Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2019-440-AC1, 2020
© Author(s) 2020. This work is distributed under
the Creative Commons Attribution 4.0 License.

Atmospheric Measurement Techniques Discussions

Interactive comment on "Unsupervised classification of snowflake images using a generative adversarial network and K-medoids classification" *by* Jussi Leinonen and Alexis Berne

Jussi Leinonen and Alexis Berne

jussi.leinonen@epfl.ch

Received and published: 15 April 2020

We thank Referee 1 for their comments that helped clarify a number of key points in the paper. Please find our response below, with original comments quoted in *italics* and our answers and explanations of the changes made to the manuscript in normal font.

Line 66: Can you clarify the meaning of "deep" CNNs?

"Deep" in neural network terminology refers to a network with many successive layers. We have now defined this above in the paragraph beginning with "The development of convolutional neural networks..."

Line 78: Can you clarify the meaning of "latent variables"?

In GANs that recover the noise variables using the discriminator, the noise becomes semantically a set of latent variables encoding the properties of the input dataset. We now explain this in some more detail.

Line 108: How does a simple constraint on the diameter of an imaged snowflake ensure that the entire snowflake is within the image frame? Couldn't a snowflake that intersects the edge of the frame have a diameter in this range?

The images that the MASC takes are actually very big compared to the size of the snowflakes. The P17 processing crops snowflakes from these images such that the snowflakes are completely contained in the frame (snowflakes intersecting the edge of the original image are rejected). We neglected to mention this in the submitted version, it is now explained in this paragraph.

Line 115: What is the specific purpose for downsampling? Is it simply to make the classification processing more computationally tractable?

Yes, the main reason is to reduce the computing burden. We believe that this can be done without losing too much information because the MASC images are usually at least slightly blurry and thus the true resolution of the images is not quite as good as the pixel resolution. We have added this explanation here.

Line 153-155: It's not clear how "neighborhoods" are defined in the context of a set of snowfall image inputs. Can you elaborate?

This explanation is perhaps confusing so we have reworded it, avoiding the use of the ambiguous term neighborhoods: Pooling layers reduce the spatial dimensions of their input by dividing it into $M \times N$ (typically 2×2) rectangular region arranged in a grid, then applying a pooling operation such that each rectangle is mapped to a single value in the output image. Usually, either the average or the maximum of the rectangle is used as the pooled value. Pooling operations can sometimes be replaced by strided convolutions, which skip some points (e.g. every other point) of the input to reduce the spatial dimensionality of the output.

To be clear, the input to a pooling layer is usually not a snowflake image as such, but instead an intermediate stage of processing in the CNN.

Lines 183-184: What is "z"? (see also the comment regarding line 264 below)

The variable $\mathbf{z}$, the input to the generator (i.e. the noise / latent variable), is defined at the end of the previous paragraph. We have added a note there that $\mathbf{z}$ denotes the noise.

Lines 221-222: Is it actually true that the distance between each point and its nearest centerpoint is minimized? I don't believe that is what is imposed by equation 14. But what does lowercase "n" represent in equations 14 and 15?

The meaning of this sentence was a bit ambiguous and it has been reworded.

The lowercase n represents “nearest”. $\mathbf{c}_{\mathrm{n},i}$, as defined in Eq. 15, is the centerpoint closest to the data point $\mathbf{y}_i$. This has been reworded for clarity.

Lines 260-262: It would help here to have some additional context describing the purpose of a styling block. What is achieved by upscaling the image and processing it through the AdaIN, activation and convolution layers? What is gained by upscaling the image?

The original feature maps that the generator starts with are 4×4 pixel size and encode the image semantically on a feature level, as learned by the network. To transform the feature maps into a 128×128 image, we need to:

1. Increase the image size; this is achieved by upscaling the image by a factor of two at a time.
2. Process the deep feature-level representation into an image through a series of

intermediate-level representations; the activation–convolution–AdaIN operations are responsible for these transformations.

We have rewritten the paragraph in a way that hopefully makes this somewhat clearer. A further explanation of the convolution and activation layers can be found in Sect. 3.1.

Line 264: Earlier (line 211), z is described as the latent distribution. Here it is described as noise, and this seems inconsistent. Can you clarify? What is the difference between "style" and "latent variable"?

Admittedly, the use of $\mathbf{z}$ in the manuscript was somewhat inconsistent. We have tried to clarify this in the revised version as follows: The generator input, as a whole, is called $\mathbf{z}$; this consists of a latent vector $\mathbf{z}_{\mathrm{l}}$ (which is recovered by the discriminator) and additive noise $\mathbf{z}_{\mathrm{a}}$ (which is not recovered). We have modified this paragraph to specify this and also made changes throughout the article such that this notation is now used consistently.

Line 340: Does it not appear that there is a threshold near K=3, 4 or 5? There seems to be a substantial change in the slope of the loss function near these K values. Why would this not be seen as an indication of the actual number of medoids?

There is a change in the slope at nearly every K, but we do not think any of these deviate significantly enough from normal to say unambiguously that they indicate a “correct” number of medoids.

That said, there does seem to be a slightly larger change at $K = 3$ in both the K-medoids and in the hierarchical clustering started at $K = 16$ (orange line in Fig. 4). We can get some insight to the nature of that change in the slope by examining Fig. 5. As mentioned in Sect. 5.2.2, the hierarchy in this case consists of three main branches. Therefore, the smaller change in slope at $K > 3$ is likely a result of diminishing returns after the clustering has identified these three main groups. Regardless, the discussion at the end of Sect. 5.2.2 demonstrates that there is a significant benefit in going from

$K = 6$ to $K = 16$, and therefore while $K = 3$ may be optimal in the sense that it provides a large change in slope that indicates the presence of three main groups of snowflakes, it is not a particularly suitable choice for our purposes.

Finally, I have two technical comments:

Figure 1: I believe the caption is wrong. Panel (a) appears to be the discriminator, and panel (b) the generator.

Thank you, this has been corrected.

Line 232: Should this be "understood as a variant"?

Yes, this was also corrected.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2019-440, 2019.