APPENDIX

A. Mnli Results

We provide different models' validation results on MNLI dataset under a different learning rate 1e-5 in Fig. 8, compared to 3e-5 in Fig. 3. It can be observed that under new hyperparameter settings, the training dynamics of the models have changed. The performances of the top two models did not continuously decline with further training, suggesting a less severe overfitting issue. This indicates that the training process of models is highly sensitive to the setting of hyperparameters. In addition, we use our two-phase model selection method for the model training process under the new hyperparameters, and the performance and efficiency are consistent. Despite the changes in the training process, the variation in model performance was not significant enough to impact the effectiveness of our method. Therefore, our approach is robust to different hyperparameter settings in model training and is applicable across various model training scenarios.

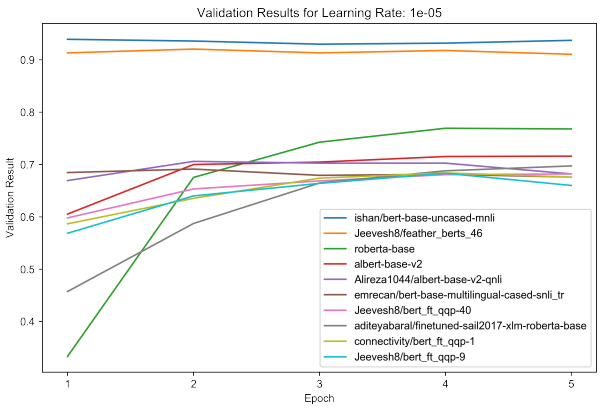

Fig. 8. Top-10 models validation and test results on MNLI dataset. Learning rate is 1e-5, which is different than 3e-5 in the Fig. 3.

B. Model Details

The pre-trained models we use are all from Huggingface's model hub [1]. We list the full names of all the NLP and CV models used in our work in Table VIII. Note that we sometimes use incomplete model names in the main text to save space by removing the name of the repository to which the model belongs. After removing the repository name prefix, the model names are still uniquely summarized in the list of models we use, so partial model names can also be used to pinpoint the corresponding model.

NLP models and CV models are listed in Table IX in total. All models are available using "https://huggingface.co/" as prefix.

[1]https://huggingface.co/models

C. Dataset Details

NLP datasets and CV datasets are listed in Table IX. Some datasets contains multiple subsets. All datasets are available using "https://huggingface.co/" as prefix. GLUE and SuperGLUE are the most common benchmark datasets in NLP. Cifar10 and MNIST are the most common benchmark datasets in CV. Other NLP datasets are described below:

- **LysandreJik/glue-mnli-train** This datasets contain labelled MNLI dataset. The original MNLI dataset in GLUE does not have label, and the label is necessary for our experiment. This task is to predict the relation between the premise and the hypothesis. The result could be entailment, contradiction, or neutral. The labels of this dataset are balanced.
- **SetFit/qnli** This datasets contain labelled qnli dataset. The original qnli dataset in GLUE does not have label, and the label is necessary for our experiment. This task is to predict whether or not the paragraph contains the answer to the question. The labels of this dataset are balanced.
- **xnli** This dataset contains part of MNLI dataset after translated into different languages. The labels of this dataset are balanced.
- **stsb_multi_mt** This task is to score the similarity between two sentences on the scale of 0 to 5. The labels of this dataset are not balanced.
- **anli** This task is the same as MNLI dataset. However, the dataset is collected in an adversarial procedure. The labels of this dataset are not balanced.
- **tweet_eval** This is a sentiment analysis task. The dataset is collected from Tweeter. The labels of this dataset are not balanced.
- **paws** This is a paraphrase identification task. The labels of this dataset are not balanced.
- **financial_phrasebank** This is a sentiment analysis task in the realm of finance. The dataset is collected from financial news. The labels of this dataset are not balanced.
- **yahoo_answers_topics** This is a classification task. The dataset is collected from Yahoo. The labels of this dataset are balanced.

Other CV datasets are described below:

- **food101** This dataset contains 101 kinds of food that need to predict. The size of the image is not the same. The labels of this dataset are balanced.
- **nelorth/oxford-flowers** This dataset contains 102 kinds of flowers that need to predict. The size of the images is not the same. The labels of this dataset are not balanced.
- **Matthijs/snacks** This dataset contains 20 kinds of snacks that need to predict. The size of the images is not the same. The labels of this dataset are slightly unbalanced.
- **beans** This dataset contains 3 kinds of leaves that need to predict. The size of the images is the same. The labels of this dataset are balanced.
- **cats_vs_dogs** This dataset contains images of cats or dogs and is a subset of Asirra dataset. The size of the images

TABLE VIII
NLP AND CV MODELS

NLP model name	CV model name
18811449050/bert_finetuning_test	facebook/deit-base-patch16-224
aditeyabaral/finetuned-sail2017-xlm-roberta-base	facebook/deit-base-patch16-384
albert-base-v2	facebook/deit-small-patch16-224
aliosm/sha3bor-metre-detector-arabertv2-base	facebook/dino-vitb16
Alireza1044/albert-base-v2-qnli	facebook/dino-vitb8
anirudh21/bert-base-uncased-finetuned-qnli	facebook/dino-vits16
aviator-neural/bert-base-uncased-sst2	facebook/vit-msn-base
aychang/bert-base-cased-trec-coarse	facebook/vit-msn-small
bert-base-uncased	google/vit-base-patch16-224
bondi/bert-semaphore-prediction-w4	google/vit-base-patch16-384
CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment	google/vit-base-patch32-224-in21k
CAMeL-Lab–bert-base-arabic-camelbert-mix-did-nadi	lixiqi/beit-base-patch16-224-pt22k-ft22k-finetuned-FER2013-6e-05
classla/bcms-bertic-parlasent-bcs-ter	lixiqi/beit-base-patch16-224-pt22k-ft22k-finetuned-FER2013-7e-05
connectivity/bert_ft_qqp-1	lixiqi/beit-base-patch16-224-pt22k-ft22k-finetuned-FER-5e-05-3
connectivity/bert_ft_qqp-17	microsoft/beit-base-patch16-224
connectivity/bert_ft_qqp-7	microsoft/beit-base-patch16-224-pt22k
connectivity/bert_ft_qqp-96	microsoft/beit-base-patch16-224-pt22k-ft22k
dhimskyy/wiki-bert	microsoft/beit-base-patch16-384
distilbert-base-uncased	microsoft/beit-large-patch16-224-pt22k
DoyyingFace/bert-asian-hate-tweets-asian-unclean-freeze-4	mrgiraffe/vit-large-dataset-model-v3
emrecan/bert-base-multilingual-cased-snli_tr	sail/poolformer_m36
gchhablani/bert-base-cased-finetuned-rte	sail/poolformer_m48
gchhablani/bert-base-cased-finetuned-wnli	sail/poolformer_s36
ishan/bert-base-uncased-mnli	shi-labs/dinat-base-in1k-224
jb2k/bert-base-multilingual-cased-language-detection	shi-labs/dinat-large-in22k-in1k-224
Jeevesh8/512seq_len_6ep_bert_ft_cola-91	shi-labs/dinat-large-in22k-in1k-384
Jeevesh8/6ep_bert_ft_cola-47	Visual-Attention-Network/van-base
Jeevesh8/bert_ft_cola-88	Visual-Attention-Network/van-large
Jeevesh8/bert_ft_qqp-40	oschamp/vit-artworkclassifier
Jeevesh8/bert_ft_qqp-68	nateraw/vit-age-classifier
Jeevesh8/bert_ft_qqp-9	-
Jeevesh8/feather_berts_46	-
Jeevesh8/init_bert_ft_qqp-24	-
Jeevesh8/init_bert_ft_qqp-33	-
manueltonneau/bert-twitter-en-is-hired	-
roberta-base	-
socialmediaie/TRAC2020_IBEN_B_bert-base-multilingual-uncased	-
Splend1dchan/bert-base-uncased-slue-goldtrascription-e3-lr1e-4	-
XSY/albert-base-v2-imdb-calssification	-
Guscode/DKbert-hatespeech-detection	-

is not the same. The labels of this dataset are balanced.

- **trpakov/chest-xray-classification** This dataset contains images of chest x-ray. The size of the images is the same. The labels of this dataset are not balanced.
- **alkzar90/CC6204-Hackaton-Cub-Dataset** This daatset contains images of birds. The size of the images is not the same. The labels of this dataset are not balanced.
- **albertvillanova/medmnist-v2** This dataset contains images about biomedical. The size of the image is the same. The labels of this dataset are not balanced.

D. Experiment on the Number of Dimensions for Max Average Error

As discussed in Eq. 1 and Section V.B., we use top-k maximum average error to measure the model similarity and the parameter k may influence the performance of the model selection algorithm. Thus, we test different values of k while fixing other items. Due to the number of datasets, we choose $k = 5, 10, 15$ for NLP clustering evaluation and $k = 3, 4, 5$ for CV clustering evaluation. The result is shown in Table X. We can find that the influence of parameter k is limited since the silhouette coefficient fluctuates within an acceptable range. Considering that the parameter k in Eq. 1 should be able to filter noise and retain valid information, we choose $k = 5$ in

TABLE IX
NLP AND CV DATASETS

NLP dataset name	CV dataset name
glue	food101
super_glue	nelorth/oxford-flowers
LysandreJik/glue-mnli-train	Matthijs/snacks
SetFit/qnli	beans
xnli	cats_vs_dogs
stsb_multi_mt	trpakov/chest-xray-classification
anli	cifar10
tweet_eval	MNIST
paws	alkzar90/CC6204-Hackaton-Cub-Dataset
financial_phrasebank	albertvillanova/medmnist-v2
yahoo_answers_topics	-

TABLE X
PARAMETER K SELECTION

	NLP			CV		
K Value	5	10	15	3	4	5
Silhouette Coefficient	0.543	0.503	0.535	0.850	0.828	0.821

both tasks.

E. Model cards

A model card is given in Fig. 9. A model card contain the general description of the model, such as structure and training information.

F. K-means Clustering Results

The result of K-means clustering is shown in Table XI. This table is related to Table II in section V. B. Model Clustering. In that section, we explain the result of hierarchical clustering in detail. We conclude that the result of hierarchical clustering is effective since the in-cluster models share the same model structure or training dataset while the silhouette coefficient is high. Here we give the result of K-means clustering to better prove our conclusion. Both the NLP clustering result and CV clustering result of the K-means clustering algorithm show less connection between in-cluster models. In the NLP part, the 2 biggest clusters, C_2 and C_8, consist of a mix of models that have different structures and training datasets. In the CV part, there is a cross mixing in C_6 and C_7, and the biggest cluster, C_4, does not show consistency in either model structure or training dataset. Thus, we take the method of hierarchical clustering as the main line of this paper.

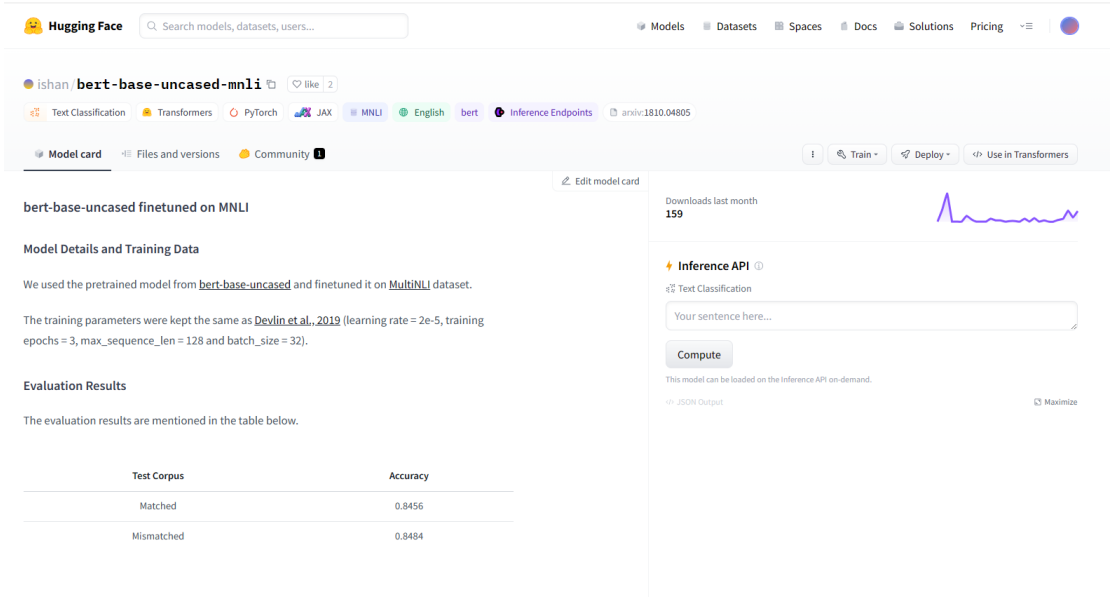

Fig. 9. Model card of bert-base-uncased-mnli. Each model on HuggingFace has a model card to describe the model.

TABLE XI
MODEL CLUSTERING RESULTS USING K-MEANS

Model Clusters of Natural Language Processing		
Cluster	**Size**	**Pre-trained Models**
C_1	2	gchhablani–bert-base-cased-finetuned-rte, anirudh21–bert-base-uncased-finetuned-qnli
C_2	5	Jeevesh8–bert_ft_cola-88, DoyyingFace–bert-asian-hate-tweets-asian-unclean-freeze-4, bert-base-uncased , aditeyabaral–finetuned-sail2017-xlm-roberta-base, Jeevesh8–512seq_len_6ep_bert_ft_cola-91
C_3	2	manueltonneau–bert-twitter-en-is-hired, aychang–bert-base-cased-trec-coarse
C_4	2	XSY–albert-base-v2-imdb-calssification, distilbert-base-uncased
C_4	4	ishan–bert-base-uncased-mnli, Alireza1044–albert-base-v2-qnli, albert-base-v2, Jeevesh8–feather_berts_46 :
C_5	2	CAMeL-Lab–bert-base-arabic-camelbert-mix-did-nadi, aliosm–sha3bor-metre-detector-arabertv2-base
C_6	3	socialmediaie–TRAC2020_IBEN_B_bert-base-multilingual-uncased, jb2k–bert-base-multilingual-cased-language-detection, emrecan–bert-base-multilingual-cased-snli_tr
C_7	2	dhimskyy–wiki-bert, bondi–bert-semaphore-prediction-w4
C_8	5	Jeevesh8–init_bert_ft_qqp-33, Jeevesh8–bert_ft_qqp-68, Jeevesh8–bert_ft_qqp-40, connectivity–bert_ft_qqp-1, Jeevesh8–bert_ft_qqp-9
C_9	4	connectivity–bert_ft_qqp-96, connectivity–bert_ft_qqp-7, connectivity–bert_ft_qqp-17, Jeevesh8–init_bert_ft_qqp-24
C_{10}	2	Splend1dchan–bert-base-uncased-slue-goldtrascription-e3-lr1e-4, Jeevesh8–6ep_bert_ft_cola-47
Model Clusters of Computer Vision		
Cluster	**Size**	**Pre-trained Models**
C_1	6	shi-labs/dinat-large-in22k-in1k-224, shi-labs/dinat-large-in22k-in1k-384, microsoft/beit-base-patch16-224-pt22k-ft22k, lixiqi/beit-base-patch16-224-pt22k-ft22k-finetuned-FER2013-7e-05, lixiqi/beit-base-patch16-224-pt22k-ft22k-finetuned-FER2013-6e-05, lixiqi/beit-base-patch16-224-pt22k-ft22k-finetuned-FER-5e-05-3
C_2	2	nateraw/vit-age-classifier, facebook/dino-vitb16
C_3	3	sail/poolformer_m48, sail/poolformer_m36, sail/poolformer_s36
C_4	7	facebook/vit-msn-small, facebook/vit-msn-base, facebook/deit-base-patch16-384, google/vit-base-patch32-224-in21k, Visual-Attention-Network/van-large, facebook/deit-base-patch16-224, facebook/dino-vits16
C_5	4	Visual-Attention-Network/van-base, microsoft/beit-large-patch16-224-pt22k, facebook/deit-small-patch16-224, shi-labs/dinat-base-in1k-224
C_6	2	microsoft/beit-base-patch16-384, google/vit-base-patch16-384
C_7	2	microsoft/beit-base-patch16-224, google/vit-base-patch16-224