Supplementary Material

Proof of Proposition 1

Denote $\text{vec}(\cdot)$ vectorization of a matrix. It follows that

$$\kappa(i,j) \triangleq KA(\mathbf{F}_i, \mathbf{F}_j) = \frac{\langle \text{vec}(\mathbf{F}_i^\top \mathbf{F}_i), \text{vec}(\mathbf{F}_j^\top \mathbf{F}_j) \rangle}{\|\text{vec}(\mathbf{F}_i^\top \mathbf{F}_i)\| \cdot \|\text{vec}(\mathbf{F}_j^\top \mathbf{F}_j)\|},$$

which is inner product between normalized $\text{vec}(\mathbf{F}_i^\top \mathbf{F}_i)$ and $\text{vec}(\mathbf{F}_j^\top \mathbf{F}_j)$. Hence κ is the Gram matrix of

$$\left[\frac{\text{vec}(\mathbf{F}_1^\top \mathbf{F}_1)}{\|\text{vec}(\mathbf{F}_1^\top \mathbf{F}_1)\|} \quad \cdots \quad \frac{\text{vec}(\mathbf{F}_n^\top \mathbf{F}_n)}{\|\text{vec}(\mathbf{F}_n^\top \mathbf{F}_n)\|} \right],$$

which is PSD.

Proof of Proposition 2

Recall the definition: a set function $f(\mathcal{S})$ is submodular, if for any subsets $\mathcal{S} \subseteq \mathcal{S}' \subseteq \mathcal{Z}$, and $i \in \mathcal{Z} - \mathcal{S}'$,

$$f(\mathcal{S} \cup \{i\}) - f(\mathcal{S}) \geq f(\mathcal{S}' \cup \{i\}) - f(\mathcal{S}').$$

Referring to Eq. (3), we realize that the left side equals $H(\{i\}|\mathcal{S}) - H(\{i\}|\bar{\mathcal{S}}_i)$, and the right side equals $H(\{i\}|\mathcal{S}') - H(\{i\}|\bar{\mathcal{S}}_i')$. Since conditioning on more variables reduces entropy, we have $H(\{i\}|\mathcal{S}) \geq H(\{i\}|\mathcal{S}')$ and $H(\{i\}|\bar{\mathcal{S}}_i) \leq H(\{i\}|\bar{\mathcal{S}}_i')$. It therefore holds that $H(\{i\}|\mathcal{S}) - H(\{i\}|\bar{\mathcal{S}}_i) \geq H(\{i\}|\mathcal{S}') - H(\{i\}|\bar{\mathcal{S}}_i')$.

Tasks that Huggingface Checkpoints were Trained on

1. **albert-base**, **albert-large**: masked language modeling + sentence order prediction
2. **bart-base**, **bart-large**, **bart-large-cnn**: text denoising
3. **bert-base-cased**, **bert-base-uncased**, **bert-large-cased**, **bert-large-uncased**: masked language modeling + next sentence prediction
4. **distilbert-base-cased**, **distilbert-base-uncased**, **distilbert-base-multilingual**: knowledege distillation on bert (matching representatations of bert)
5. **gpt**, **gpt2**, **gpt-medium**, **gpt-large**: causal language modeling
6. **longformer-base**: masked language modeling
7. **roberta-base**, **roberta-large**: masked language modeling
8. **roberta-large-mnli**: masked language modeling + entailment (finetuned on MNLI dataset)
9. **t5-3b**, **t5-base**, **t5-small**, **t5-large**: text-to-text generation
10. **xlm-clm-ende-1024**, **xlm-mlm-100-1280**, **xlm-mlm-17-1280**, **xlm-mlm-ende-1024**, **xlm-mlm-enfr-1024**, **xlm-mlm-enro-1024 xlm-roberta-base**, **xlm-roberta-large**: crosslingual masked language modeling
11. **xlnet-base-cased**, **xlnet-base-large**: permutation language modeling

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Details on Training

For experiments in section 5.1, we use a batch size of 32 sentences, adam optimizer with a learning rate of 1e-3. We run for 40 epochs and report the test metric at the "best" validation epoch.

For experiments in section 5.2, all checkpoints are instances of resnet-50. They are trained by a batch size of 128, and an initial learning rate of 0.1. We run for 200 epochs, with learning rate decay at the 60th, 120th and 160th epoch. A typical validation accuracy from these checkpoint (on its own task) is about 83% (reasonably good). For the 20 new tasks, we experiment with a softmax classifier on top of selected checkpoints. The learning rate is kept at 0.1. We report the best validation accuracy for each of the 20 tasks. For each task, its validation set is standard cifar100 validation split, but only includes the classes that are involved in this task.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See the last sentence of the 1st paragraph in section 1.
   (b) Did you describe the limitations of your work? [Yes] See Conclusion
   (c) Did you discuss any potential negative societal impacts of your work? [No] There does not seem to be any as far as the authors can tell
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] These should be very clear for Proposition 1 and 2.
   (b) Did you include complete proofs of all theoretical results? [Yes] See supplementary material
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] They are in supplementary material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See supplementary material
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See fig. 3
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] They are not detailed in the paper. But the computational resource is an internal cluster.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See section 5.1, reference [1]
   (b) Did you mention the license of the assets? [No] The data has to be obtained by emailing the author(s) of [1]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] These are not discussed in the main draft. But the data is obtained by emailing the authors, see question (b).
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] These concerns do not exist for this paper, as far as the authors can tell.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

References

[1] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, N. A. Smith. Linguistic Knowledge and Transferability of Contextual Representations. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019.