Pitfalls of conditional computation for multi-modal learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Humans have perfected the art of learning from multiple modalities, through sensory organs. Despite impressive predictive performance on a single modality, neural networks cannot reach human level accuracy with respect to multiple modalities. This is a particularly challenging task due to variations in the structure of respective modalities. A popular method, Conditional Batch Normalization (CBN), was proposed to learn contextual features to aid a deep learning task. This uses the auxiliary data to improve representational power by learning affine transformation for Convolution Neural Networks. Despite the boost in performance by using CBN layer, our work reveals that the visual features learned by introducing auxiliary data via CBN deteriorates. We perform comprehensive experiments to evaluate the brittleness of a dataset to CBN. We show the sensitivity of CBN to the dataset, suggesting that learning from visual features could often be superior for generalization. We perform exhaustive experiments on natural images for bird classification and histology images for cancer type classification. We observe that the CBN network, learns close to no visual features on the bird classification dataset and partial visual features on the histology dataset. Our experiments reveal that CBN may encourage shortcut learning between the auxiliary data and labels.

1 Introduction

Learning about complex relationships between real-world variables has been the key to progression of human kind [10]. While some of these relationships are now intuitive to us, some are more complex mathematical relationships. With advancement of deep learning in the field of computer vision [12], natural language processing [26], and speech recognition [2], learning complex relationships across modalities is still an open research question. Most of the current state-of-the-art algorithms, learn from single modality [7, 21] which might limit the representation learnt.

Processing data from multiple modalities can provide richer context to respective modality[28]. For examples, video action recognition can be benefited by learning from the audio along with the visual features[25]. Such factors motivate *context-based learning*, where auxiliary features are used to enhance the primary feature representation. To learn such contextual information, designing of methods that can learn from each individual modality along with their fusion is critical but a challenging task. In computer vision, Conditional Batch Normalization (CBN) based models that perform feature-wise affine transformations have shown to achieve strong visual reasoning[20].

CBN[6] relies on modulating the visual feature stream with auxiliary features in the intermediary layers of a CNN. The normalized activations produced by auxiliary network has been considered as the optimal method to compute contextual information. Recent works have shown unreliability of batch normalization during distributional shifts, adversarial attacks and dependency on the dataset[13, 9]. Lack of robustness of BN proposes an immediate challenge for its conditional version, CBN. With

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

adoption of CBN in various applications[1, 4, 20], it is important to study the quality of visual features learnt in presence of auxiliary data.

In this work, we empirically and visually show that while CBN network which learns auxiliary data may surpass the performance of standard Batch Normalization(BN) network, they often fail to learn visual features altogether or partially. We devise a set of experiments to perform surgery into the CBN network. We observe that the CBN network is brittle to the *type* and *quality* of the contextual information and fails to generalize realistic scenarios by learning shortcuts.

2 Background

Since the introduction of Batch Normalization[14], they have become an integral part of CNNs for faster convergence. The intermediary convolutional layers stabilise the variation of activations by normalising the input, x_i characterised by the following equation:

$$x_i^N = \frac{x_i - \mu_i}{\sigma_i} \tag{1}$$

Batch normalization introduces learnable parameters, β_i and γ_i for each normalised layer L_i^N :

$$x_i^N = (\frac{x_i - \mu_i}{\sigma_i + \epsilon})\gamma_i + \beta_i \tag{2}$$

Conditional batch normalization[6], aims to learn the learnable parameters from meta-data to "ground" visual features. It directly predicts affine scaling parameter of the affine transformation from meta-data. A shallow multi-layer perceptron, we call it auxiliary network, allows the meta-data to predict the β' and γ' that shifts and scales the normalised batch features. Formally, for meta-data a_i and auxiliary network for $\mathcal{H}$ and $\mathcal{O}$ respectively for affine parameter prediction, $\mathcal{H}(a_i) = \beta'$ and $\mathcal{O}(a_i) = \gamma'$.

$$\hat{\beta}_i = \beta_i + \mathcal{H}(a_i) \tag{3}$$

$$\hat{\gamma}_i = \gamma_i + \mathcal{O}(a_i) \tag{4}$$

These updated $\hat{\beta}_i$ and $\hat{\gamma}_i$ are used as parameters for batch normalization. Modulation of a neural network activations using conditional information through CBN has been proven successful in multi-modal setting [20]. While CBN is a computationally efficient method to fuse different modalities, the robustness of this affine layer is unexplored. The addition of β' and γ', might seemingly be small yet showing a boost in performance of the task, we observe that it degrades in the quality of visual features learnt.

3 Related Works

Batch normalization[14] layers were introduced to alleviate the problem of internal covariate shift, also leading to faster model convergence. However, santurkar2018does have argued that BN might not alleviate the internal covariate shift. Recent studies [5, 30] have shown standard BN training is detrimental to performance in domain shift task. CBN layer was introduced to provide contextual information in form of affine transformation. perez2018film uses CBN layer to modulate visual features to answer compositional questions on CLEVR dataset[15]. Conditional Instance Normalization [8] was introduced in the context of style transfer. In reinforcement learning, [3] used CBN layer to condition NMN and LSTM to follow compositionality in a 2D grid world. In generative modeling, [4, 31] uses condition biasing through image description. [19] used CBN layer to make input distribution features robust to variations in few shot setting. While CBN has proved it efficacy through its acceptance in different applications but a thorough analysis of its potency is yet to be learnt[18]. We make the first attempt to uncover the potential risk of using CBN via a series of experiments.

4 Experiments

CBN is widely used for various applications such as multi-modal language acoustic learning[1], meta learning[22, 32] and image generation[4]. Due to its varied applications, we design our methodology for a supervised image classification task with visual and auxiliary features allowing for an easier surgery of the neural network. Aim of these experiments is to expose the potential pitfalls and impacts of fusing meta-data with images using CBN.

4.1 Experiment Setup

We design our experiments to answer some fundamental questions posed by fusing of differently structured auxiliary data via CBN. We design a small experiment for each question posed although the major question, we aim to ask is - *Does the addition of auxiliary data improve visual feature representation?*

CBN training - Baseline In this setting, auxiliary data is passed as the input to the shallow auxiliary network that predicts the affine BN parameters. This baseline experiment is mainly performed to establish effectiveness of CBN for classification.

Effect on CBN in absence of test-time auxiliary data The additional features passing through auxiliary network and its fusion with spatial features[20] has been reported to improve accuracy across tasks. We aim to look at the neural network beyond final predictive accuracy with this experiment. In medical applications, the meta-data collection differs across hospitals, and often among different medical practitioners. We simulate this scenario by masking the attributes at test-time. Due to masking, the auxiliary network is switched off and the network is normalised by affine transformations only. By reducing the dependence of network on auxiliary data, it is easier to establish its independence from visual features during test-time. We eventually completely switch off the auxiliary data stream and notice the difference in reliance of this auxiliary stream versus the spatial stream.

Effect on CBN with masked train-time auxiliary data While in previous experiment, we attempted to establish the independence of the visual stream of features, in this experiment, we aim to understand the quality of visual features learnt in absence of meta-data. We mask the attributes randomly to aid the network to learn and utilize visual features more.

Introduction of Supplementary network We hypothesize that the presence of auxiliary data challenges the learning of visual features. Therefore we introduce a supplementary network to *ground* and assist with the optimization of CBN network to learn visual features. The supplementary network has same with network architecture as the CBN network but uses standard batch normalization. Experimentally, we back propagate cross entropy loss of the CBN network along with the distance between CBN network and supplementary network along with optimization over the supplementary network. More formally, let $\mathcal{F}$ represent the CBN network that maps input x_i to its label y_i using subsidiary features, a_i and $\mathcal{G}$ represent the BN network that maps input x_i to its label y_i. Let the cross entropy loss of CBN network be represented by $\mathcal{L}^{CBN}$.

$$L = m * \mathcal{L}^{CBN}(x_i, y_i) + (1 - m) * D_{kl}(\mathcal{F}||\mathcal{G}) \tag{5}$$

where m is a hyperparameter. It is important to note that the aim of supplementary network is to simply assist in optimization and grounding the CBN network, therefore the weighting over the distance between supplementary and CBN network is very small.

Removing data from the CBN network Datasets such as CUB-200-211[27], Animals with Attributes[29] 2, which are popularly used in few shot classification, have attributes which are highly correlated with the images and label. We can easily learn about the effect of attributes on the CBN from previous experiments, in this experiment we aim to learn the impact of visual features on the CBN in isolation to auxiliary data.

4.2 Dataset

We use natural images and medical images to test the effectiveness of Conditional Batch Norm using our methodology. We use [27] dataset for the task of bird identification. It contains 11788 images taken from 200 different species of bird. Every image of the dataset contains 312 binary and continuous concept labels. This dataset has been widely used in Few Shot Classification and Conditional Image Generation, making it suitable for our analysis. The auxiliary data comes from the attribute. In our experiments we use binary attributes.

For medical images, Whole Slide Pathology images were used. The tumor-infiltrating lymphocytes (TILs) [23] contains histological mappings based on each type of cancer. We use all 13 subsets of TCGA dataset, therefore constituting 13 cancer types. Although the popular task for such a dataset is necrosis classification[17] but since it is a binary classification task and is prone to over fitting without truly emphasising the effectiveness of our analysis, we modify the task to be a classification of different (here, thirteen) tumor types. Advantage of medical images is their meta-data is readily available from diagnosis. A doctor uses whole slide images along with the meta-data for diagnostics indicating importance of fusion of different modalities. The meta-data includes information such as origin of tumor, age, gender, size of tumor cells.

4.3 Results and Analysis

In this section we evaluate the performance of networks that perform affine transformations using Conditional Batch Normalization and standard Batch Normalization. Due to BN's sensitivity to the batch size[16], we perform all our experiments on batch size of 64. We perform our experiments on ResNet-18[11] backbone. For CBN network, we modify the ResNet-18 architecture to replace batch normalization layers with CBN layers. The auxiliary network that uses auxiliary data is a two layer MLP network to predict the affine parameters for CBN networks. The experiments in this paper were carried out using the V100SXM2 GPU. The results reported are averaged over 10 runs. Table 1summarises the results on test set but we breakdown the analysis according to the set of different experiments below.

CBN Baseline We train both CBN network and BN network to perform classification on the CUB and TIL dataset. In principle we expect the CBN network to learn from the spatial feature along with the features learnt from the other auxiliary modality. This suggests that, ideally, the representative power of this baseline model is greater than the representation from the standard vision only model. This should in turn lead to better generalization. From, figure 1 we observe that the validation accuracy of the CBN model is greater than the BN model by a great margin of 6.9% in CUB and over 10% in TIL dataset. This proves the effectiveness of modulating visual features by the auxiliary data. We interestingly also notice that the CBN network achieves convergence much faster than the BN network. Although convergence is a desired property, it is worrisome if the network converges in just 4 epochs achieving 100% accuracy in the case of CUB dataset! Although, the same is not observed in TIL dataset, suggesting the difference in dataset and/or auxiliary data type. This particularly motivates the upcoming experiments to make an attempt at uncovering the black-box of conditional learning from auxiliary information stream.

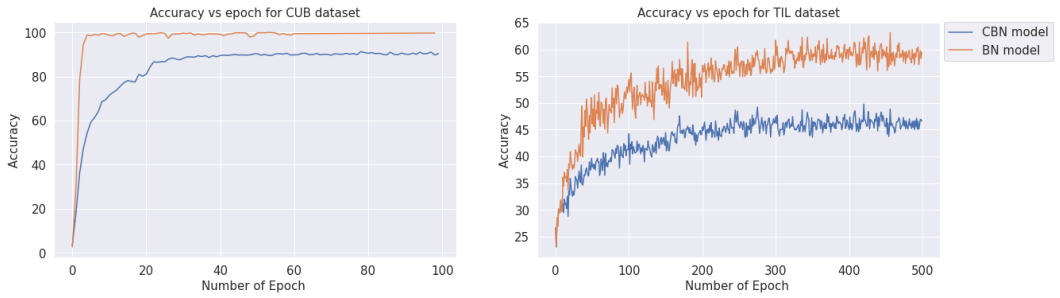

Figure 1: Validation accuracy versus epoch plotted for CUB dataset and TIL dataset respectively. Models with CBN layer perform better

Dependence of CBN on test-time auxiliary data In a realistic scenario, due to data privacy and differences in data acquisition, meta-data collected for each patient maybe varied. If we were to leverage the CBN model in such a setting, it is likely that some of the auxiliary information is missing. To test this we limit the auxiliary information during validation and test-time in increments from 0 to 1 (full CBN network) of 10% supplementary data. From figure 2 we observe that in the case of CUB, with less than 20% of auxiliary data, the network performs as well as in the presence of all of the auxiliary data. While in the case of TIL dataset, there is a sharp increase in the accuracy with greater fraction of meta-data. We hypothesize that this is due to the difference in structure and diversity of meta-data. Naturally histopathology data is very diverse suggestive of its difficulty in learning. In addition it must be noted that in complete absence of auxiliary information during testing time, the CBN network has very low accuracy, 0.5% in the case of CUB and 22.3%. This suggestive of the network *not learning visual* feature in the case of CUB dataset. While, simultaneously the CBN network does achieve greater than random accuracy in TIL dataset, its performance deteriorates highly in absence of the meta-data.

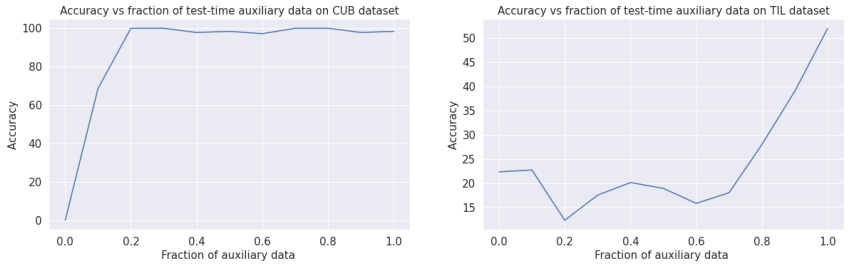

Figure 2: Validation accuracy versus the fraction of auxiliary data masked during validation.

Masking of train-time auxiliary data From figure 3, the dependence of accuracy on the auxiliary features is apparent. We therefore attempt to establish the relationship between model's performance and the auxiliary data during training phase. In this experiment, we incrementally increase the fraction of auxiliary data fed in to the CBN model. With no auxiliary data at train-time, the network behaves similar to a BN model since affine prediction is turned off, where as with all of the meta-data at train-time, the network behaves and performs similar to the baseline CBN model. In figure 3, we observe the general trend of achieving higher accuracy in presence of higher auxiliary data, although in the experiment with CUB dataset, the model converges much earlier with TIL data. Intuitively 3 also signifies the inverse change in dependency of visual feature representation with increase in auxiliary data. It is observed that, with affine parameter prediction switching on, the neural network attempts to use the auxiliary data for label prediction.

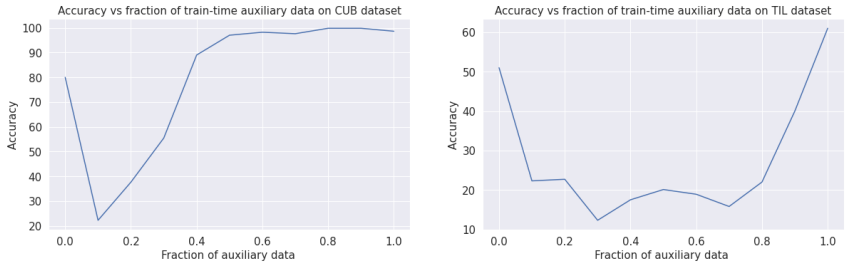

Figure 3: Validation accuracy versus the fraction of auxiliary data masked during training.

Introduction of supplementary network We hypothesized that supplementary network may help to ground and bring the CBN network closer to BN network, implying learning of visual features. Hyperparamaters m equals 0.9. We observe that the hypothesis is proven incorrect and supplementary network is not able to reduce the dependence of auxiliary data on CBN network. In figure 6 we observe that the validation accuracy curves on both CUB and TIL datasets are very similar to CBN validation curve, although more stable. Due to ineffectiveness of the supplementary network on baseline experiment, we perform all our experiments is CBN network.

(a) CBN model (b) CBN with 0 attributes at test-time (c) BN model

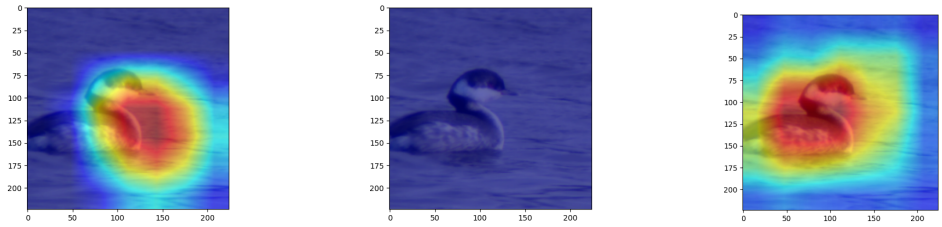

Figure 4: A visualization of GradCAM[24] features for *Horned Grebe* class. Although the CBN model achieve 100% accuracy, the feature of BN model are visually more descriptive.

Removing spatial data Table 1 present that even in the absence of image data, the network is able to perform as well as the baseline CBN network for CUB dataset. This is highly suggestive of presence of shortcuts in data and the visual features not contributing to the CBN model's predictive performance. On the contrary, CBN network on the TIL dataset learns some feature in absence of visual features, therefore indicating partial learning of vision features.

Discussion From an exhaustive set of experiments, we observe that although the attenuating the affine batch normalization parameter via a second modality leads is higher predictive performance as compared to standard batch normalization training of vision modality, it leads to deterioration of visual features. In the case of both TIL and CUB, there is an increased dependency on the auxiliary data. This suggests that the affine parameters of CBN network allow for a shortcut learning of auxiliary data leading to poor generalisation. Visually we compare the features learnt by CBN network with all attributes, CBN network with no attributes and BN network for both CUB and TIL dataset. We observe that the features learnt in each of the cases are very different signifying the potential difference in their respective minima. It is particularly interesting to note that in the absence of visual features in CUB, the network has random predictive performance signified by 4b. We observe that the visual features presented in 4a is less explanatory of the bird features as compared to 4b.

(a) CBN model (b) CBN with 0 attributes at test-time (c) BN model

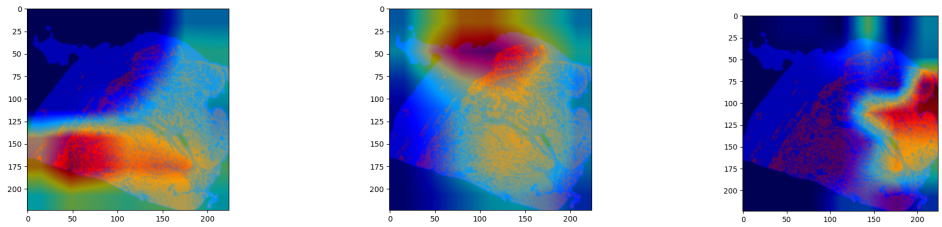

Figure 5: A visualization of GradCAM[24] features for *BLCA* cancer type class.

Ablation study - Continuous vs Discrete attributes in CUB The CUB dataset has contains continuous attribute values between 0 and 1. We threshold the values greater than 50 to 1 and rest to 0. Although in the case of continuous attributes, we observe that the neural network converges within 2 epochs which suggests that the continuous attributes leads to memorization, highly undesirable for generalization. Therefore we use binary attributes in all of our experiments.

5 Conclusion

In this work, we explored the efficacy of a popular fusion method, Conditional Batch Normalization to fuse an auxiliary network with a vision only model. We observe that the CBN network motivates learning of highly undesirable shortcuts learning between the auxiliary data and the label. This leads to a boost in performance but a drop in visual features. We observed that the effectiveness of CBN is dependent on the dataset. We hope that this work inspires research in the direction of learning generalizable visual features while conditioning in presence of differently structured modality.

References

[1] Jerome Abdelnour, Giampiero Salvi, and Jean Rouat. Clear: A dataset for compositional language and elementary acoustic reasoning. *arXiv preprint arXiv:1811.10561*, 2018.

[2] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4945–4949. IEEE, 2016.

[3] Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Pushmeet Kohli, and Edward Grefenstette. Learning to follow language instructions with adversarial reward induction. *CoRR*, abs/1806.01946, 2018.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[5] Jia-Ren Chang and Yong-Sheng Chen. Batch-normalized maxout network in network. *arXiv preprint arXiv:1511.02583*, 2015.

[6] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. *Advances in Neural Information Processing Systems*, 30, 2017.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.

[9] Angus Galloway, Anna Golubeva, Thomas Tanay, Medhat Moussa, and Graham W Taylor. Batch normalization is a cause of adversarial vulnerability. *arXiv preprint arXiv:1905.02161*, 2019.

[10] Ben Goertzel, Matt Iklé, and Jared Wigmore. The architecture of human-like general intelligence. In *Theoretical foundations of artificial general intelligence*, pages 123–144. Springer, 2012.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning. *Image Recognition*, 7, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Lei Huang, Yi Zhou, Tian Wang, Jie Luo, and Xianglong Liu. Delving into the estimation shift of batch normalization in a network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 763–772, 2022.

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[15] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[16] Xiangru Lian and Ji Liu. Revisit batch normalization: New understanding and refinement via composition optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3254–3263. PMLR, 2019.

[17] Emily Pei-Ying Lin, Tzu-Hung Hsiao, Jo-yang Lu, Siao-Han Wong, Tzu-Pin Lu, Konan Peck, Takashi Takahashi, and Pan-Chyr Yang. Translating gene signatures into a pathologic feature: Tumor necrosis predicts disease relapse in operable and stage i lung adenocarcinoma. *JCO Precision Oncology*, (2):1–13, 2018.

[18] Vincent Michalski, Vikram Voleti, Samira Ebrahimi Kahou, Anthony Ortiz, Pascal Vincent, Chris Pal, and Doina Precup. An empirical study of batch normalization and group normalization in conditional computation. *arXiv preprint arXiv:1908.00061*, 2019.

[19] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018.

[20] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[22] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

[24] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[25] Zhensheng Shi, Ju Liang, Qianqian Li, Haiyong Zheng, Zhaorui Gu, Junyu Dong, and Bing Zheng. Multi-modal multi-action video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13678–13687, 2021.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[27] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[28] Yikai Wang, Fuchun Sun, Ming Lu, and Anbang Yao. Learning deep multimodal feature representation with asymmetric multi-layer fusion. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3902–3910, 2020.

[29] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[30] Moslem Yazdanpanah, Aamer Abdul Rahman, Muawiz Chaudhary, Christian Desrosiers, Mohammad Havaei, Eugene Belilovsky, and Samira Ebrahimi Kahou. Revisiting learnable affines for batch norm in few-shot transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9109–9118, 2022.

[31] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[32] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*, pages 7693–7702. PMLR, 2019.

Appendix

Training Model	Testing Model	CUB	TIL
BN	BN	79.9 ±1.1	51.1 ±2.3
CBN-Baseline	CBN-Baseline	99.2 ±0.6	62.9 ±3.4
CBN-Baseline	CBN-No attributes	0.5 ±0.1	19.3 ±3.8
CBN + Supplementary-Baseline	CBN + Supplementary-Baseline	99.0 ±0.5	63.1 ±2.9
CBN-No data	CBN-No data	99.1 ±0.6	20.8 ±10.7

Table 1: Summary of results

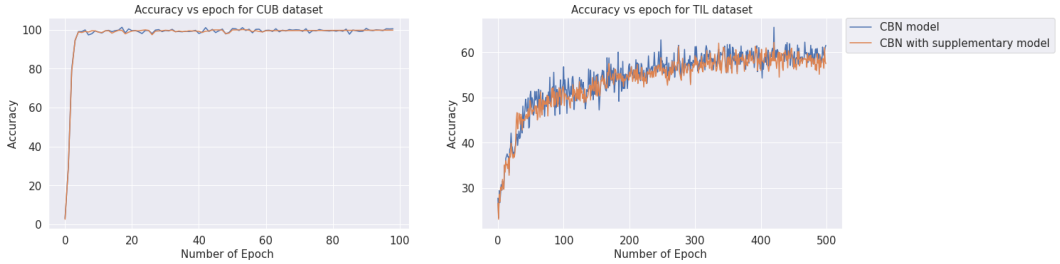

Figure 6: Validation accuracy versus epoch for CBN hybrid in presence of supplementary network.

A Limitation of the work and Future work

In order to perform easier surgery of network, we designed experiments around supervised image classification. It shall be interesting to extend the work across other multi modal settings.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Yes
   (b) Did you describe the limitations of your work? [Yes] Yes, it is added in the appendix
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? The code will although be made public upon acceptance
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Table 1 include variance of error and the other graphs are averaged over 10 seeds
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Work has been appropriately sited
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The medical data used could be sensitive but we have used publicly available dataset where data point are anonimized
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]