Is this hard for you? Personalized human difficulty estimation for skin lesion diagnosis

Peter Johannes Tejlgaard Kampen[1][0009−0004−4656−946X], Anders Nymark Christensen[1][0000−0002−3668−3128], and Morten Rieger Hannemose[1][0000−0002−9956−9226]

Department of Applied Mathematics and Computer Science,
Technical University of Denmark, Kgs. Lyngby, Denmark
{pjtka, anym, mohan}@dtu.dk

Abstract. Predicting the probability of human error is an important problem with applications ranging from optimizing learning environments to distributing cases among doctors in a clinic. In both of these instances, predicting the probability of error is equivalent to predicting the difficulty of the assignment, e.g., diagnosing a specific image of a skin lesion. However, the difficulty of a case is subjective since what is difficult for one person is not necessarily difficult for another. We present a novel approach for personalized estimation of human difficulty, using a transformer-based neural network that looks at previous cases and if the user answered these correctly. We demonstrate our method on doctors diagnosing skin lesions and on a language learning data set showing generalizability across domains. Our approach utilizes domain representations by first encoding each case using pre-trained neural networks and subsequently using these as tokens in a sequence modeling task. We significantly outperform all baselines, both for cases that are in the training set and for unseen cases. Additionally, we show that our method is robust towards the quality of the embeddings and how the performance increases as more answers from a user are available. Our findings suggest that this approach could pave the way for truly personalized learning experiences in medical diagnostics, enhancing the quality of patient care.

Keywords: Learning · Difficulty estimation · Sequence modelling.

1 Introduction

Doctors are dedicated to delivering optimal care to their patients, for which accurate diagnoses are paramount. Gaining experience is essential for proficiency, yet acquiring expertise in skin cancer diagnostics typically requires several years [16]. An alternative approach involves learning within a controlled environment, where exposure to cases with increasing difficulty can facilitate accelerated learning [12]. Prior research has concentrated on assessing a general case difficulty [6], or personal skill inferred purely from user answers [2]. However, in reality, the difficulty of a specific case will be different from person to

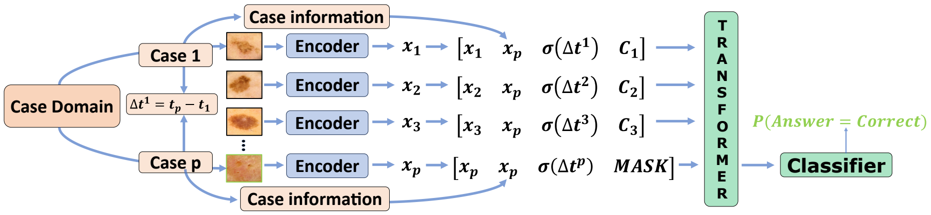


Fig. 1: Overview of our method. We predict the probability of a user answering correctly on a new case, given their history in the form of embeddings of the previous cases and which of these they answered right and wrong.

person and depend on their individual experience, which is the focus of this paper. Namely, we deal with the problem of estimating personalized difficulties in unseen cases. The ability to predict the correctness of a doctor's diagnosis has wide-ranging applications, such as optimizing the allocation of diagnostic cases between multiple doctors in a clinic, to maximize the probability of correct diagnoses. Additionally, personalized models for difficulty could assist in uncovering biases. The problem of predicting if a doctor will diagnose correctly is equivalent to predicting the correctness of a student's answer in a testing environment and we focus on the latter due to data availability. We use a transformer-based neural network that, based on previous cases and the doctor's answers, can estimate their probability of answering correctly on a new unseen image. An overview of our method is in Figure 1. We compare our method to the difficulty estimation method by Hannemose *et al.* [6], the Elo rating system [4], Bayesian Knowledge Tracing (BKT) [2] and the expected difficulty. We present results on two datasets, one consisting of images of skin lesions accompanied by diagnoses from medical students and the publicly available Duolingo SLAM dataset [14]. We show the efficacy of our method across domains and through the simulation of real-world scenarios. Put together, this opens the possibility of truly personalized learning tailored to each doctor. In summary, our contributions are:

- We combine the domain understanding of pre-trained models with sequence modeling, obtaining state-of-the-art personalized difficulty predictions.
- We can handle any number of previous answers from a user, with increasing performance for more answers, yet still outperform state-of-the-art for users with very few answers.
- Our method generalizes without modification to cases not present in the training data, with a wider performance gap compared to previous work.

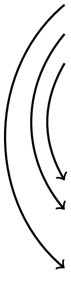
2 Related work

Estimating how difficult a given task is for a specific person is of great interest. Multiple prior works have tackled this issue. Tudor *et al.* [17] use the time taken for a visual search task as a proxy for the difficulty. They estimate this difficulty with a convolutional neural network, enabling them to predict the difficulties of new natural images. Hannemose *et al.* [6] estimated the diagnostic difficulty of medical images and demonstrated their method on dermoscopic and otoscopic images. Varshney *et al.* [18] estimate the difficulty of instances for machine learning models in natural language processing. Finally, Settles *et al.* [15] estimate the difficulty of cases in English tests using natural language processing. However, these methods do not account for user variation and seek to model a 'global' difficulty. Multiple prior works have also attempted to estimate how the skills of a user can develop over time. Among these is Bayesian Knowledge Tracing (BKT) [2]. Many versions exist that vary in terms of, e.g., the possibility of forgetting [1] and individual learning parameters [21]. The Elo rating system [4], widely used to rank chess players, can be used to handle the problem of users and cases in an environment similar to ours [9]. Both doctors and cases can be treated as players, and a doctor diagnosing a case correctly or incorrectly constitutes a win or loss, which then allows the difficulty of the case and the skill of the user to be updated. Klinkenberg *et al.* [9] implement this in an online learning setting and incorporate the time taken to answer. Hofman *et al.* [8] expanded on this to allow for statistical inference. We note that methods that estimate a single difficulty per case [17,6,18,15] utilize information obtained directly from the case to ease the estimation. In contrast, methods that can estimate individual difficulties that change over time do not [4,9,21]. Our proposed approach combines both advantages by using information about the case in the form of pre-trained embeddings and estimating individualized difficulties.

3 Method

Most methods for difficulty estimation for continuously learning users e.g., Bayesian Knowledge Tracing [2] and the Elo rating system rely on a single difficulty estimate for a specific case. This parameter is then governed by an underlying latent variable that describes the actual features of the case. However, these are generally not explicitly modeled. Therefore, these methods cannot model underlying similarities between cases, only how users usually respond to these. In this paper, we seek to model the tasks or problems directly by leveraging deep neural networks to yield explicit representations. Consider a set of test items belonging to a domain $\mathcal{D}$, e.g. skin lesions, where the items are images of these. Let $f : \mathcal{D} \to \mathbb{R}^n$ be a neural network mapping from this domain onto a vector representation. Suppose a user is repeatedly tested over time while improving based on feedback or teaching. We denote the sequence of test items $i = 1, \ldots, k$ for user u by $\boldsymbol{d}_i^u \in \mathcal{D}$. The corresponding latent representation of the sequence is then $f(\boldsymbol{d}_i^u) = \boldsymbol{x}_i^u$. We employ a transformer encoder-based model [19], with an objective similar to the masked language modeling [3]. We, therefore create sequences of vector representations of the test items along with additional information. For the i^{th} test case presented to the user u, $\boldsymbol{d}_i^u$, we represent whether the user answered correctly as a one-hot encoded vector $\boldsymbol{c}_i^u$. We are interested in predicting the probability of the user answering correctly $P(c_p^u)$ on a new case

Table 1: Description of the model. Each transformer uses four attention heads, and MLP denotes multi-layer perceptrons. h is 512 and 1224 for the Skin lesions and Duolingo datasets respectively. m is the sequence length.

Long skip connection	Module	Input dim.	Output dim.
	MLP	$(m,\, w)$	$(m,\, h)$
	Transformer	$(m,\, h)$	$(m,\, h)$
	Transformer	$(m,\, h)$	$(m,\, h)$
	Transformer	$(m,\, h)$	$(m,\, h)$
	Transformer	$(m,\, h)$	$(m,\, h)$
	Transformer	$(m,\, h)$	$(m,\, h)$
	Transformer	$(m,\, h)$	$(m,\, h)$
	Transformer	$(m,\, h)$	$(m,\, h)$
	Last Token	$(m,\, h)$	$(1,\, h)$
	MLP	$h + w$	2

with representation $\mathbf{x}_p$, given the sequence of preceding case representations and answers. Due to the temporal nature of learning, we employ temporal encodings rather than positional encodings based on the time since the answers. When predicting for $\mathbf{x}_p$ answered at time t_p, conditioned on $\mathbf{x}_i^u$ at time t_i^u, the temporal encoding for $\mathbf{x}_i^u$ is then $\sigma(t_p - t_i^u)$, where $\sigma(\cdot)$ is the sigmoid function. In most testing scenarios, additional information on the case e.g. the time taken to answer, may be available. We encode any additional information in a vector $\boldsymbol{l}_i^u$. We loosely employ the terminology from Natural Language Processing (NLP). Hence, we shall refer to each input that forms our sequence as tokens, which are $\mathbf{v}_i^u = [\mathbf{x}_i^u, \mathbf{l}_i^u, \mathbf{c}_i^u, \sigma(t_p - t_i^u)] \in \mathbb{R}^{w-n}$. Since the model predicts on $\mathbf{x}_p$, we allow it to focus on this by concatenating it to all tokens in the sequence. Hence, the model is given sequences of the form

$$\Big([\mathbf{v}_1^u, \mathbf{x}_p] \,, [\mathbf{v}_2^u, \mathbf{x}_p] \,, \ldots , [\mathbf{v}_k^u, \mathbf{x}_p] \,, \Big[\mathbf{x}_p, \mathbf{l}_p, [0 \quad 0], \sigma(0), \mathbf{x}_p \Big] \Big) \,. \tag{1}$$

At training time, sequences are generated by randomly sampling a sequence length and then sampling cases from a user. We employ a transformer-encoder architecture as described by Vaswani *et al.* [19]; however, we modify it using 'long' skip connections along with the skip connections of the original model architecture, see Table 1. As the computational complexity grows with the square of the maximum sequence length m in transformers, we present two approaches to improve performance when more than m answers are available. Our base model approach uses the previous $m - 1$ answers from the sequence. The first is to sample multiple random subsets of the user's previous answers and to average the predictions. Finally, we propose sampling the $m - 1$ cases with the highest cosine similarity to $\mathbf{x}_p$ and only including these in the sequence. Each model is trained for a maximum number of 400 epochs, with a batch size of 200, using early stopping on the validation set to prevent overfitting with balanced accuracy as the metric.

Table 2: Training, validation, and test splits. Splits are done on a user basis. Training and validation numbers are medians over the five cross validation splits.

		Num. users	Mean num. answers	Data points	Mean correct
Skin Lesions	Training	60	565	33,900	47.8%
	Validation	11	565	6,215	47.8%
	Test	11	570	6,270	46.5%
Duolingo	Training	2074	145	300,730	62.4%
	Validation	519	145	72,255	62.4%
	Test	2568	19	48,792	59.7%

3.1 Baseline methods

We compare with four baseline results, where the first is an *Expected difficulty* estimate. Let $\mathcal{D}_p$ be the set that contains all answers on case p in the training set, where each element is either 1 for a correct answer or 0 for an incorrect answer. Then the probability of a user u, in either the validation or test set, answering correctly on case p is given by $P\left(c_p^u\right) = \sum_{c \in \mathcal{D}_p} \frac{c}{|\mathcal{D}_p|}$. For cases not present in the training set, we sample predictions according to the general class distribution. As the second method, we use the Elo rating system [4], where each case and user is treated as a player in a tournament as described in Section 2. The method works by iteratively updating the ratings of users and cases. Let $s_{u,t}$ and $s_{p,t}$ be the ratings of a user and case at time t. Then the probability that the user answers correctly on the case is $P\left(c_p^u\right) = \frac{1}{1+10^{\delta/T}}$ where $\delta = s_{u,t} - s_{p,t}$ and the temperature, T, is a hyperparameter. Using the actual outcome, c_p^u, and the second hyperparameter k, the ratings are updated

$$s_{u,t+1} = s_{u,t} + k\big(c_p^u - P(c_p^u)\big), \quad s_{p,t+1} = s_{p,t} + k\big(P(c_p^u) - c_p^u\big). \tag{2}$$

The third model is based on Hannemose *et al.* [6] (*Predicted global difficulty*), which utilizes embeddings similar to ours and trains an ExtraTrees regressor [5] on the embeddings concatenated with a class label. This allows for predicting a constant difficulty estimate for unseen cases. Finally, we compare with Bayesian Knowledge Tracing (BKT) [1]. BKT models the student's knowledge as a latent variable in a hidden Markov model. Specifically, we use KT-IDEM [10], with a non-zero probability of forgetting. This version of BKT models both item difficulty and user skill. Hyperparameters for all baseline methods are found using a grid search to maximize the balanced accuracy on the validation sets.

4 Data

We demonstrate our method on two datasets from different domains, see Table 2.

Skin lesions imaged with dermoscopy were diagnosed into eight types of diagnoses by 82 medical students. Each medical student attempted to diagnose,

Table 3: Accuracy and balanced accuracy (B. accuracy) on the test set for the baselines and our base model with maximum sequence length $m = 200$ and our two methods for handling sequences longer than m. The Duolingo test set has no sufficiently long sequences.

	Skin lesions		**Duolingo**	
	Accuracy ↑	B. accuracy ↑	Accuracy ↑	B. accuracy ↑
Predicted global difficulty [6]	0.652 ± 0.02	0.641 ± 0.02	0.657 ± 0.00	0.620 ± 0.00
BKT [1,10]	0.566 ± 0.01	0.562 ± 0.00	0.601 ± 0.00	0.544 ± 0.00
Expected difficulty	0.654 ± 0.00	0.648 ± 0.00	0.652 ± 0.00	0.612 ± 0.00
Elo Rating [4]	0.610 ± 0.01	0.605 ± 0.01	0.610 ± 0.00	0.547 ± 0.00
Our base model	0.696 ± 0.01	0.692 ± 0.01	**0.703 ± 0.03**	**0.681 ± 0.03**
+ 10×random sampling	0.705 ± 0.02	0.704 ± 0.02	N/A	
+ cosine-similarity	**0.717 ± 0.02**	**0.715 ± 0.02**		

on average, 566 dermoscopic images randomly sampled from a pool of 1723 images. We have access to all student and ground truth diagnoses and refer to Hannemose *et al.* [6] for further information about the dataset. We randomly split the data on a person level, such that answers from one student only appear either in the training, validation, or test set. This was done to simulate a real-world setting, where such a model needs to generalize across users. Furthermore, the splits were obtained independently of the number of correct and incorrect answers to ensure limited bias, particularly for the constant difficulty baseline method. We make five train/validation splits, keeping a constant test set. For the encoder, we follow Hannemose *et al.* [6] and train a ResNet50 [7] with a multi-similarity loss function [20]. $\boldsymbol{l}$ contains the time to respond, the ground truth diagnosis, and the diagnosis from the user.

Duolingo SLAM is an open-source dataset [14]. We use the 'reverse translate' task with native English speakers translating from Spanish to English, with the original task being to predict errors on a word basis. However, since we consider overall problems, we collapse the labels such that if a single error is made, the entire translation is wrong. We subsume the allocated development set into the dataset and make five random train/validation splits. In total 3226 different cases are in the test set. The given test set is kept. We employ a SentenceBERT encoder with a DistilBERT base model [11,13]. $\boldsymbol{l}$ contains the time to respond.

5 Results

We present our main results in Table 3. Our method significantly outperforms the baseline methods on both datasets. Furthermore, of the two approaches for handling longer sequences, selection based on the cosine similarity between cases also yields a further increase in performance.

In almost any learning environment, new cases will continuously be added to the curriculum. Estimating the difficulty of such cases is difficult even for

Table 4: Balanced accuracy score for cases not present in the training set.

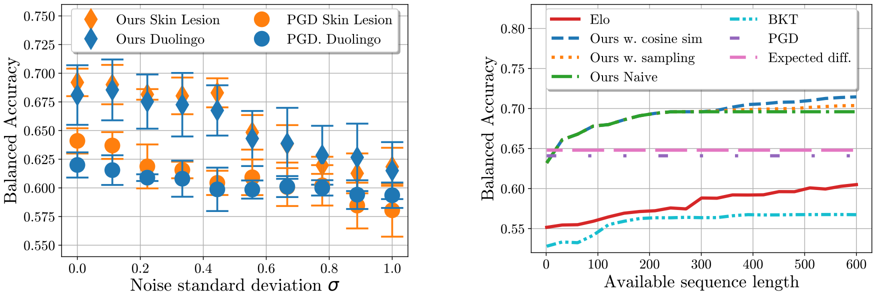
	Ours	Predicted global difficulty	BKT	Elo
Skin Lesions	**0.602 ± 0.03**	0.516 ± 0.02	0.558 ± 0.02	0.505 ± 0.00
Duolingo	**0.624 ± 0.04**	0.585 ± 0.01	0.517 ± 0.00	0.512 ± 0.01

Table 5: Test results on the skin lesions for variations of our model, $m = 200$. The `CLS` token denotes a learned representation added to the sequence [3].

	Encoder layers	Num. heads	Encoder dim.	Pooling	Balanced accuracy	Accuracy
Base model	8	4	512	Last token	**0.692±0.012**	**0.696±0.013**
Vary num. encoder layers	*2*	4	512	Last token	0.650 ± 0.012	0.655 ± 0.012
	4	4	512	Last token	0.662 ± 0.028	0.664 ± 0.028
	6	4	512	Last token	0.686 ± 0.020	0.687 ± 0.020
Vary num. attention heads	8	*2*	512	Last token	0.676 ± 0.018	0.678 ± 0.018
	8	*8*	512	Last token	0.689 ± 0.016	0.691 ± 0.017
Vary encoder dimensionality	8	4	*256*	Last token	0.674 ± 0.020	0.677 ± 0.020
	8	4	*1024*	Last token	**0.692±0.015**	**0.696±0.016**
Vary pooling	8	4	512	*Average*	0.676 ± 0.014	0.680 ± 0.013
	8	4	512	`CLS` *token*	0.675 ± 0.016	0.685 ± 0.015
No long skip conn.	8	4	512	Last token	0.684 ± 0.018	0.691 ± 0.018

experts. We have split our data such that there are cases in both test sets that are not present in training or validation. In the Skin lesions, these are 192 cases with 2112 responses, for Duolingo these are 509 cases with 9608 responses, and our performance on these are in Table 4. For Elo and BKT, we initialize their scores/difficulties as the mean of the training cases and let them update throughout the predictive process. All methods experience a notable drop in performance compared to Table 3, however, our method still significantly outperforms the baselines. Expected difficulty is not included, as it is undefined in this case.

In Figure 2a we test the reliance on the embedding quality for the two embedding-based methods, ours and predicted global difficulty [6]. They both use pre-trained encoders to compute the embeddings they use for prediction. We add increasing levels of noise to these embeddings and allow both models to fine-tune on the perturbed embeddings. We observe that our method is more robust for low noise levels, with only a slight drop in performance for the first four levels. At higher levels, most information in the embedding space is lost, and the performance reverts to approximately that of expected difficulty. In Figure 2b we test the performance of the models as a function of the number of previous cases available. The performance of our model increases rapidly for the first few cases, and it requires only 20 cases to outperform the two constant difficulty baselines



(a) Performance as a function of added noise to the embeddings. PGD denotes predicted global difficulty [6].

(b) Performance as a function of user data availability for the skin lesion dataset. For our method $m = 200$.

Fig. 2: Performance evaluation of the methods with noise and long sequences.

and always yields higher performance than Elo and BKT. To verify the choice of architecture, we present results for different hyperparameters in Table 5.

6 Discussion and conclusion

We observe that having a measure of similarity between cases, be it from the attention mechanisms in the transformer model or the cosine similarity-based case selection, significantly improves the accuracy of the difficulty estimation. Methods such as Elo and BKT that rely solely on low-dimensional measures of skill and difficulty do not adequately capture the relationships between cases. For example, Figure 2b, demonstrates that our method requires significantly fewer cases to accurately encode the skill level compared to both Elo and BKT. This is useful in practical applications, especially early in the learning setting when few previous answers are available, as one would still want to provide examples with the right difficulty. However, our model also has a limitation in this aspect. As it contains no explicit term for the user's skill, initializing a new user at a specific skill level is not currently possible.

When presented with unseen cases, most baseline methods revert to almost random guessing. Our method is more robust in this setting, maintaining significantly higher performance. The performance exhibited by our model on unseen cases is encouraging for applications beyond the learning domain, such as in a clinical setting where all cases are unseen and should be distributed among doctors to diagnose. The model could be used to assign cases to those with the highest probability of diagnosing correctly. Future work could also investigate including a neural network as a user in our model. This would provide a measure of the trust that should be placed on human and AI diagnoses respectively.

In conclusion, we have established a new state-of-the-art for individualized human difficulty estimation. By leveraging information from both cases and users, we achieve superior performance with fewer samples than other methods.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Badrinath, A., Wang, F., Pardos, Z.: pybkt: An accessible python library of bayesian knowledge tracing models (2021)
2. Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction **4**, 253–278 (2005), https://api.semanticscholar.org/CorpusID:19228797
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
4. Elo, A.E.: The rating of chessplayers, past and present. Arco Pub. (1978)
5. Geurts, P., Louis, W.: Extremely randomized trees. Machine Learning (2006). https://doi.org/10.1007/s10994-006-6226-1, https://doi.org/10.1007/s10994-006-6226-1
6. Hannemose, M.R., Sundgaard, J.V., Ternov, N.K., Paulsen, R.R., Christensen, A.N.: Was that so hard? estimating human classification difficulty. Applications of Medical Artificial Intelligence **13540**, 88 (2022). https://doi.org/https://doi.org/10.1007/978-3-031-17721-7_10
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
8. Hofman, A.D., Brinkhuis, M.J., Bolsinova, M., Klaiber, J., Maris, G., van der Maas, H.L.: Tracking with (un) certainty. Journal of Intelligence **8**(1), 10 (2020)
9. Klinkenberg, S., Straatemeier, M., van der Maas, H.: Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. Computers & Education **57**(2), 1813–1824 (2011). https://doi.org/https://doi.org/10.1016/j.compedu.2011.02.003, https://www.sciencedirect.com/science/article/pii/S0360131511000418
10. Pardos, Z., Heffernan, N.: Kt-idem: Introducing item difficulty to the knowledge tracing model. pp. 243–254 (01 1970). https://doi.org/10.1007/978-3-642-22362-4_21
11. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2020), https://arxiv.org/abs/2004.09813
12. Roads, B.D., Xu, B., Robinson, J.K., Tanaka, J.W.: The easy-to-hard training advantage with real-world medical images. Cognitive Research: Principles and Implications **3**(1), 1–13 (2018)
13. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2020)
14. Settles, B., Brust, C., Gustafson, E., Hagiwara, M., Madnani, N.: Second language acquisition modeling. In: Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications (BEA). ACL (2018)
15. Settles, B., T. LaFlair, G., Hagiwara, M.: Machine learning–driven language assessment. Transactions of the Association for computational Linguistics **8**, 247–263 (2020)
16. Ternov, N.K., Vestergaard, T., Hölmich, L.R., Karmisholt, K., Wagenblast, A., Klyver, H., Hald, M., Schøllhammer, L., Konge, L., Chakera, A.: Reliable test of

clinicians' mastery in skin cancer diagnostics. Archives of Dermatological Research **313**, 235–243 (2021)

17. Tudor Ionescu, R., Alexe, B., Leordeanu, M., Popescu, M., Papadopoulos, D.P., Ferrari, V.: How hard can it be? estimating the difficulty of visual search in an image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2157–2166 (2016)
18. Varshney, N., Mishra, S., Baral, C.: ILDAE: Instance-level difficulty analysis of evaluation data. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3412–3425. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.240, https://aclanthology.org/2022.acl-long.240
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
20. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning (2020)
21. Yudelson, M.V., Koedinger, K.R., Gordon, G.J.: Individualized bayesian knowledge tracing models. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) Artificial Intelligence in Education. pp. 171–180. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)