**NOTE: This analysis is not covered in the paper**

**ANALYSIS OF EXPERIMENTS BY TYPE**

To explore to what extent the limitations observed vary across experiments with different objectives, we classified the experiments and reassessed them by type. Table 1 provides an overview of the classification of the 194 experiments into three groups: evaluation, generalization, and optimization.

Table 1. Types of experiments found

|  |  |  |
| --- | --- | --- |
| **Category** | **Count** | **Percentage** |
| Optimization | 67 | 35% |
| Evaluation | 90 | 46% |
| Generalization | 25 | 13% |
| Optimization+Evaluation | 10 | 5% |
| Evaluation+Generalization | 2 | 1% |

**Evaluation experiments** aim at comparing the proposed DNN with a baseline of expected-values set by the researchers, DNNs proposed by other researchers, other techniques that do not rely on DNNs, or human performance. We find that 46% of the experiments aim to perform such an evaluation. **Generalization experiments** aim to assess a DNN under a different dataset, most commonly under a different test set but also sometimes under a different training set used to define the model parameters, or a new user context. We find that 13% of the experiments fall into this category. For example, [AP6] proposes a DNN for predicting developer actions (represented as a sequence of image regions). One of the experiments runs the DNN for predicting actions for developers and programming languages different from the ones in the training set. **Optimization experiments** aim at exploring and eventually identifying the best DNN configuration, within some allocated resources, through the manipulation of a large number of variables, from the model hyperparameters to the deep learning algorithm. We find a large number of optimization experiments (35%). We also find **combinations of optimization and evaluation experiments** (the same experiment compares other approaches and variants of the proposed approach) in 5% of the cases. Finally, 1% of experiments **combine evaluation and generalization** (the same experiment compares the DNN with other approaches while it is being assessed under a different dataset, or new user context).

We find that the limitations observed earlier remain mostly the same across experiment types. One noticeable difference, however, is that the optimization experiments are the ones with the most missing response variables (18% vs. 0% for the other types of experiments) and factors and treatments specification (10% vs. 0% and 4% for evaluation and generalization experiments). Since the response variable is often associated with accuracy, this is an easily fixed oversight. However, missing factors and

treatments seem more problematic since it undermines these experiments’ objective to identify the best model configuration, and the factors and treatments are key in defining such configuration. We provide the full breakdown of the data.