Multimodal Transformers for Detecting Bad Quality Ads on YouTube

Vijaya Teja Rayavarapu
rvteja24@gmail.com
Northeastern University
Boston, MA, USA

Bharath Bhat
bharathbhat@google.com
Google
Mountain View, CA, USA

Myra Nam
myranam@google.com
Google
Mountain View, CA, USA

Vikas Bahirwani
vikasbahirwani@google.com
Google
Mountain View, CA, USA

Shobha Diwakar
shobhad@google.com
Google
Mountain View, CA, USA

ABSTRACT

An ads ecosystem needs robust, scalable mechanisms to safeguard users from bad quality ads. Contemporary ad creatives typically contain different combinations of modalities like text, images and video, and as such, any system that flags bad quality ad content needs a holistic multimodal representation of the ad. In this paper, we demonstrate that modern Transformer based neural network models are effective multimodal learners. We report significant performance gains in YouTube video ads on the task of content quality prediction by transitioning to Transformer based models from simpler feed-forward neural networks. We provide ablation studies to understand the impact of each input modality, and compare various flavors of Transformer architectures. We hope that our experiments help practitioners looking to incorporate these powerful multimodal models into other parts of the ads ecosystem.

CCS CONCEPTS

• **Computer systems organization** → *Neural networks*; • **Information systems** → **Computational advertising**; **Online advertising**.

KEYWORDS

Multimodal Embeddings, Ad Quality, Natural Language Processing, Computer Vision, Attention, Transformers

ACM Reference Format:
Vijaya Teja Rayavarapu, Bharath Bhat, Myra Nam, Vikas Bahirwani, and Shobha Diwakar. 2022. Multimodal Transformers for Detecting Bad Quality Ads on YouTube. In *Proceedings of The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2022 (AdKDD '22).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
AdKDD '22, August 14–18, 2022, Washington, DC
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Contemporary ad landscape is very diverse with different types of ads (text, image, video ads etc.) shown in different placements (e.g. video search page or watch page etc.). Each of these ads has a different combination of modalities like text, image, audio and video, and developing a holistic understanding of the ad content is a prerequisite for ads targeting and policy enforcement systems. In this work, we look at the ad representation learning problem, particularly from the standpoint of their application to content policy enforcement systems.

A responsible ads ecosystem strives to safeguard its end-users and advertisers from bad quality content. Doing so at scale requires the deployment of machine learning models that incorporate the latest advances in Computer Vision and Natural Language Processing to understand ad content and determine quality. Specifically, this is a multimodal learning problem, and calls for techniques that can produce a joint representation that is more informative than the unimodal components.

Deep neural networks, and specifically, Transformer models with Attention modules are the current state of the art for both text [4, 11, 13, 16, 42, 43] and visual modalities [6, 9, 14]. In addition, a number of recent studies have found success applying Attention-based networks to model interactions between modalities [31, 32]. In this work, we conduct extensive experiments on applying these techniques to build ad representations for video ads on YouTube, with particular focus on supervised training with the content quality prediction objective.

The main contributions of this paper are as follows:

- We demonstrate that fusing information from multiple modalities like text and video while building ad representation models yields significant gains on the content quality prediction task.
- We demonstrate that state of the art Transformer models lend themselves well for such multimodal representation learning, and
- We conduct extensive experiments to establish best practices for multimodal fusion using Transformers on a real-world ads dataset collected from YouTube.

In the following, Section 2 describes related work. We introduce the problem and data in Section 3 and present our methods for multimodal fusion using Transformers in Section 4. Our experiments and key results are in Section 5 and Section 6 respectively.

We conclude with some qualitative observations on model behavior in Section 7 and avenues for future work in Section 8.

2 RELATED WORK

2.1 Unimodal Representation Learning

There is rich literature of modern deep neutral networks in learning single modality representations.

Transformers [13, 35, 36], which use an Attention mechanism [43], are currently the de-facto standard for text representation. These models are typically comprised of layers of self-Attention blocks and capture global dependencies between input and output, that is, they are able to relate tokens and positions within the same input sequence.

For supervised image understanding, this includes the family of CNNs [5]: AlexNet [28], VGG [37], ResNet [25], Inception [40], etc. In addition, recent studies show that the unsupervised image representation learning, such as via contrastive learning, presents promising results (e.g., simCLR [8], MoCo [24]). More recently, Transformer models have been applied to various vision applications for their powerful capabilities to learn visual representations [15, 23, 27].

A similar effort has been devoted to learning video representations. There have been supervised learning with video captioning [48], YouTube knowledge graph entities [2], or relational graph clustering with user clicks [29], etc. Recent works utilize temporal information and multi-view points for objects [22, 44] or the similarity between video frames [21, 45] in self-supervised video representation learning. In addition, the Transformer architecture has been popularly used to learn the spatio-temporal information in video classification problems [3, 20, 33, 47].

2.2 Transformer Models for Multimodal Representation Learning

Various models have been proposed to extend the Transformer model for multimodal learning [10, 17–19, 30, 32, 38, 39, 41]. Their goal is to learn representations that can explain relationships between different modalities. Cross modal Self-Attention [46] encodes multi-modal dependencies between linguistic and visual features, by aligning the image regions with the text inputs. The models can be grouped by the fusion strategies; early fusion [10, 38, 39], mid-fusion [30, 32, 41], or late fusion.

These models typically rely on a pre-trained language model such as BERT [13] for text, and pre-trained vision models for visual features, which are fixed throughout. This learning setup with pre-trained input features is to help accelerate convergence without introducing excessive memory requirements in training. A recent work [30] introduced a memory-efficient approach to train end-to-end, which can be effectively applied to different modalities or different distributions from their audio-visual representation.

Transformer models take the pre-training approach on a large-scale (unlabelled) dataset, with subsequent fine-tuning to various downstream tasks. The effectiveness of pre-training for large-scale Transformers has been advocated in both the language and vision domain, proving significant improvement in the downstream tasks [14, 34]. However, research shows that data size does not matter as much when the dataset is noisy [26]. In addition, similarity between the language of pre-training and evaluation datasets is important.

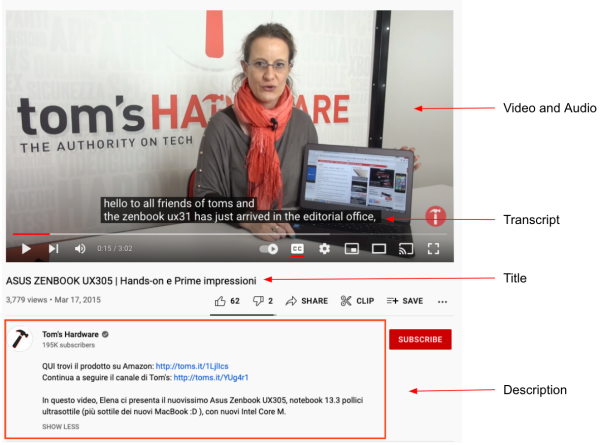


Figure 1: A video ad contains multimodal signals: video frames, audio, title, description.

Our dataset is unique because the texts are provided by the advertisers and they may not always be narrative-oriented or describe the visual modalities as the public datasets [2, 12] do. We take a practical approach to directly train with our content quality prediction task on our YouTube ads dataset.

3 PROBLEM

Given the video and texts associated with an ad, our goal is to learn multi-modal ad representations that are optimized for a specific task which, in our case, is the *quality* of the ad. We describe our definition of quality and dataset collection methodology below.

3.1 Quality Score

A video ad can be bad quality for a variety of reasons - it may be fraudulent, offensive, unethical, or it may include depictions of pornographic imagery, to name a few. For our purposes, we consider a single measure that captures any and all of these characteristics, and ask trained human evaluators to assign a badness quality score ranging from 0 to 100 for ad videos, where 100 means lowest quality. We ask three evaluators to rate each video, and consider the average of the provided scores to be the final *quality score* for a video.

3.2 Data

We use the annotation methodology described above to collect a labeled set of 600k ad videos from YouTube. Each sample in the dataset is a multimodal ad video, containing video frames, audio, titles and description, with its associated ground truth quality score. Most ad videos tend to be on the shorter side, with an average length of 30 seconds. Figure 1 shows an example of a video ad.

Among the modalities, we chose the video frames for the visual modality and the concatenation of the title and description for the textual modality, in order to learn the embedding that discovers the relationships between two modalities. We use a 80-20 split for training and testing respectively.

4 METHODS

4.1 Unimodal Representations

This section discusses how we model our input modalities - text and video. We use unimodal encoders, which are pre-trained on large-scale datasets.

4.1.1 Text Features. The text features associated with ads are typically free-form text shown in the ad. These may be the title and description that go with ad videos, or could be complementary to the video content. To encode the text, we use the pre-trained Bidirectional Encoder Representation From Transformers (BERT) model [13]. BERT models tokenize the text and produce dense float embeddings for *each* token, obtained by propagating the input tokens through a series of self-Attention layers. The outputs of this BERT model are 1024-dimensional float vectors per text token, and we consider up to 512 tokens per ad.

$$f_{text} = M_{BERT}(text) = [f^1_{token}, f^2_{token}, ..., f^{|N_{tokens}|}_{token}] \quad (1)$$

where token features $f^i_{token} \in \mathbb{R}^{1024}$ and $N_{tokens} \in [0, 512)$. We include the special token *CLS* to learn the ad-level textual representation.

4.1.2 Video Features. For ad videos, we limit ourselves to the visual content of video frames, and use a model from the ResNet family of networks [25] to encode each frame. We take the penultimate hidden layer output, averaged over pixels to obtain a 64-dimensional frame representation.

$$f_{video} = [f^1_{frame}, f^2_{frame}, ..., f^{|N_{frames}|}_{frame}] \quad (2)$$

where $f_{frame} = M_{ResNet}(frame) \in \mathbb{R}^{64}$ and $N_{frames} \in [0, 60)$. We also include the special token VID to learn the video-level representation.

We sample one frame per second, and consider up to 60 frames for each video. If a video has fewer than 60 frames, zero-padding is applied to provide the fixed number (=60) of the frames. For models that operate on this sequence, we add a sinusoidal positional encoding to f_{frame}, following similar work in [43].

Note that our unimodal encoders are fixed pre-trained encoders. We choose not to fine-tune them, since our focus is mainly on learning multimodal representations. This also follows from similar approaches adopted for task specific fine-tuning, where it is common to train smaller task specific heads on top of large scale pre-trained embeddings.

4.2 Quality Score Prediction

The overall architecture for quality score prediction is comprised of two sub-networks connected to each other in a cascading setup (Figure 2). The first sub-network $M_{encoder}$ aims to learn an ad representation E_{ad} that effectively encodes cross-modal relationships (Eq. 3). E_{ad} is then fed to the regression model M_{task} to predict the quality score (Eq. 4). We use a multi-layered dense feed forward neural network for M_{task}. We focus most of our experiments on $M_{encoder}$, experimenting with various flavors of self-Attention and co-Attention [32] modules and embedding fusion techniques (Figure 3).

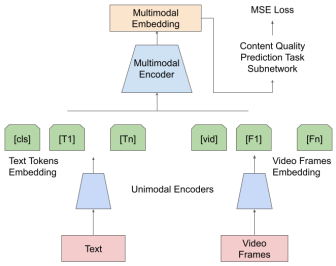


Figure 2: Model Architecture

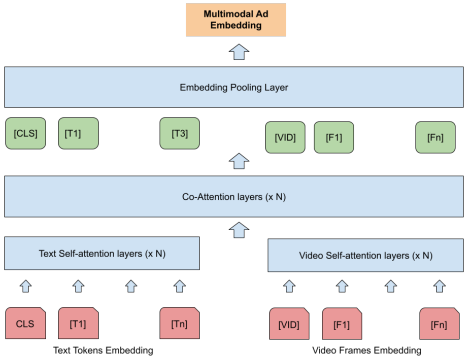


Figure 3: Multimodal Encoder Architectures

$$E_{ad} = M_{encoder}(f_{text}, f_{video}) \quad (3)$$

$$S_{quality} = M_{task}(E_{ad}) \quad (4)$$

4.3 Multimodal Transformer Encoders

Transformer models operate on sequential data, and learn representations through successive layers of Attention blocks. Given our two input sequences f_{text} and f_{video}, several choices exist for multimodal learning (Figure 3).

4.3.1 Early Fusion. The first technique is early-fusion, where we concatenate f_{text} and f_{video} into a single input sequence, and train the model to automatically discover relationships between the two domains [31, 39]. In this case, the text and video self-Attention blocks in Figure 3 get merged into one, and there are no co-Attention blocks. The average pooled embedding over the non-padded tokens is used in the second sub-task to predict the score.

4.3.2 Mid Fusion. The second technique is mid fusion, where we have independent Transformers with self-Attention for each individual modality first and later learn the cross-modal representations using another co-attentional Transformer [32]. We add special tokens $[CLS]$ and $[VID]$ to the text and visual sequences respectively, following [32]. The final outputs corresponding to these special tokens are combined via concatenation or dot-product operation to get the final ad embedding.

4.3.3 Late Fusion. The third technique is late-fusion, where we use the individual self-Attention Transformers that take the sequential embeddings for each of the modalities. In this case, there are no co-Attention blocks, and multimodal fusion happens in the final pooling block. We use the same approach to obtain the joint embedding as the mid fusion technique.

4.4 Supervised Learning

Our models are trained directly on the supervised regression task of quality score prediction. This is in contrast to many state-of-the-art Transformers that use pre-training with self-supervision to predict the next word in caption [7], followed by downstream applications such as Question Answering. The self-supervision (e.g., contrastive cross-modality matching loss) is useful to train on large-scale dataset with no labels. However, it is not appropriate for YouTube ads, where the text modality (title & description) is not narrative-oriented. It may describe the visual domain, but it is often entirely complementary, and includes substantial noise (e.g., url links, special characters, etc.).

For practicality, we directly learn the multimodal embedding to predict the quality score without pre-training or self-supervision. We show in our results that our approach can explain correctly the relationships of the bad quality content found in both modalities.

5 EXPERIMENTS

5.1 Baseline Models

In addition to the Transformer models mentioned above, we consider the following baseline models.

5.1.1 Baseline with Fixed Prediction Scores. This is a baseline whose prediction is the average of the quality score.

5.1.2 Pooled Baseline. In this model, we average pool the sequence of text token and video frame embeddings, concatenate the two and pass them through a feed forward network to obtain our multimodal ad embedding.

5.1.3 Text Only Transformer. This is a Transformer model that takes only the text as input, and calculates the final ad embedding through a block of self-Attention layers.

5.1.4 Video Only Transformer. Same as the text-only Transformer, but operates on video frames instead. These unimodal models are used in ablation studies to understand the impact of each modality.

5.2 Common Parameters

The number of hidden layers and width of each of the model architectures above is configured to ensure that the trainable parameter count is comparable (60M parameters in each). Each model produces a 1024 dimensional ad embedding. M_{task} is a dense feed-forward network with a single layer, followed by sigmoid activation to produce the quality score from the ad embedding. We normalize the quality score labels to be in the range $[0, 1]$. All models are trained end to end with the Mean Square Error (MSE) loss, using the Adam optimizer, with a constant learning rate of $1e-5$ for 20 epochs, and a batch size of 512. For models that use Attention, we use 8 Attention heads, with a hidden layer dimension of 2048. We use a dropout of 0.2 for each Attention layer in the Transformer models. Our primary evaluation metric is MSE on our test dataset.

6 RESULTS

Table 1: Summary of Experiment Results

Model	% Reduction in MSE
Comparing Transformers to Fixed Prediction Baseline	
Mid Fusion (Co-Attention) Over Fixed Prediction Baseline	44.67%
Comparing Transformers to Pooled Baseline	
Early Fusion Over Pooled Baseline	3.25%
Comparing Multimodal Fusion Methods	
Mid Fusion (Co-Attention)	
Over Early Fusion	3.84%
Over Late Fusion	0.52%
Without Self-Attention	3.52%
Ablation of Input Modalities	
Mid Fusion (Co-Attention) Over Text Only Transformer	18.94%
Over Video Only Transformer	1.80%

We look at our key experimental findings below.

6.1 How well do Transformers fare against simpler baselines?

We first validate our Transformer model by comparing with the baseline with the fixed prediction scores as the average of the actual quality scores. We observe a significant improvement by 44.67% (Table 1 - first row).

We then compare the performance of our simplest Transformer model (*Early Fusion*) with the *pooled baseline* model, and see that the multimodal representation learned by the Transformer improves MSE on the task of quality score prediction by 3.25% (Table 1 - second row). This demonstrates that Transformers are indeed effective at condensing multimodal sequential input data into a useful ad representation.

6.2 Comparison of Multimodal Fusion Methods

We experimented with several different choices for multimodal fusion within the context of Transformer models, and found that co-Attention layers from Mid Fusion provide an additional 3.84% boost in task performance over early fusion models with self-Attention only (Table 1 - third row).

To better understand the role of co-Attention, we experimented with the placement of the co-Attention layers at different levels of the encoder stack. We found that co-Attention was most effective when placed after a few layers of self-Attention blocks (mid fusion). While using co-Attention layers, we obtained 3.52% MSE reduction when we had the self-Attention blocks prior to co-Attention

layers, compared to having no self-Attention blocks. This shows that fine-tuning the uni-modal representations using independent Transformer blocks prior to co-training across modalities is important.

6.3 Ablation of Input Modalities

We find that multimodal ad representations are indeed better than the unimodal components (Table 1 - forth row). In addition, we find that the video modality is the primary driver of quality in our data, demonstrated by the 18.94% increase in MSE when we drop the video modality from the input.

6.4 Special Tokens for Regression Tasks

We compare approaches to provide the joint embeddings from Transformers into the supervised regression task. We found that using the special tokens ([CLS] and [VID]) was superior to average pooling the final outputs across all tokens. This confirms that the special tokens summarize content quality well across modalities. Further, we looked at the Attention weights leading to these special tokens in the co-Attention layers, and found that high weights were placed on tokens that contained bad quality content in the alternate modality. That is, the [CLS] token focused on video frames that contained undesirable content, and the [VID] token focused on the undesirable word tokens.

Finally, we experimented with concatenation vs. dot-product of the special token embeddings into the regression task. The concatenation of the special token embeddings showed improvement of MSE by 1.15% compared to the dot-product of the special token embeddings.

7 ANALYSIS OF BAD QUALITY ADS

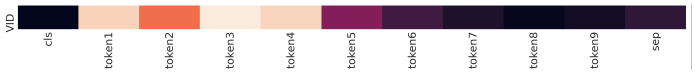
We used Attention Rollout [1] to understand how the content quality information propagated to the multimodal embeddings. The Attention weights were averaged across all heads, and the weight matrices were multiplied recursively from all layers.

Figure 4 shows the Attention rollout maps on the special tokens [VID] and [CLS] of a low quality ad from our best performing model (mid-fusion co-Attention). We see that the [CLS] token places high weight on the visual frame $F4$, which, in this case, is a frame with undesirable imagery. Similarly, the [VID] token looks at $token3$ of the undesirable word.

Interestingly, we have observed that late fusion with unimodal self-Attention performs well when the text is short. However, late fusion also amplifies noise in each modality, and we have seen that model performance deteriorates when the text is long with lots of noise (e.g., urls, random symbols). The early-fusion Transformer, on the other hand, predicts the quality of the ad correctly by overemphasizing the visual information for such cases. This seems to be the result of concatenating both modalities together at the very beginning and treating them as a single modality.

8 CONCLUSION & FUTURE WORK

In this work, we have established that Transformer models are effective at learning multimodal representations for video ads. Some promising directions for future work include extension to more modalities, such as audio signals, or in-video text, via Optical Character Recognition (OCR). Multi-task learning with tasks such as Click-Through Rate (CTR) prediction, or using additional contrastive learning objectives from user co-click data are other interesting extensions.



(a) Attention Rollout Map of the visual special token [VID] attended to the texts. High Attention weights (bright colors) were given to the text tokens with undesirable content.

(b) Attention Rollout Map of the text special token [CLS] attended to the video frames. High Attention weights (bright colors) were given to the undesirable video frames.

Figure 4: Attention Visualization from Rollout Map. High Attention weights (bright colors) were propagated onto the undesirable information in both special tokens.

ACKNOWLEDGMENTS

We are grateful to Thales Filizola Costa and Harsh Agarwal for helpful discussions on the dataset collection methodology, and the requirements for the content quality prediction task in general.

REFERENCES

[1] Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928* (2020).
[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).
[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 6836–6846.
[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/ARXIV.2005.14165
[5] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. 2016. An Analysis of Deep Neural Network Models for Practical Applications. *arXiv preprint arXiv:1605.07678* (2016).
[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. https://doi.org/10.48550/ARXIV.2005.12872
[7] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. 2021. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 5 (2021), 1–32.
[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *International conference on machine learning*. PMLR, 1597–1607.
[9] Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. 2021. Pix2seq: A Language Modeling Framework for Object Detection. https://doi.org/10.48550/ARXIV.2109.10852
[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. arXiv:cs.CV/1909.11740
[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin,

Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. https://doi.org/10.48550/ARXIV.2204.02311

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[15] 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. arXiv:cs.CV/2010.11929

[16] Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathy Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2021. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. https://doi.org/10.48550/ARXIV.2112.06905

[17] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. 2021. MDMMT: Multidomain Multimodal Transformer for Video Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 3354–3363.

[18] Valentin Gabeur, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2022. Masking Modalities for Cross-Modal Video Retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1766–1775.

[19] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*. Springer, 214–229.

[20] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. 2019. Video Action Transformer Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. 2020. Watching The World Go By: Representation Learning from Unlabeled Videos. *arXiv preprint arXiv:2003.07990* (2020).

[22] Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. 2015. Unsupervised Learning of Spatiotemporally Coherent Metrics. In *Proceedings of the IEEE international conference on computer vision*. 4086–4093.

[23] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence* (2022).

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[26] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics* 9 (2021), 570–585.

[27] Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1439–1449.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

[29] Hyodong Lee, Joonseok Lee, Joe Yue-Hei Ng, and Paul Natsev. 2020. Large Scale Video Representation Learning via Relational Graph Clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6807–6816.

[30] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. 2020. Parameter efficient multimodal transformers for video representation learning. *arXiv preprint arXiv:2012.04124* (2020).

[31] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:cs.CV/1908.03557

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[33] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. 2021. Video Transformer Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 3163–3172.

[34] 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. *CoRR* abs/2103.00020 (2021). arXiv:2103.00020 https://arxiv.org/abs/2103.00020

[35] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-training. (2018).

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[37] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[38] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* (2019).

[39] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. arXiv:cs.CV/1904.01766

[40] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-V4, Inception-Resnet and the Impact of Residual Connections on Learning. In *Thirty-first AAAI conference on artificial intelligence*.

[41] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. arXiv:cs.CL/1908.07490

[42] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. https://doi.org/10.48550/ARXIV.2201.08239

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:cs.CL/1706.03762

[44] Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised Learning of Visual Representations Using Videos. In *Proceedings of the IEEE international conference on computer vision*. 2794–2802.

[45] Haiping Wu and Xiaolong Wang. 2021. Contrastive Learning of Image Representations with Cross-Video Cycle-Consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10149–10159.

[46] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-Modal Self-Attention Network for Referring Image Segmentation. *CoRR* abs/1904.04745 (2019). arXiv:1904.04745 http://arxiv.org/abs/1904.04745

[47] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. 2021. VidTr: Video Transformer Without Convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13577–13587.

[48] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end Dense Video Captioning with Masked Transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8739–8748.