Explaining the Effectiveness of Multi-Task Learning for Efficient Knowledge Extraction from Spine MRI Reports

Arijit Sehanobish, McCullen Sandora, Nabila Abraham,
Jayashri Pawar, Danielle Torres, Anasuya Das, Murray Becker,
Richard Herzog, Benjamin Odry, Ron Vianu
Covera Health, New York City, New York
{arijit.sehanobish, mccullen.sandora, nabila.abraham,
jayashri.pawar, danielle.torres, anasuya.das, rherzog,
murray.becker, benjamin.odry, ron.vianu}@coverahealth.com

Abstract

Pretrained Transformer based models finetuned on domain specific corpora have changed the landscape of NLP. However, training or finetuning these models for individual tasks can be time consuming and resource intensive. Thus, a lot of current research is focused on using transformers for multi-task learning (Raffel et al., 2020) and how to group the tasks to help a multi-task model to learn effective representations that can be shared across tasks (Standley et al., 2020; Fifty et al., 2021). In this work, we show that a single multi-tasking model can match the performance of task specific models when the task specific models show similar representations across all of their hidden layers and their gradients are aligned, i.e. their gradients follow the same direction. We hypothesize that the above observations explain the effectiveness of multi-task learning. We validate our observations on our internal radiologist-annotated datasets on the cervical and lumbar spine. Our method is simple and intuitive, and can be used in a wide range of NLP problems.

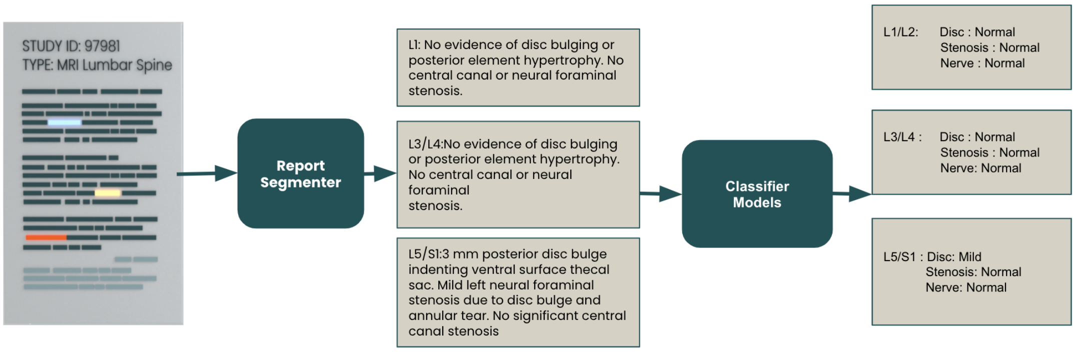
1 Introduction

Since the seminal work by (Vaswani et al., 2017), Transformers have become the main architecture for almost all Natural Language Processing (NLP) tasks. Self-supervised pretraining of massive language models like BERT (Devlin et al., 2019) and GPT (Brown et al., 2020) has allowed practitioners to use these large language models with little or no finetuning to various downstream tasks. Multi-task learning (MTL) in NLP has been a very promising approach and has shown to lead to performance gains even over task specific fine-tuned models (Worsham and Kalita, 2020; Raffel et al., 2020; Aribandi et al., 2021). However, applying these large pre-trained Transformer models to downstream medical NLP tasks is quite difficult. Medical NLP has its unique challenges ranging from domain specific corpora, noisy annotation labels and scarcity of high quality labeled data. Despite these challenges, a number of researchers and practitioners have successfully finetuned these large language models for various medical NLP tasks. However, there is not much literature that uses multi-task learning in medical NLP to classify and extract diagnoses from clinical text (Peng et al., 2020; Crichton et al., 2017). Moreover, there is almost no work in predicting spine pathologies from the radiologists' notes (Azimi et al., 2020).

In this article, we are interested in extracting information from radiologists' notes on the cervical and the lumbar spine. In a given note, the radiologist discusses the specific, and often multiple pathologies, present in the medical images and grade their severity. Extracting relevant pathologies from these reports can facilitate the creation of structured databases that can be used for a number of downstream use-cases, such as cohort creation, quality assessment and outcome tracking. Single-task learning for information extraction in medical NLP has enjoyed much success in deep learning (Kanakarajan et al., 2021).

However, an ultimate NLP system for a complete understanding of the medical report must be able to perform many diverse information extraction and classification tasks simultaneously and efficiently. Such a system can be enabled by MTL, where one model shares weights across multiple tasks and makes multiple inferences in one forward pass. Such networks can not only be trained with limited resources, but are more scalable and deployable when compared to several single-task models. Moreover, the shared features within these MTL networks can induce more robust regularization and boost performance. Thus there is a lot of interest in the academic and industry research communities to understand when multi-task learning improves performance over single-tasking models (Crawshaw, 2020), and how to group a diverse set of tasks to

Proceedings of NAACL-HLT 2022: Industry Track Papers, pages 130 - 140
July 10-15, 2022 ©2022 Association for Computational Linguistics



L1: No evidence of disc bulging or posterior element hypertrophy. No central canal or neural foraminal stenosis.

L3/L4:No evidence of disc bulging or posterior element hypertrophy. No central canal or neural foraminal stenosis.

L5/S1:3 mm posterior disc bulge indenting ventral surface thecal sac. Mild left neural foraminal stenosis due to disc bulge and annular tear. No significant central canal stenosis

L1/L2: Disc : Normal
Stenosis : Normal
Nerve : Normal

L3/L4 : Disc : Normal
Stenosis : Normal
Nerve: Normal

L5/S1 : Disc: Mild
Stenosis: Normal
Nerve: Normal

Figure 1: Figure showing how a report looks as it goes through our pipeline.

encourage the model to learn a representation that can be shared across tasks (Standley et al., 2020; Fifty et al., 2021; Bingel and Søgaard, 2017; Zamir et al., 2020). Some of the aforementioned works, most notably in (Shui et al., 2019), define a notion of task similarity via the Wasserstein distance and show that a small Wasserstein distance between tasks aids in MTL.

This work is an extension of our earlier work (Sehanobish et al., 2022) where we used parameter efficient MTL models to extract information from cervical spine. In that work, we defined tasks as a conditional distribution over the classes, and we attributed our success of MTL to smaller Wasserstein distance between tasks. However, computing Wasserstein distance is expensive and suffers from the curse of dimensionality (Cuturi, 2013), which requires the number of samples to be significantly larger than the dimension of the representation (768 for many transformer models) in order for the distance to be accurately estimated. This prevents us from being able to estimate Wasserstein distance for some of our minority classes, which have about 200 examples. Even for majority classes where we have about 5k samples, our work suffers from large error rates. Thus, to alleviate the above drawbacks, in this work, we sought to use methods that are applicable to small data regimes that lie in high dimensional space.

Inspired by the work of (Yu et al., 2020; Chen et al., 2020) and (Kornblith et al., 2019), we hypothesize if the single-task models show similar representations across their hidden layers and the task specific gradients are *aligned* (see Definition 1 in Section 4.2), the multi-task model can match or outperform the task-specific, single-task models. We validate this hypothesis on two multi-task settings on our internal datasets: **(a)** Four of the most common pathologies in the cervical spine - central canal and foraminal stenosis, disc herniation and cord compression, and **(b)** Three pathologies in the lumbar spine - central canal stenosis, disc herniation and nerve root impingement.

In this work, we (**a**) extend our novel pipeline to extract and predict the severity of various pathologies in the lumbar and cervical spine at *each motion segment*, (**b**) compute Central Kernel Alignment (CKA) and show similarity between the transformer layers trained for individual tasks on a given dataset, (**c**) compute dot products between the gradients of the task specific loss functions with respect to various parameters and show that most of the gradients flow along a similar trajectory and (**d**) show how to leverage that information into a simple MTL framework allowing us to achieve significant model compression during deployment and also speed up our inference without sacrificing the accuracy of our predictions.

2 Datasets

We use an internal dataset consisting of radiologists' MRI reports on the cervical and the lumbar spine. Our dataset is heterogeneous and is diversely sampled from a large number of different radiology practices and medical institutions; the cervical MRI data consists of 1578 reports from 97 different radiology practices detailing various pathologies of the cervical spine and our lumbar MRI data contains 2004 reports from 170 different practices.

We annotate the cervical reports with the 4 following pathologies: spinal stenosis, disc herniation, cord compression, and neural foraminal stenosis, and the lumbar reports with the 3 pathologies: disc herniation, spinal stenosis, and nerve impingement. Each of these pathologies is accompanied by an indication of severity. In the cervical reports, the three categories for the central canal stenosis are based on gradation; none/mild are not clinically significant, moderate and severe definitions involve cord compression or flattening. The moderate versus severe gradation refers to the varying degrees of cord involvement. For disc herniation and central canal stenosis, the categories are based on a continuous spectrum and it is a standard practice in radiology for any continuous spectrum to be bucketed in mild, moderate and severe discrete categories. Cord compression severity is binary: compression/signal change versus none. This is because both cord compression and signal change can cause symptoms, and are therefore clinically relevant. Foraminal stenosis is treated as a binary task as well: severe versus non-severe, as severe foraminal stenosis may indicate nerve impingement, which is clinically significant. Similar considerations are taken into account when annotating the lumbar reports. The splits and the details of each category can be found in Table 1. The data distribution is highly imbalanced, and about 25% of these reports are OCR-ed, which leads to additional challenges stemming from bad OCR errors.

Dataset	Pathology	Training Label Distribution	Test Label Distribution
Lumbar	Disc	None/Mild : 1885 Moderate : 1998 Severe : 456	None/Mild : 1068 Moderate :1588 Severe :332
	Stenosis	None/Mild : 3787 Moderate : 350 Severe : 202	None/Mild : 2411 Moderate : 304 Severe : 273
	Nerve	Normal : 3790 Abnormal : 549	Normal : 2376 Abnormal : 612
Cervical	Disc	None/Mild : 2731 Moderate : 2699 Severe : 797	None/Mild : 401 Moderate : 378 Severe : 101
	Stenosis	None/Mild : 5488 Moderate : 561 Severe : 178	None/Mild : 793 Moderate : 68 Severe : 19
	Cord Compression	Normal : 5702 Abnormal : 525	Normal : 806 Abnormal : 74
	Neural Foraminal Stenosis	Normal : 5262 Abnormal : 965	Normal : 789 Abnormal : 91

Table 1: Table showing statistics of our datasets

For a given report, each task is to predict the severity of a pathology for each motion segment - the smallest physiological motion unit of the spinal cord (Swartz et al., 2005). Breaking information down at the motion segment level in this way enables pathological findings to be correlated with clinical exam findings, and can inform future treatment interventions.

Every report is tagged by annotators with labels for relevant pathologies and severities, along with span information indicating which part(s) of the report mentions each pathology. For example, in a report for the lumbar spine, the sentence “L1-L2: There is no disc herniation. No spinal canal or foraminal narrowing" would be given normal or 0 class for each of the 3 pathologies (central canal stenosis, disc herniation and nerve root impingement). Similarly in a cervical spine report, the sentence “ C2-3: Normal; no disc herniation or bulge. No central canal stenosis or neuroforaminal narrowing" would be given a normal or 0 class for all the 4 pathologies. An example of a full radiology report can be found in Appendix A.

3 Workflow

In this section, we will briefly describe our pipeline. The reports are first de-identified according to HIPAA regulations. Next, a Spacy (Honnibal et al., 2020) parser is used to break the report into sentences.

A BERT based NER model which we call the *report segmenter* is then used to identify the motion segment(s) referenced in each sentence, and all the sentences containing a particular motion segment are concatenated together. This report segmenter has been shown to achieve an F1 score of .9 on our internal datasets, and the same model is common across both the lumbar and the cervical datasets. More details about the NER model and the hyperparameters used to train it can be found in Appendix B and C. All pathologies are predicted using the concatenated text for a particular motion segment. Finally, the severities for each pathology are modeled as multi-label classification problem, and a pre-trained transformer is finetuned using the text for each motion segment.

For more details about our pipeline and data processing, please see Appendix B. Figure 1 breaks down how a report looks as it is processed through our spine pipeline.

4 Similarity of Representations between Task Specific Models

In this section we will describe our methodology to understand the similarity between the representations of various single task models. For all the experiments in this section, we use the PubMedBERT (Gu et al., 2020) as the backbone.

4.1 Central Kernel Alignment

We use the linear Central Kernel Alignment (CKA), introduced in (Kornblith et al., 2019). CKA is a scalar similarity index that can be used to compare representations within and across neural networks. (Linear) CKA can be defined by the following:

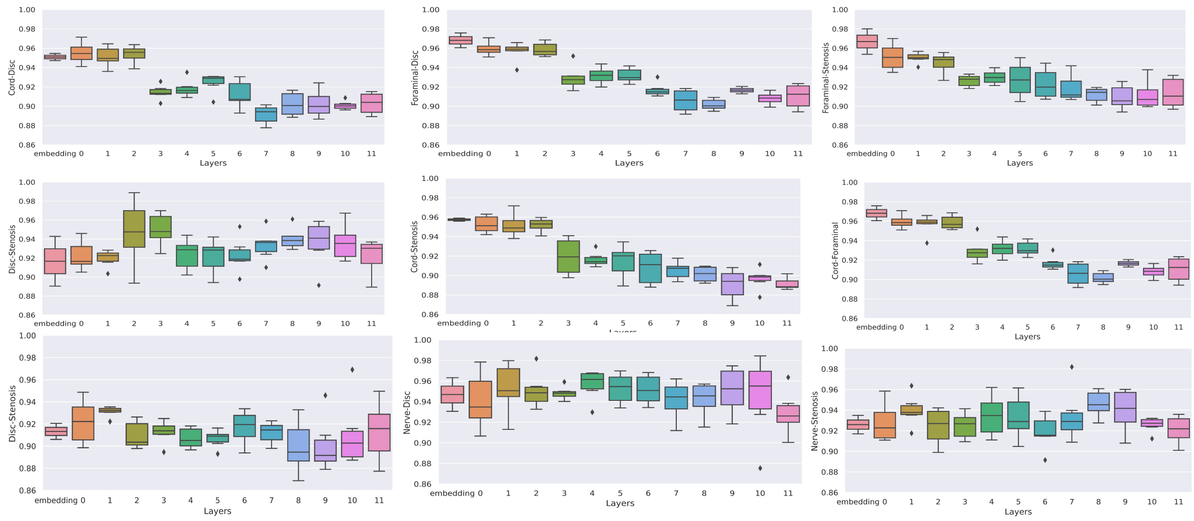


Figure 2: CKA between activation matrices between different finetuned single-task models. The top 2 rows are single-task models trained to predict specific pathologies from cervical dataset and the bottom row for the lumbar dataset. The y-axis is chosen to be between the min and the max values, i.e. in the interval (.86, 1.0)

Given N examples and two activation outputs on these examples, $R_1 \in \mathbb{R}^{N \times d_1}$ and $R_2 \in \mathbb{R}^{N \times d_2}$,

$$\text{CKA}(R_1, R_2) = \frac{||R_1^\top R_2||_F}{||R_1^\top R_1||_F ||R_2^\top R_2||_F} \quad (1)$$

where $|| \cdot ||_F$ is the Frobenius norm.

It is widely believed that similar representations lead to similar performances on downstream tasks (Nguyen et al., 2021). In this work, we compare the representations learned by various single tasking models. For two single task models trained on a specific part of a spine, the CKA between the matrix of activations for each layer of the corresponding models is computed. For illustration purposes, we collect all the CKA values for various activation matrices in a given layer and plot them in a box plot, as shown in figure 2. We observe that for various tasks on both cervical and lumbar spine, all layers of the task specific models learn similar representations.

Additional results on comparing models from the tasks from the lumbar dataset and the cervical dataset can be found in Appendix D.

However, the high value of CKA may also be attributed to the following factors : (i) larger and deeper networks converge to similar solutions (Morcos et al., 2018) and (ii) CKA values do not change drastically when models start from pretrained weights and are only trained for a few epochs (Mirzadeh et al., 2021).

Thus in addition to the above analysis of the activations with the CKA, in the next subsection we look at the gradient level information to understand the trajectory of the task specific learned activations.

4.2 Gradient Alignment

There has been a lot of work in understanding the task specific gradients in the context of MTL. Given tasks $T_1, \cdots T_n$ (for example, they can be classification tasks), one can define n loss functions $\mathcal{L}_{T_j}$ for each task T_j. In our work, all loss functions are cross-entropy losses. Then the task specific gradients are defined to be $\nabla_{\theta_j} \mathcal{L}_{T_j}$ where θ_j are the parameters of the task specific model. More specifically, it is shown in (Chen et al., 2018), that MTL is competitive with single task learners when the norms of the task specific gradients have similar magnitudes. However in (Yu et al., 2020; Javaloy and Valera, 2021), the authors show that the direction of the gradient flow is more important than the magnitude for the success of MTL. More precisely, Theorem 1 in (Yu et al., 2020), shows that the multitask objective converges to the optimum of one of the tasks or a sub-optimal minima in the presence of conflicting gradients. Furthermore, authors in (Javaloy and Valera, 2021) use a synthetic toy example to show the difficulties of optimizing a multi-task loss in the presence of conflicting gradients.

Inspired by the above works, we define the following:

Definition 1 *Two gradient vectors g_i and g_j are aligned if $g_i \cdot g_j > 0$, i.e. the vectors are pointing* *in the same direction.*

To show that the gradients get more aligned as models are trained, we store the gradients for all the parameters for all the mini-batches after every epoch. We then compute the dot products between the corresponding gradients for two tasks. We observe that as the task specific models gets trained, an overwhelming proportion of these gradients are aligned (see Table 2). To illustrate our findings, we take the proportion of these aligned parameters in a given layer and plot them using a box plot in Figure 3. Finally, we compute the proportion of weights across all layers for which the gradients are aligned which we call the Average Proportion of Aligned Gradients (APAG).

$$\text{APAG} = \frac{1}{N_{\text{layers}}}\frac{1}{N_{\text{heads}}}\sum_{\textit{layers}}\sum_{\textit{heads}}\theta\left(g_i \cdot g_j\right) \quad (2)$$

where $\theta(x)$ is the Heaviside step function. This is a scalar value that summarizes the box plot and we show the progression of alignment of the gradients as training progress and the end of the training (Table 2 and Table 6 in Appendix D respectively). Note that, in the above formula, the token embedding layer is included in the computation and it is assumed to have 1 head.

Dataset	Task Comparisons	Epoch 1	Epoch 2	Epoch 3	Epoch 4
Cervical	Cord-Stenosis	.46	.67	.75	.81
	Cord-Disc	.37	.52	.69	.74
	Cord-Foraminal	.49	.61	.77	.83
	Disc-Stenosis	.51	.62	.69	.78
	Disc-Foraminal	.47	.59	.65	.73
	Foraminal-Stenosis	.54	.66	.72	.79
Lumbar	Disc-Stenosis	.44	.53	.59	.68
	Nerve-Stenosis	.51	.57	.66	.73
	Disc-Nerve	.48	.55	.63	.71

Table 2: Results showing the Average Proportion of Aligned Gradients between various task specific models at various epochs.

To summarize: The task specific models not only show similar representations but they arrive at these representations by moving in a similar direction after starting from the pretrained weights. We would also like to point that we observe similar behavior when we run our experiments with the BERT (Devlin et al., 2019) and the Clinical BERT (Alsentzer et al., 2019) models.

5 Results on Multi-Task Models

In this section, we give empirical evidence on the success of MTL for our datasets. The results shown in this section are from our test set.

For our classification task, the PubMedBERT model is used as the backbone. This BERT model is finetuned on the the cervical tasks resulting in 4 task-specific BERT sequence classifier models which provides our baseline results. For the lumbar dataset, the PubMedBERT model is finetuned on the 3 classification tasks resulting in 3 task-specific BERT sequence classifier models.

Now, instead of finetuning the task specific models for extracting various pathology information from the cervical spine dataset, 4 classifier heads (i.e. 4 linear layers) are added to a single PubMedBERT model to create an output layer of shape $[3, 3, 2, 2]$, where the first 3 outputs correspond to the logits for the stenosis severity prediction, the next 3 for the disc severity, the next 2 for the cord severity and the final 2 logits for the foraminal severity. For the lumbar dataset, 3 classifier heads are added to the PubMedBERT model to create an output layer of shape $[3, 3, 2]$, where the first 3 outputs correspond to the logits for the stenosis severity prediction, the next 3 for the disc severity, and the final 2 logits for the nerve severity.

For the experiments, with both the datasets, a dropout of .5 is added to the BERT vectors before passing them to the classifier layers. Each of these classifier heads is trained with a cross entropy loss with the predicted logits and the ground truth targets. All the losses are added up which allows the gradients to backpropagate through the whole model and train these classifier heads jointly.

The results for our experiments are shown in Table 3 for the lumbar dataset and Table 4 for the cervical dataset.

Backbone	Model	Disc	Stenosis	Nerve
BERT BASE	Baseline (single tasker)	$.78 \pm .03$	$.79 \pm .02$	$.8 \pm .03$
	Multi-Tasking	$.77 \pm .02$	$.78 \pm .01$	$.79 \pm .02$
CLINICAL BERT	Baseline (single tasker)	$.81 \pm .03$	$.83 \pm .02$	$.82 \pm .03$
	Multi-Tasking	$.83 \pm .02$	$.8 \pm .04$	$.81 \pm .02$
MSR PubMedBERT	Baseline (single tasker)	$.82 \pm .03$	$.83 \pm .03$	$.81 \pm .04$
	Multi-Tasking	$\mathbf{.84 \pm .01}$	$\mathbf{.84 \pm .03}$	$\mathbf{.86 \pm .04}$

Table 3: Table showing the macro F1 scores over 5 trials of our Baseline and Multi-Tasking Models on the Lumbar Dataset.

For fair comparisons, we also conduct experi-

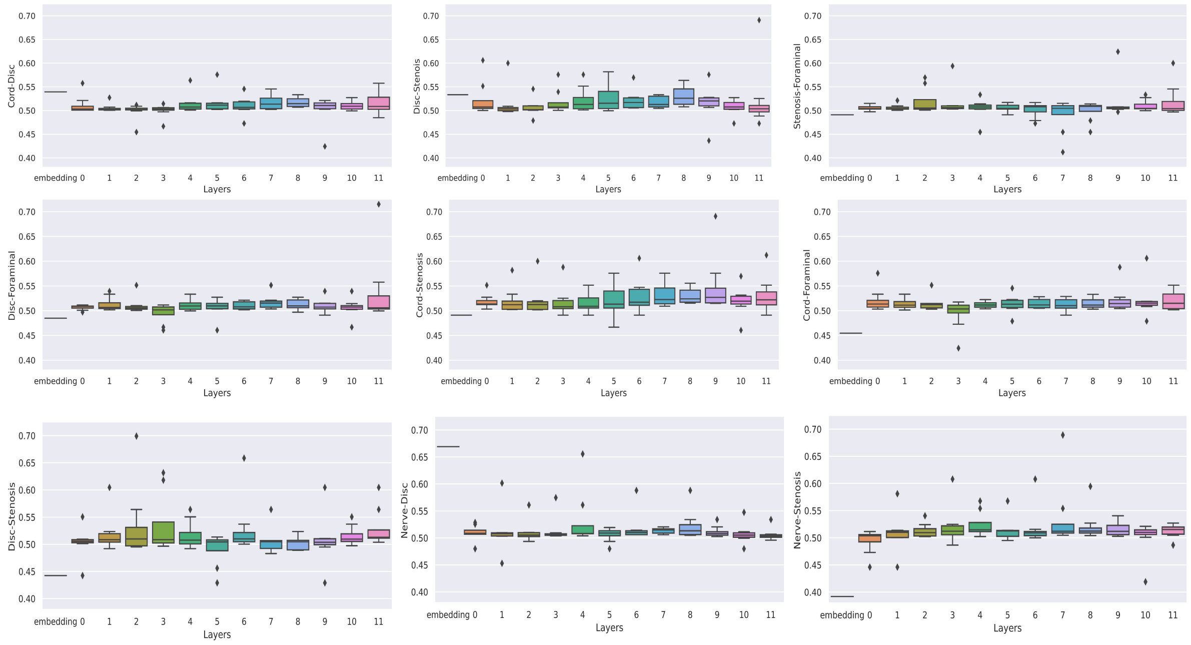


Figure 3: Box Plot showing the proportion of aligned gradients between various task specific models, after training. The top 2 rows are single tasking models trained to predict specific pathologies from the cervical dataset and the bottom row for the lumbar dataset. The y-axis is chosen to be between the min and the max values, i.e. in the interval (.38, .725).

ments with the BERT base and the Clinical BERT models as well. We notice that the PubMedBERT produces slightly better results than both the Clinical BERT and the BERT base. We believe this is due to the fact that the vocabulary for PubMedBERT is tailored for clinical text, unlike that of Clinical BERT, which uses the same vocabulary as that of BERT.

Backbone	Model	Stenosis	Disc	Cord	Foraminal
BERT BASE	Baseline (single tasker)	.62 ± .03	.64 ± .03	.70 ± .03	.79 ± .03
	Multi-Tasking	.62 ± .02	.65 ± .03	.72 ± .02	.78 ± .01
CLINICAL BERT	Baseline (single tasker)	.64 ± .05	.66 ± .02	.71 ± .02	.82 ± .01
	Multi-Tasking	.63 ± .02	.67 ± .01	.75 ± .01	.79 ± .03
MSR PubMedBERT	Baseline (single tasker)	.66 ± .03	.68 ± .04	**.73 ± .05**	**.84 ± .01**
	Multi-Tasking	**.67 ± .01**	**.69 ± .01**	.72 ± .04	.83 ± .03

Table 4: Table showing the macro F1 scores over 5 trials of our Baseline and Multi-Tasking Models on the Cervical Dataset.

The hyperparameters and other training and implementation details can be found in Appendix C.

6 Deployment

We deploy our spine pipeline system on an AWS p3.2x machine with a single NVIDIA V100 GPU. Reports are passed through the pipeline daily and first go through the report segmenter which tags sentences belonging to our set of motion segments. Post-processing is done per report to aggregate sentences belonging to each motion segment group and to filter out any reports that do not contain motion segments. Each grouping of motion segments is individually classified through our MTL model to predict a severity class per pathology. Both the report segmenter and the multi-tasking model are processed in batch mode with latencies of 31ms/report and 56ms/report, respectively. Compared to single pathology models, we observe a 3x improvement in latency per study when using the MTL pathology model. The spine pipeline is routinely evaluated in an offline setting for studies that do not produce any motion segment groupings or fail to capture any sentences for a given motion segment, per report. Our current deployment only supports the lumbar reports and we are in the process of extending our deployment to also support the cervical pathologies.

7 Conclusion and Future Work

In this work, a simple multi-tasking model is presented that is competitive with task specific models. Instead of training and deploying task specific models, only one model is trained and deployed. This allows us to save significant costs during training and faster inference during deployment while achieving significant model compression, without any loss in the quality of performance. Our work opens the possibility of using multi-tasking models to extract information over various different body parts, allowing users to leverage large transformer models using limited compute resources.

Our novel pipeline is one of the very few works that attempts to extract pathologies and their severities from a heterogeneous source of radiologists' notes on lumbar and cervical spine MRIs at the level of *motion segments*. These findings suggest that our approach may not only be more widely generalizable and applicable, but also more clinically actionable.

We believe our analysis with CKA and gradient alignment sheds more light on the success of MTL. This insight has led to our process change from single-task BERT based models to a more cost-effective MTL system. Our analysis is widely applicable for other datasets and tasks.

It is tempting to ask if one can use one multi-task model for both the lumbar and the cervical datasets. This is a work in progress and we have found strong similarity between single task models in the two datasets (most notably between the lumbar disc and the cervical disc models and the lumbar stenosis and the cervical stenosis models). However, unlike in the above analysis, we see low CKA scores between various other task specific models which may make MTL difficult (see Appendix D). We are in the process of using our analysis, along with insights borrowed from (Standley et al., 2020; Yu et al., 2020) to either group tasks from the two datasets or align different task-specific gradients to create an efficient learner.

The biggest drawback of our work is the limited amount of data on which our observations are verified. We are actively addressing this issue as we annotate more reports concerning various pathologies in different body parts.

Ethical Considerations

Because of legal and institutional concerns arising from the sensitivity of clinical data, it is difficult for the NLP community to gain access to relevant data except for MIMIC (Johnson et al., 2016). Despite its large size (covering over $58k$ hospital admissions), it is only representative of patients from a particular clinical domain (the intensive care unit) and geographic location (a single hospital in the United States). Such a sample is not representative of either larger population of patient admissions or other geographical regions/hospital systems. We have tried to address the second issue by collecting data across multiple practices in the US. However, it is impossible to predict whether our models will generalize to the entire patient population without actually evaluating on *all* the different radiology practices. Thus we have to be extra careful about out-of-distribution data since the actionable insights we generate from our models can be potentially faulty and can lead to severe consequences.

Finally, we recognize the need to minimize ethical risks of AI implementation which can include threats to privacy and confidentiality, informed consent, and patient autonomy. We strongly believe that stakeholders should be encouraged to be flexible in incorporating AI technology, most likely as a complementary tool and not a replacement for a physician. Thus, we develop our workflows, annotation guidelines and generate actionable insights by working in conjunction with a varied group of radiologists and medical professionals.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2021. Ext5: Towards extreme multi-task scaling for transfer learning.

Parisa Azimi, Taravat Yazdanian, Edward C. Benzel, Hossein Nayeb Aghaei, Shirzad Azhari, Sohrab Sadeghi, and Ali Montazeri. 2020. A Review on the Use of Artificial Intelligence in Spinal Diseases. *Asian Spine J*, 14(4):543–571.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 793–802. PMLR.

Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. 2020. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, 18(1):368.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2292–2300.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. Efficiently identifying task groupings for multi-task learning.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. ArXiv.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Adrián Javaloy and Isabel Valera. 2021. Rotograd: Gradient homogenization in multitask learning.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Dilan Görür, Razvan Pascanu, and Hassan Ghasemzadeh. 2021. Linear mode connectivity in multitask and continual learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5732–5741.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. 2021. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on BERT for biomedical text mining. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 205–214, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Arijit Sehanobish, Nathaniel Brown, Ishita Daga, Jayashri Pawar, Danielle Torres, Anasuya Das, Murray Becker, Richard Herzog, Benjamin Odry, and Ron Vianu. 2022. Efficient extraction of pathologies from C-spine radiology reports using multi-task learning.

Changjian Shui, Mahdieh Abbasi, Louis-Émile Robitaille, Boyu Wang, and Christian Gagné. 2019. A principled approach for learning task similarity in multitask learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3446–3452. ijcai.org.

Trevor Standley, Amir Roshan Zamir, Dawn Chen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9120–9132. PMLR.

Erik E. Swartz, R. T. Floyd, and Mike Cendoma. 2005. Cervical Spine Functional Anatomy and the Biomechanics of Injury due to Compressive Loading. *Journal of athletic training*, 40(3):155–161.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

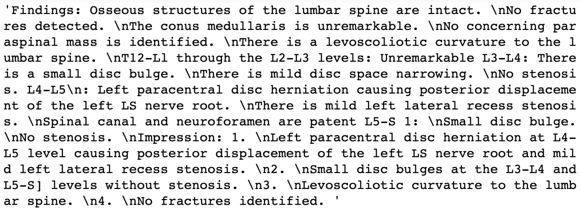
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Joseph Worsham and Jugal Kalita. 2020. Multi-Task Learning for Natural Language Processing in the 2020s: Where are we going? *Pattern Recognition Letters*, 136:120–126.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Amir Roshan Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J. Guibas. 2020. Robust learning through cross-task consistency. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11194–11203. IEEE.

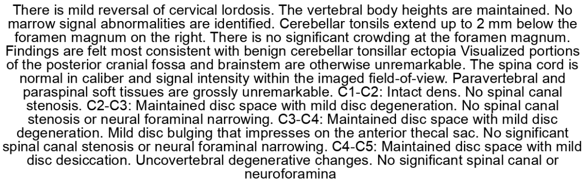
A Example of our Dataset



```
'Findings: Osseous structures of the lumbar spine are intact. \nNo fractu
res detected. \nThe conus medullaris is unremarkable. \nNo concerning par
aspinal mass is identified. \nThere is a levoscoliotic curvature to the l
umbar spine. \nT12-Ll through the L2-L3 levels: Unremarkable L3-L4: There
is a small disc bulge. \nThere is mild disc space narrowing. \nNo stenosi
s. L4-L5\n: Left paracentral disc herniation causing posterior displaceme
nt of the left LS nerve root. \nThere is mild left lateral recess stenosi
s. \nSpinal canal and neuroforamen are patent L5-S 1: \nSmall disc bulge.
\nNo stenosis. \nImpression: 1. \nLeft paracentral disc herniation at L4-
L5 level causing posterior displacement of the left LS nerve root and mil
d left lateral recess stenosis. \n2. \nSmall disc bulges at the L3-L4 and
L5-S] levels without stenosis. \n3. \nLevoscoliotic curvature to the lumb
ar spine. \n4. \nNo fractures identified. '
```

Figure 4: An example of a report from our Lumbar Dataset.

In this section, we will show some examples of lumbar and cervical reports from our dataset.



There is mild reversal of cervical lordosis. The vertebral body heights are maintained. No marrow signal abnormalities are identified. Cerebellar tonsils extend up to 2 mm below the foramen magnum on the right. There is no significant crowding at the foramen magnum. Findings are felt most consistent with benign cerebellar tonsillar ectopia Visualized portions of the posterior cranial fossa and brainstem are otherwise unremarkable. The spina cord is normal in caliber and signal intensity within the imaged field-of-view. Paravertebral and paraspinal soft tissues are grossly unremarkable. C1-C2: Intact dens. No spinal canal stenosis. C2-C3: Maintained disc space with mild disc degeneration. No spinal canal stenosis or neural foraminal narrowing. C3-C4: Maintained disc space with mild disc degeneration. Mild disc bulging that impresses on the anterior thecal sac. No significant spinal canal stenosis or neural foraminal narrowing. C4-C5: Maintained disc space with mild disc desiccation. Uncovertebral degenerative changes. No significant spinal canal or neuroforamina

Figure 5: An example of a report from our Cervical Dataset.

B More Details about our Workflow

In this section, we give a more detailed description of our novel workflow. Our main goal is to detect pathologies at the *motion segment* level from radiologists' MRI reports. The motion segments in the cervical reports that we are interested in are C2-C3, C3-C4, C4-C5, C5-C6, C6-C7 and C7-T1 and the motion segments of interest in the lumbar reports are L1-L2, L2-L3, L3-L4, L4-L5 and L5-S1. We first make sure that the reports are de-identified and then use a Spacy (Honnibal et al., 2020) parser to break the report into sentences. Then each sentence is tagged by annotators and they are given

Hyperparameter Type	Single Tasking Models on Cervical Dataset	Multi-Tasking Models on Cervical Dataset	Single Tasking Models on Lumbar Dataset	Multi-Tasking Models on Cervical Dataset	NER
Epochs	5	12	6	11	5
Batch Size	16	16	16	16	16
Sequence Length	512	512	512	512	256
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Learning Rate	2e-5	3e-5	2e-5	3e-5	1e-5
Weight Decay	1e-4	1e-4	1e-4	1e-4	1e-3
Gradient Clip	2	5	2	5	2
Early Stopping	Yes	Yes	Yes	Yes	Yes
Learning Rate Scheduler	Linear	Linear	Linear	Linear	Linear

Table 5: Hyperparameters used for all our experiments

labels of various pathologies and their severities if the sentence mentions that pathology. To detect pathologies at a motion segment level, we use our BERT based NER system to tag the locations present in each sentence. Our BERT based NER model is a binary classifier model (Location Tag vs the Other Tag). It is is trained on both lumbar and cervical MRI reports that can predict the location tags in those reports. Our NER model achieves an F1 score of .9.

We then use an appropriate body part specific rule based system to group all sentences to the correct motion segment. If a sentence does not explicitly have a motion segment mentioned in it, we use a rule based method to assign the sentence to one of the above mentioned motion segments or to a generic category ”No motion segments found". Given the disparate source of our data and due to typos and OCR errors, for example, L23, L2L3, L@L3, $L2\_L3$ all may refer to the motion segment L2-L3 and thus our systems are mindful of this diversity of the clinical notes. Finally to use our BERT based models for pathology detection on the level of motion segments for a given report, we concatenate all sentences for a given motion segment and use the [CLS] token for the segment that is used for the downstream classification task.

Since we are interested in predictions at the motion segment level, we do not use the sentences that are grouped under ”No motion segments found" to train the classifier models, nor do we evaluate our classifier models on those sentences.

C Hyperparameters and Other Training Details

We create a validation set using 20% of the samples of the training set where the samples are drawn via stratified samples so the data distribution is maintained across splits. The hyperparameters used for training the NER model and various classification models can be found in Table 5.

PyTorch (Paszke et al., 2019) and the HuggingFace library (Wolf et al., 2020) is used to conduct our experiments which are run on 1 NVIDIA V100 16GB GPU.

D Additional CKA Results and Gradient Alignment Results

In this section, we present some additional results on comparing representations between our various models.

We present the average proportion of aligned gradients (APAG) at the end of training in Table 6. We also compute the cosine similarity between the gradients. We then take the average of them for a given layer, thus yielding a scalar value per layer. This yields cosine similarity values which are over 90 % positive. For simplicity, we average those numbers to produce a scalar value that measures the cosine similarity between the gradients of two models. Model level statistics can be found in Table 6.

Dataset	Task Comparisons	Cosine Similarity	Average Proportion of Aligned Gradients
Cervical	Cord-Stenosis	.013	.89
	Cord-Disc	.005	.87
	Cord-Foraminal	.011	.91
	Disc-Stenosis	.012	.88
	Disc-Foraminal	.007	.87
	Foraminal-Stenosis	.005	.84
Lumbar	Disc-Stenosis	.008	.75
	Nerve-Stenosis	.01	.86
	Disc-Nerve	.002	.86

Table 6: Results showing the Cosine Similarity and the Average Proportion of Aligned Gradients between various task specific models after the end of training.

Given some similarities between certain label

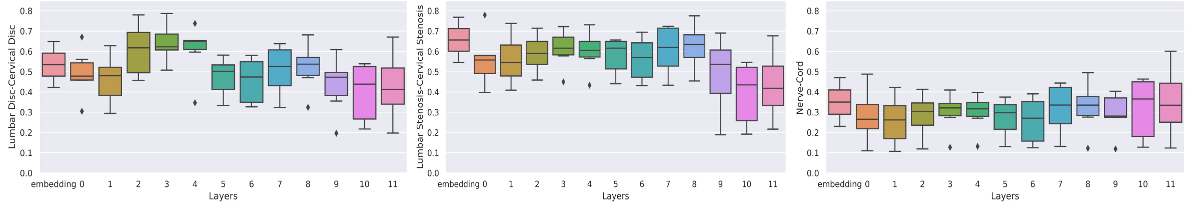


Figure 6: Box plot showing CKA scores between models trained on tasks on the lumbar and the cervical dataset. The y-axis is chosen to be (0,.85). The figure shows the low CKA scores between the cord and the nerve models and high scores between the stenosis models and the disc models.

spaces in the lumbar and the cervical dataset (particularly for the disc herniation and the central canal stenosis labels), we believe that some task specific models between tasks across datasets may show similar representations. To validate this hypothesis, we computed the CKA between lumbar stenosis and the cervical stenosis models and the lumbar disc and the cervical disc models. The natural question is : what happens to the single tasking models that are trained on label spaces that are semantically different? Fig 6 shows low CKA scores between the cord and the nerve models. This is an active work in progress to be able to group similar tasks (Standley et al., 2020) to create a MTL framework that works for both the cervical and the lumbar spine. Another future direction is to use realign gradients using the techniques in (Yu et al., 2020). However to realign the gradients, one has to save the entire computation graph after the backward pass via *loss.backward(retain_graph=True)* which becomes a bottleneck for large transformer models. To mitigate this issue, one can use parameter efficient methods like adapters which we have shown to work in these MTL settings in our previous work (Sehanobish et al., 2022).

E Annotation Process

All data are annotated by our team of inhouse annotators with clinical expertise. All annotators are trained for the given task and provided clear guidelines on the task and performance is measured periodically on a benchmark set and feedback is provided.