Response to Submission 1424

Learning of Colors from Color Names: Distribution and Point Estimation

Lyndon White, Roberto Togneri, Wei Liu, Mohammed Bennamoun

We sincerely thank the reviewers and the editor for their constructive and well-reasoned feedback on the paper. The manuscript has been revised to address the concerns raised. The remainder of this letter responds to the feedback from the review letter. We quote liberally from the review letter, these sections we highlight in green. We begin by responding to the key issues summarised by the editor, as several of these were raised by multiple reviewers. To avoid repetition in our responses to the individual reviewers, we refer back to this section, adding any pertinent details to address particular concerns.

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.
All references to it in the text have been replaced (The term operational upper bound should no longer occur). The Section 4.3 "Non-compositional baseline" where it is introduced 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. This is an important point, as 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.

Additional analysis has been included. In particular results on the training set have been included in Figures 9 and 10. Both the final results and plots are shown for the fitting of the models on the full training set. Additional discussion has been added in Section 6.3: "Training set results". The discussion added explains how the training set error curves which supports the explanation of the relative performance, in terms of the error gradient.

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

For completeness, we have included results on LSTM. It performs very similarly to the GRU, as expected. On the consideration of the the training results, discussed above we believe the models are fitting adequately. As we are using early stopping, we have some confidence that we are training for close to an optimal number of epochs to get the best fit for these model with this set of hyper-parameters
While the RNNs do not perform as well as the other models on the training set, it was within a small margin. It is always possible to increase training set results by increasing the number of parameters of the model. But as it is the LSTM has 4x as many as SOWE and 2x as many as the CNN. The GRU has 3x as many as SOWE and 1.5x as many as the CNN. Further increasing the number of model parameters on the RNNs would put them notably out of proportion to the nonrecurrent models, and likely result in overfitting and earlier stopping.

- Include corpus statistics after preprocessing.

A variety of additional corpus statistics have been added. We believe they are informative. They have been added to each of the subsections under Section 4.4 "Evaluation Strategies and Data".

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.

We have addressed these concerns, they are well made. We refer the reviewer to our above responses to the editor's summary. We have renamed Operation Upper Bound to Noncompositional Baseline, and we have highlighted why we expect it the be a very hard baseline to beat — the composition is the most difficult part of the natural language understanding, and this bypasses it.
We include additional training set results and analysis Figure 9 and 10 and section 6.3. Also discussed in our responses to the editor's summary. Indeed as the reviewer predicted the optimisation difficulties for the RNNs are readily apparent, in their poorer training set loss. We feel the inclusion of these results are a solid enhancement to the the strength of the paper, and thank the reviewer for the recommendation.

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.

We have made this revision to Section 4.3. As discussed 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 (Figures 9 and 10), and corresponding discussion. In brief, it can be seen that during training the loss decreases much faster for the SOWE (Section 6.3) model, which shows that it is a "easer" error surface to gradient descend. A more difficult to fit problem would have a overall lower

slope and more regions with in that where the training loss flattens further (as seen in Figure 9, for the GRU and RNN around epoch 75). These flattened regions correspond to “near saddle points” where it is harder for to determine how to modify the parameters to decrease the error. The final loss when early stopping terminates training, is worst for the RNN models which we claimed to be harder to optimise.

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

On further consideration, the difference in performance of the CNN on the unseen combinations task, with improvement when trained on the restricted training set, vs the the full training set, is too small to make any claims about the reasoning (A difference of MSE 0.02). We have removed this sentence.

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. The relevant section of LeCunn, and Bottoui, 1998, is their 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.

This point is well made. It may indeed be that training using one hot improves the smoothness by forces that knowledge deeper into the network, allowing the effective smoothing effect to be more specific to the color name in question. We have not included additional results on this as it would require substantial new experiments to properly investigate and confirm the effects. We believe such experiments are better discussed in a separate work where they can be covered in full detail. We have revised that claim in the new section 6.2.4 to discuss in more detail the possible outcomes of such experiments.

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. A mutual information test would indeed be more suitable were this not the case.
The assumption of conditional independence of colour channels has been made in prior works (see Section 3.3.1). In this case it does not invalidate the assessment of the models, it merely limits their absolute performance.
Our appendix contains only a simple limited test to support the assumption not being completely unreasonable. We agree more through investigation of this property would be worthwhile, however we believe such detailed work would be better in its own publication in a venue less dedicated to natural language, rather than in our appendix to this paper which focuses on the natural language understanding.

5 Minor Revisions Required :

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 revised.

The claim that colour 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

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

Reviewer C:

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 transformation, 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 mixing weights, means and variances predicted by the neural network is a more elegant approach, but not one expected to pragmatically work better based on the loss of information.
Indeed the a GMM with less than an equivalent number of mixture components as we used histogram bins will lose more information, as arbitrary distribution shape can not be estimated via GMM. An obvious difficulty is when the number of modes in a multimodal distribution exceeds the number of mixture components — however given a reasonably high number if component this can be avoided. A less obvious, but more problematic case is asymmetric distributions. Fitting to an arbitrarily asymmetrically sloped curve requires arbitrary numbers of mixture components. When such a quantity of components is not available, then information will be lost.

A further concern we have is difficulties actually fitting such a model. While fitting softmax via gradient decent is completely standard, fitting a GMM is normally via expectation maximisation. It is unclear as to how difficult such a fitting would be. We can envision that in a multimodal case, all the mixture components may get stuck on a single peak neglecting the other. As such we feel this additional complication would detract from the papers focus on the input stage (SOWE vs CNN, vs RNNs). The use of a GMM in this way is an approach we've been considering and we feel deserves exploration in a separate work, where that would be the primary focus.

With regards to our claim that word-order is not very significant for colour names.
We find it to be closely related to the statement that the dataset 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.

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, and additional experiments using LSTM have been included. We refer the reviewer to our responses to the editors summary and to Reviewer A.

- Include corpus statistics after preprocessing.

We have included a variety of statistics on the corpus and its various subsets after preprocessing.

Reviewer E:

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, in Section 2; to more clearly explain McMahan and Stone (2015), and Winn and Muresan (2018). Additional passes of editing have been completed.

5 Minor Revisions Required :

* 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

We have not done this, as the point estimates shown form part of a qualitative examples section (Section 6.1).

This section serves to highlight that while the quantitive evaluations (Section 6.2) show differences between the model's performances, in practicals examples as show to the naked eye the differences are often imperceptible. The introductiory paragraphs of section 6.1 have been amended to highlight this fact.
Including such numbers would subtract from the point, that they are visibly very similar. We feel the the quantitative results in Section 6.2, which are based on the test dataset and thus the population's perception are a better indicative of quantitative differences between the models.

* 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?

We currently highlight a number of interesting patterns based not on individual best and worst colours but on classes of colours that perform better and worse.
In particular the difficulties with word-order and more generally with multiword colour names, which are common to the RNNs. The general commonness of particular types of mistakes in the completely unseen color terms discussed in 6.4. As well as the break-downs with the restricted training set and unseen color combinations in Section 6.2.2
We do not feel direct discussion of the best and worst colours would add much insight, as the dataset are highly unbalanced — the most common color occurs with several orders of magnitude more frequency than the rarest, and thus the best and worst would simply reflect this.

* 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

A grey background has been added to the white histograms.

* 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

Kawakami is already cited in Section 1.1 (near the bottom of page 2). An additional sentence has been added to future comment on their use of a character rather than word sequence model.

* 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.

We have now specified in Section 1.2 "Contributions" that the input modules use pertained FastText embeddings.

* 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

An example has been added. For practical purposes this actually matters very little given how small our bins are. The implementation of such tabulation is a standard part of kernel density estimation software. In our case, as discussed in Section 5.1 “Implementation”, we use the tabulation from KernalDensityEstimation.jl for this.

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

The latter pages have received some additional revision to the grammar. (As has the rest of the document.)

6 Typographic Errors

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.