I have included the inference code now and also a new README section called 'Post competition findings', where I thrash myself for having forgotten to train the model in actual train mode after the first round. Seems my original contribution to the solution involving paired attention doesn't outperform the vanilla UNITER model as I previously thought. That leaves all credits to be going to the UNITER guys.

Please see section 3 answers below.
1. Who are you (mini-bio) and what do you do professionally?
You can find me at http://www.vladsandulescu.com/about/

2. What motivated you to compete in this challenge?
I am doing some research on multimodal models and the competition provided a good opportunity to learn more about this through a practical task.

3. High level summary of your approach: what did you do and why?
I adapted a couple of multimodal pretrained architectures and did simple ensembling to smooth out the predictions. I also tried LXMERT and VLP, because I wanted to see how the architectures and pretraining procedures impacted the finetuning for a downstream task.

5. I used mostly an NVIDIA V100 32GB RAM GPU. But I do expect this to work on 16GB RAM GPUs also, by changing the batch size.
Memory(GB): 32
Ubuntu 16.4
Train duration: depending on the number of training steps, it took max 1 hour.

6. They are mentioned in the README file

7. I tried using the MMHS (https://gombru.github.io/2019/10/09/MMHS/) dataset to pretrain/finetune the model, but it didn't work well. Seems the dataset is highly biased towards certain offensive keywords, while the competition dataset has many more subtleties.

8. No

9. I tracked the AUC on validation set, as well as Accuracy

10. No, it should be self explanatory and straightforward to replicate

11. Not really

12.
* I would look at the impact of the pretrained dataset on my downstream task. These guys make some very good points regarding this: https://arxiv.org/pdf/2004.08744.pdf. It seems to benefit UNITER to be pretrained on multiple datasets: CC, COCO, VG and SBU, compared to LXMERT

for example which is not trained on CC for example. In general I noticed CC was key to finetuning large pretrained models for this task, probably because it's the largest dataset.
* I would also try to improve the paired attention technique I tried in the competition but with captions from a newer image captioning model, e.g. VLP.
* I'd also try data augmentation to make the model more robust and increase the dataset size
* I would look into different image understanding and ROI extraction techniques, besides the bottom-up-attention.

13. https://github.com/vladsandulescu/hatefulmemes

14. https://arxiv.org/abs/2012.13235