Response to Review #1:

Reply to weakness argument 1:  
Discretising a kernel density estimate, rather than blurring the bins would be reasonable. Though a Von Mises kernel rather than Gaussian kernel would be needed for hue.  
This is a good idea with simply did not occur to us.  
Perplexity on the validation set was used when selecting Ω.

Reply to weakness argument 2:  
From a theoretical stand point, this is correct. All bens should contain some nonzero amount of probability mass.  
However, for bins sufficiently far from any observations, that value is smaller than can be represented on a 32bit float. (The paper fails to mention this, thank you for bring it to our attention)

Reply to weakness argument 3:  
Yes, it should. (For reference it is the TensorFlow.jl default, Gaussian, mean 0, std dev 0.01)

Thank you additionally for your close reading of the text and your additional comments.

Response to Review #2:

Reply to Weakness 1/Question 1:  
It does have the limit that all tokens used must appear in the training data in some combination.

Question 2:  
Multimodal and multi-token are unrelated. Multimodal refers to the distribution in HSV space. Having two or more peaks (or “modes”). “Green-ish blue” is multi-token (3 tokens in fact). “Green-ish” is multimodal – a peak to the left and the right of green. (“Green-ish” is also multi-token)

Weakness 2/Question 3a:  
The reverse task is covered by existing works discussed in section 2. In particular McMahan and Stone (2015) and Kawakami et al. (2016).

Weakness 3:  
HSV was chosen because of its channels coming close to meeting the conditional independence assumption. This is (indirectly) discussed in section 3.1

Weakness 4/Question 3b:  
Indeed, we model the fact that differences exist within the population, but do not attempt to model individuals within the population. That would be very interest future work, using demographic information.

Weakness 5:  
This is standard usage of RNN for NLP.

Question 4:   
It would indeed be good to include additional figures showing each term in insolation. That would better show compositional effects. We chose these figures to highlight the multimodal nature of some colors.

Reviewer 3

For multimodal colors, i.e. with a distribution has two or more peaks:  
a naïve point estimate would result in a color in the “valley” between.  
Selecting the highest peak would be better, but it would be significantly wrong for the other peak(/s).

For wide variance colors, a point estimate gives no degree of confidence.

For asymmetrically distributed colors: a point estimate would give no indication that most of the probability mass was to one side. (a naïve system will also miss the peak)

All colors are at least one of multimodal, wide variance or asymmetric.  
Including point estimate based evaluations would mislead the reader as to the true nature of the problem.

Weakness 1:  
Lines 083-096 give an example of a possible application. But the focus of this work is on the pure task.

Weakness 2:  
An evaluation based on point estimate (e.g. MSE) would be misleading. Not all errors are equal.  
E.g. A point estimate that is on the second highest peak is much better than a closer estimate in a low probability valley. Errors need to be weighed using the true probability (as is done in perplexity).

Weakness 3:  
A point estimate system is performing a different task.

Reviewer #3 states that as they found a non-anonymised arxiv version of the paper, and thus that the ACL author guidelines have not been followed.

The non-anonymised arxiv version was posted well before the anonymity period. I declared its existence at declare its existence at submission time.

Did I violate the guidelines?