A Literature Review: Effectiveness of Grad-CAM and Latent Spaces in Deep Learning

Vedant Devank Patadia
Lakehead University - Computer Science
email: vpatadia@lakehead.ca

***Abstract*— Deep learning has come to the new horizons in the past decade, but the understanding and development of models that require low level precision is still quite complex and thus, require a model intuition and long experience to complement themselves with. There comes Latent Spaces Approaches and Grad-CAM(Gradient -weighted Class Activation Mapping) Visualizations to the rescue. Latent Space Approaches are usually always constructive and provide differently angled insights whilst Visualizing Grad-CAMs always provide some idea into the model's working and they both are helpful in Visualizing the internal working of a Deep Learning model. This paper focuses upon a travel through the general concept of the topic media, i.e., Latent Spaces and Grad-CAMs as well as the various trending methodologies that they both have been employed on.**

I. INTRODUCTION

Deep Learning itself isn't a new coin to the world, and neither is the understanding of how it actually works and approaches a problem. But that's practically only true on the Higher Levels of the concept and further at lower levels the application and transparency of the internal variables as such is quite the complex business. And that is exactly where I want to focus on.

Working on application in Latent Spaces is always a challenge, as every single instance contains not one, or two but even several independently moving parts that aren't quite evident on the first glance. The way they work is actually defined using statistics and probability, but that's when we are working with say hundreds of parameters or I'd dare say Thousands. When Deep Learning Comes into play, we aren't talking about small numbers, the usual models even scale upto millions of trainable parameters, which just puts the latent variables or codes in a black box.

More, the building of any AI model, be it based a Machine Learning Algorithm, Deep Learning Algorithm or an ensemble is not just getting the inputs, outputs and then fitting it to the model. It is far more complicated than that, a good model is fine tuned to the problem it approaches, it may not always give out the best of the best evaluation but an assurance that the model is generalized enough and is correctly approaching the sweet spot. That is, basically it being balanced over all.

Such bold assurances are usually made with the help evaluation metrics and running through the evaluation or testing datasets. But there's always a risk of the model breaking down on some unseen scenarios, even when such faults can be avoided using the dataset at avail. Such, errors may just be because of us just evaluating the results and reports based on the higher level metrics and never investigating the lower level variables, practically making a black box. Generally it is irrelevant but, on some critical matters, a requirement should be added such that one should be able understand the output and the path it took from the input. The reason being, as the learning process is unsupervised and even smallest optimization, loss minimization, etc can improve the scores but add vulnerabilities.

Such endeavors can be performed using a variety of techniques like, Saliency Maps, Class Activation Mapping(CAM), Gradient-weighted Class Activation Mapping(Grad-CAM), Latent Spaces Exploration, etc. There are even better more complex variants of them and one of the more recent advancements have introduced us with Grad-LAMs, it is very well intriguing and currently a new concept to explore.

A. Latent Spaces

Latent space is a very abstract term, and it has been in the play ever since we have been embedding data in the smaller compressed forms, some references in the field of probability and statistics even go as early as 1959 by Anderson, T. W [1].

As described in [2], latent space is a "reduced-dimensionality vector space embeddings of data, fit via machine learning". Which also means, vector embeddings of data, where the distance between similar vector is minimized.

The term latent space is quite broad, and is literally relevant to almost every aspect of machine learning, let alone deep learning. Thus, The review is particularly focusing on how the analysis of latent space helps in understanding models as well as the methods or models that employ the latent space/variables at their very core algorithm.

B. Saliency Map

Saliency Maps are typically used in image recognition and analysis. First stated by Laurent Itti, Christof Koch and Ernst Niebur [3] as, "An image-specific class saliency map, highlights the areas of the given image, discriminative with respect to the given class"

Simply, it derives visual features like, resolution, pixel intensity, colors, contours, orientations,etc and then combine them to create feature maps. Well the concept is quite

straightforward, but the employment of the same in our work is not quite.

Practically, to help interpreting the model Saliency Maps derive the features of an image that pitch-in for generating the particular output. Some of the methods are, deconvolutional network proposed by Zeiler, Matthew D., and Rob Fergus[4], Backpropagation method proposed by Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman [5] and then combining both we get guided backprapogation method proposed by Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin Riedmiller [6].

Although the method works very good in interpretation of a model and is still used to this date for various applications. But, recent studies suggests that they are not very reliable [7] and are very much vulnerable to adversarial attacks/attentions [8].

C. Class Activation Map - CAM

When Lin, Min, Qiang Chen, and Shuicheng Yan, [9] proposed the "Network in Network" architectures, the deep learning world introduced several wide and long architectures that the Saliency Maps simply failed to precisely interpret and out came the CAM, Class Activation Map [10], where the authors of the paper replaced the last stacking layer with the Global Average Pooling,

$$y_c = \sum_k w_k^c \frac{1}{Z} \sum_i \sum_j A_{ij}^k$$

which are simply Average Pools of Previous Feature Activations and then the weighted sum,

$$L_{ij}^c = \sum_k w_k^c A_{ij}^k$$

is then transferred to the softmax layer. Interpreting the Model is just one step far, i.e., just projecting the weights of the final output on the convolutional feature map. Which is shown quite handsomely in [10] and [11].

D. Gradient-weighted Class Activation Map - Grad-CAM

Grad-CAM is just an extended or precisely more versatile version of CAM but quite a similar approach to saliency map, where we can theoretically take any other CNN and generate the Mapping for the respective outputs.

It was first proposed by Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, Dhruv Batra [12], stating it to be a generalization of the CAM.

With the GAP,

$$y_c = \sum_k w_k^c \frac{1}{Z} \sum_i \sum_j A_{ij}^k$$

And as the Gradient-weighing in the Grad-CAM suggests, the gradients for y_c are then calculated and then global-average-pooled to get,

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k}$$

The results are thus passed through a ReLU,

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k)$$

The results were then demonstrated by R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra in [13], which gives a more detailed architecture, results as well as working of the method proposed in [12].

The results obtained in [13] had a localization issue, due to the upsampling step which is then discussed and resolved in Grad-CAM++ by Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, Vineeth N Balasubramanian [14]. Though, Grad-CAM++ is more accurate on localizing the important areas of the output decision but, it adds quite a few steps and computation load on the system. Further, the older Grad-CAM approach is quite accurate enough for the purpose and has been used quite a bit so as to understand and review the interpretation and developments based off on it.

II. BODY

The review process has majorly been benefited from the IEEE, ACM and ACL based publications.

The majority of the definitions and domain knowledge is covered in the Section I. The Usage of Latent Space is quite broad in terms of methodologies it can be used for, and Saliency Maps, CAMs, Grad-CAMs itself are considered one of the applications of the same.

Further the embedding in the spatial realm itself is considered an Application of latent space, this makes it's domain so wide that almost the whole deep learning taxonomy can be based off on it. Limiting its meaning to our theme is the first step to take, thus we will majorly consider classifiers and generators whose domain revolve around images which can be defined by,

$$Image(x, y) = (r, g, b) \ where \ x, y, r, g, b \ \epsilon \ \mathbb{R}$$

And, Natural Language Representation tasks that employ the concept of Latent Space Also, the main aim of the review is to state the way Latent Spaces and Grad-CAM have affected the results and the methodologies of the publications reviewed.

The selection is quite straight forward, any publication tags matches one or more of the following keywords and is inclined towards helping the review topic is chosen:-

- Latent Space, Latent Dirichlet Allocation, Latent semantic analysis
- Autoencoders
- Generative Adversarial Networks
- Representation Learning, Manifold Learning
- Interpolation, Latent Space
- Interpretation, Deep Learning
- CAM, Grad-CAM, etc
- Visualizations, Latent Space
- Visualizations, Deep Learning Model

As it can be seen that the keywords are all over the place, and our goal is to gather how Latent Space and Grad-CAM has affected methodologies under the paper themes. Thus

bifurcating them under more general themes and then weakly chronologically ordering them is the deduced way, as strictly chronologically ordering won't give us concrete conclusions and methodology of each publication will be quite similar to differentiate.

The following themes have been chosen to review upon,

- Autoencoders and GANs
- Representation Learning
- Visualizations of Latent Space and Grad-CAM

A. Autoencoders and GANs

Autoencoders and GANs are quite literally playgrounds based off on Latent Space of a Deep Learning model. Where in Autoencoder, the model compresses the data into a latent space representation and then extrapolates the latent space back into data and in a Generative Adversarial Network, a Generator and Discriminator(or Classifier) pass on data back and forth, usually through the latent space in-order to train the generator on the problem at hand.

Like in, "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram [20]" The authors train a GAN to generate spectrograms based on the adversarial input function provided to it, practically making a text to speech network.

The WaveGAN, is accurate an fast but needs quite the training data and time which is the general rule of thumb for training GANs, they require data and time to train for accuracy to prevail, but the authors of the publication "Training generative adversarial networks with limited data. [21]" thought differently, while also dealing with the discriminator overfitting, by attaching an adaptive non-leaking invertible augmentation pipeline to the generator and discriminator, i.e., on the latent space itself and they were able to control divergence and stablize the training process through it.

While adaptive augmentation is quite novel, StyleGANs are currently heated to crisp how else would you describe a network that added more noise to remove noise. Well not exactly, but the StyleGAN Proposed in "A Style-Based Generator Architecture for Generative Adversarial Networks [22]", is able to take in noise from one end and then result in detail on the other, practically merging the inputs, with the noise being described as style and the phenomenon called style mixing using latent codes.

Under the mist of StyleGAN lies the concept called, "Self-Attention Generative Adversarial Networks [23]", SAGAN is said to be the base of StyleGAN where self attention simply means drawing out context based off on the models own output, this enables long range dependency in attention driven tasks to generate images and practically beats the state of the art GAN models while doing so.

It just shows how much uncharted territory the latent space, the space of infinite dimensionality has towards it and it is not just GANs, Autoencoders are quite the show as well like, "3D MRI brain tumor segmentation using autoencoder regularization. [24]" where the autoencoder is used as an regularization mechanism in training the CNN, Although the CNN itself is structured as an encoder-decoder network, taking in scans and bringing out tumor locations.

B. Representation Learning

Representation learning is quite a recent perspective, it is as simple as comparing them to the usual deep learning models and finding a few new layers on it. Though internally it is not as simple as it sounds. If we take in any complex task, say speech recognition the very first step will be to extract the data representation from the raw audio excerpts, this step is as important as saying that one needs to train a model inorder for it to function. But is it though? Do we really need to train the model? do we really need to extract features beforehand?.

Well the researchers have also asked the same questions again and again in different scenarios to hit upon the Representation Learning domain, and it has been beautifully hypothesized by, Y. Bengio, A. Courville and P. Vincent, in "Representation Learning: A Review and New Perspectives. [25]" that, "There is a need for preprocessing steps because different representations can entangle and hide more or less the different explanatory factors of variation behind the data[25]".

But we already know about latent spaces and its applications like autoencoders, GANs, etc, it just practically behind the imaginary wall when instead of MFCCs or TF-IDF, we'd just use the latent space representation instead, which is what is done in, "metapath2vec: Scalable representation learning for heterogeneous networks. [26]", well not to the word, but basically a step towards it. As the name suggests, the network learns to disambiguate the metapath of raw data, employ representation learning and vectorize the input to generate an output and is also scalable.

Moving away from the traditional meaning we have, "Unsupervised Representation Learning by Sorting Sequences. [27]", though the base concept is the same, the authors build a model that learns unlabeled sequences from raw data, here it is a video. Practically, they send in a video through the model, then shuffle the frames in it and send it again wanting for the model to train itself on how to determine the order of frames. Thus, showing long term evolution. Adding to it, we have "Representation Learning by Learning to Count. [28]", where the authors, feed an input image for training, then feed a transformed input image for training where the image supervision signal should be invariant to the transformation applied.

There is another like such in, "Unsupervised Representation Learning by Predicting Image Rotations. [29]" where the authors feed in the image rotated by some value repetitively to understand the attention maps that the ConvNets generate, this method is slightly different from [28] as it is self-supervised.

Moving on to, "Large Scale Adversarial Representation Learning. [30]", the authors combine the ethics of Representation Learning with GANs, proposing their model BigBiGAN which is buit upon BigGAN, while addin an encoder as an inference model and a joint discriminator

which complements the encoder added before, thus not only discriminating the generator but also the latent space distribution and thus achieve state of the art results on ImageNet.

"Momentum Contrast for Unsupervised Visual Representation Learning [31]", also known as MoCo, basically trains a visual representation encoder (Image Encoder), by first encoding it and then querying it with a dictionary of such distinct encodings, and the word momentum comes from the slow progression of the encoder in learning the embeddings.

C. Visualization of Latent Space and Grad-CAM

Visual Analysis is the first step in solving any given problem, especially in the field of deep learning. This is practically the driving force behind the intuition of any experienced researcher, but is it the limit to build the model using the probabilistic or statistic measures. Absolutely not, that may be positive for machine learning but deep learning is an another realm of complexity where such approaches which treat the model as a black box isn't always correct. Which is exactly why, we have,

"Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks [14]", an improved version of its predecessor, with better localization of the object while corresponding it as a whole. Further, at the core it is an algorithm that traces the activations that the input have had taken to generate the particular output.

While talking about inputs and outputs, "Latent space cartography: Visual analysis of vector space embeddings [15]", talks about something in the middle where practically every hidden layer in a deep learning model is vector space embedding of data. Here, the author precisely demonstrate the proposed workflow in various case studies and in every case an additional insight explains the inconsistency.

The importance of such insights are further solidified in "ClusterGAN: Latent Space Clustering in Generative Adversarial Networks [16]", where the authors achieve K-means clustering in the latent space of the Generative Adversarial Networks, the visuals representing the clusters latent space are surprisingly smooth and distinct. Further they showed that "GANs can preserve latent space interpolation across categories, even though the discriminator is never exposed to such vectors.[16]"

Getting back to CAM, in Applied Science J. Kim and J.-M. Kim in their paper "Bearing Fault Diagnosis Using Grad-CAM and Acoustic Emission Signal [17]" referred to GANs as the determining factor in assessing their Critical Application Problem, where it is needed to ensure that the predictions are based on correct importance of the features.

These Visualizations and computations even help in building more compact and faster approaches to a problem, as described in "A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images [18]", here the the authors are able to provide predictions within 2 seconds of getting the inputs for a situation clearly critical to one's life, whilst using Grad-CAM as the complement, stating that, "we have used the Grad-CAM based color visualization approach in order to clearly interpret the detection of radiology images and taking further course of action [18]"

Such advancements and applications in interpretations of deep learning models have proven that, visualizations of the inner variables is equally as the exploratory visualization performed at initial steps of every single deep learning project. And various approaches have shown that these Visualizations of latent spaces and Grad-CAMs are consistent in interpreting the model better. Which is clearly shown in, "EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by Grad-CAM [19]", where the authors specifically state that, "The pixels of the localization map correspond to the spatial regions where the electrodes are placed, we select the channels that are more important for decision-making [19]", which comes to be one of the most difficult part of their process while also helping them, reducing the channels into half whilst maintaining accuracy and improving decision speeds.

III. CONCLUSION

From all the reviews and results discussed in this paper, it is quite evident that the different applications of latent spaces are quite efficient in their performance and the scalability in terms of the methodologies used as well as the multimodality of each is not just for show. Several researchers have produced state-of-the-art results out of the approach chosen and at some places there is no competition to adhere. Further, the use of Grad-CAM is quite extensive in several cases and in representation learning model interpretation is something unfathomable. The usage of Grad-CAM has not just improved the model accuracy, but even the turnaround time, training time, decreased complexity and provided assurance of the results produced even at quite critical tasks. All at once, the effectiveness of operating in he latent space is unquestionable and so the improvements that Grad-CAM brings to a research, there are several research methodologies that still revolve in the black box of deep learning, and some of them are even stated in the Section III.B

REFERENCES

[1] Grenander, Ulf. Probability and statistics: the Harald Cramér volume. Almqvist & Wiksell, 1959.

[2] Y. Liu, E. Jun, Q. Li, and J. Heer, "Latent Space Cartography: Visual Analysis of Vector Space Embeddings," Computer Graphics Forum, vol. 38, no. 3, pp. 67–78, 2019, doi: 10.1111/cgf.13672

[3] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, Nov. 1998, doi: 10.1109/34.730558.

[4] Zeiler, Matthew D., and Rob Fergus, "Visualizing and understanding convolutional networks," European conference on computer vision, Springer, Cham, pp. 818-833, 2014, doi: 10.1007/978-3-319-10590-1_53.

[5] Simonyan, K., et al. "Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR, pp. 1–8, 2014.

[6] Springenberg, Jost Tobias, et al. "Striving for simplicity: The all convolutional net.", arXiv preprint arXiv:1412.6806, 2014.

[7] Kindermans, Pieter-Jan, et al. "The (un) reliability of saliency methods.", Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, Cham, 2019. 267-280.

[8] A. Ghorbani, A. Abid and J. Zou, "Interpretation of Neural Networks Is Fragile", Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3681-3688, 2019, doi: 10.1609/aaai.v33i01.33013681.

[9] Lin, Min, Qiang Chen, and Shuicheng Yan, "Network in network.", arXiv preprint arXiv:1312.4400, 2013.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921-2929, doi: 10.1109/CVPR.2016.319.

[11] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu and J. Jiao, "Soft Proposal Networks for Weakly Supervised Object Localization," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1859-1868, doi: 10.1109/ICCV.2017.204.

[12] Selvaraju, Ramprasaath R., et al. "Grad-cam: Why did you say that?." arXiv preprint arXiv:1611.07450 (2016).

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

[14] A. Chattopadhay, A. Sarkar, P. Howlader and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839-847, doi: 10.1109/WACV.2018.00097.

[15] Liu, Yang, et al. "Latent space cartography: Visual analysis of vector space embeddings." Computer Graphics Forum. Vol. 38. No. 3. 2019.

[16] Mukherjee, Sudipto, et al. "Clustergan: Latent space clustering in generative adversarial networks." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

[17] J. Kim and J.-M. Kim, "Bearing Fault Diagnosis Using Grad-CAM and Acoustic Emission Signals," Applied Sciences, vol. 10, no. 6, p. 2050, Mar. 2020 [Online]. Available: http://dx.doi.org/10.3390/app10062050.

[18] H. Panwar, P. Gupta, M. Siddiqui, R. Morales-Menendez, P. Bhardwaj and V. Singh, "A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images", Chaos, Solitons & Fractals, vol. 140, p. 110190, 2020, doi: 10.1016/j.chaos.2020.110190.

[19] [5]Y. Li, H. Yang, J. Li, D. Chen and M. Du, "EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by Grad-CAM", Neurocomputing, vol. 415, pp. 225-233, 2020, doi: 10.1016/j.neucom.2020.07.072.

[20] R. Yamamoto, E. Song and J. -M. Kim, "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6199-6203, doi: 10.1109/ICASSP40776.2020.9053795.

[21] Karras, Tero, et al. "Training generative adversarial networks with limited data." arXiv preprint arXiv:2006.06676, 2020.

[22] T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4396-4405, doi: 10.1109/CVPR.2019.00453.

[23] Zhang, Han, et al. "Self-attention generative adversarial networks." International conference on machine learning. PMLR, 2019.

[24] Myronenko, Andriy. "3D MRI brain tumor segmentation using autoencoder regularization." International MICCAI Brainlesion Workshop. Springer, Cham, 2018, doi: 10.1007/978-3-030-11726-9_28.

[25] Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, Aug. 2013, doi: 10.1109/TPAMI.2013.50.

[26] Dong, Yuxiao, Nitesh V. Chawla, and Ananthram Swami, "metapath2vec: Scalable representation learning for heterogeneous networks." Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017, doi: https://doi.org/10.1145/3097983.3098036.

[27] H. Lee, J. Huang, M. Singh and M. Yang, "Unsupervised Representation Learning by Sorting Sequences," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 667-676, doi: 10.1109/ICCV.2017.79.

[28] M. Noroozi, H. Pirsiavash and P. Favaro, "Representation Learning by Learning to Count," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5899-5907, doi: 10.1109/ICCV.2017.628.
[29] Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." arXiv preprint arXiv:1803.07728 (2018).
[30] Donahue, Jeff, and Karen Simonyan. "Large scale adversarial representation learning." arXiv preprint arXiv:1907.02544 (2019).
[31] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9726-9735, doi: 10.1109/CVPR42600.2020.00975.