{p}^*) \quad & \text{such that} \\ \mathbf{p}^* = \arg\min_{\mathbf{p}} & \mathbb{E}_{\mathbf{m}\sim Ber(\mathbf{p})} \left[\|\mathbf{A}G(\boldsymbol{\theta}_{in} \circ \mathbf{m}, \mathbf{z}) - \mathbf{y}\|_2^2 \right] \\ & + \lambda KL(Ber(\mathbf{p})||Ber(\mathbf{p}_0)). \end{aligned} \tag{1}$$

In Figure 6, we show the 4 MRI scans that are used in the following experiment. $\mathbf{x}$ denotes the ground truth MRI image (obtained from a full set of k-space measurements), $\mathbf{M}_{4\times}$ and $\mathbf{M}_{8\times}$ denote the $4\times$ and $8\times$ undersampling masks for k-space or Fourier space (white lines are sampled), respectively. $A^H(\mathbf{M}_{4\times})\mathbf{y}$ and $A^H(\mathbf{M}_{8\times})\mathbf{y}$ denote the conventional zero-filling MRI reconstructions that produce aliasing artifacts. We will denote the set of the forward operator and measurement pair as $(\mathbf{A}_i(\mathbf{M}_{4\times}), \mathbf{y}_i)$ for data index $i = 1, 2, 3, 4$ for $4\times$ undersampling rate. For $8\times$ undersampling rate, we denote the pair as $(\mathbf{A}_i(\mathbf{M}_{8\times}), \mathbf{y}_i)$. In our

experiments, we train the OES mask using the pair $(\mathbf{A}_1(\mathbf{M}_{4\times}), \mathbf{y}_1)$, and then use the mask subnetwork to reconstruct MRI in four different scenarios across various network sparsity levels:

1. **Self + same undersampling**: The target reconstruction pair is $(\mathbf{A}_1(\mathbf{M}_{4\times}), \mathbf{y}_1)$. We denote this experiment as $P(\mathbf{A}_1(\mathbf{M}_{4\times}), \mathbf{y}_1)$.
2. **Self+higher undersampling**: The target reconstruction pair is $(\mathbf{A}_1(\mathbf{M}_{8\times}), \mathbf{y}_1)$. We denote this experiment as $P(\mathbf{A}_1(\mathbf{M}_{8\times}), \mathbf{y}_1)$
3. **Cross + same undersampling**: The target reconstruction pair is $(\mathbf{A}_i(\mathbf{M}_{4\times}), \mathbf{y}_i)$ for $i = 2, 3$ and 4. We denote this experiment as $P(\mathbf{A}_i(\mathbf{M}_{4\times}), \mathbf{y}_i)$
4. **Cross + higher undersampling**: The target reconstruction pair is $(\mathbf{A}_i(\mathbf{M}_{8\times}), \mathbf{y}_i)$ for $i = 2, 3$ and 4. We denote this experiment as $P(\mathbf{A}_i(\mathbf{M}_{8\times}), \mathbf{y}_i)$.

Note that transfer to a higher undersampling rate demonstrates the capability of transferring to a different level of degradation (Reviewer GGcd).

Once the mask $\mathbf{m}^*(\mathbf{A}_1, \mathbf{y}_1)$ is obtained, the subnetwork at initialization is further trained to convergence with the following optimization. Similar notations extend to $8\times$ undersampling rate.

$$\min_{\theta} \|\mathbf{A}_i(\mathbf{M}_{4\times})G(\boldsymbol{\theta} \circ \mathbf{m}^*(\mathbf{y}_1, \mathbf{A}_1), \mathbf{z}) - \mathbf{y}_i\|_2^2 \qquad (\mathrm{P}(A_i(\mathbf{M}_{4\times}), \mathbf{y}_i)\text{: Sparse-DIP})$$

We make the following observations from the PSNR curves in Figure-7.

- *Sparse-DIP reduces overfitting:* Vanilla Dense DIP produces artifact-affected images in all the cases. This is due to the nullspace of the forward operator that does not offer any control over nonsampled frequencies. Sparse DIP has very less overfitting.
- *Sparse-DIP is robust to higher undersampling rate:* For higher undersampling factor, i.e, $8\times$ undersampling, vanilla dense DIP overfits much more. Sparse DIP at higher sparsities (above 90%) seems to be robust to overfitting even at $8\times$ undersampling.
- *Moderate overfitting at moderate sparsity*: With moderate sparsity level $(50\%, 80\%)$, subnetwork overfits artifacts when cross transfer tasks take place (different image's measurements) or when the undersampling rate is $8\times$. However, overfitting (at moderate sparsity levels) takes place to much less extent when self transfer takes place with the same undersampling rate $4\times$.
- *Limited representation capability at very high sparsity* For higher sparsity levels (90% or higher), overfitting rarely happens in any of the scenarios (cross-transfer or higher undersampling rate). For very high sparsity level 97%, the PSNR curve fails to rise very high, denoting that the network has already reached its representation capability.

3 Reviewer MpG7, Rating: 6, Confidence: 2

3.1 Weaknesses

1. However, there are a couple of shortcomings worth mentioning. Firstly, the text in the images is too small, making it difficult for readers to discern details. Moreover, in Figure 7, although four colours are used, the legend only describes three, which can be confusing for readers trying to interpret the data accurately.

   Answer: Thank you for the careful suggestion and advice to modify the figure texts. We updated the figures and the legend in Figure-7 in the manuscript based on your suggestions. Please let us know if you have any further questions.

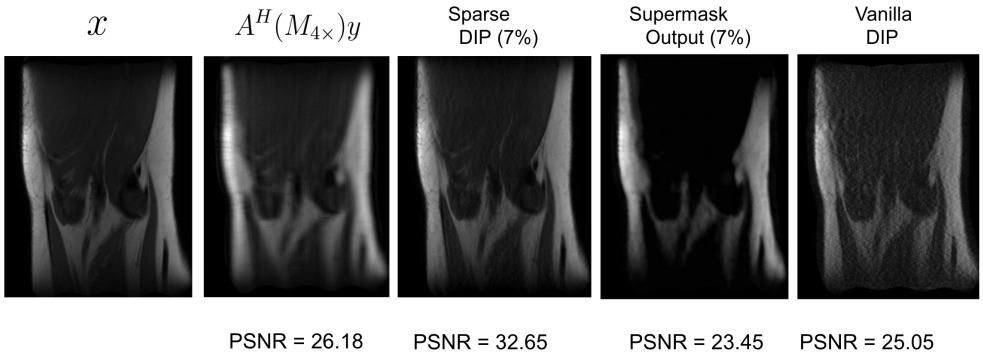


Figure 5: MRI reconstruction comparison with Sparse-DIP and Vanilla Dense DIP without early stopping. Sparse-DIP removes aliasing artifacts and preserves the important details of the images when compared to the ground-truth **x**. Vanilla dense DIP overfits to the aliasing artifacts and requires careful early stopping (See Figure 7a). Supermasked output at network initialization still manages to capture some important image details.

4 Reviewer GGcd, Rating: 7, Confidence: 4

4.1 Weaknesses

1. All experiments address only for one inverse problem- denoising, with some inpainting experiments in the appendix. Some discussion is needed about how the performance of the proposed method varies across various noise levels, across degradations etc.

   Answer: Thank you for the comment. In Table-2 of appendix section of the submitted manusrcipt, we provided denoising result across three different noise levels $\sigma = 25, 12, 17dB$. Here we note that even though OES used noisy images with $\sigma = 25dB$, it still generalized well to lower noise levels $\sigma = 12dB$ and $\sigma = 17dB$ (performed better than deep decoder and GP-DIP).

   To extend our framework, where measurement is acquired in a different domain (Fourier domain), we perform MRI reconstruction from undersampled k-space measurements (refer to MRI-OES section) and generalize OES in this setting. We learn the OES network mask with $4\times$ undersampling measurements and perform image reconstruction on measurements with $8\times$ undersampling.

   With moderate sparsity level $(50\%, 80\%)$, subnetwork overfits to aliasing noise when cross transfer tasks take place (different image measurements) or when the undersampling rate is $8\times$. However, overfitting (at moderate sparsity levels) takes place to very less extent when self transfer takes place with the same undersampling rate $4\times$.

   For higher sparsity levels (90% or higher), overfitting rarely happens in any of the scenarios (cross-transfer or higher undersampling rate). For very high sparsity level 97%, the PSNR curve fails to rise very high, denoting that the network has already reached it's representation capability.

2. For which reconstructions tasks is this method suitable, for e.g. can this method work where the measurement is very different from the clean image e.g. MRI and compressed sensing?

   Answer: We perform extensive experiments for MRI reconstruction from undersampled k-space measurements in Section. OES framework extends well when measurements are acquired in the Fourer space (as opposed to pixel-space). Furthermore, when measurement across diverse image slices are used (Figure-6), we get reasonable reconstruction even at higher undersampling rates. However, moderate sparsity level $(50\%, 80\%)$ seems to overfit for cross transfer tasks.

3. Some discussion is needed about why this method performs better than previous methods despite the overlap (i.e. the use of bernoulli mask based optimization objective in eq2.), beyond appendix H which is focused on the choice of regularization and appendix D.3 *Sufficiently addressing this point can increase the rating of the paper.*

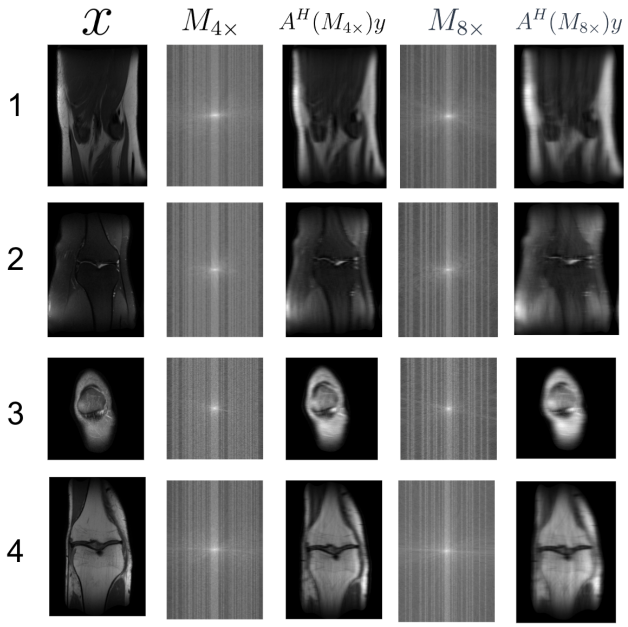


Figure 6: The 4 MRI ground-truth and measurement used in this experiment. $\mathbf{x}$ denotes the ground-truth image or full-kspace reconstruction. $\mathbf{M}_{4\times}$ and $\mathbf{M}_{8\times}$ denote the k-space undersampling masks. $A^H(\mathbf{M}_{4\times})\mathbf{y}$ and $A^H(\mathbf{M}_{8\times})\mathbf{y}$ denotes the zero-filling reconstruction that produces aliased artifacts.

Answer: We have provided an extensive comparison with KL regularization and previously used ℓ_1 regularization. We observe that KL regularization (Figure-2), prevents OES from collapsing the p's to a single point and this helps in preserving the relative ranking of the importance scores. Across 2 different images, we observe that the middle layers (which are larger) suffer from layer collapse when ℓ_1 regularization is used. We also state that having parameter-wise importance score makes it flexible to use non-uniform priors. Although further improvements can be made in carefully chosing such regularizers, but we believe that the element-wise KL regularization is the most generic type of regularizer that is resilient to layer collapse by preserving rank importance throughout.

4.2 Limitations

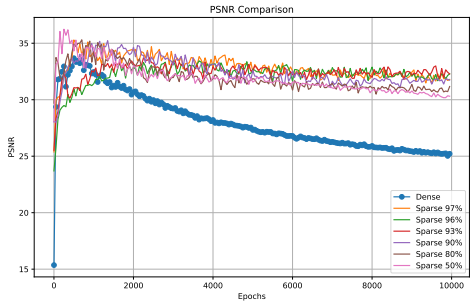
1. There is no explicit discussion of the limitations of the proposed method.

   Answer: We will add the following limitations to the appendix:

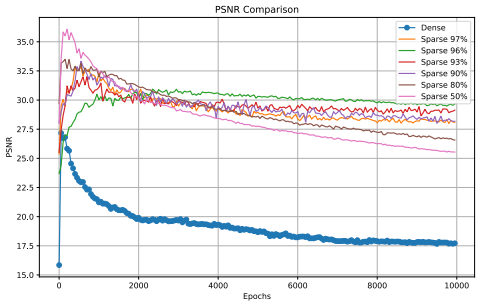
   (a) For transfer experiments, when the target measurement is from a different domain, sparse networks seem to overfit slightly.

   (b) There is an added cost of finding the mask which takes slighlty more wall time than training a network for similar iterations. This is because of the inner loop of Gumbel Softmax reparameterization trick. Although training the subnetwork once from one image would suffice in many cases unless the target image is very different from the image that was used to learn the mask.

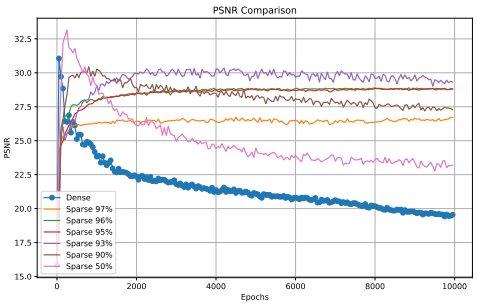
   (c) Certain tasks require specialised network architecture. For example MRI images are usually complex, and two channel Unet (separate channels to process real and complex part) are used in this case. So, transferring OES subnetworks to natural image restoration tasks can be limited in certain cases.



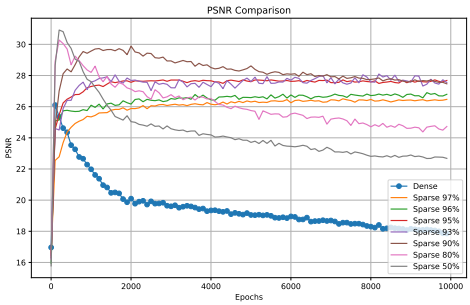
(a) Self: $P(A_1(M_{4\times}), y_1)$



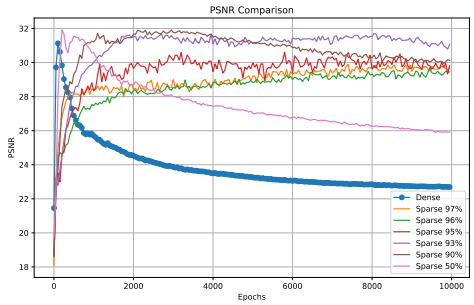
(b) Self: $P(A_1(M_{8\times}), y_1)$



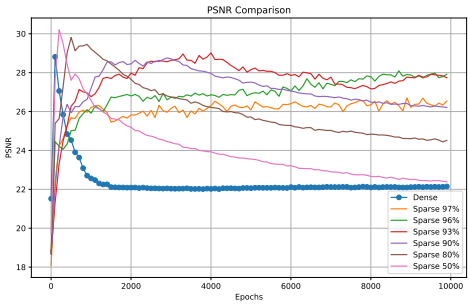
(c) Cross: $P(A_2(M_{4\times}), y_2)$



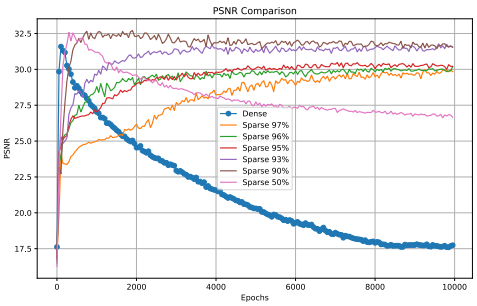
(d) Cross: $P(A_2(M_{8\times}), y_2)$



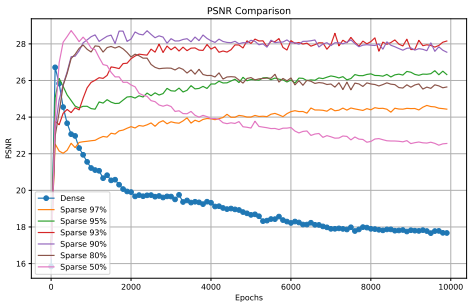
(e) Cross:$P(A_3(M_{4\times}), y_3)$



(f) Cross: $P(A_3(M_{8\times}), y_3)$



(g) $P(A_4(M_{4\times}), y_4)$



(h) Cross: $P(A_4(M_{8\times}), y_4)$

Figure 7: Performance of OES subnetworks for MRI reconstruction from 4× (left column) and 8× (right column) undersampled k-space measurements. In Figures(a,b) (self), the OES network mask m^* was learned from pair $((A_1(M_{4\times}), y_1))$ and then the subnetwork was used to reconstruct image from $((A_1(M_{4\times}), y_1))$.

5 Reviewer ySfU, Rating: 5, Confidence: 5

5.1 Weaknesses

1. The paper could be more well-written and the sections better organized. a) Sections 2 and 3 could be better organized as separate related works, background, and methods sections instead of combining all three for DIP and pruning, respectively. b) The sections are mostly comprised of singular paragraphs with no breaks or separation between different logical concepts. Splitting these large paragraphs into several smaller ones would greatly improve the readability of the paper. c) Ordering and numbering of figures is awkward. Some figures (e.g. 1, 2, and 3) are never referenced in the main text. The authors place several figures that are integral for understanding key findings in the appendix instead of the main text (e.g. Figures 15 and 16, referenced in section 4.1).

   Answer: Thank you for the suggestions on improving the readability of the paper. In the revised version of the paper, we accomodate the following changes based on your suggestion:

   - We will add the paragraphs and the logical breaks between the logical concepts to improve the readability.
   - We will add the necessary references to the first three figures. We will move the sections consisting Figures-15 and 16 ahead in the appendix for easy reference, since we have more relevant figures in the main text and limited space.
   - For point-a), we mention the related section/methodologies for DIP in Section-2 and that of pruning in section-3 (separately) in the submitted manuscript.

2. Only one specific type of corruption, additive Gaussian noise, is considered in the manuscript. While small experiments are also conducted on inpainting in the Appendix, it would have been interesting to see how the proposed technique performs under different corruptions, e.g. compressed sensing from random Fourier coefficients. As it stands, it is difficult to claim that the proposed method will generalize beyond the limited cases presented in the paper.

   Answer: Thank you for this suggestion that can demonstrate the broader applicability of our method. We perform extensive experiments for MRI reconstruction from undersampled k-space measurements in Section. OES framework extends well when measurements are acquired in the Fourier space (as opposed to pixel-space). We refer to the discussion and experiments in section for the detailed methodology and findings. Please let us know if you have any questions regarding this section or the new experiments.

3. The reduction in run time and computational complexity claimed by the authors (e.g. finding 2 on line 250) should include the complexity of finding the mask and not just the optimization time for the network weights.

   Answer: Thank you for pointing it out. We will add this in the finding-2 section. The wall time for computing the mask using OES is slightly more than training a normal dense network (which we will also include in the limitation section). The wall time for finding a mask on average is $(2\times)-(3\times)$ more than training a dense network because of the inner loop in the Gumbel softmax algorithm. However, due to efficient transferability of OES masks within images in the same dataset, it may not be necessary to find the mask for each image in a dataset.

5.2 Questions

1. (Background for my question) I understand that your choice of KL regularization is superior to L1 for promoting sparsity and preserving rank importance, as you explain in Appendix H. However, I am still not convinced that KL between your distribution and a Bernoulli with uniform probability at all possible mask locations is the optimal choice of regularizer. I would think that the element-wise comparisons between your fixed prior distribution and your learned mask distribution in the closed-form KL loss (i.e. on the right column in lines 195-197) is too strict of a condition, and would dampen the magnitudes of your learned probabilities even when those probabilities correspond to very useful network parameters.

Answer: Will refer to the previous section on L1 and pointwise regularization.

2. Can you clarify the effects of your design choice for the sparsity regularizer? Specifically with respect to the element-wise comparisons between the prior and your learned distribution? Do you have any intuition about what would happen if you simply penalize the mean probability of your learned mask to be centered at your desired sparsity level

   Answer:

6 References

1) Zbontar, Jure, et al. "fastMRI: An open dataset and benchmarks for accelerated MRI." arXiv preprint arXiv:1811.08839 (2018).
2) Knoll, Florian, et al. "fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning." Radiology: Artificial Intelligence 2.1 (2020): e190007.
3) Tamir, Jonathan I., et al. "Generalized magnetic resonance image reconstruction using the Berkeley advanced reconstruction toolbox." ISMRM Workshop on Data Sampling and Image Reconstruction, Sedona, AZ. 2016.
4) Darestani, Mohammad Zalbagi, and Reinhard Heckel. "Accelerated MRI with un-trained neural networks." IEEE Transactions on Computational Imaging 7 (2021): 724-733.