We thank the reviewers for their feedback. In our paper, we show the existence of different examples that map to the same feature activation in a neural net with ReLU activations, which can be viewed as, in some sense, the *opposite* of adversarial examples, and demonstrate that there could be infinitely many such examples lying in convex polytopes. All reviewers found the paper "interesting", and various reviewers commented that "the phenomenon identified here seems interesting and worth considering", "the paper gives a well reasoned and executed solution" and "the approach is interesting and novel".

R1 and R3 asked about the possible applications of the findings, and we briefly discuss several different ways in which our observation can be used below (we will discuss these more extensively in the camera-ready):

1. **Representative Data Collection:** The size of the colliding polytope around a training example can be used to discover regions of the data space where insufficient training examples have been collected. More concretely, if the size of a colliding polytope around a training example is large, then the neural net could be over-generalizing in the neighbourhood of that example, and so the model may not be accurate in this neighbourhood. This can be used to inform the end-user whether the prediction in this neighbourhood should be trusted (as suggested by R1), or to guide data collection, so that more examples in this neighbourhood are collected in the future.
2. **Design of Regularizers:** The insight our method reveals can lead to the design of regularizers that mitigates undesirable over-generalization. For example, one could try to minimize the size of the colliding polytopes, by discouraging the hyperplanes associated with each hidden unit from being near-collinear (i.e. having highly positive cosine similarity) with other hyperplanes. This can be also used to guide architecture selection.
3. **Identification of Vulnerable Training Examples:** The proposed method can identify the training examples that a neural net depend most on, which could have large colliding polytopes around them. This can help detect outliers and training examples that could have been mislabelled or adversarially tampered with, or legitimate training examples that could be vulnerable to manipulation due to how much the neural net depends on them.

We respond to the main remarks by each individual reviewer below:

R1: We agree that the existence of feature collisions should be well-known (and it certainly was to us as authors), but found from conversations with others that it is not actually known to many people. So, the purpose of this paper is to highlight this phenomenon, which we hope will spark further investigation in this direction. It is possible that this phenomenon was pointed out in a previous paper; we weren't able to find an example of this in our literature search, but would be happy to include a reference to it if you are aware of such a paper.

For the results in Sect. 3.1, yes, we do find similar behaviour with more complex convnets. We tried the same experiment on LeNet, and found that 100% of the sampled input within the polytope are successful collisions and that they are all classified the same as the target example with a confidence of 100%. The average distance between the polytope corners is actually larger – it is 1.24 when the source and target examples are the same and 7.45 when the source and target examples are different (which is quite large, since the average distance between MNIST examples is 10.21).

Thanks for the suggestion – the findings do hold approximately for any activation function that saturates, i.e. sigmoid, tanh or ELU. We will comment on this in the camera-ready. If the polytope is around a target example, then it would not be empty by construction because the target example must be contained in the polytope. However, it is possible that when the activations for all hidden units are positive, the polytope only contains a single point, namely the target example.

R2: Thank you for your suggestion! We will formalize the observations in Sect. 2 as propositions in the camera-ready.

To respond to your questions, if all hyperplanes are parallel, the polytope would still be convex, but would not be bounded. As mentioned on L84 and L91-93, the implication is that there could be colliding examples that we would not be able to generate, but we can still generate colliding examples from a (bounded) subset.

If we were to enforce the equality constraints on the hidden units with positive pre-activations and the hidden units with non-positive pre-activations other than the current hidden unit i strictly, and assuming the hyperplane associated with hidden unit i is not parallel to the hyperplanes of other hidden units, then the solution is an exact corner of the polytope. The reason why the activations are not identical to the target example is because the equality constraints are not enforced strictly.

For Sect. 4.1, the softmax is over just over the k-nearest neighbours of the patches of the source image at each spatial location, where k-NN is performed over the dataset of 37,011,074 patches.

R3: Thanks for the suggestion! We agree feature collisions on language data could be easily distinguishable perceptually and would be interesting to explore.