LINKING MODEL PERFORMANCE TO UNCERTAINTY

Complete the analysis in the code block below. Write short (~1 sentence) answers to the following questions and include them in your Debiasing Faces Lab submission to complete the TODOs!

1. What, if any, trends do you observe comparing the samples with the highest and lowest reconstruction loss?

   ANSWER: There are mostly white people who are looking pretty directly into the camera. Their faces take up most the screen and there are no obstructions such as glasses, or hair that covers the face.

2. Based on these observations, which features seemed harder to learn for the VAE?

   ANSWER: There are a lot of faces in the high loss group who are either not looking directly at the camera or have something that blocks part of the face such as a hat, glasses, or long hair. There are also more bright colors than in the other group.

3. How does reconstruction loss relate to uncertainty? Think back to our lecture on Robust & Trustworthy Deep Learning! What can you say about examples on which the model may be more or less uncertain?

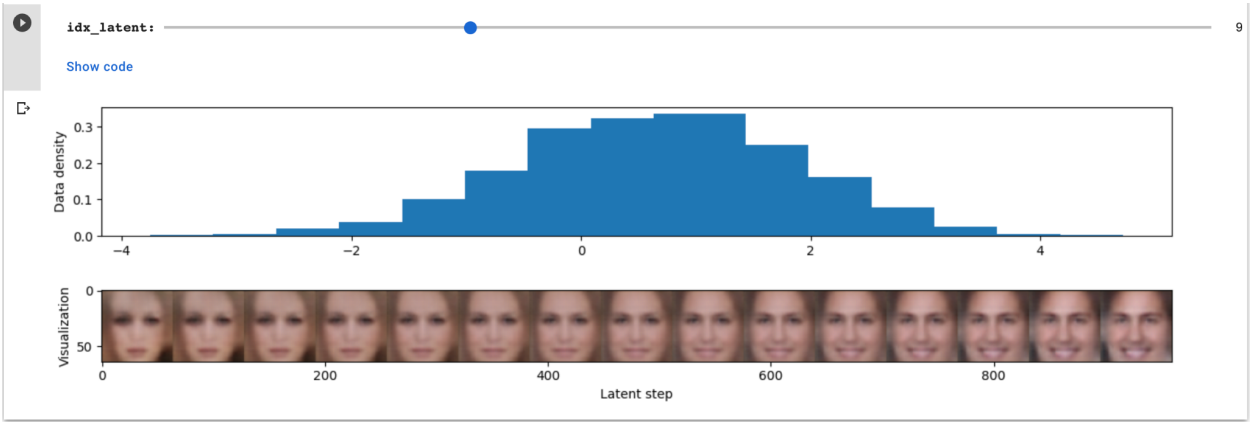
   ANSWER: High reconstruction loss indicates that the model isn't good at recognizing certain features and representing them effectively in the latent space. This could be because those features are uncommon in the training data, making the model less good at reconstructing or classifying those images.

UNCOVERING HIDDEN BIAS

Complete the analysis in the code blocks below. Carefully inspect the different latent variables and their corresponding frequency distributions. Write short (~1 sentence) answers to the following questions and include them in your Debiasing Faces Lab submission to complete the TODOs!

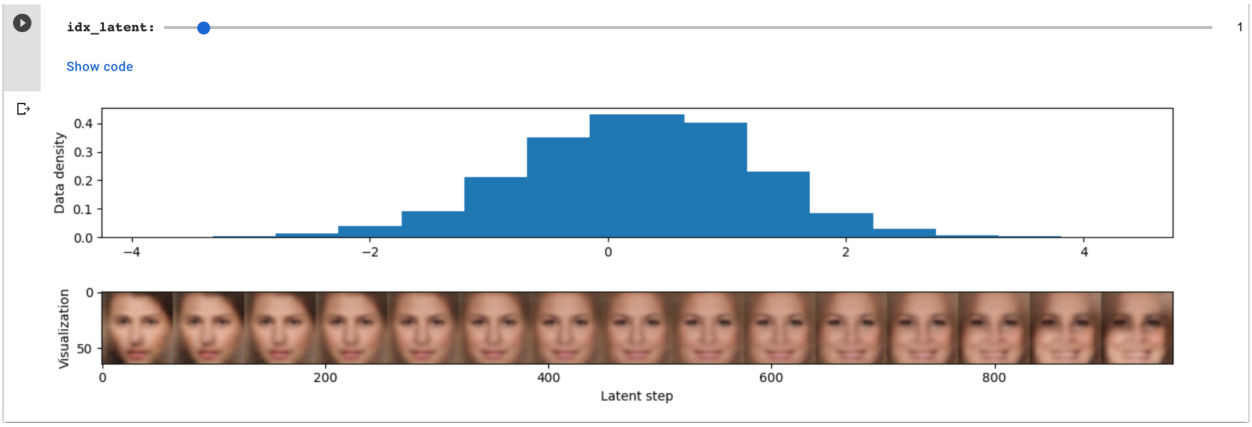
1. Pick two latent variables and describe what semantic meaning they reflect. Include screenshots of the realizations and probability Distribution for the latent variables you select.

Example 1:



This variable seems to reflect both the amount of smile in a given face (less to the left, more to the right) and the amount of feminine vs masculine features displayed (more feminine to the left, more masculine to the right).

Example 2:



I find this example really interesting. On the left, the variable appears to represent, at least partially, a beard-like structure. On the right, it seems to represent hair covering the face.

2. For the latent variables selected, what can you tell about which features are under- or over-represented in the data? What might this tell us about how the model is biased?

   ANSWER: In my first example of a latent variable, the data implies that the wide smiles appear little, because the smile associated with the variable has a low data density (it is on the far right of the distribution).

In my second example, the characteristics that have low data density are beards and long hair around the face. This similarly implies that these characteristics are not well represented in the data.

To take a skeptical view, it also seems fair to make the assumption that the features that appear as latent variables are at least somewhat represented in the data because that is how they are learned. For example, in the first latent variable I selected, it can be inferred that both men and women are represented in the dataset, and that there are images of people smiling and not smiling.

But, with that being said, features that have low density (along with features not present at all) are good indications of the potential biases in the model. By analyzing the latent space, it is clear that some features are more common in the dataset compared to others, which makes sense when we refer back to the relative accuracy of some photos fed through the model relative to others.

3. For the latent variables selected, how do these feature distribution differences affect classification performance? What, if any, general trends do you observe across the latent variables?

ANSWER: Photos with smiles seem to be more accurate than photos without, which is not what you would expect considering that my answer to #2 implies that smiles are under-represented in the dataset. Beards (and possibly also hair covering the face) however, do negatively affect performance which is what you would expect based on my response to #2.

Based on my judgment by looking at more of the outputs on the lab, there seems to be strong support that the characteristics that fall on the ends of latent variable distribution, have overall worse accuracy.

4. Based on these observations, please describe your understanding of the bias of the facial detection classifier.

ANSWER: Put simply, the idea is that if a characteristic isn’t well represented in the data, that will affect the model’s ability to classify for instances that contain that characteristic. The lab

displayed that you don’t need to manually investigate to do this. You can automate that process, even with unsupervised training.