paring no anchors and default anchors, the performance of XNLI and parsing drops only around 1 point while NER even improve 1 points averaging over three languages. Overall, these results show that we have previously overestimated the contribution of anchor points during multilingual pretraining. Concurrently, Karthikeyan et al. (2020) similarly find anchor points play minor role in learning cross-lingual representation.

4.3 Parameter sharing

Given that anchor points are not required for transfer, a natural next question is the extent to which we need to tie the parameters of the transformer layers. Sharing the parameters of the top layer is necessary to provide shared inputs to the task-specific layer. However, as seen in Figure 1, we can progressively separate the *bottom* layers 1:3 and 1:6 of the Transformers and/or the embedding layers (including positional embeddings) (**Sep Emb**; **Sep L1-3**; **Sep L1-6**; **Sep Emb + L1-3**; **Sep Emb + L1-6**). Since the prediction layer is tied with the embeddings layer, separating the embeddings layer also introduces a language-specific softmax prediction layer for the cloze task. Additionally, we only sample random words within one language during the MLM pretraining. During fine-tuning on the English training set, we freeze the language-specific layers and only fine-tune the shared layers.

The third group in Tab. 1 shows cross-lingual transfer performance under different parameter sharing conditions with "Sep" denote which layers **is not** shared across languages. Sep Emb (effectively no anchor point) drops more than No anchors with 3 points on XNLI and around 1 point on NER and parsing, suggesting have a cross-language softmax layer also helps to learn cross-lingual representations. Performance degrades as fewer layers are shared for all pairs, and again the less closely related language pairs lose the most. Most notably, the cross-lingual transfer performance drops to random when separating embeddings and bottom 6 layers of the transformer. However, reasonably strong levels of transfer are still possible without tying the bottom three layers. These trends suggest that parameter sharing is the key ingredient that enables the learning of an effective cross-lingual representation space, and having language-specific capacity does not help learn a language-specific encoder for cross-lingual representation. Our hypothesis is that the representations that the models learn for different languages are similarly shaped and models can reduce their capacity budget by aligning representations for text that has similar meaning across languages.

4.4 Language Similarity

Finally, in contrast to many of the experiments above, language similarity seems to be quite important for effective transfer. Looking at Tab. 1 column by column in each task, we observe performance drops as language pairs become more distantly related. Using extra anchor points helps to close the gap. However, the more complex tasks seem to have larger performance gaps and having language-specific capacity does not seem to be the solution. Future work could consider scaling the model with more data and cross-lingual signal to close the performance gap.

4.5 Conclusion

Summarised by Figure 3, parameter sharing is the most important factor. More anchor points help but anchor points and shared softmax projection parameters are not necessary for effective cross-lingual transfer. Joint BPE and domain similarity contribute a little in learning cross-lingual representation.

5 Similarity of BERT Models

To better understand the robust transfer effects of the last section, we show that independently trained monolingual BERT models learn representations that are similar across languages, much like the widely observed similarities in word embedding spaces. In this section, we show that independent monolingual BERT models produce highly similar representations when evaluated at the word level (§5.1.1), contextual word-level (§5.1.2), and sentence level (§5.1.3) . We also plot the cross-lingual similarity of neural network activation with center kernel alignment (§5.2) at each layer. We consider five languages: English, French, German, Russian, and Chinese.

5.1 Aligning Monolingual BERTs

To measure similarity, we learn an orthogonal mapping using the Procrustes (Smith et al., 2017) approach:

$$W^\star = \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F = UV^T$$