# MEVTR: A Multilingual Model Enhanced With Visual Text Representations

**Xiaohua Wang**[*,†,‡], **Wenlong Fei**[*,†,‡], **Min Hu**[†,‡], **Qingyu Zhang**[†,‡], **Aoqiang Zhu**[†,‡]
[†]School of Computer Science and Information Engineering, HeFei University of Technology
[‡]Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine
xh_wang@hfut.edu.cn, feiwenlong@mail.hfut.edu.cn

**Abstract**

The goal of multilingual modelling is to generate multilingual text representations for various downstream tasks in different languages. However, some state-of-the-art pre-trained multilingual models perform poorly on many low-resource languages due to the lack of representation space and model capacity. To alleviate this issue, we propose a Multilingual model Enhanced with Visual Text Representations (MEVTR), which complements textual representations and extends the multilingual representation space with visual text representations. First, the visual encoder focuses on the glyphs and structure of the text to obtain visual text representations, and the textual encoder obtains textual representations. Then, multilingual representations are enhanced by aligning and fusing visual text representations and textual representations. Moreover, we propose similarity constraint, a self-supervised task to prompt the visual encoder to focus on more additional information. Prefix alignment and multi-head bilinear module are designed to acquire an improved integration effect of visual text representations and textual representations. Experimental results indicate that MEVTR benefits from visual text representations and achieves significant performance gains in downstream tasks. In particular, in the zero-shot cross-lingual transfer task, MEVTR achieves results that outperform the state-of-the-art adapter-based framework without the target language adapter.

**Keywords:** multilingual representation, visual text representation，multilingual language model

## 1. Introduction

Pre-trained multilingual language models (MLLMs), such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mDeBERTa (He et al., 2023), aim to generate multilingual text representations that can be used for downstream tasks in different languages. However, even the state-of-the-art MLLM still performs poorly on cross-lingual transfer tasks for many low-resource languages. The reason behind this is the current lack of capacity in the model to represent most languages equally in vocabulary and representation space (Bapna and Firat, 2019; Artetxe et al., 2020). This problem limits the development of multilingual models towards more language coverage and better multilingual text representations.

In this paper, we propose a Multilingual model Enhanced with Visual Text Representations (MEVTR), a novel architecture for multilingual representation learning. As shown in Figure 1, we render the text into a pixel image and pay attention to the glyphs and structure of the text to obtain the visual text representations. The visual text representations are used to supplement the textual representation and expand the textual representation space, allowing the model to cover a wider range of languages and obtain more effective multilingual representations. In effect, we transform the NLP problem into a bimodal problem of language and vision.

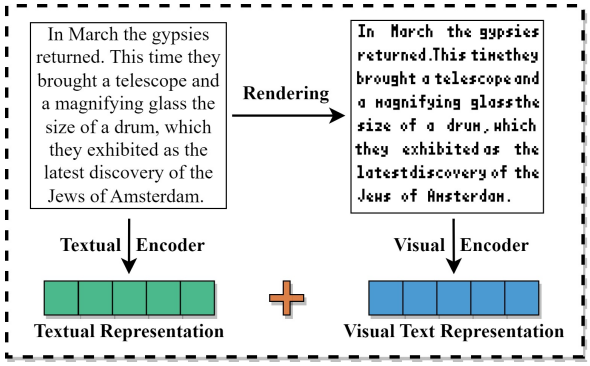


Figure 1: A brief illustration of rendering text into images, obtaining visual text representations, and supplementing multilingual text representations.

In most language models, each text is converted into a sequence of tokens by vocabulary lookup. This approach ignores the glyphs and the structure of the words. However, different languages may have some similarities in their writing systems. For example, the Chinese word “天气” and the Japanese word “天気” both refer to the weather and have similar writing systems, as do the Spanish word “Ángel” and the English word “Angel”. We believe that in multilingual representation learning, this similarity of writing systems can be seen as a potential data augmentation, and words in

[*] Both authors contributed equally to this work.



*LREC-COLING 2024*, pages 11247–11261

low-resource languages will benefit from additional training data from high-resource languages. This similarity can be easily captured from a visual perspective, whereas subword-based approaches require considerable effort in vocabulary construction to ensure that similar semantic words in different languages receive similar word embeddings.

Some previous works have focused on the importance of visual features. To exploit visual information in text, some works use visual features to construct word embeddings (Broscheit, 2018), or combine them with word embeddings as joint embeddings (Meng et al., 2019). Other works treat the textual tasks as computer vision tasks, building language models based on the images of text (Mansimov et al., 2020; Rust et al., 2023). Unlike previous works, we use word embeddings and the images of text as parallel inputs to obtain textual and visual representations, and use the visual text representation as a supplement to the multilingual text representation. In principle, the architecture we propose is suitable for most pre-training MLLMs.

As far as we know, we are the first to use the bimodal method for multilingual text representation learning, so we make a preliminary exploration on the fusion method of *textual bimodal representation learning*. We propose similarity constraint, a self-supervised task to prompt visual perspectives to pay attention to more semantic information, and design prefix alignment and multi-head bilinear module to achieve representation alignment and fusion, respectively.

We validate the performance of MEVTR in multiple languages and on multiple tasks, and MEVTR significantly outperforms the baselines. In particular, for cross-lingual tasks, MEVTR significantly outperforms state-of-the-art adapter-based frameworks (Pfeiffer et al., 2021; Ansell et al., 2021) without the target language adapter[1].

**Our contributions are threefold:**

• We introduce MEVTR, a novel language vision architecture for learning multilingual text representations. It extends the text representation space with visual text representations and improves the effective performance of multilingual models.

• We propose similarity constraint for richer semantic representations; prefix alignment and multi-head bilinear module are used for representation alignment and fusion, respectively.

• We conduct a series of experiments to demonstrate the performance improvements that MEVTR brings to multilingual models on a variety of downstream tasks, including named entity recognition, part-of-speech tagging, and structured sentiment analysis.

[1]We briefly introduce the adapter approach in section 2. The target language adapter refers to an adapter that has been trained in the target language.

## 2. Related work

**Multilingual representation learning** involves learning and understanding the shared and specific semantic and syntactic structures of multilingualism to perform downstream tasks in different languages. To effectively learn features for cross-lingual transfer learning, some MLLMs rely on monolingual corpora in different languages without additional cross-lingual supervision, such as XLM-R (Conneau et al., 2020) and mDeBERTa (He et al., 2023). Other models learn multilingual targets during pre-training using parallel data in multiple languages, such as XLM (Conneau and Lample, 2019) and UniCoder (Huang et al., 2019). Moreover, there are also adapter-based methods (Pfeiffer et al., 2021; Ansell et al., 2021). Adapters allow adding new parameters to the pre-trained model as additional layers. By training on different languages and tasks, task adapters and language adapters can be obtained. When working on different tasks and languages, language-specific representations can be obtained by stacking task and language adapters. However, these approaches ignore the visual importance of text and fail to take advantage of potential visual representations in multilingual text.

**Visual Language Modelling** unlike subword-based vocabularies, focuses on the visual text representation. Meng et al.(2019) enriched the representation of Chinese characters by combining their glyph features with the corresponding character embedding. Mansimov et al.(2020) proposed an end-to-end neural network model that effectively translates images with text from one language to another while preserving the same semantics. Rust et al.(2023) proposed a pixel-based pre-trained language model, where input text is transformed into pixel images, to achieve competitive results against the vocabulary-based language model BERT in various tasks. Instead, we obtain visual text representations and multilingual representations separately, and then supplement the multilingual representations with visual text representations.

**Vision-Language Models** aim to perceive, understand and integrate visual and textual information in our complex multimodal world, and then create cross-modal representations to overcome difficult cross-modal challenges. Most vision-language models, such as ViLT (Kim et al., 2021), METER (Dou et al., 2022), and BridgeTower (Xu et al., 2023) apply the TWO-TOWER architecture, consisting of a visual encoder, a textual encoder and a cross-modal encoder. The visual and textual

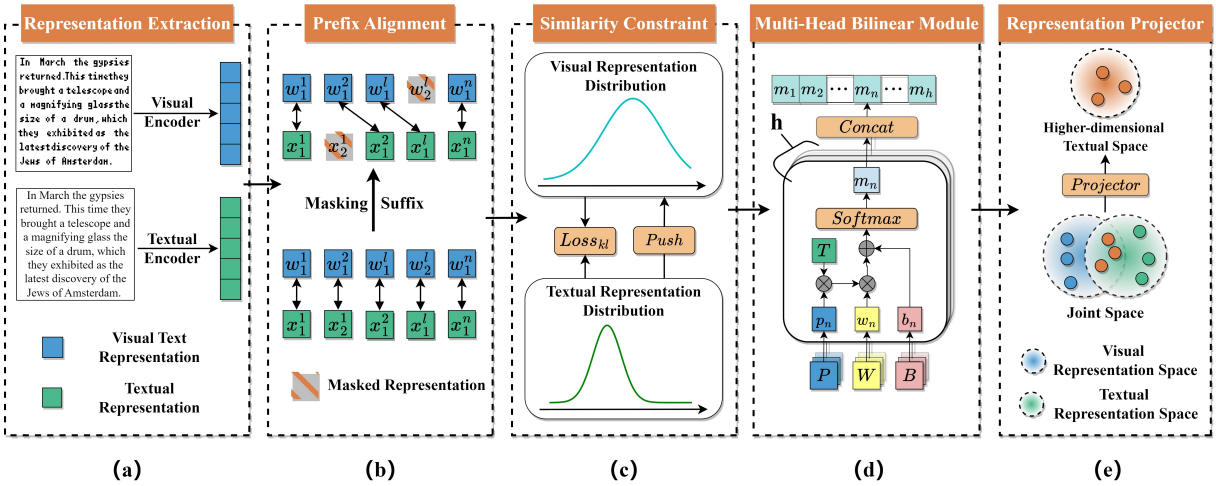


Figure 2: **An overview of our proposed MEVTR.** MEVTR consists of five components as shown above: (a) visual encoder and text encoder extract visual text representations and textual representations respectively; (b) prefix alignment achieves alignment of two modal representations; (c) similarity constraint is used to constrain the two encoders to focus on different semantic representation; (d) multi-head bilinear module is used for modal fusion; (e) representation projector is used to map the fused representations onto the high-dimensional textual representation space.

encoders capture visual and textual semantics respectively. The cross-modal encoder fuses the semantics of the two modalities. We take advantage of the TWO-TOWER architecture and use a visual encoder to obtain visual text representations, with the aim of enriching the semantic representations of the textual encoder.

## 3. MEVTR: The Proposed Method

As shown in Figure 2, our model consists of a visual encoder, a textual encoder, and subsequent components for aligning and fusing the visual and textual representations. Our objective is to acquire a visual text representation by means of a visual encoder that concentrates on the glyphs and structure of the text. By aligning and fusing the visual and textual representations, a more efficient textual representation is attained.

### 3.1. Textual Encoder

Since our goal is to improve the semantic representations of MLLMs through visual text representations, we adopt the pre-trained language model XLM-R (Conneau et al., 2020) as our textual encoder. It is based on the Transformer architecture (Vaswani et al., 2017) and uses a large amount of multilingual text for pre-training. It uses the Byte-Pair Encoding (BPE) (Sennrich et al., 2016; Radford et al., 2019) for tokenization, which splits some words into multiple subwords. A tokenized word can be represented as: $C_\ell = [c_1; c_2; \cdots; c_\alpha]$, where $\alpha$ denotes the number of subwords.

Assuming that the length of the text sequence is $n$ and the tokenized sequence is $i$, the tokenized sequence can be expressed as: $S = [C_\ell^1; C_\ell^2; \cdots; C_\ell^{n-1}; C_\ell^n]$, where $S \in \mathbb{R}^i$,the superscript of $C$ indicates the position of the text sequence. $[< s >]$ token and $[< /s >]$ token are added to the sequence as the start and end tokens, respectively. The output textual representations can be represented as:

$$T_{output} = [x_{[<s>]}; X_\ell^1; X_\ell^2; \cdots; X_\ell^{n-1}; X_\ell^n; x_{[</s>]}], \quad (1)$$

where $T_{output} \in \mathbb{R}^{(i+2)\times d_t}$, $d_t$ is the dimension of the textual encoder, $X_\ell = [x_1; x_2; \cdots; x_\alpha]$ and $x$ is the last layer representation of the textual encoder.

### 3.2. Visual Encoder

The purpose of the visual encoder is to extract the glyphs and structural features of the text. While typical visual encoders mainly tackle images with intricate semantics, we focus on the matrix generated by text rendering, which generally conveys less semantic information than regular images. Therefore, we employ the 12-layer encoder of PIXEL (Rust et al., 2023) as our visual encoder. It is a pre-trained language model based on pixel modelling and adopts the pre-training architecture of ViT-MAE (He et al., 2022).

Visual encoder employs the patch embedding strategy identical to Vision Transformer (Dosovitskiy et al., 2021). The pixel image rendered from the text is sliced into a sequence of patch embeddings with the same pixel resolution. Since

the size of each patch cannot be flexibly adjusted based on the length of the word, a single word may necessitate several patch embeddings to represent it, and can be denoted as :$P_\ell = [p_1; p_2; \cdots; p_\beta]$, where $\beta$ denotes the number of patch embeddings. The length $j$ of the patch embedding sequence will be longer than the length $n$ of the original text. The visual encoder also adds a special $[cls]$ token to the patches sequence. The output visual representations can be represented as:

$$V_{output} = \left[w_{[cls]}; W_\ell^1; W_\ell^2; \cdots; W_\ell^{n-1}; W_\ell^n\right], \quad (2)$$

where $V_{output} \in \mathbb{R}^{(j+1)\times d_v}$, $d_v$ is the dimension of the visual encoder, $W_\ell = [w_1; w_2; \cdots; w_\beta]$ and $w$ is the last layer representation of the visual encoder.

### 3.3. Prefix Alignment

Both the BPE employed in the textual encoder and the patch embedding approach implemented in the visual encoder may make the length of the representation sequence larger than the original text, resulting in patch tokens and word tokens being unaligned.

Therefore, we employ prefix alignment, a simple method of alignment. When a word is split into several tokens or patches, we choose the token or patch at the first position to represent the word, as can be seen in Figure 2 (b). As identical suffixes are frequently used by several different words, it is widely presumed that these suffixes provide little representational information, while the main information is concentrated at the first position (Lin et al., 2022). After the prefix alignment, the textual and visual representations are as follows:

$$T_{rep} = \left[x_{cls}; x_1^1; x_1^2; \cdots; x_1^{n-1}; x_1^n\right], \quad (3)$$

$$V_{rep} = \left[w_{cls}; w_1^1; w_1^2; \cdots; ; w_1^{n-1}; w_1^n\right], \quad (4)$$

where $T_{rep} \in \mathbb{R}^{(n+1)\times d_t}$, $V_{rep} \in \mathbb{R}^{(n+1)\times d_v}$.

### 3.4. Similarity Constraint

We propose similarity constraint to learn better unimodal representations before fusion. Specifically, we constrain parallel text-image pairs to have low similarity between textual representations $x_{cls} = [t_1; \cdots; t_i; \cdots; t_{d_t}]$ and visual representations $w_{cls} = [v_1; \cdots; v_i; \cdots; v_{d_v}]$, thus constraint two encoders to focus on as different information as possible. To avoid affecting the representation of the textual encoder, we utilize the Kullback-Leibler (KL) divergence with asymmetry as our similarity score, where a high similarity score indicates a significant difference in representation. The similarity function is defined as follows:

$$g_t(t_i) = a\frac{t_i - \mu(x_{cls})}{\sigma(x_{cls})} + b, \quad (5)$$

$$g_v(v_i) = a\frac{v_i - \mu(w_{cls})}{\sigma(w_{cls})} + b, \quad (6)$$

$$S = \mathsf{KL}\left(g_t(t_i) \parallel g_v(v_i)\right), \quad (7)$$

where $S$ denotes the similarity score, $g_t$ and $g_v$ are linear transformations that map $x_{cls}$ and $w_{cls}$ to normalized representations, $\mu$ and $\sigma$ represent the calculation of the average and variability respectively, and $a$ and $b$ are trainable parameters. The loss function for the similarity constraint is defined as follows:

$$L_{sc} = \frac{\omega}{S}, \quad (8)$$

where $\omega$ is the scaling weight of the similarity score, and we set $\omega$ to 10,000. Throughout the training process, we expect to see an increase in similarity scores as well as an increase in the divergence between the semantic representations of the two modalities.

### 3.5. Multi-head Bilinear Module

To integrate textual and visual representations, we construct the multi-head bilinear module to compute attention scores in the representation space. We are inspired by the multi-head attention mechanism (Vaswani et al., 2017) that uses multi-head structures to focus on information in different representation subspaces. Our multi-head bilinear module differs from the multi-head attention mechanism. The multi-head attention mechanism focuses on the associations between different tokens. On the other hand, the multi-head bilinear module focuses on the relations within the representation space, and the size of the attention score matrix is related to the dimension of the representation space. Furthermore, since two modal representations of the same text contain less semantics than the Vision-Language representation, the requirements for modal fusion are relatively uncomplicated. The bilinear model (Lin et al., 2015) is a simple feature fusion method and can well consider the correlation of two modal representations, so we propose the multi-head bilinear module based on the bilinear model.

We introduce the bilinear model and the multi-head bilinear module, respectively, and illustrate the computation of the attention score matrix for visual representations, as shown in Figure 2 (d). We utilize textual representations $T = [t_1; \cdots; t_i; \cdots; t_{d_t}]$ and visual representations $P = [v_1; \cdots; v_i; \cdots; v_{d_v}]$ as inputs. The bilinear model can be expressed as:

$$M = P^T T W + b = \sum_{j=1}^{d_v}\sum_{k=1}^{d_t} p_j t_k w_{j,k} + b, \quad (9)$$

| Task | Datasets | Languages |
|---|---|---|
| Named Entity Recognition (NER) | MasakhaNER (Adelani et al., 2021)<br>CoNLL 2003 (Sang and Meulder, 2003) | Hausa (hau), Igbo (ibo), Luo (luo), Swahili, Luganda (lug), Wolof (wol), Yorùbá (yor), Nigerian-Pidgin (pcm), Kinyarwanda (kin) |
| Part-of-Speech Tagging (POS) | Universal Dependencies 2.10 (Zeman et al., 2022) | Arabic (ar), Bambara (bm), Cantonese, Livvi, Erzya, Uyghur (ug), Faroese, Komi Zyrian (kpv), Upper Sorbian, Buryat |
| Structured Sentiment Analysis (SSA) | MPQA (Wiebe et al., 2005)<br>$\text{MultiB}_{\text{CA}}$ and $\text{MultiB}_{\text{EU}}$ (Barnes et al., 2018)<br>$\text{NoReC}_{\text{Fine}}$ (Øvrelid et al., 2020) | English, Catalan, Basque, Norwegian |

Table 1: Details of the tasks, datasets and languages involved in our experiments. Abbreviations for some languages are given in brackets. Further details of all the language and datasets used are provided in Appendix B.

where $W \in \mathbb{R}^{d_t \times d_v}$ is a trainable parameter, $b \in \mathbb{R}^{d_v \times d_v}$ is a bias. To focus on different subspaces within the visual representation space, we first divide the visual representation space into $h$ sections. We then compute the bilinear model separately for each representation subspace, normalise it via the *softmax* function, and connect them to obtain the attention score matrix. Finally, the new visual representation is obtained by weighting the $d_v$ dimensional semantic representation of the original visual representation:

$$P = \text{concat}\left(P_1, \cdots, P_n, \cdots, P_h\right), \quad (10)$$

$$m_n = \text{softmax}\left(P_n^T T W_n + b_n\right), \quad (11)$$

$$P_{new} = P \cdot \text{concat}\left(m_1, \cdots, m_n, \cdots, m_h\right), \quad (12)$$

where $m_n, b_n \in \mathbb{R}^{(d_v/h) \times d_v}$ and $W_n \in \mathbb{R}^{d_t \times d_v}$.

### 3.6. Representation Projector

To achieve a more extensive semantic representation space, we combine the textual and visual representation spaces to create a joint representation space. Then, we employ the non-linear mapping to map the joint spatial representation to a higher-dimensional textual representation space, as shown in Figure 2 (e). Specifically, we use two consecutive linear layers with the activation function *ReLu* as the projector.

We aim to use visual representations to supplement the information neglected in the textual representation without changing the original semantic information. Therefore, we build a joint spatial representation using weighted visual representations $P_{new}$ and original textual representations $T$ . The calculation of the high-dimensional textual representation space is defined as follows:

$$Z = \text{projector}\left(\text{concat}\left(T, P_{new}\right)\right), \quad (13)$$

where the dimension of $Z$ is determined by the size of the hidden layer in the projector. In section 6, we conduct extensive experiments on the selection of different dimensions in the high-dimensional textual representation space.

## 4. Experiments

### 4.1. Datasets and Settings

To evaluate MEVTR, we use named entity recognition (NER) and part-of-speech (POS) tagging for zero-shot cross-lingual transfer, and structured sentiment analysis (SSA) [2] (Barnes et al., 2021) for multilingual fine-tuning. Table 1 summarises our datasets for the experiments. These datasets contain more than 30 languages, including not only high-resource languages but also a large number of low-resource languages.

For our experimental setup, we use XLM-RoBERTa-base as the textual encoder and PIXEL-base as the visual encoder. The maximum input text sequence length is 512 for the textual encoder and 529 for the visual encoder. The patch size in our visual encoder is *16* × *16*, and the number of heads in the multi-head bilinear module is 6. In the representation projector, we set the textual representation space dimension to 1,536.

In zero-shot cross-lingual transfer tasks, we use English as the source language. We first fine-tune the model on the English dataset and then transfer it directly to the target language for performance evaluation. During fine-tuning, we freeze the first 6 layers of both encoders and use the *AdamW* optimiser (Loshchilov and Hutter, 2019) with a base learning rate of 2e-5 and a weight decay of 0.05. For the NER and POS tagging tasks, we train

[2] SSA aims to predict all opinion tuples in a text. Each opinion $O$ is a tuple $(t, h, e, p)$, where $h$ is a holder expressing a polarity $p$ towards a target $t$ through a sentiment expression $e$, and is an NLP task that combines syntactic and semantic complexity.

| | NER (F1 score) | | | | | POS (Accuracy) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | hau | luo | pcm | yor | **avg** | ar | bm | kpv | ug | **avg** |
| XLM-R | 77.0 | 33.8 | 77.1 | 49.6 | 51.1 | 75.3 | 23.6 | 37.5 | 74.9 | 58.7 |
| Concat | 75.9 | 37.3 | 76.3 | 46.7 | 54.2 | 74.9 | 24.3 | 29.7 | 69.8 | 50.0 |
| Cross-modal Attention | **77.2** | 35.1 | 77.4 | 49.5 | 54.1 | 43.4 | 22.8 | 23.9 | 44.0 | 33.8 |
| MAD-X TA | 44.0 | 33.0 | 71.0 | 66.6 | 52.4 | 70.8 | **37.2** | 35.8 | 36.8 | 56.2 |
| LT-SFT TA | 46.5 | 37.7 | 74.4 | **69.3** | 55.3 | 70.6 | 34.2 | 37.1 | 34.0 | 55.0 |
| MEVTR | **77.2** | **44.1** | **77.7** | 50.6 | **57.8** | **78.3** | 31.9 | **38.9** | **75.0** | **59.7** |

Table 2: Results of the NER and POS tagging tasks in cross-lingual transfer, F1 score and accuracy are used as the corresponding evaluation metrics. TA indicates no target language adapter. The results of MAD-X TA and LT-SFT TA are from Ansell et al. (2022). **Bold** denotes best-performing method per language. The table shows the results for a subset of the languages, and avg indicates the average results for all the languages in Table 1. The complete experimental results are provided in Appendix C.

15,000 steps on each dataset with a batch size of 64. We select the best checkpoint for evaluation based on the validation performance of the model on the English dataset.

In multilingual fine-tuning, we fine-tune our model directly on each language dataset. The experimental setup is the same as for zero-shot cross-lingual transfer, except we use a batch size of 48 for SSA.

## 4.2. Zero-shot Cross-lingual Transfer

We compare MEVTR with two state-of-the-art adapter-based frameworks: 1) MAD-X (Pfeiffer et al., 2020) : which uses an additional invertible adapter for cross-language transfer tasks; and 2) LT-SFT (Ansell et al., 2022) : which is based on MAD-X and introduces sparse fine-tuning. In addition, to verify the effectiveness of our proposed modal fusion method, we also set up for comparison the fusion methods of Cross-modal Attention and Concat, which also introduce visual text representations. We present the results for the NER [3] and POS tagging tasks in Table 2.

MEVTR outperforms XLM-R, with average performance gains of 6.7 F1 score in NER and 1 accuracy in POS tagging. These gains are driven by visual text representations, demonstrating the importance of focusing on the visual features of text. MEVTR also has significant advantages over state-of-the-art adapter-based frameworks. It is important to emphasise that we compare the results without applying the target language adapter. This is because the target language adapter is trained on the target language, whereas the target language is completely invisible to MEVTR. We argue that this result is due to the fact that MEVTR focuses on visually written representations of text and benefits from visual similarities between languages when performing zero-shot cross-lingual transfer tasks.

Compared to the Cross-modal Attention and Concat methods, which also use visual text representations, MEVTR has the best performance. In some cases, the more complex Cross-modal Attention method, which is commonly used in multimodal domains, performs much worse than the simple Concat method. This result supports our view to some extent. In multimodality, to obtain a better multimodal representation, it is necessary to let the representations of two modalities interact sufficiently to achieve alignment and fusion. However, visual text representations have less semantic information, and excessive interaction leads to an undesired loss of textual semantics. Therefore, the complex Cross-modal Attention method sometimes produces a worse result than the Concat method. The multi-head bilinear module used in MEVTR is a modal fusion method that performs simple interactions and thus achieves the best performance.

For POS tagging, we also find that the visual text representations brought less gains. The Cross-modal attention method and the Concat method perform much worse than the unimodal method. We attribute this to the poor quality of the extracted visual text representations, which instead of providing an additional representational complement to the textual representation, compromise it. Indeed, the average accuracy of using only visual text representations in POS tagging is 24.3%. It is worth noting that MEVTR can still extract effective semantics from visual text representations with superior results via the multi-head bilinear module. This also demonstrates the superiority of the multi-head bilinear module in representations fusion.

[3] MasakhaNER and CoNLL 2003 datasets use different tags, with $DATE$ and $MISC$ used uniquely by each; therefore we replace them with the $O$ tag during both training and testing.

| | MPQA (English) | MultiB$_{CA}$ (Catalan) | MultiB$_{EU}$ (Basque) | NoReC $_{Fine}$ (Norwegian) | avg |
|---|---|---|---|---|---|
| XLM-R | 27.9 | 61.5 | 58.4 | 42.4 | 47.6 |
| Concat | 24.9 | 61.0 | 56.5 | 33.7 | 44.0 |
| Cross-modal Attention | 24.0 | 57.2 | 51.9 | 36.7 | 42.5 |
| MEVTR | **31.5** | **61.6** | **60.4** | **42.7** | **49.1** |

Table 3: Results of the SSA task in multilingual fine-tuning, sentiment graph F1 score (Barnes et al., 2021)as the evaluation metric.

| | NER F1 score | SSA F1 score |
|---|---|---|
| Visual encoder | 31.1 | 34.7 |
| Textual encoder | 51.1 | 47.6 |
| MEVTR w/o SC & RP | 55.2 | 47.9 |
| MEVTR w/o SC | 55.6 | 48.0 |
| MEVTR | **57.8** | **49.1** |

Table 4: Results of the ablation studies on MEVTR for NER and SSA. w/o denotes 'without'.

### 4.3. Multilingual Fine-tuning

We further evaluate MEVTR in the multilingual fine-tuning task. Table 3 shows the performance of MEVTR in structured sentiment analysis. Using visual text representation, MEVTR significantly outperforms XLM-R on four datasets in different languages, with an average improvement of 1.5 sentiment graph F1 score [4]. This further demonstrates the validity of MEVTR in using visual text representations to complement textual representations and obtain better multilingual semantic representations.

In comparison with Cross-modal Attention and Concat methods, which also use visual text representations, MEVTR shows significant advantages. Furthermore, we find that Cross-modal Attention and Concat methods show a similar pattern to the POS tagging task. They both suffer significant performance degradation due to their inability to capture meaningful semantic information within visual text representations, instead of introducing large amounts of irrelevant or erroneous information. However, MEVTR achieves an improvement with the visual text representation, demonstrating its ability to focus on valuable information.

### 4.4. Ablation Studies

To evaluate the impact of each component, we perform ablation studies of MEVTR on the NER task in zero-shot cross-lingual transfer and the SSA task in multilingual fine-tuning. 1) Visual encoder denotes that only visual text representations are used; 2) Textual encoder denotes that only textual representations are used; 3) SC denotes the similarity constraint; 4) MBM denotes the multi-head bilinear module; 5) RP denotes the representation projector.

As shown in Table 4, MEVTR significantly outperforms MEVTR w/o* (* indicates components) in both tasks, demonstrating the importance of each component in MEVTR. Specifically, the performance of the visual text representation obtained by the visual encoder is significantly lower than the others, which is actually expected. The semantics of text obtained from a visual perspective is often not superior. Using only the multi-head bilinear module (MEVTR w/o SC and RP) to fuse the two representations gives better results than either one alone. This demonstrates the feasibility of using visual text representations to complement textual representations and the effectiveness of the multi-head bilinear module for modal fusion.

Compared to the representation projector, the similarity constraint has a more significant improvement over MEVTR. Because the representation projector is a simple non-linear mapping for remapping visual textual joint representations into a higher dimensional textual representation space. Nevertheless, the similarity constraint can force the two encoders to focus on different textual information, resulting in a more complete representation.

## 5. Case Study

For a more in-depth analysis, we further analyze the effect of introducing word glyphs into the model. We first construct two pairs of cases using the four languages, each pair having the same textual semantics, and in addition each pair containing morphologically similar and semantically similar word pairs, as shown in Figure 3.

We then fine-tune MEVTR and XLM-R on the same English dataset, making them both 97% accurate on the POS tagging task. Finally, we use

[4]The sentiment graph F1 score is a metric that attempts to quantify the extent to which the model captures fully structured sentiment.

今天天气很好。——Chinese
今日いい天気だ。——Japanese

The angel flapped its wings. ——English
El ángel batió las alas. ——Spanish

Figure 3: The case we designed contains four sentences in Chinese, Japanese, English, and Spanish. Among them, the Chinese word ” 天气” and the Japanese word ” 天気” are morphologically and semantically similar as a pair of similar words, and the English word ”Angel” and the Spanish word ”Ángel” are also a pair of similar words.

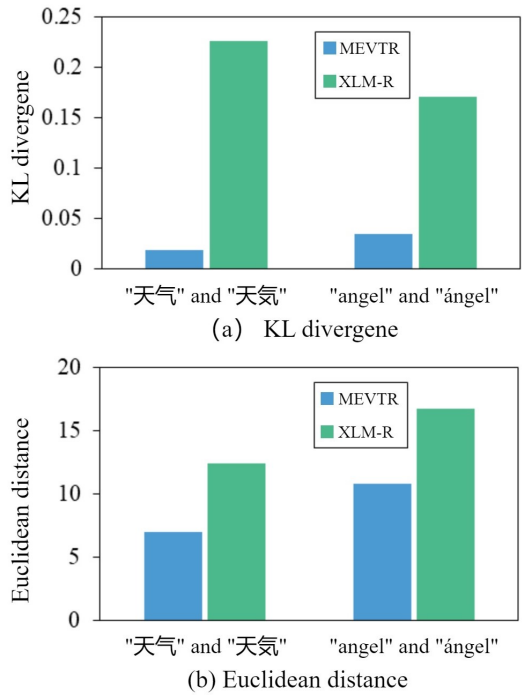


Figure 4: KL scatter and Euclidean distance between representation vectors of similar words in different languages.

the fine-tuned MEVTR and XLM-R to obtain representation vectors for two pairs of similar words in different languages but in the same sentence meaning. We compare the representation vectors obtained by MEVTR and XLM-R when processing similar words in different languages, and the results are shown in Figure 4.

In Figure 4 (a), the KL scatter between the representation vectors obtained by MEVTR is significantly smaller than that of XLM-R, indicating that the distribution of representation vectors of similar words is closer in MEVTR. In Figure 4 (b), the Euclidean distance between the representation vectors obtained by MEVTR is again smaller than that of XLM-R. We argue that MEVTR introduces the glyph information of the text, so that

| | NER (F1 score) | |
|---|---|---|
| | mBERT | mDeBERTa |
| Textual encoder | 57.6 | 67.7 |
| MEVTR | **58.6** | **69.6** |

Table 5: Results of MEVTR in the NER task using mBERT and DeBERTa as text encoders, respectively.Textual encoder indicates that only the original MLLM is used.

even words that do not belong to the same language but have similar glyphs often get closer representation vectors, whereas the XLM-R model completely ignores the glyph features and treats words in different languages as completely different words, regardless of whether they have similar glyphs or not. We believe that the fact that MEVTR can exploit this similarity is an important reason why it outperforms general language models in multilingual representation tasks.

## 6. Further Analysis

**Apply Different Textual Encoders.** In principle, most pre-trained language models can be used as our text encoder. We use mBERT (Devlin et al., 2019) and mDeBERTa [5] (He et al., 2023) as our textual encoder for the NER task, respectively, to further analyse the effect of introducing visual text representations via MEVTR. As shown in Table 5, both pre-trained multilingual models, mBERT and DeBERTa, perform better after the introduction of visual text representation via MEVTR. This further demonstrates the generalisability of our proposed MEVTR and the effectiveness of the approach that uses visual text representations to complement textual representations.

**Explore the Dimension of Textual Representation Space.** We conduct experiments on the NER task in 6 languages using different dimensions of the representation projector to analyse the effect of dimensions on textual representations. As shown in Figure 5, there is a general tendency that too large or too small a dimension of the representation space affects the effectiveness of multilingual text representations, and the dimension of the mapped high-dimensional textual representation space should be between 1,024 and 2,048. This suggests that when we use visual text representations to complement textual representations, we need to expand the original representation space in order to obtain better multilingual text representations.

[5]We use mDeBERTa-v3-base, which is a multilingual version of DeBERTa, using the same structure as DeBERTa and trained with CC100 multilingual data.

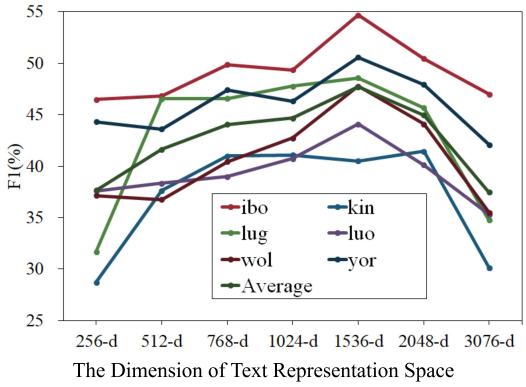


Figure 5: Performance of MEVTR with different dimensions of the textual representation space for NER.

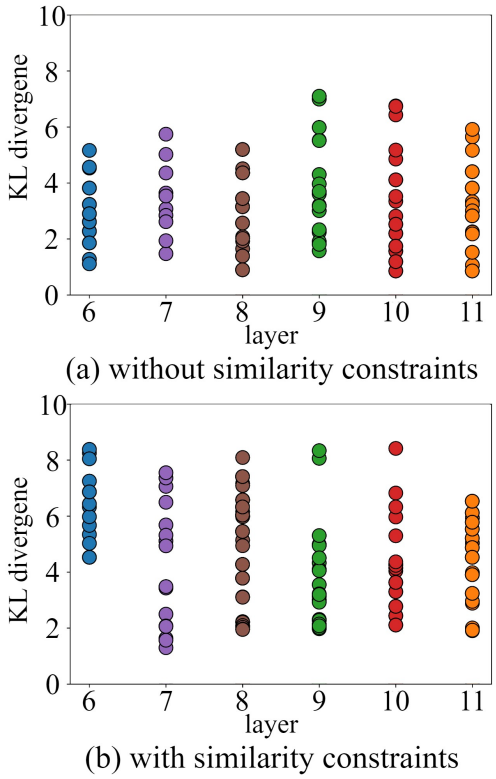


Figure 6: The KL divergene between the attention distributions of the different heads in the last 6 layers of the visual encoder with and without similarity constraint.

**Visualise the effect of similarity constraint on models.** Analyzing attention weights is intuitive because it measures how much attention each token pays to other tokens. We analyse the effect of similarity constraint on the visual encoder by calculating the attention weight distributions for different attention heads in the last 6 layers. (We freeze the first 6 layers of the model, so we only calculate the KL divergence for the last 6 layers). KL divergence can be seen as the diversity of attentional heads. Higher/lower KL divergence means that different attention heads focus on different/similar tokens. As shown in Figure 6, by comparing the attention weight distributions for different heads in each layer, we find that the diversity of attentional heads increases significantly with similarity constraint. We attribute this to the fact that with similarity constraint the visual encoder focuses on more information and has access to richer semantic representations.

## 7. Conclusion and Future Work

We propose MEVTR, a language vision architecture for improving multilingual text representation. MEVTR obtains visual text representations by focusing on the glyph and structure of the text through the visual encoder. The visual text representations are then used to complement the textual representations, resulting in a larger representation space and more effective multilingual representations. We also propose similarity constraint, prefix alignment and multi-head bilinear module for better complements. In addition, MEVTR can use most pre-trained multilingual models as the textual encoder. We experimentally demonstrate the effectiveness of MEVTR, which achieves significant performance on a wide range of tasks.

In future work, we will continue to focus on the impact of visual text representations on multilingual text representations, further exploring and exploiting visual features. Furthermore, we argue that texts express information in a multimodal way, and that not only the visual aspect of the writing system, but also the auditory aspect of pronunciation deserves attention.

## 8. References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D' souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Alan Ansell, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulic. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1778–1796. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavas, Ivan Vulic, and Anna Korhonen. 2021. MAD-G: multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4762–4781. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4623–4637. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1538–1548. Association for Computational Linguistics.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. Multibooked: A corpus of basque and catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3387–3402. Association for Computational Linguistics.

Samuel Broscheit. 2018. Learning distributional token representations from visual features. In *Proceedings of The Third Workshop on Representation Learning for NLP, Rep4NLP@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 187–194. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. 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 7057–7067.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. 2022. An empirical study of training end-to-end vision-and-language transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18145–18155. IEEE.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 2022. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2485–2494. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. 2015. Bilinear CNN models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1449–1457. IEEE Computer Society.

Yangkun Lin, Chen Liang, Jing Xu, Chong Yang, and Yongliang Wang. 2022. ZHIXIAOBAO at semeval-2022 task 10: Apporoaching structured sentiment with graph parsing. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 1343–1348. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Elman Mansimov, Mitchell Stern, Mia Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. 2020. Towards end-to-end in-image neural machine translation. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 70–74, Online. Association for Computational Linguistics.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. 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 2742–2753.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for norwegian. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5025–5033. European Language Resources Association.

Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: an adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7654–7673. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. 2021. Unks everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10186–10203. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer 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. 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.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Lang. Resour. Evaluation*, 39(2-3):165–210.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 10637–10647. AAAI Press.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1–9. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Erik Andersen, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Yifat Ben Moshe, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Daniela Corbetta, Marine Courtin, Mihaela Cristescu, Philemon Daniel, Elizabeth Davidson, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak,

Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Shafi Sourov, Carolyn Spadine, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2022. Universal dependencies 2.10. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

# Appendix

## A. Code

We release the code and models at https://github.com/wenlong1019/MEVTR. Currently we only provide a simple version of the method implementation, we will provide the full version later.

## B. Datasets

| Task | Language | ISO Code | Family | Datasets | UD Treebank |
|---|---|---|---|---|---|
| NER | English | en | Indo-European, Germanic | CoNLL 2003 (Sang and Meulder, 2003) | |
| | Hausa | hau | Afro-Asiatic, Chadic | MasakhaNER (Adelani et al., 2021) | N/A |
| | Igbo | ibo | Niger-Congo, Volta-Niger | | |
| | Kinyarwanda | kin | Niger-Congo, Bantu | | |
| | Luganda | lug | Niger-Congo, Bantu | | |
| | Luo | luo | Nilo-Saharan | | |
| | Nigerian-Pidgin | pcm | English Creole | | |
| | Swahili | swa | Niger-Congo, Bantu | | |
| | Wolof | wol | Niger-Congo, Senegambian | | |
| | Yorùbá | yor | Niger-Congo, Volta-Niger | | |
| POS | Arabic | ar | Afro-Asiatic, Semitic | Universal Dependencies 2.10 (Zeman et al., 2022) | PUD |
| | Bambara | bm | Mande | | CRB |
| | Buryat | bxr | Mongolic | | BDT |
| | Cantonese | yue | Sino-Tibetan | | HK |
| | Erzya | myv | Uralic, Mordvin | | JR |
| | Faroese | fo | Indo-European, Germanic | | FarPaHC |
| | Komi Zyrian | kpv | Uralic, Permic | | Lattice |
| | Livvi | olo | Uralic, Finnic | | KKPP |
| | Upper Sorbian | hsb | Indo-European, Slavic | | UFAL |
| | Uyghur | ug | Turkic, Southeastern | | UDT |
| SSA | English | en | Indo-European, Germanic | MPQA (Wiebe et al., 2005) | N/A |
| | Catalan | ca | Indo-European | $\text{MultiB}_{\text{CA}}$ and $\text{MultiB}_{\text{EU}}$ (Barnes et al., 2018) | |
| | Basque | eu | Basque | | |
| | Norwegian | no | Indo-European | $\text{NoReC}_{\text{Fine}}$ (Øvrelid et al., 2020) | |

Table 6: Details of the languages and datasets we used in our experiments.

## C. Results

| | hau | ibo | kin | lug | luo | pcm | swa | wol | yor | avg |
|---|---|---|---|---|---|---|---|---|---|---|
| XLM-R | 77.0 | 49.8 | 20.1 | 26.8 | 33.8 | 77.1 | 81.7 | 44.4 | 49.6 | 51.1 |
| Concat | 75.9 | 54.1 | 33.7 | 42.4 | 37.3 | 76.3 | 79.1 | 42.3 | 46.7 | 54.2 |
| Cross-modal Attention | 77.2 | 54.0 | 27.3 | 32.7 | 35.1 | 77.4 | 82.3 | 51.8 | 49.5 | 54.1 |
| MAD-X TA | 44.0 | 54.5 | 50.2 | 53.3 | 33.0 | 71.0 | 69.6 | 29.8 | 66.6 | 52.4 |
| LT-SFT TA | 46.5 | 56.8 | 52.9 | 53.8 | 37.7 | 74.4 | 69.5 | 37.1 | 69.3 | 55.3 |
| MEVTR | 77.2 | 54.7 | 40.5 | 48.6 | 44.1 | 77.7 | 79.3 | 47.8 | 50.6 | 57.8 |

Table 7: The complete results of the NER task in cross-lingual transfer and F1 score is used as the evaluation metric.

| | ar | bm | bxr | yue | myv | fo | kpv | olo | hsb | ug | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| XLM-R | 75.3 | 23.6 | 61.0 | 54.6 | 48.9 | 75.3 | 37.5 | 62.7 | 72.7 | 74.9 | 58.7 |
| Concat | 74.9 | 24.3 | 46.1 | 35.8 | 38.0 | 65.6 | 29.7 | 53.6 | 62.1 | 69.8 | 50.0 |
| Cross-modal Attention | 43.4 | 22.8 | 31.6 | 21.7 | 33.5 | 36.6 | 23.9 | 41.1 | 39.1 | 44.0 | 33.8 |
| MAD-X TA | 70.8 | 37.2 | 62.0 | 64.1 | 48.5 | 74.1 | 35.8 | 63.4 | 69.6 | 36.8 | 56.2 |
| LT-SFT TA | 70.6 | 34.2 | 59.5 | 64.5 | 45.7 | 72.9 | 37.1 | 62.2 | 69.2 | 34.0 | 55.0 |
| MEVTR | 78.3 | 31.9 | 61.7 | 50.0 | 49.4 | 75.2 | 38.9 | 63.9 | 72.9 | 75.0 | 59.7 |

Table 8: The complete results of the POS tagging task in cross-lingual transfer and accuracy is used as the evaluation metric.