Adversarial Reprogramming of Text Classification Neural Networks

Supplementary Material

Table 1: Hyper-parameter Details of Adversarial Reprogramming Experiments. Same White-box settings were used for both pretrained and untrained networks. Epochs Annealed is the number of epochs over which temperature is linearly annealed from Temp Max to Temp Min. All the adversarial programs are trained for 500 epochs by default and the model with the highest validation accuracy is chosen for testing. We use Adam optimizer for all our experiments.

			White-box					Black-box	
Model	**Original Task**	**Adversarial Task**	**Batch Size**	**Learning Rate**	**Temp Max**	**Temp Min**	**Epochs Annealed**	**Batch Size**	**Learning Rate**
LSTM	Names-18	Questions	8	0.01	2	0.1	5	8	0.005
LSTM	Names-18	Arabic Tweets	8	0.01	2	0.5	5	8	0.01
LSTM	Questions	Names-5	8	0.01	2	0.5	5	8	0.01
LSTM	Questions	Arabic Tweets	8	0.01	2	0.5	5	8	0.01
LSTM	IMDB	Arabic Tweets	8	0.01	2	0.5	5	8	0.005
Bi-LSTM	Names-18	Questions	8	0.005	2	0.2	15	8	0.005
Bi-LSTM	Names-18	Arabic Tweets	8	0.005	2	0.5	15	8	0.005
Bi-LSTM	Questions	Names-5	8	0.005	2	0.2	15	8	0.005
Bi-LSTM	Questions	Arabic Tweets	8	0.005	2	0.5	15	8	0.005
Bi-LSTM	IMDB	Arabic Tweets	8	0.005	2	0.5	15	4	0.005
CNN	Names-18	Questions	8	0.005	2	0.2	15	8	0.01
CNN	Names-18	Arabic Tweets	8	0.005	2	0.5	15	8	0.005
CNN	Questions	Names-5	8	0.005	2	0.2	15	8	0.005
CNN	Questions	Arabic Tweets	8	0.005	2	0.5	15	8	0.005
CNN	IMDB	Arabic Tweets	8	0.005	2	0.5	15	4	0.005

Table 2: Hyper-parameter Details of the classifiers. The column *Hidden Units/Num filters* corresponds to hidden units in case of the LSTM, Bi-LSTM and number of filters in each layer for the CNN. We use a 3-layer CNN with filter widths 3,4 and 5. All the models are optimized using Adam optimizer.

Model	**Dataset**	**Hidden Units**	**Embedding Size**	**Learning Rate**
LSTM	Names-18	256	256	0.0001
LSTM	Names-5	256	256	0.0001
LSTM	Questions	256	256	0.0001
LSTM	Arabic Tweets	256	256	0.0001
LSTM	IMDB	256	256	0.0001
Bi-LSTM	Names-18	256	256	0.0005
Bi-LSTM	Names-5	256	256	0.0005
Bi-LSTM	Questions	256	256	0.0005
Bi-LSTM	Arabic Tweets	256	256	0.0005
Bi-LSTM	IMDB	256	256	0.0005
CNN	Names-18	100	100	0.0005
CNN	Names-5	100	100	0.0005
CNN	Questions	100	100	0.0005
CNN	Arabic Tweets	100	100	0.0005
CNN	IMDB	100	100	0.0005