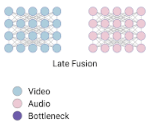
Research

Multimodal Bottleneck Transformer (MBT): A New Model for Modality Fusion



March 15, 2022

Posted by Arsha Nagrani and Chen Sun,
Research Scientists, Google Research,
Perception Team

People interact with the world through multiple sensory streams (e.g., we see objects, hear sounds, read words, feel textures and taste flavors), combining information and forming associations between senses. As real-world data consists of various signals that co-occur, such as video frames and audio tracks, web images and their captions and instructional videos and speech transcripts, it is natural to apply a similar logic when building and designing multimodal machine learning (ML) models.

Effective multimodal models have wide applications — such as multilingual image retrieval, future action prediction, and vision-language navigation — and are important for several reasons; robustness, which is the ability to perform even when

Research

dominant paradigm for multimodal fusion, called *late fusion*, consists of using separate models to encode each modality, and then simply combining their output representations at the final step, investigating how to effectively and efficiently combine information from different modalities is still understudied.

In "Attention Bottlenecks for Multimodal Fusion", published at NeurIPS 2021, we introduce a novel transformer-based model for multimodal fusion in video called *Multimodal Bottleneck Transformer* (MBT). Our model restricts cross-modal attention flow between latent units in two ways: (1) through *tight fusion bottlenecks*, that force the model to collect and condense the most relevant inputs in each modality (sharing only necessary information with other modalities), and (2) to later layers of the model, allowing early layers to specialize to information from individual modalities. We demonstrate that this approach achieves state-of-the-art results on video classification tasks, with a 50% reduction in FLOPs compared to a vanilla multimodal transformer model. We have also released our code as a tool for researchers to leverage as they expand on multimodal fusion work.

A Vanilla Multimodal Transformer Model

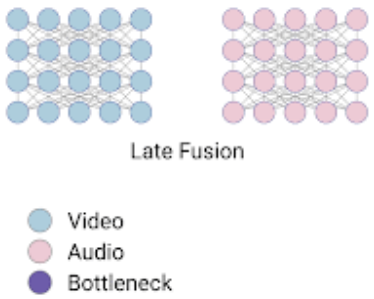
Transformer models consistently obtain state-of-the-art results in ML tasks, including video (ViViT) and audio classification (AST). Both ViViT and AST are built on the Vision Transformer (ViT); in contrast to standard convolutional approaches that process images pixel-by-pixel, ViT treats an image as a sequence of *patch tokens* (i.e., tokens from a smaller part, or patch, of an image that is made up of multiple pixels). These models then perform self-attention operations across all pairs of patch tokens. However, using transformers for multimodal fusion is challenging because of their high computational cost, with complexity scaling quadratically with input sequence length.

Because transformers effectively process variable length sequences, the simplest way to extend a unimodal transformer, such as ViT, to the multimodal case is to feed the model a sequence of both visual and auditory tokens, with minimal changes to the transformer architecture. We call this a vanilla multimodal transformer model, which allows free attention flow (called vanilla cross-attention) between different spatial and temporal regions in an image, and across frequency and time in audio inputs, represented by spectrograms. However, while easy to implement by concatenating audio and video input tokens, vanilla cross-attention at all layers of the transformer model is unnecessary because audio and visual inputs contain dense, fine-grained information, which may be redundant for the task — increasing complexity.

Research

The issue of growing complexity for long sequences in multimodal models can be mitigated by reducing the attention flow. We restrict attention flow using two methods, specifying the fusion layer and *adding attention bottlenecks*.

- **Fusion layer (early, mid or late fusion)**: In multimodal models, the layer where cross-modal interactions are introduced is called the fusion layer. The two extreme versions are *early fusion* (where all layers in the transformer are cross-modal) and *late fusion* (where all layers are unimodal and no cross-modal information is exchanged in the transformer encoder). Specifying a fusion layer in between leads to *mid fusion*. This technique builds on a common paradigm in multimodal learning, which is to restrict cross-modal flow to later layers of the network, allowing early layers to specialize in learning and extracting unimodal patterns.
- **Attention bottlenecks**: We also introduce a small set of latent units that form an attention bottleneck (shown below in purple), which force the model, within a given layer, to collate and condense information from each modality before sharing it with the other, while still allowing free attention flow within a modality. We demonstrate that this bottlenecked version (MBT), outperforms or matches its unrestricted counterpart with lower computational cost.

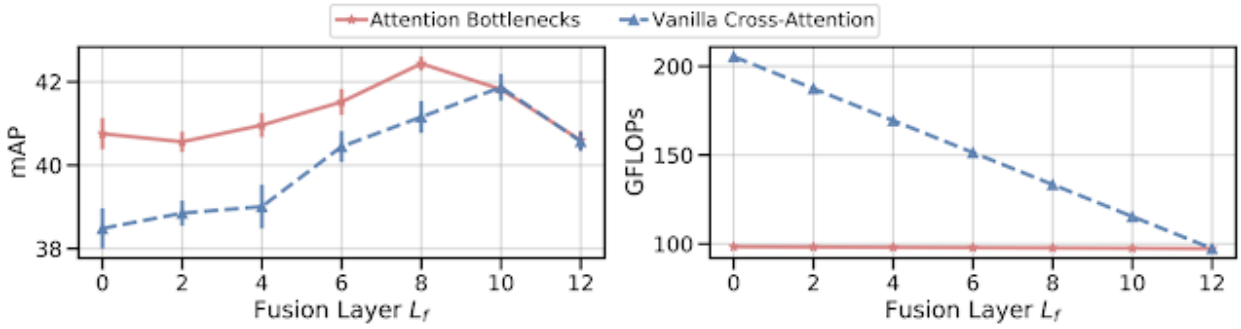


***The different attention configurations in our model**. Unlike late fusion (**top left**), where no cross-modal information is exchanged in the transformer encoder, we investigate two pathways for the exchange of cross-modal information. Early and mid fusion (**top middle**, **top right**) is done via standard pairwise self attention across all hidden units in a layer. For mid fusion, cross-modal attention is applied only to later layers in the model. Bottleneck fusion (**bottom left**) restricts attention flow within a layer through tight latent units called attention bottlenecks. Bottleneck mid fusion (**bottom right**) applies both forms of restriction in conjunction for optimal performance.*

Bottlenecks and Computation Cost

We apply MBT to the task of sound classification using the AudioSet dataset and investigate its performance for two approaches: (1) vanilla cross-attention, and (2) bottleneck fusion. For both approaches, mid fusion (shown by the middle values of the x-axis below) outperforms both early (fusion layer = 0) and late fusion (fusion

flow. We find that adding attention bottlenecks (bottleneck fusion) outperforms or maintains performance with vanilla cross-attention for all fusion layers, with more prominent improvements at lower fusion layers.



*The impact of using attention bottlenecks for fusion on mAP performance (**left**) and compute (**right**) at different fusion layers on AudioSet. Attention bottlenecks (**red**) improve performance over vanilla cross-attention (**blue**) at lower computational cost. Mid fusion, which is in fusion layers 4-10, outperforms both early (fusion layer = 0) and late (fusion layer = 12) fusion, with best performance at fusion layer 8.*

We compare the amount of computation, measured in GFLOPs, for both vanilla cross-attention and bottleneck fusion. Using a small number of attention bottlenecks (four bottleneck tokens used in our experiments) adds negligible extra computation over a late fusion model, with computation remaining largely constant with varying fusion layers. This is in contrast to vanilla cross-attention, which has a non-negligible computational cost for every layer it is applied to. We note that for early fusion, bottleneck fusion outperforms vanilla cross-attention by over 2 mean average precision points (mAP) on audiovisual sound classification, with less than half the computational cost.

Results on Sound Classification and Action Recognition

MBT outperforms previous research on popular video classification tasks — sound classification (AudioSet and VGGSound) and action recognition (Kinetics and Epic-Kitchens). For multiple datasets, late fusion and MBT with mid fusion (both fusing audio and vision) outperform the best single modality baseline, and MBT with mid fusion outperforms late fusion.

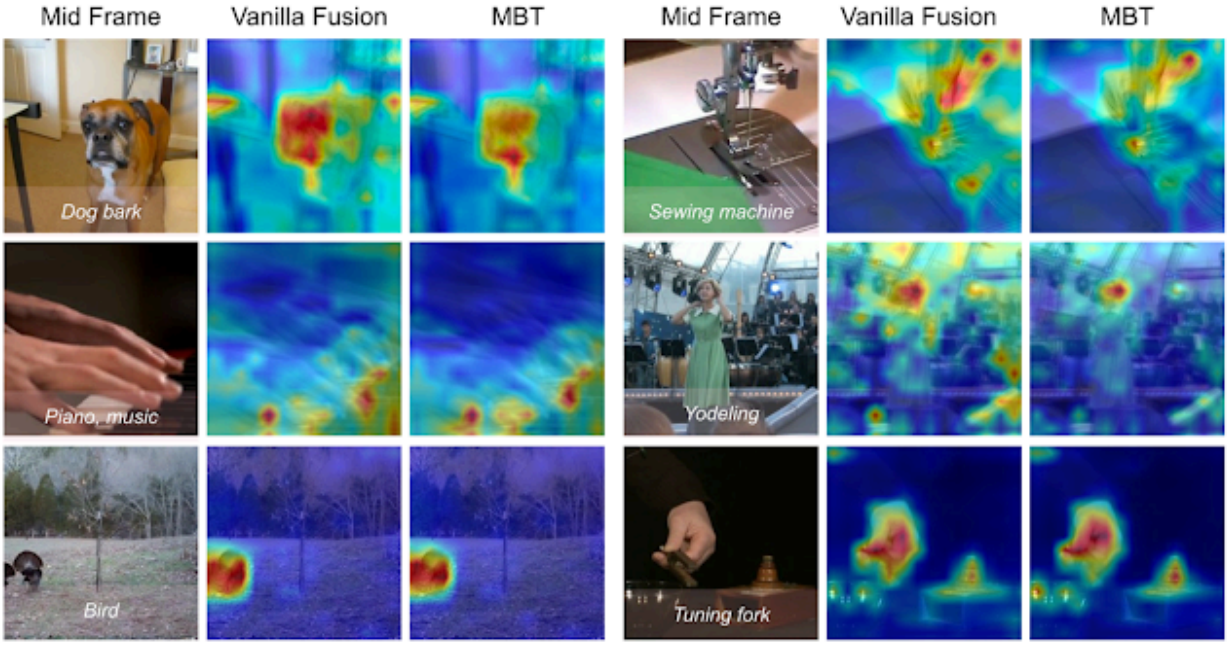
Dataset	mini-Audioset	Epic-Kitchens	VGGSound	Moments in Time	Kinetics
Best single modality baseline	31.3	40.7	52.3	36.3	79.4
Late fusion	41.8	37.90	63.3	36.5	77.0
MBT with mid fusion	**43.9**	**43.40**	**64.1**	**37.3**	**80.8**

Research

VGGSound, Moments-in-Time and Kinetics: Top-1 classification accuracy.

Visualization of Attention Heatmaps

To understand the behavior of MBT, we visualize the attention computed by our network following the attention rollout technique. We compute heat maps of the attention from the output classification tokens to the image input space for a vanilla cross-attention model and MBT on the AudioSet test set. For each video clip, we show the original middle frame on the left with the ground truth labels overlayed at the bottom. We demonstrate that the attention is particularly focused on regions in the images that contain motion and create sound, e.g., the fingertips on the piano, the sewing machine, and the face of the dog. The fusion bottlenecks in MBT further force the attention to be localized to smaller regions of the images, e.g., the mouth of the dog in the top left and the woman singing in the middle right. This provides some evidence that the tight bottlenecks force MBT to focus only on the image patches that are relevant for an audio classification task and that benefit from mid fusion with audio.



Summary

We introduce MBT, a new transformer-based architecture for multimodal fusion, and explore various fusion approaches using cross-attention between bottleneck tokens. We demonstrate that restricting cross-modal attention via a small set of fusion bottlenecks achieves state-of-the-art results on a number of video classification

Research

Acknowledgements

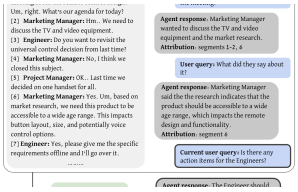
This research was conducted by Arsha Nagrani, Anurag Arnab, Shan Yang, Aren Jansen, Cordelia Schmid and Chen Sun. The blog post was written by Arsha Nagrani, Anurag Arnab and Chen Sun. Animations were created by Tom Small.

Labels:

Conferences & Events

Machine Intelligence

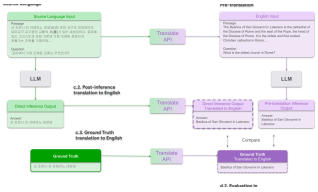
Other posts of interest



JUNE 25, 2024

Efficient data generation for source-grounded information-seeking dialogs: A use case for meeting transcripts

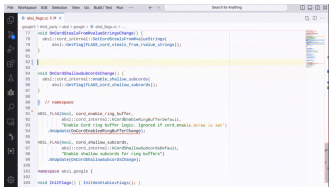
Machine Intelligence · Natural Language Processing ·



JUNE 14, 2024

Pre-translation vs. direct inference in multilingual LLM applications

Machine Intelligence · Machine Translation · Natural Language Processing



JUNE 12, 2024

Smart Paste for context-aware adjustments to pasted code

Machine Intelligence · Natural Language Processing

Google Research

Follow us