Review #1

Summary of Weaknesses*

1. **I am not sure if the paper as is really fits into the selected track ("Low-resourced and Less Studied Languages") which – in my understanding – focuses on languages that have a lack of available material and linguistic resources and solutions that address the resulting problems. This paper seems to focus strongly on "low-resourced" in the sense of low-resourced technical environments (which in itself is obviously very relevant). The sense of "not enough available data" is in the current state of the paper only mentioned in form of some side notes.**
    - We create an artificial scarcity of data by randomly selecting 500 samples with 80:20 ratios for training and testing. With a demonstration of competitive accuracy for a reduced dimension of the embedding vector. Our works are also relevant for resource-constrained development of NLP models, such as in edge devices.
    - From the technical environment aspect, we tried to keep track of the computational units, CO2 emission, and power consumption (some values are shown in Table 1). Physical tests on cutting-edge devices are part of our next work plan.

Review #2

Summary of Weaknesses*

1. **Reproducibility Concerns: There are mentions of hardware switching during experiments (Google Colab T4/P100 GPU), which may introduce variability in energy and timing measurements, raising concerns about reproducibility.**

    - Yes, we agree with the concern, and to address it, we tested the efficacy of the proposed approach over QQP and CoLA datasets trained using T4 GPU. As observed, dissimilarity is minimal compared to QNLI and SST2 datasets, as in Table X. To reduce the uncertainty in the reproducibility of results, we took the average of the results. Furthermore, most of the simulations and results stated in the paper use the T4 GPU during the simulation.

Dataset	Models	Using	Mean Accuracy	Variance	Standard Deviation
QQP	LSTM+LSTM	Fourier	60.80	1.96	1.19
		Max Pooling PCA	63.44	2.25	1.50
	CNN+LSTM	Fourier	62.15	2.02	1.42
		Max Pooling PCA	63.14	1.42	1.19
	BILSTM+LSTM	Fourier	62.59	0.83	0.91
		Max Pooling PCA	63.99	2.00	1.41
	DistilBERT		58.20	22.76	4.77
	TinyBERT		55.75	18.59	4.31
CoLA	LSTM+LSTM	Fourier	59.70	1.22	1.10
		Max Pooling PCA	59.59	4.34	2.08
	CNN+LSTM	Fourier	62.15	0.12	0.35
		Max Pooling PCA	65.20	4.36	2.09
	BILSTM+LSTM	Fourier	62.54	0.74	0.86
		Max Pooling PCA	60.65	0.63	0.79
	DistilBERT		57.65	19.83	4.45
	TinyBERT		50.40	14.24	3.77
QNLI	LSTM+LSTM	Fourier	63.10	9.07	3.01
		Max Pooling PCA	63.25	10.27	3.21
	CNN+LSTM	Fourier	62.09	8.29	2.88
		Max Pooling PCA	64.60	4.85	2.20

	BILSTM+LSTM	Fourier	63.09	2.79	1.67
		Max Pooling PCA	65.70	4.61	2.15
	DistilBERT		52.80	20.26	4.50
	TinyBERT		52.20	19.56	4.42
SST2	LSTM+LSTM	Fourier	69.69	7.61	2.76
		Max Pooling PCA	69.50	8.16	2.86
	CNN+LSTM	Fourier	65.74	7.59	2.75
		Max Pooling PCA	65.65	15.42	3.93
	BILSTM+LSTM	Fourier	66.20	6.25	2.50
		Max Pooling PCA	65.69	18.22	4.27
	DistilBERT		69.95	34.65	5.89
	TinyBERT		60.90	33.79	5.81

Table 2.1- Reproducibility test for four different datasets

2. **Lack of Dataset Diversity: The evaluation focuses only on a limited set of GLUE datasets, which may not provide a comprehensive view of the model's performance across different NLP tasks and datasets.**

    - We partially agree as the experiments were run over a subset of the GLUE datasets, but we chose different applications, for instance, QQP and QNLI, which contain a single question and sentence, and SST2 datasets based on sentiment analysis with sentiment text contained in a single sentence.
    - Additionally, we check the reproducibility (shown in Table 2.1), the accuracy, and some of the other metrics for the CoLA dataset, which also aligns with our previous findings for other datasets.

3. **Limited Explanation of Trade-offs: Although energy efficiency is discussed, the trade-offs between accuracy and model size are not sufficiently explored, especially in terms of how much accuracy is sacrificed for efficiency.**

    - Some relevant information is in Figure 3(a), and we intend to update the manuscript with the necessary discussion. The trained model size is already a few hundred kilobytes before pruning, yet it gives competitive accuracies for the merging functions with minimal power consumption.

A)

Mean: LSTM+LSTM			
Merge Embeddings	Accuracy	Power (kWh)	**Size (MB)**
(x*y)	62.957	0.026	0.656
(x/y)	65.591	0.035	0.673
(x+y)	64.327	0.033	0.657
log(x*y)	63.459	0.027	0.657
log(x/y)	63.459	0.026	0.671
log(x+y)	63.991	0.031	0.656

B)

Mean: BILSTM+LSTM			
Merge Embeddings	Accuracy	Power (kWh)	**Size (MB)**
(x*y)	62.327	0.035	0.828
(x/y)	64.791	0.044	0.834
(x+y)	64.725	0.033	0.828
log(x*y)	62.257	0.038	0.828
log(x/y)	62.791	0.043	0.834
log(x+y)	64.059	0.038	0.828

C)

Mean: CNN+LSTM			
Merge Embeddings	Accuracy	Power (kWh)	**Size (MB)**
(x*y)	63.660	0.016	0.440
(x/y)	62.191	0.029	0.451
(x+y)	64.525	0.013	0.440
log(x*y)	63.593	0.017	0.440
log(x/y)	63.193	0.020	0.451
log(x+y)	63.327	0.017	0.439

Table 2.3- Comparisons among blended embedding Accuracies vs. Power Consumption vs. Model size (before pruning) shown in A, B, C

4. **Spectral Analysis Underdeveloped: The spectral analysis step, which reduces the dimensionality of contextual embeddings, is mentioned as crucial but lacks an in-depth explanation or visual representation of how it benefits the overall model performance.**

    - We agree that spectrally reduced contextual embedding may be represented visually and actively considering it—for instance, an amplitude of insignificant strength is analogous to minimal contextual variations. However, an optional threshold to differentiate signifies contextual vs. static variation remains an open question.

5. **No Clear Baseline Comparisons: Although the paper compares blended embeddings with models like DistilBERT, it could benefit from a clearer comparison with purely static or purely contextual embeddings in similar experimental setups to quantify the exact gain.**

    - We have added similar comparisons in Table 6 (Blend PCA 75, Contextual PCA 75, Glove PCA 75, Purely Static Glove 300). However, it could be more extensive with additional GLUE datasets.
    - We also showed for n = 75, 100, 150, and 300 dimensions for log(x/y), log(x+y), that compare performance with purely contextual and purely static in Appendix Table 9, Table 10, and Table 11.

Review #3

Summary of Weaknesses*

1. **While the paper is generally easy to read, it is sometimes difficult to see how everything hangs together. The GLUE dataset is mentioned, but it is unclear exactly how these datasets are tested, and how the models are trained.**

    - We create an artificial scarcity of data by randomly selecting 500 samples with 80:20 ratios for training and testing—such selection mimics artificial data scarcity. With a demonstration of competitive accuracy for a reduced dimension of the embedding vector. Our works are also relevant for resource-constrained development of NLP models, such as in edge devices.
    - Then, we train the model with 50 trails where the base learning rate is $1e^{-3}$ and the max learning rate is $6e^{-3}$ and perform a random search from where we get the best accuracy for each model. Training datasets are added in the appendix. **Detailed Training procedure** mentioned in reviewer 3 answer 3.

2. **The role of the spectral analysis and the frequency component could be made more concrete. It is not tied to well to other parts of the paper. I am not so familiar with the technique, so for those who are more familiar, this might be clearer, but a few sentences making this more explicit would be nice for the general reader, I think. For example, it is not immideately clear to me how Fourier series and embeddings are related, but again, this might just be me.**

    - We agree that spectrally reduced contextual embedding may be represented visually and actively considering it—for instance, an amplitude of insignificant strength is analogous to minimal contextual variations. However, an optional threshold to differentiate signifies contextual vs. static variation remains an open question.
    - Overall, we will improve the relevance of Fourier Transformation and embedding in the final version of the manuscript.

3. **The training and experiments are a bit difficult to understand. I am not sure I understand how it was done exactly. (Full Training procedure)**

    - For this experiment, we create an artificial scarcity of data by randomly selecting 250 samples from each class, a total of 500 samples from three unique datasets, and encode them with 1 and 0. We removed the null values. Then, we split these 500 samples into 80-20 ratios for training and testing.
    - Then, we tokenized the text input to turn the token into an integer vector sequence, and padding sequence 50 ensures the same length for all sequences.
    - After that, we encoded the label columns as 1 and 0 in place of Entailment and Not Entailment for the QNLI dataset. QQP and SST2 did not require encoding.

- In the next step, we generate the embedding matrix using Glove/Fasttext, Blended embedding, where each row represents each word and its vector representation.
- During model initialization, we set the weights of the Embedding layer to a pre-trained GloVe embedding matrix, which converts word indices into dense vectors. The output from this layer is fed into a series of convolutional layers (CNN, BILSTM, LSTM). The output of each convolutional layer is passed to the next, culminating in a global max-pooling layer (GlobalMaxPooling1D).
- Global max-pooling extracts the maximum value from the feature map, producing a fixed-length representation. This output is connected to a dense layer with a softmax activation function, which generates class probabilities. The number of units in the dense layer matches the number of classes, ensuring valid probability scores for each class.
- The model uses categorical cross-entropy loss and the Adam optimizer, with accuracy as the performance metric during training and validation.
- We also used 50 new samples to test the inference time and took the average of 5 times.

4. **I cannot help but miss some discussion of the semantics of the new embeddings, not just in terms of model results, but a more direct comparison of their semantics.**

- We are not removing any words but reducing the insignificant values from each word vector to reduce the overall embedding size. Additionally, we used PCA to reduce the number of dimensions in large datasets to principal components that retain most of the original information. We did not perform in-depth semantics tracking in this part of the work.

Review #4

Summary of Weaknesses*

1. **Unfortunately, the paper does not provide any hypotheses about why the combination of static and contextualized embeddings would be informative and what it would mean to combine them. I find it very difficult to understand how the combined embeddings relate to other, contextualized-only solutions in terms of the information they can provide.**

    - Combined embedding can capture static and context-specific meanings and may be useful in NLP applications that require static and contextual information.
    - It may be helpful to reinforce the out-of-vocabulary (OOV) words with static embedding relying on the contextual information of the words.

2. **It is not quite clear to me how the proposed solution addresses low-resource scenarios, as the method explained in Section 2.2 seems to rely on pre-trained BERT embeddings after all. It is possible that I misunderstood something.**

    - We are not making any new embedding. Upon tokenization of the sentence, we would rather extract the word embedding from the pre-trained BERT for different datasets than use the BERT in the inference process. Such an arrangement reduces inference computation, which is feasible with low computation resources.
    - Besides, the low resource scenario also implies data scarcity, where we artificially create data scarcity by randomly selecting 250 samples from each class, a total of 500 samples from the used GLUE datasets.

3. **The low-resource solution proposed in the paper has only been evaluated on English datasets. As far as I understand, no low-resource scenarios have been simulated. I find it difficult to draw conclusions about low-resource scenarios based on such experiments.**

    - We create an artificial scarcity of data by randomly selecting 250 samples from each class, a total of 500 samples with 80:20 ratios for training and testing. With a demonstration of competitive accuracy for a reduced dimension of the embedding vector. Our works are also relevant for resource-constrained development of NLP models, such as in edge devices.
    - Some relevant information is in Figure 3(a), and we intend to update the manuscript with the necessary discussion. The trained model size is already a few hundred kilobytes before pruning, yet it gives competitive accuracies for the merging functions with minimal power consumption. (Detailed tables are shown in Reviewer 2 - Answer 3).
    - From the technical environment aspect, we tried to keep track of the computational units, CO2 emission, and power consumption (some values are shown in Table 1). Physical tests on cutting-edge devices are part of our next work plan.