Multimodal Embeddings

Jan Kels
Heinrich Heine Universität
Düsseldorf, Germany
Jan.Kels@hhu.de

Omar Hassan
Heinrich Heine Universität
Düsseldorf, Germany
Omar.Hassan@hhu.de

1 Introduction

Multimodal learning involves combining different channels of information to understand our environment. Humans possess five basic senses that enable us to perceive and comprehend the world. Similarly, AI researchers aim to train deep learning models that can effectively integrate different modalities. Two key challenges arise in multimodal learning. Firstly, there is a need to represent unstructured data numerically (embeddings or representations), this has been researched extensively in the last decade in two main modalities, text and image. Secondly, how to combine the representations of these different modalities effectively. (Akkus et al., 2023)

2 Language Embeddings

Representing language numerically has developed largely over the last decade. Starting with learning word embeddings which allow words to be encoded as dense vectors, capturing their semantic meaning (Mikolov et al., 2013). Then Encoder-decoder architectures were used to map input sequences to output sequences of varying lengths (Bahdanau et al., 2016). They prove useful in complex tasks like machine translation, as they are capable of handling different word orders and active or passive voice. Transformers (Vaswani et al., 2017) rely solely on attention and do not require sequential processing like traditional RNNs. Transformer architectures such as BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), and GPT-3 (Brown et al., 2020) are pre-trained on large corpora and can be fine-tuned for specific language tasks. With these breakthroughs, deep learning networks have achieved success in representing semantic content in text data numerically.

3 Visual Embeddings

Images representation research started with a long race to solve the task of image classification. CNNs were studied and experimented for so long and resulted in a long list of architectures that could represent images in lower dimensional spaces as ResNet (He et al., 2015). Another approach used is contrastive learning in the latent space, it has shown promise, focusing on reducing the distance between representations of augmented views from the same image (positive pairs) while increasing the distance between representations of augmented views from different images (negative pairs) (van den Oord et al., 2019). Inspired by their success in NLP, researchers have attempted to combine CNN-like architectures with self-attention, sometimes replacing convolutions entirely.

4 Multimodal Embeddings

Models that were introduced for Image2Text tasks used templates based on object detection or attribute prediction (Socher and Li, 2010). Then RNNs and their variants, like LSTMs, were commonly used for sequence generation, with visual information encoded in the output of Convolutional Neural Networks (CNNs) (Yao et al., 2018). Also, Graph convolutional neural networks and attention mechanisms have been proposed to model relationships between image regions and words. Fully-attentive models, based on the Transformer architecture or BERT, have emerged as alternatives to RNN-based models. Other variations include combining transformers with LSTMs or incorporating geometric relations and context-based gating mechanisms.

On the other hand, models that were introduced for Text2Image tasks, have shown great success recently, Dall-E (Ramesh et al., 2021) is one of those models. Essentially, it is training a Discrete Varia-

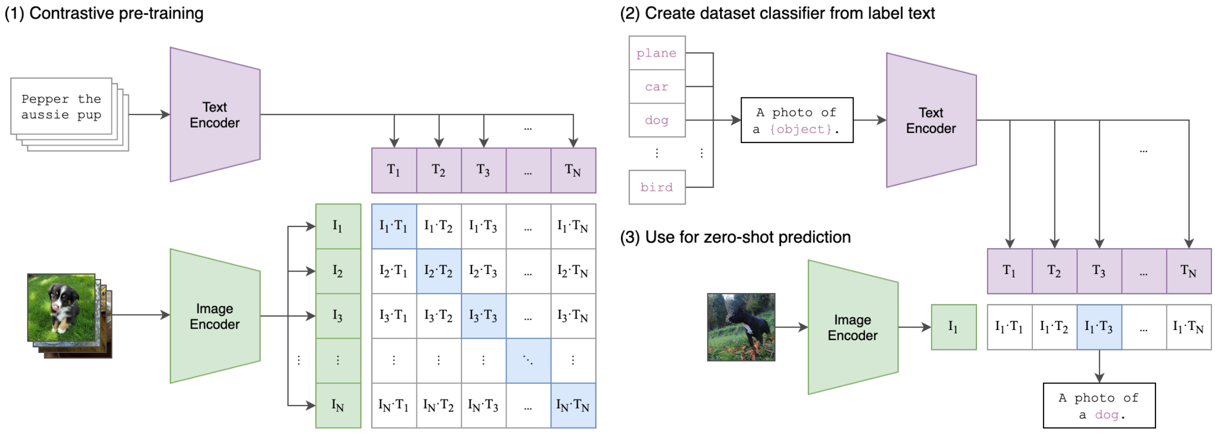

Figure 1: on the left an overview of the pretaining objective, on the right the zero shot classification task

tional Autoencoders (dVAE) to compress 256x256 images into a 32x32 grid of tokens. Then the model learns the prior distribution of text-image pairs. The text is byte-pair (Sennrich et al., 2015a) encoded into a maximum of 256 tokens. And the image representation encoded by previously trained dVAE is unrolled (from 32x32 grid to 1024 tokens) and concatenated to the text tokens. This sequence (of 256+1024 tokens) is used as an input for a huge transformer-like architecture. Its goal is to autoregressively model the next token prediction. During inference time, the text caption is again encoded into 256 tokens at most. The generation process starts with predicting all of the next 1024 image-related tokens. They are later decoded with the dVAE decoder that was trained in the first step. Its output represents the final image.

5 CLIP

The CLIP (Radford et al., 2021) architecutre uses contrastive loss to learn a common embedding space for image, text pairs. It's based on consine similarity and can be used with a variety of text encoder as well as image encoders. If needed, a projection layer ensures common dimensionality. Figure 1 details the main task for CLIP, zero shot image calssifcation. By constructing the classes using various prompts or ensembles, a single pre-trained model can act as a image classifier for a variety of datasets. It outperforms several state of the art models that were fine tuned on the specific task. CLIP models do underperform in tasks such as statellite image classification. The other way around is also an option, having a caption and multiple images to rank which has the higehst similarity. Such an approach was use by several submissions to the Visual Word Sense Diambiguation (Raganato et al., 2023) shared task. Models can be found on Huggingface [1] under the *Zero-Shot Image Classification* task.

5.1 following research

CLIP provides image-caption similarities and has therefore enabled text conditional image generation such as DALL-E 2 (Ramesh et al., 2022), by providing feedback to the model of how well a generation matches the input prompt. Building upon the idea of CLIP are models like BLIP(Li et al., 2022) which focus to sovle visual question answering by using a generative text decoder.

References

Cem Akkus, Luyang Chu, Vladana Djakovic, Steffen Jauch-Walser, Philipp Koch, Giacomo Loss, Christopher Marquardt, Marco Moldovan, Nadja Sauter, Maximilian Schneider, Rickmer Schulte, Karol Urbanczyk, Jann Goschenhofer, Christian Heumann, Rasmus Hvingelby, Daniel Schalk, and Matthias Aßenmacher. 2023. Multimodal deep learning.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish,

[1] https://hf.co/tasks/zero-shot-image-classification

Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

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

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. SemEval-2023 task 1: Visual word sense disambiguation. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2227–2234, Toronto, Canada. Association for Computational Linguistics.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation.

Richard Socher and Fei-Fei Li. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. pages 966–973.

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

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning.