Uncertainty Aware Review Hallucination for Science Article Classification

Korbinian Friedl[1,*], **Georgios Rizos**[1,*], **Lukas Stappen**[2], **Madina Hasan**[3], **Lucia Specia**[1], **Thomas Hain**[3], **Björn W. Schuller**[1,2]
[1]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[2]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]SpandH – Speech and Hearing Research Group, University of Sheffield, UK
friedlkorbinian@gmail.com, gr912@ic.ac.uk

Abstract

The high subjectivity and costs inherent in peer reviewing have recently motivated the preliminary design of machine learning-based acceptance decision methods. However, such approaches are limited in that they: a) do not explore the usage of both the reviewer and area chair recommendations, b) do not explicitly model subjectivity on a per submission basis, and c) are not applicable in realistic settings, by assuming that review texts are available at test time, when these are exactly the inputs that should be considered to be missing in this application. We propose to utilise methods that model the aleatory uncertainty of the submissions, while also exploring different loss importance interpolations between area chair and reviewers' recommendations. We also propose a modality hallucination approach to impute review representations at test time, providing the first realistic evaluation framework for this challenging task.

1 Introduction

An analysis (Langford and Guzdial, 2015) of the NeurIPS 2014 experiment shows that 60% of the selected accepted papers were rejected by a second, independent review committee. Such significant reviewer disagreement makes the task of the area chair harder, and may even invite questioning of their decision. Software tools have been piloted in an effort to aid the human reviewers with a computational recommendation on aspects like absence of bias and proper statistical reporting in scientific submissions (Sizo et al., 2019).

Natural Language Understanding (NLU) could also offer decision *support* to the area chair, as argued in (Ghosal et al., 2019; Stappen et al., 2020). Such systems jointly model the entire or part of the article and one (Kang et al., 2018; Wang and Wan, 2018; Ghosal et al., 2019), or a variable number of potentially contradicting reviews (Stappen et al., 2020). We adopt the latter, review-aggregating approach, that resembles the editorial process more.

1.1 Contributions

In this short paper, we offer solutions to three particularities of this task that the above approaches do not address: a) Often, the recommendations given by the area chair and the reviewers are in disagreement. Whereas previous studies have used either the former (Kang et al., 2018; Wang and Wan, 2018; Ghosal et al., 2019) or a soft label average of the latter (Stappen et al., 2020) for supervision, we show that both signals comprise complementary information. b) Whereas soft labels de-emphasise subjective articles with disagreeing reviews during training (Stappen et al., 2020), we manage to outperform the latter study by explicitly modelling *aleatory uncertainty* as an auxiliary prediction task. c) A model that aims to support the editorial decision process should only assume the availability of human review text during training, and be able to make recommendations in their absence. Inspired by missing modality hallucination methods (Hoffman et al., 2016; Tang et al.; Pérez et al., 2020)), we propose a *realistic* system that uses all available data for training, but imputes review representations at test time based on the abstract text.

1.2 Purpose & Ethical Statement

We sincerely believe that human peer reviews should continue to be the main component of the paper acceptance selection process, and this work in no way attempts to replace the human reviewers; instead, we believe an NLU model can serve as an additional reviewer, aiding an area chair's decision-making process by slot-filling a missing reviewer, or providing a data-driven, tie-breaking perspective to the editor in cases of borderline reviews. The motivation behind this proposal is that NLU models trained on large-scale data, can learn to robustly

*KF and GR contributed equally to this work.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5004–5009

cancel out individual human biases – in a similar way neural networks are robust to non-systematic label noise (Rolnick et al., 2017). Admittedly, such a model can still learn and reflect systematic biases, but we leave an approach to this problem by means of methods that learn with biased data (Kim et al., 2019) for future work.

2 Related work

Kang et al. (2018) compiled the PeerRead dataset of submissions, and proposed NLU baselines for binary acceptance decision and score prediction such as novelty and technical correctness. Wang and Wan (2018) explored the acceptance task by modelling the abstract via a memory mechanism (Weston et al., 2015), along with one review. Ghosal et al. (2019) improved performance on the PeerRead dataset by utilising sentiment information using the VADER tool (Hutto and Gilbert, 2014) and universal sentence embeddings (Cer et al., 2018). Unfortunately, PeerRead is imbalanced in that the NeurIPS rejected submissions are not included, despite the fact that 90% of the accepted submissions with reviews in PeerRead are from NeurIPS. Furthermore, around 80% of the submissions are from arxiv, thus having no reviews attached to them.

Stappen et al. (2020) worked on the largest such dataset – the Interspeech 2019 submission corpus – and fused the variable number of text reviews per submission. On incorporating reviewer disagreement information, they showed the simple label average to be better than the adapted version proposed in (Ando et al., 2018), and also approached the score prediction tasks via deep quantile regression (Rodrigues and Pereira, 2020). Direct modelling of a label disagreement value, instead of using soft labels, has been utilised in areas such as affective computing (Han et al., 2017) and medical image modelling (Raghu et al., 2019). Alternatively, Kendall and Gal (2017) devised a method for aleatory uncertainty modelling that is learnt from the data, instead of requiring ground truth disagreement "labels". Both Han et al. (2017); Kendall and Gal (2017) have shown regularisation benefits of learnt uncertainty prediction.

3 Submission-level modelling

Following Wang and Wan (2018); Stappen et al. (2020), we focus on abstract x_i^{abs} and review texts $x_{i,r}^{rev}$ (numbering R_i) for the i-th submission, the acceptance classification labels given by the area chair y_i^{ac}, as well as by the reviewers $y_{i,r}^{rev}$. We use a model $\mathcal{M}$ that: a) learns abstract h_i^{abs} and review h_i^{rev} representations using corresponding *modules*, b) fuses the aforementioned into a submission representation h_i^{sub}, and generates the class probability distribution $\hat{y}_i$ via a *prediction module* and softmax. We then calculate the cross entropy (CE) loss with the *true* probability distribution y_i^{true}:

$$\mathcal{L}_{pred} = CE(\hat{y}_i, y_i^{true}). \quad (1)$$

The most straightforward way to do this is by using a *hard label*, i. e., assuming $y_i^{true} \equiv y_i^{ac}$, with all the probability concentrated at the final recommendation given by the area chair. This way, however, we withhold information about the reviewer uncertainty for the particular submission. Stappen et al. (2020) have successfully used the simple *soft label*:

$$y_i^{true} \equiv \frac{1}{R_i} \sum_{r=1}^{R_i} y_{i,r}^{rev}. \quad (2)$$

The value of soft labels becomes clear when one considers that, in their absence, true acceptance probabilities of .51 and .89 would receive the same treatment. Occasionaly, the area chair may disagree with the reviewers' aggregate decision, which motivates the interpolation of the two factors:

$$\mathcal{L}_{pred} = \lambda^{soft}\mathcal{L}_{pred}^{soft} + \lambda^{hard}\mathcal{L}_{pred}^{hard}, \quad (3)$$

where $\mathcal{L}_{pred}^{*}, \lambda_{*}$ refers to prediction loss and regularisation parameter for either hard or soft labels.

3.1 Modelling recommendation subjectivity

We add a second "head" in our prediction module that outputs a predictive uncertainty estimate $\hat{\sigma}_i$. We now require a supervision signal to train it, either by: a) treating label disagreement as ground truth uncertainty (**GTU**), or b) learning a heteroscedastic loss attenuation score (**HLA**). Inspired by (Han et al., 2017; Raghu et al., 2019), we define our approach to GTU as a multi-task loss:

$$\mathcal{L}_{pred}^{GTU} = \gamma^{pred}\mathcal{L}_{pred} + \gamma^{unc}MSE(\hat{\sigma}_i, \sigma_i), \quad (4)$$

where σ_i is the standard deviation among reviewer recommendations; MSE is mean squared error.

For HLA we use the method proposed in (Kendall and Gal, 2017), where $\hat{\sigma}_i$ is the standard deviation of a normal distribution centred at the mean denoted by the main head logits. By sampling T logits, and calculating a corresponding

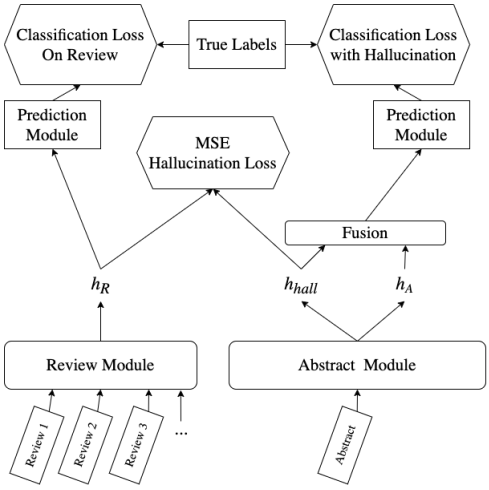

Figure 1: The peer review machine support system, including the hallucination mechanism. In a first training stage, only the *Classification Loss on Review* is used, to learn review representations. In the second training stage, the *MSE Hallucination Loss* and the *Classification Loss with Hallucination* are used.

class distribution $\hat{y}_i^t$, the loss function becomes:

$$\mathcal{L}_{pred}^{HLA} = \frac{1}{T}\sum_{t=1}^{T} CE(\hat{y}_i^t, y_i^{true}). \quad (5)$$

A larger $\hat{\sigma}_i$ relaxes the loss value for a sample that is difficult to predict correctly.

3.2 Imputing reviews at test time

Figure 1 depicts an overview of our architecture. Inspired by modality hallucination studies (Hoffman et al., 2016; Pérez et al., 2020), we use the abstract module to predict both the abstract h_i^{abs} and a review hallucination h_i^{hall} representation; *we do not generate review text*. In the first training stage, we train a model to predict based *only* on the review texts to learn meaningful h_i^{rev} representations. In the second stage, we use the true h_i^{rev} only as an auxiliary supervision target for training the h_i^{hall} representations by minimising MSE. Thus, we avoid *teacher forcing* (Bengio et al., 2015) by training the model to predict based only on the hallucinated representations, also available at test-time. The total loss, then, is:

$$\mathcal{L}_{hall} = \zeta^{hall} MSE(h_i^{rev}, h_i^{hall}) + \zeta^{pred}\mathcal{L}_{pred}. \quad (6)$$

4 Experiments

Small available dataset size is a limitation known to the community working on this domain (Kang et al., 2018; Ghosal et al., 2019) – we use the largest database of its kind (Stappen et al., 2020), i. e., the 2 179 preprocessed academic submissions, 5 842 reviews, with corresponding acceptance decisions and reviewer scores from the submission system of Interspeech 2019, shared with us by the technical chairs of the conference. After data cleaning and removal of corrupt entries, the accepted and rejected classes are well-balanced: 50.2 % acceptances, and 49.8 % rejections. The dataset is shuffled and split into 80-10-10 train-validation-test set percentages. We monitor the validation performance in terms of Macro-averaged F1 score, and also report the Macro-averaged Area Under Receiver Operator Characteristic (AU-ROC), averaged across 20 trials. We use the Adam optimiser (Kingma and Ba, 2014) with learning rate of 1e-3, and represent words using FastText (Bojanowski et al., 2017). Our abstract and review modules comprise a stacked 1D convolutional network with kernel sizes 4-4, interleaved by max pooling with rates 2-2, followed by a recurrent layer with gated recurrent unit cell and 100 hidden units, and attentional sequence pooling. The prediction module consists of two dense layers of 50-2 units, with a ReLu activation between them.

4.1 To model the reviewer or the area chair?

The interpolation weights $\lambda^{soft}, \lambda^{hard}$ for the prediction error (cf. Eq. 3) are dataset-based and should be set based on validation performance. We experiment with a grid, ranging from $[1.0, 0.0]$ to $[0.0, 1.0]$ using a step of 0.2. $\lambda^{soft} \equiv 0.0$ *denotes the simple hard label case.* The results using the GTU loss are summarised in Table 1.

We find that the area chair and the reviewers' recommendations carry *complementary* information, and the best results of this study are at $\lambda^{soft} \equiv 0.8$. Interestingly, the agreement/accuracy between the editorial labels and the reviewer soft averages (rounded to 0 or 1) is 78.901%. The disagreements occur on close-to-borderline papers, in which cases the additional supervision is the most informative.

4.2 Are soft labels enough?

A comparison among the different loss functions, without hallucination, is summarised in Table 2. We report the best soft loss interpolation per case.

λ^{soft}	Metric	No Review		With Review	
		Mean	±	Mean	±
0.0	AU-ROC	.589	.026	.683	.065
	F1	.559	.019	.625	.061
0.2	AU-ROC	.590	.021	.724	.031
	F1	.558	.023	.651	.030
0.4	AU-ROC	**.608**	.027	.750	.029
	F1	**.565**	.033	.673	.030
0.6	AU-ROC	.605	.026	.745	.050
	F1	.553	.027	.668	.040
0.8	AU-ROC	.601	.034	**.776**	.044
	F1	.512	.066	**.694**	.043
1.0	AU-ROC	.564	.047	.730	.084
	F1	.396	.102	.629	.108

Table 1: Results with **GTU**, for classification with and without reviews, for different relative weightings or (soft) reviewer ratings / (hard) editorial decisions.

Loss	Metric	No Review		With Review	
		Mean	±	Mean	±
BL	AU-ROC	-	-	.550	.130
	F1	-	-	.652	.100
Soft	AU-ROC	.597	.022	.772	.029
	F1	.558	.024	.683	.024
GTU	AU-ROC	**.608**	.027	**.776**	.044
	F1	**.565**	.033	**.694**	.043
HLA	AU-ROC	.578	.044	**.776**	.020
	F1	.407	.082	*.688*	.020

Table 2: Results on using the abstract with or without the reviews, using different kinds of losses. **BL** denotes the baseline by Stappen et al. (2020).

In the case of *GTU* we found that the choice of $\gamma^{unc} \equiv \gamma^{pred} \equiv 0.5$ works best. The additional complexity of explicit uncertainty modelling is shown to be beneficial when compared to the simple soft labels, and GTU is better than the self-learnt uncertainty method HLA. Our model implementation with soft-hard loss mixing is also shown to greatly outperform a baseline (**BL**), i. e., the best result found in (Stappen et al., 2020).

We also performed statistical significance testing, using Welch's unequal variances t-test. No significance was found in improvement brought by uncertainty-aware methods compared to hard labels in the abstract-only experiments. However, GTU with hallucinated review representations is significantly better than abstract-only with $p < 0.1$ for AU-ROC and $p < 0.05$ for F1, and HLA with $p < 0.05$ for both measures. In the experiments using both abstract and reviews, the simple soft labels as well as HLA are both significantly better than hard labels in terms of AU-ROC with $p < 0.05$. GTU was significantly better than hard labels with $p < 0.1$ for F1 and $p < 0.05$ for AU-ROC.

Loss	Metric	Abstract		Hallucination	
		Mean	±	Mean	±
Hard	AU-ROC	.589	.026	**.592**	.022
	F1	.559	.019	**.562**	.024
Soft	AU-ROC	.611	.035	**.612**	.029
	F1	.535	.030	**.544**	.028
GTU	AU-ROC	.601	.025	**.608**	.030
	F1	.512	.054	**.532**	.076
HLA	AU-ROC	.557	.034	**.636**	.033
	F1	.358	.042	**.575**	.091

Table 3: We report the review hallucination results; for the uncertainty-aware methods, we used $\lambda^{soft} \equiv 0.8$.

4.3 Can we impute reviews?

Table 3 summarises the improvement brought by hallucinated reviews over the abstract-only case. Even though we only report a specific hard-soft interpolation weight $\lambda \equiv 0.8$, we observe this improvement universally. The HLA method with hallucination achieves both the best performance in this experiment, and the largest relative improvement (t-test, $p < .05$) upon the abstract-only case, as shown in Table 4. Lacking the true reviews, we have high label variance for the same abstract input, i. e., high aleatory uncertainty. HLA (Kendall and Gal, 2017) is designed for such cases, and guides the learning of hallucinated review representations through regularisation, allowing for a significant fraction of the performance gap to be covered. Hard-labels with hallucination is the method that performs relatively closest to its ceiling performance, but this can be explained by the ceiling being comparatively low in the hard-labels case.

The additional label uncertainty information, whether explicit or learnt, informs not just the classification capacity of the model, but also its ability to generate review representations. These hallucinated representations should be placed in embedding space such that they inform the model regarding the label, however not in an overconfident manner, given that the actual reviews are missing – this is exactly where knowledge of uncertainty contributes. In terms of a final method recommendation: we recommend the learnt attenuation based HLA, due to its better performance along with modality hallucination and the fact that it does not require the presence of multiple reviewer recommendations even at training-time.

4.4 Can we learn model disagreement?

The Pearson Correlation Coefficient (PCC) between the predicted uncertainty and the standard deviation of reviewer recommendations is .25 and

Loss	Metric	+ Relative	- Relative
Hard	AU-ROC	+0.5	-13.3
	F1	+0.5	-10.1
Soft	AU-ROC	+0.2	-19.7
	F1	+1.7	-19.8
GTU	AU-ROC	+1.2	-21.5
	F1	+3.8	-23.3
HLA	AU-ROC	+12.4	-17.9
	F1	+37.8	-16.4

Table 4: Relative improvements (in %) brought by hallucinated reviews compared to using only the abstract, and relative reductions compared to the performance ceiling in the case the reviews are available at test time. In cases where the true reviews cannot be assumed to be present in test/deployment, our hallucination approach allows for improvement of results compared to excluding reviews altogether.

.08 for GTU and HLA respectively in the abstract plus review case. The former indeed learns on actual disagreement labels, although high uncertainty prediction fidelity may not be necessary for high predictive performance, shown by the competitive HLA. When using only abstracts, PCC drops to .08 and .04 respectively, whereas by using hallucination we observe .08 and .05, indicating that the true review representations are required for good uncertainty prediction.

5 Conclusion

We have proposed a machine learning framework for automatic peer review support that makes better use of the available information, and is also realistic with respect to the limitations set by the task[1]. We have found that the the area chair and reviewer recommendations comprise independent supervision signals that should be used *in conjunction* to train the system. Furthermore, in order to relax the penalty for mispredicting subjective submissions, it is not enough to use a simple soft label average of the reviewer recommendations; one has to directly model an aleatory uncertainty score as an auxiliary task, either using ground truth "uncertainty labels", or through learnt attenuation of the loss. Finally, we utilise review representation hallucinations at test-time to best utilise available review texts in a realistic manner, and find that this approach works well with and *benefits* from the regularisation introduced by direct uncertainty modelling.

Even with the application of our review representation hallucination, the performance gap from the ceiling set by using the true review representations is still high. We intend to approach the task via self-supervision methods (He et al., 2020) that focus on multimodal data (Nagrani et al., 2020). Requesting additional reviewers based on inference-time uncertainty, similar to (Raghu et al., 2019), is another promising future work step, as is an analysis of the uncertainties predicted by our model using the different losses. Finally, we have shown in our study that only a representation of the abstract is required as the input both for acceptance and hallucination modelling. Since previous work (Ghosal et al., 2019) has shown that modelling an article based on the entire paper can be beneficial, we also intend to explore the impact of using such a highly expressive article representation for hallucinating review representations.

[1] https://github.com/glam-imperial/Uncertainty-Aware-Machine-Paper-Reviewing

Acknowledgments

GR is funded by the Engineering and Physical Sciences Research Council (EPSRC) Grant No. 2021037.

References

Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Roy Masumura, Yusuke Ijima, and Yushi Aono. 2018. Soft-Target Training With Ambiguous Emotional Utterances For DNN-Based Speech Emotion Classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174. ACL.

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. 2019. DeepSentiPeer: Harnessing Sentiment in Review Texts to Recommend Peer Review Decisions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1120–1130. ACL.

Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty. In *Proceedings of the 2017 ACM on Multimedia Conference - MM '17*, pages 890–897. ACM Press.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Judy Hoffman, Saurabh Gupta, and Trevor Darrell. 2016. Learning with Side Information through Modality Hallucination. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834. IEEE.

Clayton J Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *18th International AAAI Conference on Weblogs and Social Media*. AAAI.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1647–1661.

Alex Kendall and Yarin Gal. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances In Neural Information Processing Systems*.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

John Langford and Mark Guzdial. 2015. The arbitrariness of reviews, and advice for school administrators. *Commun. ACM*, 58(4):12–13.

Arsha Nagrani, Joon Son Chung, Samuel Albanie, and Andrew Zisserman. 2020. Disentangled speech embeddings using cross-modal self-supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6829–6833. IEEE.

Andrés F. Pérez, Valentina Sanguineti, Pietro Morerio, and Vittorio Murino. 2020. Audio-Visual Model Distillation Using Acoustic Images. *The IEEE Winter Conference on Applications of Computer Vision*.

Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2019. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, pages 5281–5290.

Filipe Rodrigues and Francisco C. Pereira. 2020. Beyond expectation: Deep joint mean and quantile regression for spatio-temporal problems. *IEEE Transactions on Neural Networks and Learning Systems*.

David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. 2017. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.

Amanda Sizo, Adriano Lino, Luis Paulo Reis, and Álvaro Rocha. 2019. An overview of assessing the quality of peer review reports of scientific articles. *International Journal of Information Management*, 46:286 – 293.

Lukas Stappen, Georgios Rizos, Madina Hasan, Thomas Hain, and Björn W. Schuller. 2020. Uncertainty-Aware Machine Support for Paper Reviewing on the Interspeech 2019 Submission Corpus. In *Interspeech*, pages 1808–1812.

Yongyi Tang, Lin Ma, and Lianqiang Zhou. Hallucinating Optical Flow Features for Video Classification. *Proceedings of the Twenty-Eight International Joint Conference on Artificial Intelligence*, pages 926–932.

Ke Wang and Xiaojun Wan. 2018. Sentiment Analysis of Peer Review Texts for Scholarly Papers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 175–184. ACM.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory Networks. *arXiv:1410.3916*.