Editors summery

- Rename the operational upper bound and change the interpretations of its results to reflect that it does not have any unfair advantage over the other models under evaluation (assuming that the first reviewer did not mis-interpret your paper).

We have renamed the operational upper bound to non-compositional baseline.
The section introducing it has been updated to reflect further details of why one would expect that this “baseline” would be very hard to beat.
Reviewer A’s point that it does have overall outright less information is well made.
However, the non-compositional model does have a sizeable advantage, as it effectively bypasses the natural language processing part of the task, and is left with only the point estimation or distribution estimation modelling problem for which it uses the well established methods.
The determining of how to combine the words is a core part of this problem and is the difference between the RNN, CNN, and SOWE models.
It is as we found, a difficult problem.
A section has also been added to the conclusion, highlighting that indeed none of the models are able to perform this composition part well enough to beat not using composition. Of course due to the combinoric natural of language, composition is required, as is shown in the unseen combinations task.

- Carry out a deeper analysis to back up your explanations of the relative performance of the various models.

- Include results on the training set. If GRU is not enough to fit the data, LSTMs are recommended.

Results on the training set have been included. Both final results and plots.
Additional discussion has been added based on this, which supports the explanation of the relative performance.
LW: Do I need more
Considering the training results, we believe the GRU fit adequately and have thus not added additional results for an LSTM. While the RNN did not perform as well as the other models on the training set, it was within a small margin.
LW: Take a look at the results on training set, and tell me if I should train an LSTM also. It would be annoying to run more experiments but not impossible.

- Include corpus statistics after preprocessing.
A variety of additional corpus statistics have been added. We believe they are informative.

Detailed comments on revising the paper

Reviewer A:

Weaknesses:
:

The most consequential flaw in this paper surrounds its treatment of the "operational upper bound", the model that constructs a smoothed distribution of the data for each color term without considering the compositional structure of the color terms. The paper presents its results as being a kind of oracle (or ceiling) performance, an aspirational goal for the other models to chase. I claim it is no such thing: Sec. 4.3 makes it clear that the OUB has access only to the training data. A correct oracle evaluation would involve a histogram built from the *test* data. This would establish a lower bound on the perplexity of any model with a particular histogram resolution.

In fact, the OUB's input representations are strictly less informative than the CNN and RNN models. As the authors point out in Sec. 4.3, giving the model additional information risks overfitting, so it isn't a contradiction that the OUB outperforms the other models. I see this as remarkable, though, while the paper treats it as expected! A better characterization of whether the models are overfitting or underfitting would also serve as a test of the claim, in Sec 6.2, that the poor performance of the RNN and the success of the SOWE can be attributed to difficulties in learning via gradient descent. How do the RNN, SOWE, and OUB compare on their training set loss? Optimization difficulties would predict a poor training set loss in addition to poor perplexity on the test set.

See Above

Revisions to be Required:
:

First, assuming my assessment of the operational upper bound accurately describes the model that was used in the paper's experiments and is not a misinterpretation, this model should be renamed (perhaps calling it a non-compositional or atomic baseline), and the interpretations of its results should be changed to reflect the fact that it does not have any unfair advantage over the other models under evaluation.

See Above

Second, the authors should do a deeper analysis to back up their explanations of the relative performance of the various models--in particular, the claim that the SOWE is "easier to learn...from a gradient descent perspective" (Sec. 6.2),

We have added a plot of the training performance of the models, and corresponding discussion. In brief, it can be seen that during training the loss decreases much faster for the SOWE model, which shows that it is a “easer” error surface to gradient descend.

but also e.g. that the CNN is "better at capturing the shared information about term usages" (Sec. 6.2.2),

TODO: NEED TO SAY SOMETHING ABOUT THIS IN THE PAPER.

and that training with random minibatches results in "increased resilience to local minima" (Sec. 6.2.3).

We have added a supporting citation for this claim (LeCunn, and Bottoui, 1998, Efficient Back Propagation. (Section 4.1)).

Revisions to be Encouraged:
:

The result that the neural models learn smooth functions of the output space is interesting, and could use some more analysis. The authors speculate (Sec. 6.2.3) that training the distribution models using one-hot outputs would harm the models' ability to capture this smoothness. I would be interested in seeing this prediction tested, since my bet would be that the smoothness comes from representations shared between different inputs (e.g.: "pink-purple" may have an anomalous spike at saturation 130, but the mapping from input to output layer reflects contributions from other inputs using "pink", "-", and "purple", which do not have that spike). Performing this experiment would also eliminate the differences between the point estimation and distribution models mentioned in that section, which seem non-essential, and make them more comparable.

TODO: NEED TO SAY SOMETHING ABOUT THIS IN THE PAPER.
I think I generally agree with the first part of the statement, using one-hot should not make a difference. Possibly that section is unclear and should be deleted.

Additionally, an investigation of the mutual information between the three color dimensions could be a better test of the conditional independence assumption (Sec. 3.3.1, Appendix 1.1) than Spearman correlation.

To the best of our knowledge it is not possible to perform a mutual information test, without assuming a distribution for the data.
An alternative would be distance correlation. This would not have the hue related issues.
Should I Do that. Do I have software for that?

5 Minor Revisions Required [Help]
:
The second paragraph mentions that something is "indispensable for executable semantic parsing"--either color understanding itself, which is only really true in the most extreme visual grounding tasks, or "natural language understanding", which I would argue is a superset of executable semantic parsing rather than prerequisite for it.

This sentence was is indeed unclear and has been reworked.

The claim that color
distributions "are almost always multimodal or asymmetrical" (middle of Sec. 1.1) is also misleading--multimodality seems to be rather uncommon in the Munroe data, and it could arguably be attributed to pragmatic interference from a more specific term, as in the case of "greenish" with "green".

While multimodality is rare in hue (though very interesting where it occurs),
it is much more common in the saturation and value channels.
Furthermore the asymmetry is ubiquitous in the saturation and value channels.
This section has been reworded to clarify this.

Later in the introduction, the phrases "address these linguistic phenomena around the short-phrase descriptions of a color" (last paragraph of Sec. 1) and "qualify our estimate of the distribution" (smoothness paragraph in Sec 1.2) should be reworded to be clearer.

This has been reworded to be clearer.

In Sec. 3.4.1, the meaning of "basic RNN" should be clarified--is this an LSTM? To say that the LSTM has "longer short term memory" than the GRU would be a misinterpretation of the terminology: both can learn moderately long-distance dependencies, at least partially because of the fact that both have the additive cell update term that prevents vanishing gradients (see

e.g. Chung, Gulcehre, Cho & Bengio 2014).
This has been reworded to be clearer. By Basic RNN we mean an RNN without any form of gating.

More detail about how the dev set for early stopping (Sec. 4.4.2) is
constructed (does it consist of the next 100 rarest descriptions?) would
clarify how the dev set is isolated from the test set.

More detail has been added. It was not constructed like the test set, but rather like the training set. Constructing a dev set using the next 100 rarest descriptions would be a in interesting alternative, but not one we pursued.

Also, tables 5 & 6 mention an "extrapolation subset". Is this the same as
the unseen combinations test set?

It indeed should have read unseen combinations tests set. This has been corrected.

Finally, the mention of document classification and sentiment detection in
the second-last paragraph of the conclusion feels like a
non-sequitur--surely color terms are rare enough in e.g. movie reviews that
they would be of little help in predicting sentiment.

This indeed was a non-sequitur, and has been removed.

6 Typographic Errors [Help]

All typographic errors have been corrected. Thank you for your attention to detail.

Reviewer C:

2 What is this paper about? [Help]
:

Since a colour description does not correspond to a single point in colour
space, this paper proposes models to predict distributions in colour space
conditioned on colour descriptions. A colour description is represented as a
sequence of word embeddings and different types of neural networks (MLP,
CNN, RNN) are applied to predict a distribution in HSV space. The model
predicts H, S, V values independently. Instead of learning distributions in

continuous space, they discretized the HSV space with histogram and predict a discrete value for each output. The models were evaluated in terms of learned predictive distributions.

3 Strengths and Weaknesses [Help]

Strengths:
:

- The motivation and implementations are presented clearly.
- Focused analysis on word order is interesting.
- Although there are many pieces of evidence that fine-tuning word embeddings improve down steam performance, it might be a good idea to fix pre-trained word vector to test the model on unseen word types.

Weaknesses:
:

- Novelty is somewhat limited. They discretize the continuous color space with histogram but it loose some information about the color space. If the aim is to learn distributions over continuous colour space, it will be better to use continuous distributions such as a mixture of Gaussians model where mean and variance of Gaussians are predicted by neural networks.
- Based on the fact that most of the previous works used RNNs, it is quite surprising that RNN underperform bag-of-words model especially when word orders have some effect (e.g. Table 4). Since model comparison is one of the main components in the paper, it should be addressed properly. To avoid potential problem in their implementation, they should include performance on training data. The GRU is not capable of fitting training data? Or it's an generalization issue? If GRU is not enough, LSTM is an obvious candidate to try.
- They conclude that word-order is not very significant for colour names (6.2.1) because SOWE model performed better than order sensitive models. However, it's probably because the dataset do not contain enough examples to learn the effect.
- Dataset statistics (after preprocessing) is not provided. How many colors per descriptions?

With regards to the discretisation of the continuous colour space losing information. We discretise to 3x256 bins (HSV space). The original data was collected using colors displayed on RGB monitors using 3x256 bits. While there can be some loss of information in the RGB to HSV tranformation, the information lost in discretisation is negligible, each unique HSV as a triple of floating point values, corresponds to exactly one tuple of three bins. Using a truly continuous distribution such as a GMM with means and variances

predicted by the neural network is a more elegant approach, it is not one that we expect to pragmatically work better based on the loss of information. As well as elegance it has other advantages such as not being as memory expensive as constructing 3D histograms, which would thus allow us to remove the conditional independence assumption.
It is an approach we've been considering and feel deserves exploration in a future work. (We suspect 5-10 mixture components may be enough to capture all the difficulties in the unusual shape of the distribution.)

With regards to our claim that word-order is not very significant for colour names.
We find it to be closely related to your statement that the dataset that does not contain enough examples to learn the effect.
The dataset is itself a good sample of natural colour language as it exists in the wild.
The fact that it does not contain many training examples where word order is significant, is a strong argument for such examples being rare in the wild. Which in turn suggests that word order is not significant.

4 Substantive Revisions Required [Help]

Complete this section if either 'Revise and Resubmit' or 'Reject' has been recommended.

Revisions to be Required:
:
- Include results on the training set. If GRU is not enough to fit the data, LSTMs are recommended.

The training set results have been included, the GRU is fitting.
See Above

- Include corpus statistics after preprocessing.

See Above

Reviewer E:

2 What is this paper about? [Help]
:

Generating estimations of multi-word color terms through three methods of word embedding processing: sum of embeddings, CNN, and RNN. Colors are represented in two ways: point estimations, and distributional estimations as histograms of each HSV channel (assumed independent). The distributional estimation is considered to be closer to truth as color terms are ambiguous, and in combination with the sum of embeddings processing generally perform best.

3 Strengths and Weaknesses [Help]

Strengths:
:

Overall, very well written.
* Strong motivations/arguments are given for essentially every decision made choosing the parameters of the model.
* Solid background of color understanding and ambiguity provided; great job demonstrating its representation of language in general
* Full representations of distributions are visualzed well, and diagrams in general are very informative and clear

Weaknesses:
:

* Some citations are slightly inaccurate:
1) McMahan and Stone (2015) do in fact map color terms to probability representations in HSV space. They are not using discretized histograms, nor considering different methods of handling multi-word terms, but it is inaccurate to say that they only look at likely color names given a colour point.
2) Winn and Muresan (2018) are generating vectors in colorspace representing the comparatives; not providing a point estimation of a new color compared to another.
* Minor general points; the paragraphs of the contribution section feel slightly disorganized, and the latter pages of the paper have some grammatical errors and couple paragraphs have mildly repetitive sentences. Given the overall quality of the paper, it is assumed that a few more passes of edits will resolve these issues.

The related works have been clarified.
Additional editing for flow/organisation has not been done yet

5 Minor Revisions Required [Help]
:

* Would recommend showing HSV values in the point estimation (or provide a couple examples); allow the reader to have both a qualitative and quantitative analysis of at least a few data points

I have not done this, I can't actually regenerate these plots, without rerunning the experiments. But I could do this by hand.

* Could have a slightly more in-depth analysis of results, at least in comparing the best and worst results: is there a difference in performance given the number of words in the color term? Are there patterns to the best/worst results? Per input/output method?

I think there are patterns across everything, but I did not examine this carefully or record notes. I would have to rerun the training to examine this more.

* The white histograms are difficult to see - perhaps find a way to outline (maybe just a line at the top of the bar) or for the white only have a gray/black background square, or just present those bars in gray

I have not done this, I can't actually regenerate these plots, without rerunning the experiments. But I could do this by hand.

* Perhaps cite Kawakami in 1.1 as they generate color from text even though they do so through a character sequence model: interesting to contrast the underlying linguistics being examined

I do not feel this is needful. What do you think?

* 1.2 p1: Helpful here to specify that the input to SOWE, CNN, and RNN are existing word embeddings; it is ambiguous until FastText is mentioned whether you are creating the embeddings or using existing ones.

I have not done this, yet.

* 3.3.2 p3: As I am unfamiliar with uniform weight attribution, it is unclear to me what the "adjacent midpoints" would be; perhaps describing this through an example would be clearer

I have not done this, yet. It is really hard to explain this without taking up a lot of space.

* As stated earlier, some grammatical revisions to the latter pages

I expect this has been fixed by fixing the typos identified by reviewer A.

6 Typographic Errors [Help]

:

1.1 p2 s1: Comma not needed (this extra comma near the beginning of the sentence happens a few times throughout)

1.1 p3: s3: "To our knowledge The generation "

All typographical errors have been corrected. Thank you.
