	Single input sequence tasks						Dual input sequence tasks					
	SST	**IMDB**	**Yelp**	**Amazon**	**20News**	**Tweets**	**SNLI**	**QQP**	**Babi1**	**Babi2**	**Babi3**	**CNN**
<u>Model configuration</u>												
Model LSTM variation	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla
Attention type	tanh	tanh	tanh	tanh	tanh	tanh	tanh	tanh	tanh	tanh	tanh	tanh
Embedding dim	300	300	300	300	300	300	300	300	50	50	50	300
Embedding voc	13.826	12.487	63.328	49.883	6.515	6.845	20.981	26.635	24	38	39	70.190
Pre-embed	FastText	FastText	FastText	FastText	FastText	FastText	GLOVE	GLOVE	None	None	None	FastText
LSTM hidden dim	256	256	256	256	256	256	256	256	64	128	128	256
Output size	1	1	1	1	1	1	3	2	36 (6)*	36 (6)*	36 (6)*	584 (26)*
<u>Optimizer hyperparameters</u>												
Diversity weight (if applic	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.5	0.5	0.5	0.5	0.2
Batch size	32	32	32	32	32	32	128	128	32	64	64	90
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
LR	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Weight decay	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05
Epochs	8	8	8	8	8	8	25	25	100	200	200	12
<u>Trainable weights, including fine-tuning of embeddings</u>												
Pencoder	4.719.192	4.317.492	19.569.792	15.536.292	2.525.892	2.624.892	6.865.692	8.480.892	30.896	94.060	94.110	21.628.392
Qencoder	0	0	0	0	0	0	6.865.692	8.480.892	30.896	94.060	94.110	21.628.392
Decoder	33.281	33.281	33.281	33.281	33.281	33.281	132.099	131.970	9.540	35.428	35.428	207.048
Total	4.752.473	4.350.773	19.603.073	15.569.573	2.559.173	2.658.173	13.863.483	17.093.754	71.332	223.548	223.648	43.463.832
<u>Trainable weights, without fine-tuning of embeddings</u>												
Pencoder	571.392	571.392	571.392	571.392	571.392	571.392	571.392	571.392	n/a	n/a	n/a	571.392
Qencoder	0	0	0	0	0	0	571.392	571.392	n/a	n/a	n/a	571.392
Decoder	33.281	33.281	33.281	33.281	33.281	33.281	132.099	131.970	n/a	n/a	n/a	207.048
Total	604.673	604.673	604.673	604.673	604.673	604.673	1.274.883	1.274.754	n/a	n/a	n/a	1.349.832

* Output size is numer of total entities in the dataset, part of which is masked in each datapoint (numer of categories used on average per data point)