A EMNLP 2018 Checklist Survey

Checklist item	Percentage of EMNLP 2018 papers
Reports train/validation/test splits	92%
Reports best hyperparameter assignments	74%
Reports code	30%
Reports dev accuracy	24%
Reports computing infrastructure	18%
Reports empirical runtime	14%
Reports search strategy	14%
Reports score distribution	10%
Reports number of hyperparameter trials	10%
Reports hyperparameter search bounds	8%

Table 1: Presence of checklist items from §5 across 50 randomly sampled EMNLP 2018 papers that involved modeling experiments.

B Hyperparameter Search Spaces for Section 4.2

Computing infrastructure	GeForce GTX 1080 GPU
Number of search trials	50
Search strategy	uniform sampling
Best validation accuracy	40.5
Training duration	39 sec
Model implementation	`http://github.com/allenai/show-your-work`

Hyperparameter	**Search space**	**Best assignment**
number of epochs	50	50
patience	10	10
batch size	64	64
embedding	GloVe (50 dim)	GloVe (50 dim)
encoder	ConvNet	ConvNet
max filter size	*uniform-integer*[3, 6]	4
number of filters	*uniform-integer*[64, 512]	332
dropout	*uniform-float*[0, 0.5]	0.4
learning rate scheduler	reduce on plateau	reduce on plateau
learning rate scheduler patience	2 epochs	2 epochs
learning rate scheduler reduction factor	0.5	0.5
learning rate optimizer	Adam	Adam
learning rate	*loguniform-float*[1e-6, 1e-1]	0.0008

Table 2: SST (fine-grained) CNN classifier search space and best assignments.

Computing Infrastructure	3.1 GHz Intel Core i7 CPU
Number of search trials	50
Search strategy	uniform sampling
Best validation accuracy	39.8
Training duration	1.56 seconds
Model implementation	`http://github.com/allenai/show-your-work`

Hyperparameter	**Search space**	**Best assignment**
penalty	*choice*[L1, L2]	L2
no. of iter	100	100
solver	liblinear	liblinear
regularization	*uniform-float*[0, 1]	0.13
n-grams	*choice*[(1, 2), (1, 2, 3), (2, 3)]	[1, 2]
stopwords	*choice*[True, False]	True
weight	*choice*[tf, tf-idf, binary]	binary
tolerance	*loguniform-float*[10e-5, 10e-3]	0.00014

Table 3: SST (fine-grained) logistic regression search space and best assignments.

C Hyperparameter Search Spaces for Section 4.3

Computing Infrastructure	GeForce GTX 1080 GPU
Number of search trials	50
Search strategy	uniform sampling
Best validation accuracy	87.6
Training duration	1624 sec
Model implementation	`http://github.com/allenai/show-your-work`

Hyperparameter	**Search space**	**Best assignment**
number of epochs	50	50
patience	10	10
batch size	64	64
gradient norm	*uniform-float*[5, 10]	9.0
embedding dropout	*uniform-float*[0, 0.5]	0.3
number of pre-encode feedforward layers	*choice*[1, 2, 3]	3
number of pre-encode feedforward hidden dims	*uniform-integer*[64, 512]	232
pre-encode feedforward activation	*choice*[relu, tanh]	tanh
pre-encode feedforward dropout	*uniform-float*[0, 0.5]	0.0
encoder hidden size	*uniform-integer*[64, 512]	424
number of encoder layers	*choice*[1, 2, 3]	2
integrator hidden size	*uniform-integer*[64, 512]	337
number of integrator layers	*choice*[1, 2, 3]	3
integrator dropout	*uniform-float*[0, 0.5]	0.1
number of output layers	*choice*[1, 2, 3]	3
output hidden size	*uniform-integer*[64, 512]	384
output dropout	*uniform-float*[0, 0.5]	0.2
output pool sizes	*uniform-integer*[3, 7]	6
learning rate optimizer	Adam	Adam
learning rate	*loguniform-float*[1e-6, 1e-1]	0.0001
learning rate scheduler	reduce on plateau	reduce on plateau
learning rate scheduler patience	2 epochs	2 epochs
learning rate scheduler reduction factor	0.5	0.5

Table 4: SST (binary) BCN GloVe search space and best assignments.

Computing Infrastructure	GeForce GTX 1080 GPU
Number of search trials	50
Search strategy	uniform sampling
Best validation accuracy	91.4
Training duration	6815 sec
Model implementation	`http://github.com/allenai/show-your-work`

Hyperparameter	**Search space**	**Best assignment**
number of epochs	50	50
patience	10	10
batch size	64	64
gradient norm	*uniform-float*[5, 10]	9.0
freeze ELMo	True	True
embedding dropout	*uniform-float*[0, 0.5]	0.3
number of pre-encode feedforward layers	*choice*[1, 2, 3]	3
number of pre-encode feedforward hidden dims	*uniform-integer*[64, 512]	206
pre-encode feedforward activation	*choice*[relu, tanh]	relu
pre-encode feedforward dropout	*uniform-float*[0, 0.5]	0.3
encoder hidden size	*uniform-integer*[64, 512]	93
number of encoder layers	*choice*[1, 2, 3]	1
integrator hidden size	*uniform-integer*[64, 512]	159
number of integrator layers	*choice*[1, 2, 3]	3
integrator dropout	*uniform-float*[0, 0.5]	0.4
number of output layers	*choice*[1, 2, 3]	1
output hidden size	*uniform-integer*[64, 512]	399
output dropout	*uniform-float*[0, 0.5]	0.4
output pool sizes	*uniform-integer*[3, 7]	6
learning rate optimizer	Adam	Adam
learning rate	*loguniform-float*[1e-6, 1e-1]	0.0008
use integrator output ELMo	*choice*[True, False]	True
learning rate scheduler	reduce on plateau	reduce on plateau
learning rate scheduler patience	2 epochs	2 epochs
learning rate scheduler reduction factor	0.5	0.5

Table 5: SST (binary) BCN GLoVe + ELMo (frozen) search space and best assignments.

Computing Infrastructure	NVIDIA Titan Xp GPU
Number of search trials	50
Search strategy	uniform sampling
Best validation accuracy	92.2
Training duration	16071 sec
Model implementation	`http://github.com/allenai/show-your-work`

Hyperparameter	**Search space**	**Best assignment**
number of epochs	50	50
patience	10	10
batch size	64	64
gradient norm	*uniform-float*[5, 10]	7.0
freeze ELMo	False	False
embedding dropout	*uniform-float*[0, 0.5]	0.1
number of pre-encode feedforward layers	*choice*[1, 2, 3]	3
number of pre-encode feedforward hidden dims	*uniform-integer*[64, 512]	285
pre-encode feedforward activation	*choice*[relu, tanh]	relu
pre-encode feedforward dropout	*uniform-float*[0, 0.5]	0.3
encoder hidden size	*uniform-integer*[64, 512]	368
number of encoder layers	*choice*[1, 2, 3]	2
integrator hidden size	*uniform-integer*[64, 512]	475
number of integrator layers	*choice*[1, 2, 3]	3
integrator dropout	*uniform-float*[0, 0.5]	0.4
number of output layers	*choice*[1, 2, 3]	3
output hidden size	*uniform-integer*[64, 512]	362
output dropout	*uniform-float*[0, 0.5]	0.4
output pool sizes	*uniform-integer*[3, 7]	5
learning rate optimizer	Adam	Adam
learning rate	*loguniform-float*[1e-6, 1e-1]	2.1e-5
use integrator output ELMo	*choice*[True, False]	True
learning rate scheduler	reduce on plateau	reduce on plateau
learning rate scheduler patience	2 epochs	2 epochs
learning rate scheduler reduction factor	0.5	0.5

Table 6: SST (binary) BCN GloVe + ELMo (fine-tuned) search space and best assignments.

D Hyperparameter Search Spaces for Section 4.4

Computing Infrastructure	GeForce GTX 1080 GPU
Number of search trials	100
Search strategy	uniform sampling
Best validation accuracy	82.7
Training duration	339 sec
Model implementation	`http://github.com/allenai/show-your-work`

Hyperparameter	**Search space**	**Best assignment**
number of epochs	140	140
patience	20	20
batch size	64	64
gradient clip	*uniform-float*[5, 10]	5.28
embedding projection dim	*uniform-integer*[64, 300]	78
number of attend feedforward layers	*choice*[1, 2, 3]	1
attend feedforward hidden dims	*uniform-integer*[64, 512]	336
attend feedforward activation	*choice*[relu, tanh]	tanh
attend feedforward dropout	*uniform-float*[0, 0.5]	0.1
number of compare feedforward layers	*choice*[1, 2, 3]	1
compare feedforward hidden dims	*uniform-integer*[64, 512]	370
compare feedforward activation	*choice*[relu, tanh]	relu
compare feedforward dropout	*uniform-float*[0, 0.5]	0.2
number of aggregate feedforward layers	*choice*[1, 2, 3]	2
aggregate feedforward hidden dims	*uniform-integer*[64, 512]	370
aggregate feedforward activation	*choice*[relu, tanh]	relu
aggregate feedforward dropout	*uniform-float*[0, 0.5]	0.1
learning rate optimizer	Adagrad	Adagrad
learning rate	*loguniform-float*[1e-6, 1e-1]	0.009

Table 7: SciTail DAM search space and best assignments.

Computing Infrastructure	GeForce GTX 1080 GPU
Number of search trials	100
Search strategy	uniform sampling
Best validation accuracy	82.8
Training duration	372 sec
Model implementation	`http://github.com/allenai/show-your-work`

Hyperparameter	**Search space**	**Best assignment**
number of epochs	75	75
patience	5	5
batch size	64	64
encoder hidden size	*uniform-integer*[64, 512]	253
dropout	*uniform-float*[0, 0.5]	0.28
number of encoder layers	*choice*[1, 2, 3]	1
number of projection feedforward layers	*choice*[1, 2, 3]	2
projection feedforward hidden dims	*uniform-integer*[64, 512]	85
projection feedforward activation	*choice*[relu, tanh]	relu
number of inference encoder layers	*choice*[1, 2, 3]	1
number of output feedforward layers	*choice*[1, 2, 3]	2
output feedforward hidden dims	*uniform-integer*[64, 512]	432
output feedforward activation	*choice*[relu, tanh]	tanh
output feedforward dropout	*uniform-float*[0, 0.5]	0.03
gradient norm	*uniform-float*[5, 10]	7.9
learning rate optimizer	Adam	Adam
learning rate	*loguniform-float*[1e-6, 1e-1]	0.0004
learning rate scheduler	reduce on plateau	reduce on plateau
learning rate scheduler patience	0 epochs	0 epochs
learning rate scheduler reduction factor	0.5	0.5
learning rate scheduler mode	max	max

Table 8: SciTail ESIM search space and best assignments.

Computing Infrastructure	GeForce GTX 1080 GPU
Number of search trials	100
Search strategy	uniform sampling
Best validation accuracy	81.2
Training duration	137 sec
Model implementation	`http://github.com/allenai/show-your-work`

Hyperparameter	**Search space**	**Best assignment**
number of epochs	140	140
patience	20	20
batch size	64	64
dropout	*uniform-float*[0, 0.5]	0.2
hidden size	*uniform-integer*[64, 512]	167
activation	*choice*[relu, tanh]	tanh
number of layers	*choice*[1, 2, 3]	3
gradient norm	*uniform-float*[5, 10]	6.8
learning rate optimizer	Adam	Adam
learning rate	*loguniform-float*[1e-6, 1e-1]	0.01
learning rate scheduler	exponential	exponential
learning rate scheduler gamma	0.5	0.5

Table 9: SciTail n-gram baseline search space and best assignments.

Computing Infrastructure	GeForce GTX 1080 GPU
Number of search trials	100
Search strategy	uniform sampling
Best validation accuracy	81.2
Training duration	1015 sec
Model implementation	`http://github.com/allenai/show-your-work`

Hyperparameter	**Search space**	**Best assignment**
number of epochs	140	140
patience	20	20
batch size	16	16
embedding projection dim	*uniform-integer*[64, 300]	100
edge embedding size	*uniform-integer*[64, 512]	204
premise encoder hidden size	*uniform-integer*[64, 512]	234
number of premise encoder layers	*choice*[1, 2, 3]	2
premise encoder is bidirectional	*choice*[True, False]	True
number of phrase probability layers	*choice*[1, 2, 3]	2
phrase probability hidden dims	*uniform-integer*[64, 512]	268
phrase probability dropout	*uniform-float*[0, 0.5]	0.2
phrase probability activation	*choice*[tanh, relu]	tanh
number of edge probability layers	*choice*[1, 2, 3]	1
edge probability dropout	*uniform-float*[0, 0.5]	0.2
edge probability activation	*choice*[tanh, relu]	tanh
gradient norm	*uniform-float*[5, 10]	7.0
learning rate optimizer	Adam	Adam
learning rate	*loguniform-float*[1e-6, 1e-1]	0.0006
learning rate scheduler	exponential	exponential
learning rate scheduler gamma	0.5	0.5

Table 10: SciTail DGEM search space and best assignments.

Computing Infrastructure	GeForce GTX 1080 GPU
Number of search trials	128
Search strategy	uniform sampling
Best validation EM	68.2
Training duration	31617 sec
Model implementation	`http://github.com/allenai/show-your-work`

Hyperparameter	**Search space**	**Best assignment**
number of epochs	20	20
patience	10	10
batch size	16	16
token embedding	GloVe (100 dim)	GloVe (100 dim)
gradient norm	*uniform-float*[5, 10]	6.5
dropout	*uniform-float*[0, 0.5]	0.46
character embedding dim	*uniform-integer*[16, 64]	43
max character filter size	*uniform-integer*[3, 6]	3
number of character filters	*uniform-integer*[64, 512]	33
character embedding dropout	*uniform-float*[0, 0.5]	0.15
number of highway layers	*choice*[1, 2, 3]	3
phrase layer hidden size	*uniform-integer*[64, 512]	122
number of phrase layers	*choice*[1, 2, 3]	1
phrase layer dropout	*uniform-float*[0, 0.5]	0.46
modeling layer hidden size	*uniform-integer*[64, 512]	423
number of modeling layers	*choice*[1, 2, 3]	3
modeling layer dropout	*uniform-float*[0, 0.5]	0.32
span end encoder hidden size	*uniform-integer*[64, 512]	138
span end encoder number of layers	*choice*[1, 2, 3]	1
span end encoder dropout	*uniform-float*[0, 0.5]	0.03
learning rate optimizer	Adam	Adam
learning rate	*loguniform-float*[1e-6, 1e-1]	0.00056
Adam β_1	*uniform-float*[0.9, 1.0]	0.95
Adam β_2	*uniform-float*[0.9, 1.0]	0.93
learning rate scheduler	reduce on plateau	reduce on plateau
learning rate scheduler patience	2 epochs	2 epochs
learning rate scheduler reduction factor	0.5	0.5
learning rate scheduler mode	max	max

Table 11: SQuAD BiDAF search space and best assignments.