# Contrastive Learning for Cross-lingual Pretraining

**Mitchell Donley**
mitchdonley@gatech.edu

**Seth Baer**
sbaer8@gatech.edu

## 1 Abstract

Cross lingual language models have greatly improved in overall language understanding, but a common practice in word to word translation and representation learning in vision is to constrain the latent space to ensure similarity across common features. We introduce a contrastive loss objective that can be utilized to improve sentence embeddings for cross lingual downstream tasks which can include zero-shot tasks as well. This leads to better overall understanding of the similarities/differences between languages and can be accuractly capture by the language model. By only introducing an auxiliary loss, this method is model agnostic within the cross lingual domain. Our results surpass other experimental models we did not expect, as well as reached near baseline. Even though this is not close to SOTA, the results did show that this method could be effective in promoting a shared latent space for all languages. Code is at: https://github.com/MitchDonley/nlp-project.

## 2 Introduction and Related Work

Cross lingual languages models attempt to understand semantics and syntactic meaning of each language and how these languages are similar/different within a single model. These language models can be trained in a supervised or unsupervised fashion, and are achieving state of the art performance (SOTA) on many language inference tasks such as GLUE or XNLI (Lample and Conneau, 2019; Devlin et al., 2018; Conneau et al., 2019).

In vision, contrastive learning is a method that utilizes a common image with different augmentations and attempts to align the latent representation of these augmented images. This is done through a variety of methods but most recently was done in a work called SimCLR (Chen et al., 2020).

We have integrated contrastive learning into a cross-lingual language model and is extendable to other cross-lingual language models as well (i.e. mBERT, XLM-R). We propose that a single sentence represented in two languages can be considered two different augmentations of the sentence's true semantic meaning. By aligning these two translations in the embedding space, we will explore if aligning multi-lingual sentence embeddings will improve the understanding of cross-lingual language models. We express our contrastive objective as a weighted loss that can be added to other loss functions, therefore making our method model agnostic. We will evaluate the language model's understanding using XNLI as a downstream task. In summary our goal is to:

- Integrate contrastive learning into cross lingual language models, which can be fine tuned on downstream tasks
- Understand the effects of different sentence embeddings on contrastive learning

These goals are slightly different from our previous goals, as we have removed the unsupervised machine translation through back translation and have this for future work.

### 2.1 Related Work

#### 2.1.1 Cross-lingual Language Models

Language models made a big performance jump with the introduction of transformers (Vaswani et al., 2017). Transformers utilize the idea of self-attention to capture dependencies that are close neighbors or "far away" which address the vanishing gradients that occurred in RNNs. BERT is a transformer-based language model that performs pretraining on the Cloze task (Devlin et al., 2018; Skory and Eskenazi, 2010). This lead to SOTA performance on many language inference tasks and

is still a common model for many downstream language tasks today. From here many language models have adapted this concept of training on the Cloze task including cross lingual language models such as Multi-lingual BERT and XLM. XLM is a cross lingual model that utilizes monolingual and parallel data with the Cloze task to improve sentence embeddings across languages. XLM performed SOTA on several multi-lingual tasks such as XNLI and machine translation (Lample and Conneau, 2019).

#### 2.1.2 Constrained Cross-lingual Embedding Space

When dealing with cross lingual latent spaces, (Conneau et al., 2017b) has shown that different languages' word embeddings can be trained independently and then a mapping, $W$, can found that allows for accurately moving between languages.

(Alaux et al., 2018) creates the common cross-lingual embedding space by denoting a "pivot" language where a mapping $Q_i$ is learned to map language $i$ to the pivot language. Figure **??** shows this idea. Both of these approaches are unsupervised as mono-lingual embeddings are learned independently of each other then aligned afterwards. (Artetxe and Schwenk, 2018) generates sentences embeddings by training a biLSTM encoder and decoder on one to many machine translation tasks. From here a max pool layer is applied to the output of the encoder to generate sentence embeddings.

#### 2.1.3 Contrastive Learning

Contrastive learning is a self supervised approach that has become popular to generate embeddings for images without any supervision. (Hadsell et al., 2006; Chopra et al., 2005) is a more generalized form of contrastive learning as it attempts to map samples of the same category to a similar, easily measure space (ex: L2 distance). This approach uses a loss function similar to the one above where positive examples have a small loss and negative examples have a large loss.

In a recent work (cite SimCLR), representation learning relies heavily on aligning augmented image data in an embedding space. This is done by performing a series of independent augmentations to the same image, passing these augmented images through a classifier, and utilizing the last layer of the model as the embeddings. From here, a contrastive loss called NT-xent creates positive examples of augmented data that comes from the same image, and all others are considered negative examples. This then maximizes the similarity between the positive examples and minimizes the score between negative examples. SimCLR is notably different, as it performs contrastive learning on the same image with different augmentations compared to different images in the same category.

Our work is unique in that it combines the goals of cross lingual language models and contrastive learning. To our knowledge, it is the first language model that enforces aligned multi-lingual sentence embeddings. We hypothesize that by constraining the sentence embeddings to a shared space, the representations learned will perform better than those that are unconstrained and near the level of the supervised counterparts (Chen et al., 2020; Alaux et al., 2018; Conneau et al., 2017b). We apply this hypothesis to cross-lingual language models, as they are the SOTA models for natural language inference (Conneau et al., 2018).

## 3 Methods

### 3.1 Data

We will be utilizing data that is written in 15 different languages. Specifically, we will be using the languages that are included in the XNLI tasks: Arabic (**ar**), Bulgarian (**bg**), German (**de**), English (**en**), Greek (**el**), Spanish (**es**), French (**fr**), Hindi (**hi**), Russian (**ru**), Swahili (**sw**), Thai (**th**), Turkish (**tr**), Urdu (**ur**), Vietnamese (**vi**), and Chinese (**zh**). We will use data that is mono-lingual (data that is only in one language) and parallel data (data that has been translated by humans from one language to another). Lastly, we will use the XNLI dataset for our downstream evaluation task.

#### 3.1.1 Mono-lingual Data

We are using the Wikipedia dumps that contain all Wikipedia articles in a variety of languages. In total, the Wikipedia datasets are 37GB in size with a total of 143,591,249 sentences. Each language ranges from having up to 50 million sentences (en) or as little as 282,000 sentences (sw). An example of pre-processed sentence are shown below:

**Ru:**
престол своего шестнадцатилетнего сына давида , а сам отстранился от государственных дел ( 1089 ) .

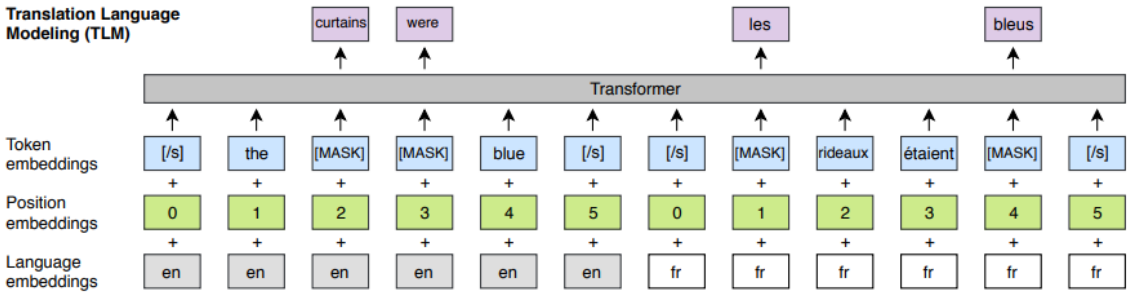


Figure 1: XLM training, setup to utilize parallel data. This image was taken from (Lample and Conneau, 2019)

#### 3.1.2 Parallel Data

The parallel datasets come from a collection of datasets called OPUS (open parallel corpus). Specifically, our datasets are a mix of MultiUN (translated documents from the United Nations) (Eisele and Chen, 2010), Open Subtitles (translated movie subtitles) (Lison and Tiedemann, 2016), TED Talks translated subtitles from 2013, and the EU Bookshop Corpus (Skadiņš et al., 2014). We only use the data that has been translated to or from English. Ideally we would incorporate translations that do not include English, but that is for future work. An example of the parallel data:

**En-Sw:**
he is truly the most forgiving , the most merciful . — hakika mwenyezi mungu ni mwenye kujua mwenye hikima .

Other statistics about the entire datasets can be seen below:

- MultiUN: 7 languages, 81.41 Million sentence fragments
- Open Subtitles 2016: 65 languages, 2.6 Billion sentence fragments
- EU Bookshop: 48 languages, 173.20 Million sentence fragments
- TED 2013: 15 languages, 3.81 Million sentence fragments

| Language | Premise / Hypothesis | Genre | Label |
|---|---|---|---|
| English | You don't have to stay there.<br>You can leave. | Face-To-Face | Entailment |
| French | La figure 4 montre la courbe d'offre des services de partage de travaux.<br>Les services de partage de travaux ont une offre variable. | Government | Entailment |
| Spanish | Y se estremeció con el recuerdo.<br>El pensamiento sobre el acontecimiento hizo su estremecimiento. | Fiction | Entailment |
| German | Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod.<br>Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an. | Travel | Neutral |

Figure 2: XNLI examples.

#### 3.1.3 Downstream Task Data

We have one downstream task: cross-lingual text entailment (XNLI) (Conneau et al., 2018). XNLI has train data that is English only while test and dev data are 7500 have been translated into the 15 languages. This is a benchmark for cross lingual sentence embeddings in which the model's task is to learn the relationship between sentence pairs. The labels for text entailment are "entailment", "neutral", and "contradiction". Examples can be seen in Figure 2.

### 3.2 Model/Analysis

#### 3.2.1 XLM Model

In XLM pretraining (Lample and Conneau, 2019), the authors show how a cross lingual language model can be trained by modifying the Cloze task (Skory and Eskenazi, 2010) used to train the BERT language model (Devlin et al., 2018). Lample and Conneau then show how training can be modified to include parallel data by concatenating two parallel input sentences and marking the tokens with a language index. Using only monolingual data during training is referred to as masked language modeling (MLM), and using parallel data is referred to as translation language modeling (TLM). As with BERT, a transformer is the core neural network architecture used. Figure 1 shows the TLM training setup.

For this project, we will be adding a contrastive learning loss to the TLM loss of a XLM model. Due to computation constraints, we apply our contrastive training to pretrained models.

#### 3.2.2 Contrastive Learning

Contrastive learning is a recent area of study in self-supervision to learn visual features. In our work, we draw inspiration from SimCLR (Chen et al., 2020), which is a simple, generalizable method for contrastive learning that we believe translates well

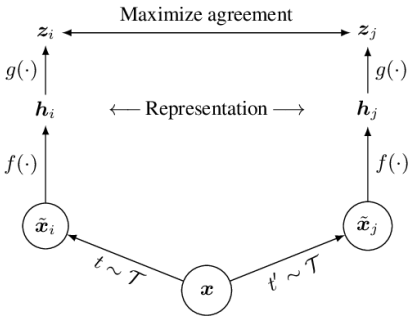


Figure 3: The SimCLR training pipeline. $t \sim \mathrm{T}$ represents a random data augmentation, $f(\cdot)$ represents a neural net encoder, and $g(\cdot)$ represents a projection head. The final outputs of the pipeline, $(z_i, z_j)$, is put into a contrastive loss function, and the encoder $f(\cdot)$ is used for downstream tasks after pretraining is finished. This image was taken from (Chen et al., 2020)

to natural language processing.

The main idea of SimCLR is to pass a datapoint, $x$, through two independent data augmentations followed by a common neural net and projection head in parallel. This creates a positive pair, $(z_i, z_j)$, which is passed into the contrastive loss. After pretraining, the neural net, $f(\cdot)$, is taken to be the encoder used in downstream tasks. The SimCLR pretraining setup is depicted in Figure 3.

We follow the SimCLR training setup, with some adjustments to account for our XLM model, then add the resulting contrastive loss to TLM loss. Rather than explicit data augmentation, we consider the two input sentences as representations of a common semantic meaning, augmented into two different languages. Our XLM model corresponds to the $f(\cdot)$ network, and we use a 2-layer neural net as the projection head, $g(\cdot)$.

One other difference in our method is that we have to split the XLM transformer output to create embeddings of the two sentences. We consider two types of sentence embeddings:

1. Taking the first element of the last transformer layer's output (Start Token Embedding)

2. Taking the max over sentence length of the last transformer layer's output (Max Embedding)

See Figure 4 for a visual representation of our training setup.

For the contrastive loss function, we follow SimCLR's contrastive predictive loss. It is defined as follows:

$$l_{i,j} = -log \frac{exp(sim(z_i, z_j)/\mathrm{T})}{\Sigma_{k=1}^{N} exp(sim(z_i, z_k)/\mathrm{T})}, k \neq i \tag{1}$$

Here, $sim(u, v)$ denotes cosine similarity between two vectors, $(z_i, z_j)$ represents the two outputs of the projection head, $N$ is the number of sentences in a training batch, and T is a temperature parameter. According to this loss function, the higher the cosine similarity between two parallel sentences in a batch, the lower the loss for that pair. Contrastive loss is averaged over parallel sentences before being added to the TLM loss.

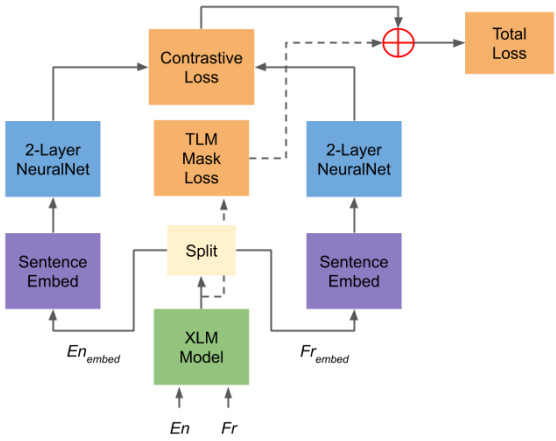


Figure 4: Our contrastive training pipeline.

### 3.2.3 Sentence Embeddings

Although BERT and XLM utilize the first token of the last transformer layer as the sentence embedding, we compare this with max over all the tokens in the sentence after the last transformer layer. We do this specifically for our model because the contrastive objective needs two sentence embeddings. Thus, we adapt this idea used in BERT and XLM to take the "start" token for each sentence after the last transformer layer. For max embedding, we take all the tokens after the last transformer layer and use the max value for each element along the embedding dimension. This results in a sentence embedding the same size as the "start" token sentence embedding. Comparing these will allow us to determine which is a better representation of the sentence for our contrastive objective.

## 3.3 Baseline Models

Our baseline models consist of three different pretrained/fine-tuned XLM models. Specifically

we compare our results with the pretrained XLM model that achieved SOTA on the XNLI dataset along with two XLM models with additional training. The specifics of the intermediate training can be seen below:

- XLM with intermediate training only using the TLM task, which uses the parallel data mentioned in Section 3.1.2
- XLM with intermediate training only with both the MLM and TLM tasks, which uses the monolingual data and parallel data mentioned in 3.1

## 4 Results

### 4.1 Experimental Setup

Our main experiment consists of two parts: extending the pretraining of XLM (dubbed *intermediate training*), and training/testing on the downstream task: XNLI (text entailment). All of our models have 12 transformer layers, 8 attention heads per layer, and an embedding size of 1024 between all layers. We found performance to be extremely sensitive to hyperparameter changes (specifically optimizer type, learning rate, and batch size), so each of our models have their own set of hyperparameters. Results are reported in Table 2.

#### 4.1.1 Intermediate Training XLM

The data splits for the intermediate training were done such that the validation set and the test set had 5,000 sentences and/or sentence pairs while the rest of the data was used for training. This was at the recommendation of Lample and Conneau. We perform intermediate training for 10 epochs each with two different architecture choices:

- Vary the intermediate training task: exclusively TLM or MLM + TLM
- Vary the objective: No Contrastive Loss, Contrastive Loss with Start Token Sentence Embeddings (SE), or Contrastive Loss with Max Pool Sentence Embeddings (ME)

#### 4.1.2 Downstream Training

The data for the downstream task was XNLI and we used their recommended splits which has 112.5K annotated pairs of which 5,000 are used for testing, 2,500 are used for validation, and the rest are used for training. For the downstream training we used the combinations mentioned above (resulting in 6 different models) as well as the baseline model with no intermediate training, totaling 7 models. From here we optimized the following hyperparameters using grid search.

- learning rate: [5e-6, 2.5e-5, 4e-5, 1.25e-4]
- optimizer: [SGD, Adam]
- batch size: [4, 8]

| Int. Training | TLM | MLM + TLM |
|---|---|---|
| Standard | 16.0 | 6.3 |
| Contrastive (SE) | 260 | 650 |
| Contrastive (ME) | 1675 | 695 |

Table 1: Perplexity of the language models after intermediate training for 10 epochs

### 4.2 Intermediate Training Results

In Table 1, we can see the different perplexities for each of the different combinations of architecture choices. We see that the contrastive loss models have a much worse perplexity. Here we had two hypotheses: the standard perplexity would outperform either language model with contrastive loss and that the Contrastive Loss (ME) would out perform the Contrastive Loss (SE). Note: lower perplexity is better.

We chose to analyze the first hypothesis because if a language model has lower perplexity, it generally understands language well and thus we wanted to see how each of the language models would fair. Our first hypothesis holds true as the standard intermediate training outperforms either contrastive loss by a wide margin. We believe this is due to the language models having to optimize for an auxilary task, causing the language to be more difficult to understand. This could be exacerbated because our auxiliary task enforces different languages to be in a similar embedding space. We believe that although the perplexities for the contrastive language models are far worse, the models could have a better understanding of the similarity and differences across languages, and thus this could be beneficial for the downstream task. This will be discussed in Section 4.3.

Our second hypothesis was rejected as we see the Contrastive Loss (SE) has a much better perplexity compared to the Contrastive Loss (ME) model. This one is interesting, as sentence embeddings are an active area of research. BERT utilizes the Start Token approach while Conneau et al. utilizes a max pooling approach. We hypothesized that the latter

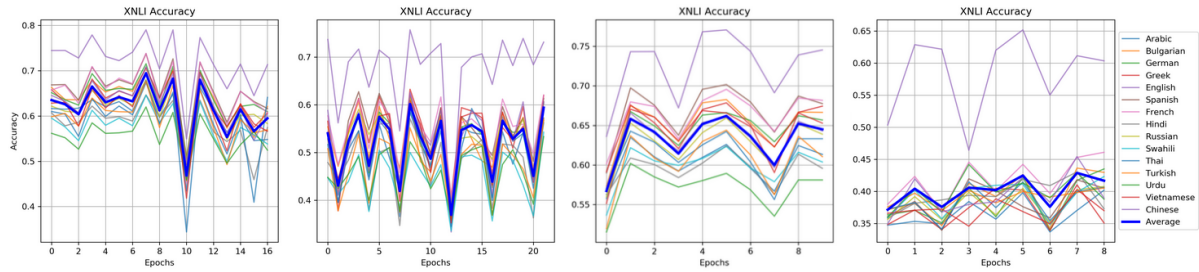


Figure 5: XNLI accuracy per epoch. From left to right: (1) Baseline, (2) Standard Intermediate Training TLM+MLM, (3) Contrastive Loss (SE) TLM+MLM, (4) Contrastive Loss (ME) TLM+MLM

would perform better, as this specific research is in the space of universal sentence embeddings compared to BERT which is trained on monolingual data. Unfortunately this was not the case. We believe this is due to the fact that our model is much more similar to BERT than to Conneau et al.'s model. Running this same experiment on the standard intermediate training would help clear up if the Start Token Embeddings are generally better.

While this is useful to analyze, we believe XNLI as a downstream is a better metric to compare how cross lingual language models understand the similarities and differences across languages.

| Int. Training | TLM | MLM + TLM |
|---|---|---|
| None | 71% | 71% |
| Standard | 53.5% | 59.5% |
| Contrastive (SE) | 60.8% | 66.25% |
| Contrastive (ME) | 40.3% | 42.4% |

Table 2: Top XNLI accuracy (avg across all languages) comparison using models with varying types of intermediate training. SE refers to start token embedding, and ME refers to max embedding.

### 4.3 XNLI

Table 2 shows the results on the test set according to the best validation scores for all models. Below we will analyze each model with differing intermediate training separately and then compare the results across all models. Figure 5 shows us the results of the baseline as well as all model trained on with TLM+MLM as the intermediate training. The accuracy curves are similar for all respective TLM intermediate training, but with scaled down results. This is likely due to the fact that MLM allows a language to independently learn about itself and thus maintains understanding of its semantic and syntactic structure. With this in mind, we will only focus on those models that used TLM+MLM for intermediate training in this section.

#### 4.3.1 Standard Intermediate Training

In Figure 5, model (2), we see results for the standard intermediate training. This model was unable to reach that of the baseline. This is an interesting finding as it was trained on the same data as the baseline with the exact same objective. This leads us to believe that the model started overfitting. This looks to be true because XNLI is trained on English data exclusively and we see the ratio of English to all other languages is much greater in model (2) compared to the baseline. In conclusion, this shows that adding extra training for XLM is detrimental and can affect the results discussed next.

#### 4.3.2 Contrastive Loss Start Token Embedding

In Figure 5, model (3), Contrastive Loss (SE) performs well given the conclusion above and the short amount of training that included the contrastive objective. This model reached near baseline performance which leads us to believe that with more training this could be a viable approach to improving the understanding between languages. Figure 5 model (3) supports this claim by having the smallest ratio between English and other languages compared to the other models with intermediate training. One downside that can be observed is that the variance between languages is generally larger, so although English is not far more accurate than other languages, we see that high resource languages have far greater accuracy than low resource languages. This is not seen in any of the other models. We hypothesize that this model is more sensitive to the amount of parallel data for each language.

#### 4.3.3 Contrastive Loss Max Pooling Embedding

Figure 5, model (4), shows that the utilization of Max Pooling Sentence Embeddings (ME) does not

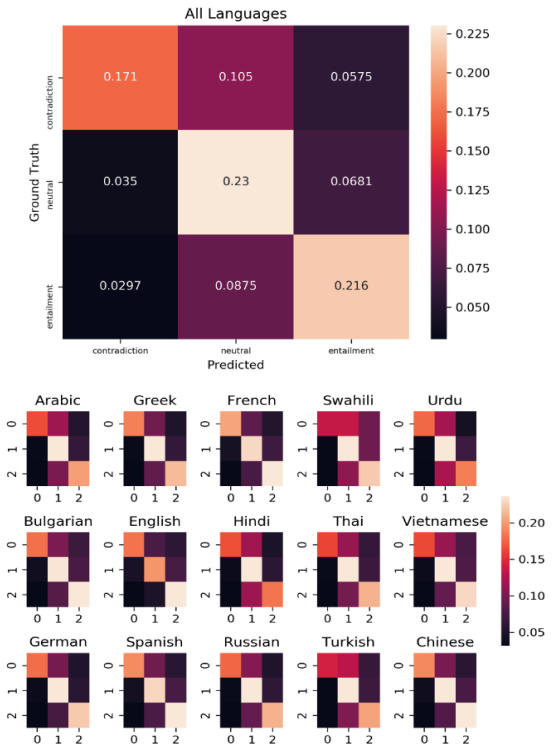


Figure 6: Confusion matrices from a contrastive XLM model. On the right we have separate confusion matrices for each XNLI language, and on the left we have all languages combined.

work well with either the contrastive loss or the XLM architecture. Exploring this is left for future work, but we hypothesize that the First Token Sentence Embeddings perform well with transformer architectures while pooling work well with RNNs. In general, we see that the model was unable to understand outside of the language it was trained on for the downstream task as English is far more accurate than any other languages (Note: XNLI training data is exclusively English). On top of this, it is the only model we examined that is unable to connect similarities and differences between languages as the non-English accuracies are all below 50%.

#### 4.3.4 Overall Comparisons

With Figure 5, we see that the baseline is superior to all adaptions we attempted. We also noticed that the XLM model was extremely sensitive to hyperparameter tuning when performing XNLI as the downstream task, many of the hyperparameters caused the model perform at random chance. Overall, we believe that more training on the baseline XLM model will lead to overfitting, but our results with Contrastive Loss First Token Sentence Embeddings have enough promise that we would be interested in investigating this method of contrastive learning further. Below we will attempt to take a closer look at what our models learned and what common errors were made.

### 4.4 Error Analysis

#### 4.4.1 Confusion Matrices

TLM + contrastive intermediate training outperformed the other TLM intermediate training models; however, by looking at the confusion matrix provided in Figure 6 we see an interesting pattern of errors. The rows of the confusion matrices represent the ground truth labels in our test set, and the columns represent what was predicted by the model. In each cell, we see the total percentage of predictions that fall in the category. Ideally we would like to see the diagonal of the matrix each have 33% of the predictions. In this model, we see a trend in the right direction, but the middle columns are more densely populated, which means that the model tends to over-predict neutral. We hypothesize that since the contrastive loss is encouraging some sentences to be close in embedding space, the line between ”contradiction” and ”entailment” could be more heavily blurred, creating more borderline cases. At the same time, the clear cases could become more pronounced, giving overall better performance.

#### 4.4.2 Attention Score Comparison

In Figure 7, attention scores from the last transformer layer are shown. On the left, the model used is the baseline (no intermediate training), and on the right is a model trained with our contrastive loss. First, generally we don’t see a degradation of the quality of attention learned in our models. We tend to see verbs attending to their modifiers, and nouns attending to their adjectives. In this particular example, we see that the attention learned with contrastive loss can potentially be more useful than in the baseline. The token for ”frustrated” is attending to the token for ”upset” rather than the immediate words surrounding it. This could be a result of the contrastive loss encouraging similar to be embedded close together.

### 4.5 Work Divison

Mitch ran the experiments for pertaining the intermediate training. Seth developed all the visualizations. We shared in analyzing the results and

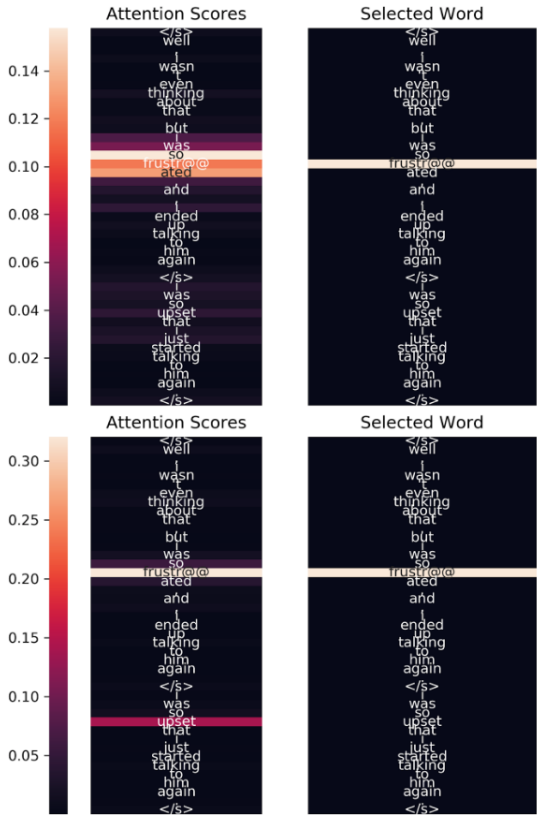


Figure 7: A comparison of attention scores for a baseline (top) and contrastive model (bottom).

running the downstream task across a varied of hyperparameters.

## 5 Conclusion

With this being said, we have found that contrastive learning as an additional objective to intermediate training does not meet the baseline, it does out perform XLM with standard intermediate training. We have also found the XLM model on its own is capable of achieving near SOTA results with the caveat of being very sensitive to hyperparameters for XNLI and potentially other cross lingual downstream tasks. We have also seen that XLM's language embeddding space may not be shared due to drastic change in results with little additional training including the contrastive objective. We believe these results could be improved by implementing some of the steps in the future work below. Lastly, when comparing the First Token Sentnece Embeddings and the Max Pooling Sentence Embeddings, we see the First Token Embeddings are far superior for the transformer architecture while Max Pool Embedding could work better with RNNs which is seen in our results as well as the related work. Analyzing past research in different domains, we believe contrastive learning could be a next step to improve Cross-lingual Language models without explicitly requiring more parallel data for improved results.

### 5.1 Model Improvements/Future Work

There are many aspects of this work we would have liked to explore if time and resources had permitted, but we leave that to future work. A few possible directions include:

1. *More extensive hyperparameter tuning, particularly lambda and batch size:* In regards to batch size, we believe that a sharp increase could have significantly boosted performance of our model as SimCLR states their optimal performance was at batch size: 2048.
2. *Exploring other data augmentations:* We chose not to try augmentations such as randomly rearranging or dropping words because the semantic meaning of the sentences would be distorted. We theorize that data augmentations such as replacing some tokens with synonyms or translating random words could work better for this model.
3. *Extending the model to a fully unsupervised setting:* With the use of iterative backtranslation (Lample et al., 2018), our language model could run without the use of any truly parallel data.
4. *Train the model from scratch:* Due to low compute resources (relative to the original XLM authors) we were not able to train our model from scratch with the contrastive loss.
5. *Evaluate on XTREME (Hu et al., 2020):* Last month, the cross-lingual language model benchmark, XTREME, was released. This benchmark uses 40 languages and 9 tasks, making it a highly compatible metric for our model. We learned of the paper too late to incorporate it into our work.

## 6 Github

https://github.com/MitchDonley/nlp-project

## References


Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2018. Unsupervised hyperalignment for multilingual word embeddings.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, page 539–546, USA. IEEE Computer Society.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017a. Supervised learning of universal sentence representations from natural language inference data.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017b. Word translation without parallel data.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based neural unsupervised machine translation.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *LREC*.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

Adam Skory and Maxine Eskenazi. 2010. Predicting cloze task quality for vocabulary training. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 49–56, Los Angeles, California. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.