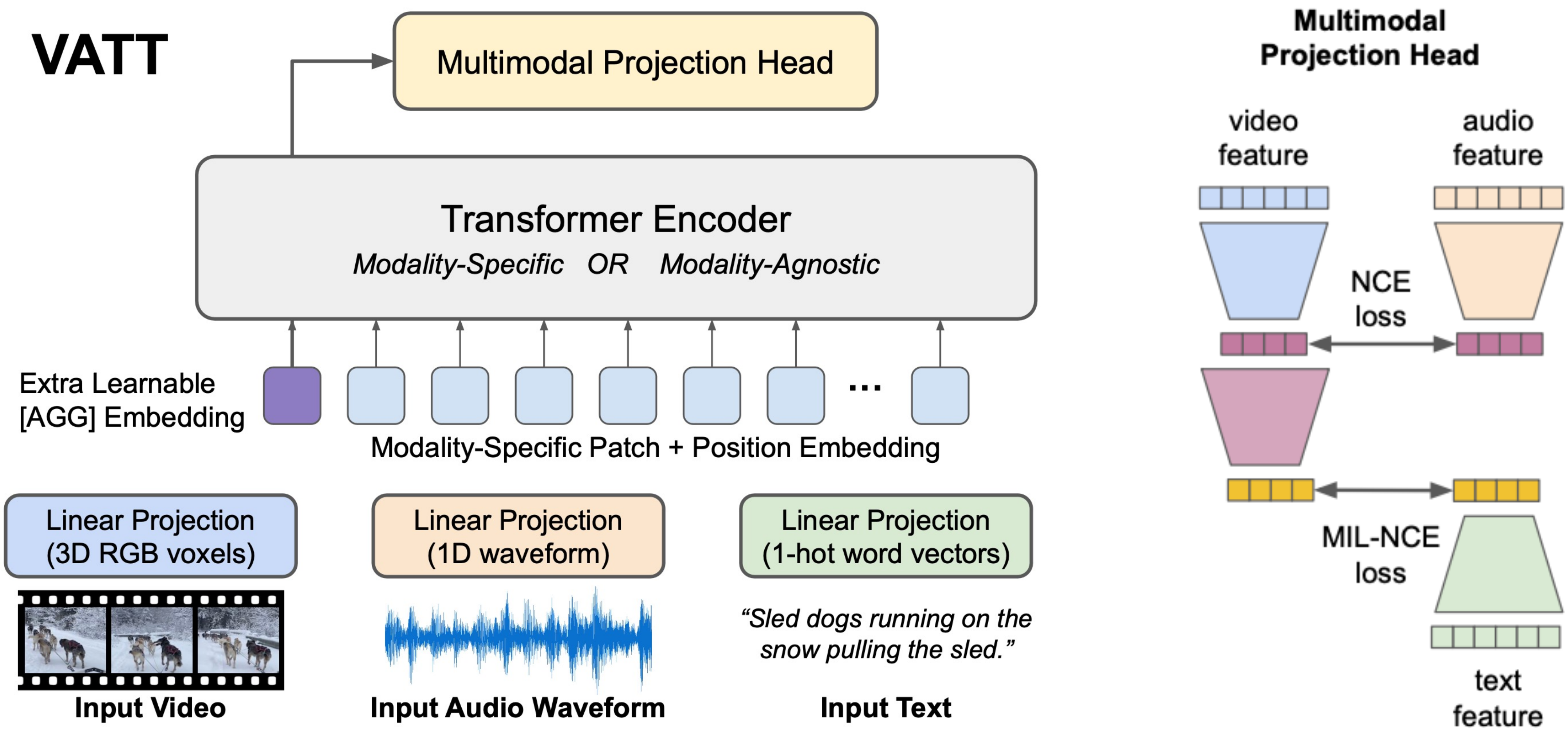
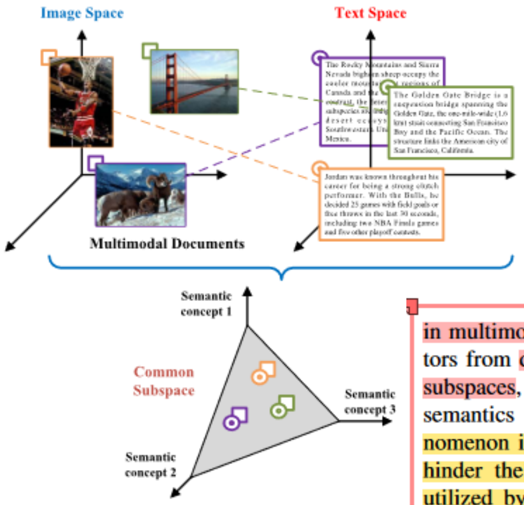
VATT :
Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and text

백승우

VATT


VATT
Multimodal Projection Head
Transformer Encoder
Modality-Specific OR Modality-Agnostic
Extra Learnable
[AGG] Embedding
...
Modality-Specific Patch + Position Embedding
Linear Projection
(3D RGB voxels)
Linear Projection
(1D waveform)
Linear Projection
(1-hot word vectors)
"Sled dogs running on the
snow pulling the sled."
Input Video
Input Audio Waveform
Input Text
Multimodal
Projection Head
video
feature
audio
feature
NCE
loss
MIL-NCE
loss
text
feature

VATT



in multimodal data. As Fig. 1 shows, since the feature vectors from different modalities originally located in unequal subspaces, the vector representations associated with similar semantics would be completely different. Here, this phenomenon is referred to as heterogeneity gap, which would hinder the multimodal data from being comprehensively utilized by the subsequent machine learning modules [4]. A popular method for addressing this problem is projecting the heterogeneous features into a common subspace, where the multimodal data with similar semantics will be represented by similar vectors [5]. Thus, the primary objective of multimodal representation learning is narrowing the distribution gap in a joint semantic subspace while keeping modality specific semantics intact.

Focus on

서로 다른 feature vector를 가지는

multi-modal의 specific semantic을 유지하며

Heterogeneity gap (멀티 모달의 이질성의 차이) 을 줄이는 것

Method

representation learning 을 통해

비슷한 sematic 을 가지는

Multi-modal data 의 heterogeneous feature 을

비슷한 feature vector로 표현

VATT

Representation learning?

특정 task에 따라 새로운 representation 에 해당하는
New feature (input feature) 을 출력하는 것,
즉 , 이런 representation 을 반영하는 feature vector 를 뽑게 학습하는 것을 말합니다.

Modality-specific representation learning

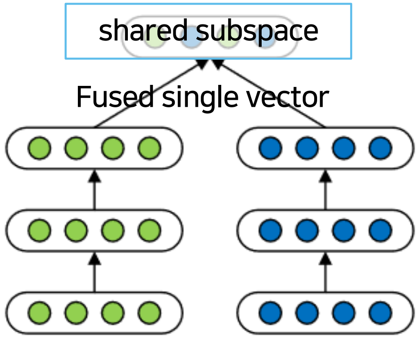
Image feature learning	Neural language processing	Video feature learning
- Convolution neural networks (CNN) such as LeNet, AlexNet , GoogleNet , VGGNet, ResNet ↓ Multi-modal model 로 통합될 수 있음 하지만 , 충분한 훈련데이터 및 계산 리소스를 고려 해야함	Word embedding RNN , BRNN LSTM,GRU CNN	Video & Audio frame encoding ↓ CNN or RNN ↓ Sequence Individual vector

VATT

Multi-modal representation learning

Joint representation

- strategy of integrating different types of features to improve the performance



(a)

Each modal
-> shared subspace
-> fused single vector

To bridge the heterogeneity gap of different modalities, joint representation aims to project unimodal representations into a shared semantic subspace, where the multimodal features can be fused [18]. As Fig. 2(a) showed, after each modality is encoded via an individual neural network, both of them will be mapped into a shared subspace, where the conceptions shared by modalities will be extracted and fused into a single vector.

각 modality는
개별 Neural Network 로 encoding
↓
공유된 하위공간에 mapping 된다.

VATT

method 1 Concatenate multi-modal features directly

각각의 modality 에 대해서 학습 후

modal의 feature concatenate

(구체적인 벡터를 추가해주는 hidden layer 병합)

problem . 따라서 각각의 modal 에서의 semantic이 결합될 것

$$z = f(w_1^T v_1 + w_2^T v_2)$$

method 2 Tensor Fusion Network - 2017

모든 modality-specific feature vector를 외적을 통해 수행된다.

TFN(Tensor Fusion Network) 에서 tensor fusion layer는

modality- embedding 에서 3-fold Cartesian product(데카르트 곱)을

사용하여 single-modal, bi-modal , tri-modal의 상호 작용을 명시적인 모델링

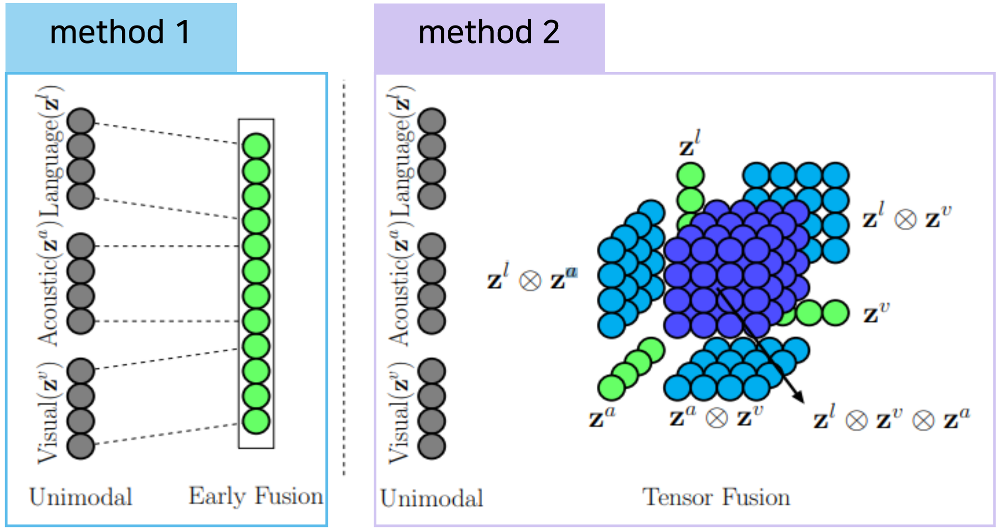
problem . 외적 계산의 연산량 많이 들음

$$\mathbf{z}^m = \begin{bmatrix} \mathbf{z}^l \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}^v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}^a \\ 1 \end{bmatrix}$$

4.2 Tensor Fusion Layer

While previous works in multimodal research has used feature concatenation as an approach for multimodal fusion, we aim to build a fusion layer in TFN that disentangles unimodal, bimodal and trimodal dynamics by modeling each of them explicitly. We call this layer Tensor Fusion, which is defined as the following vector field using three-fold Cartesian product:

$$\left\{ (z^l, z^v, z^a) \mid z^l \in \begin{bmatrix} \mathbf{z}^l \\ 1 \end{bmatrix}, z^v \in \begin{bmatrix} \mathbf{z}^v \\ 1 \end{bmatrix}, z^a \in \begin{bmatrix} \mathbf{z}^a \\ 1 \end{bmatrix} \right\}$$



[2017 Tensor Fusion Network for Multimodal Sentiment Analysis]

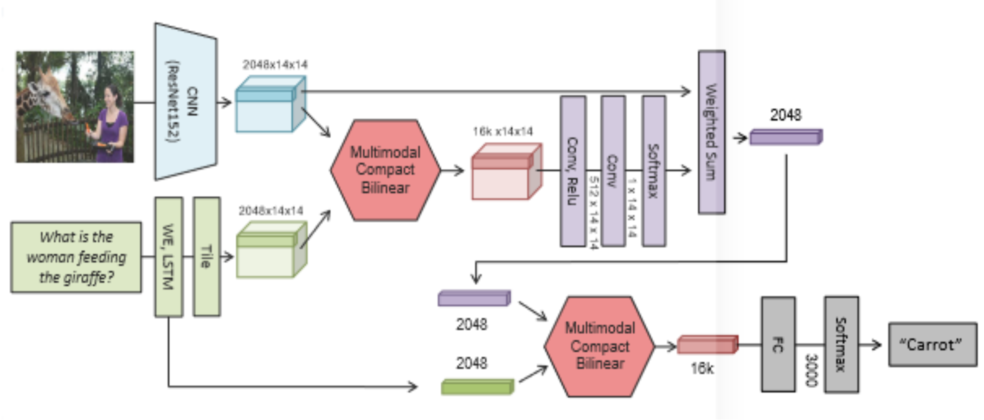
VATT

method 3 Multimodal Compact Bilinear Pooling (MCB)

Count Sketch projection function 을 통해
차원을 축소 후 외적하고 연산량을 줄임

problem . 차원을 축소하면서 일부 modal에서 데이터 일부 손실 문제로
down stream 성능에 영향

$$\begin{aligned}\Phi &= \Psi(x \otimes q)\\ &= \Psi(x) * \Psi(q)\\ &= \mathrm{FFT}^{-1}\left(\mathrm{FFT}\left(\Psi(x)\right) \odot \mathrm{FFT}\left(\Psi(q)\right)\right)\end{aligned}$$



[Multimodal compact bilinear pooling for visual question answering and visual grounding – 2016]

method 4 Statistical Regularization

누락 데이터 처리에 사용되는 트레이닝 트릭으로
Cross-modal CNN을 **정규화**를 통해
modality agnostic(modality에 구애 받지 않은)
representation 방법을 제시

Modality 간의 분포의 유사성을 가질 수 있도록 하는 hidden layer의
활성 함수로 Statistical Regularization 을 이용

Modality - invariant (모달의 불변성)이 증대

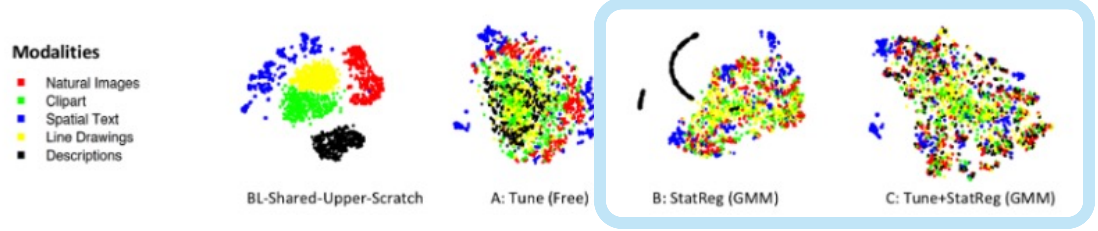
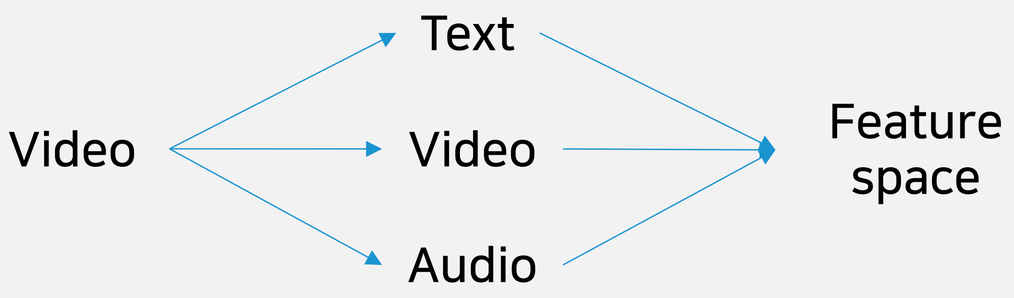


Fig. 7: **t-SNE Embedding of Cross-Modal Representation:** We visualize the embedding for `fc7` of representations from different networks using t-SNE [27]. Colors correspond to the modality. If the representation is agnostic to the modality, then the features should not cluster by modality. These visualizations suggest that our full method does a better job at discarding modality information than baselines.

[Cross-Modal Scene Networks – 2018]

VATT



Modality-specific Backbone

Problem .
Computing cost 가 증가
Data integrated 도 어려움

1.연산량을 효과적으로 줄임

2. raw data 에 대해 Modality agnostic

3. 성능이 좋음

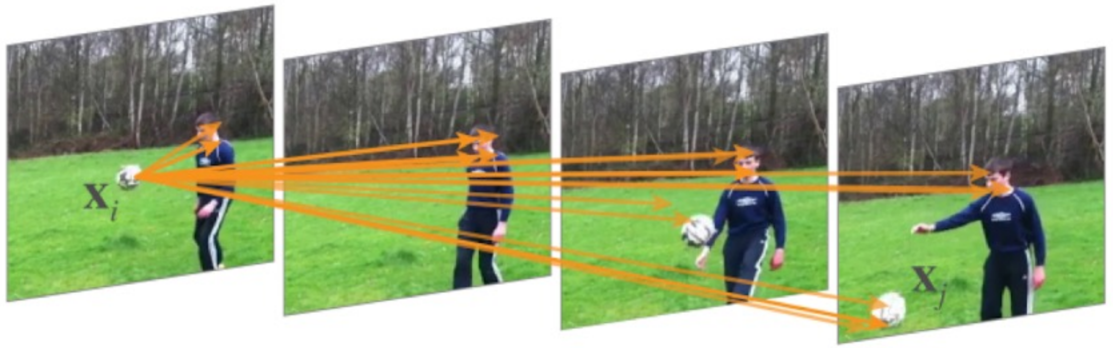
↓

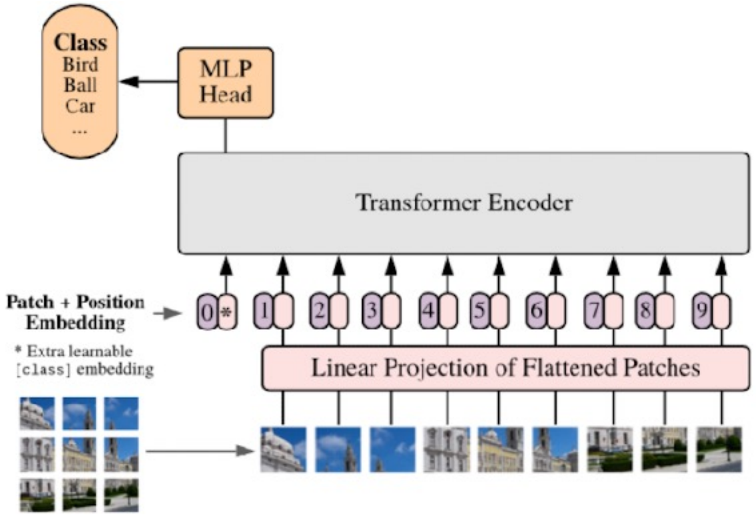
'VATT'

VATT

Background: Transformers in the Visual Domain

- Vision분야에서 weak relational inductive bias를 가진 backbone architecture에 관한 연구가 많이 진행
ex) Non-local neural networks, ViT(Transformer backbone architecture)





- 문제점

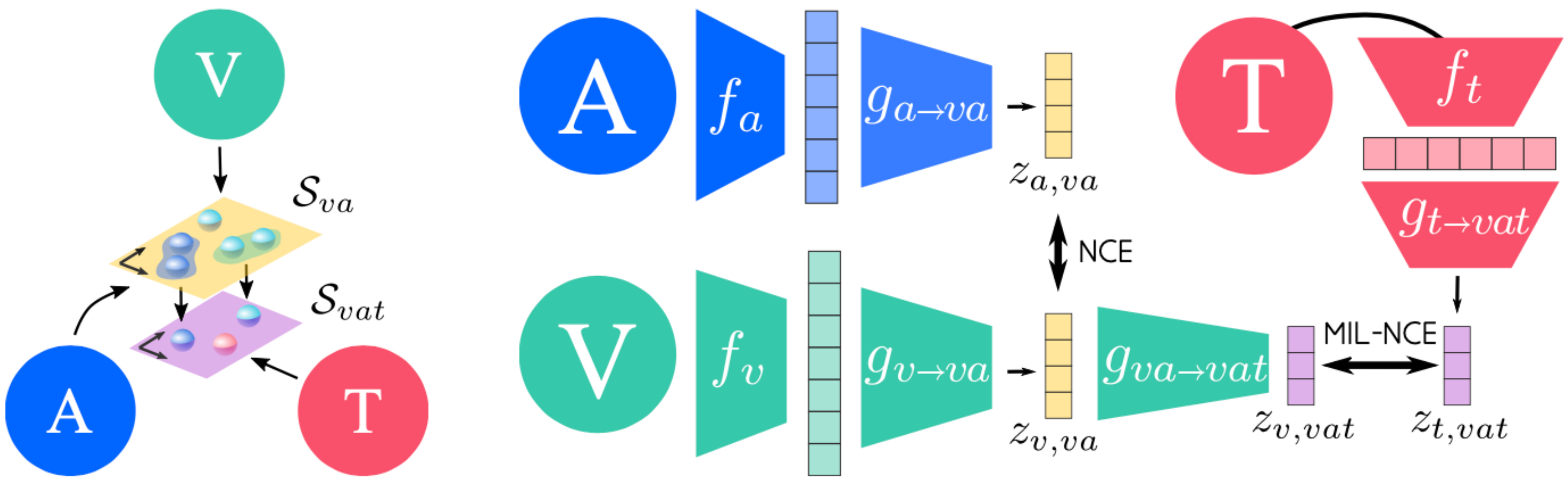
- Visual transformer는 large scale supervised training에 의존한다.
- Train 과정 중 Unlabeled visual data를 배제한다.
- Labeled data를 수집하는 과정에서 시간과 비용이 극도로 든다.

VATT

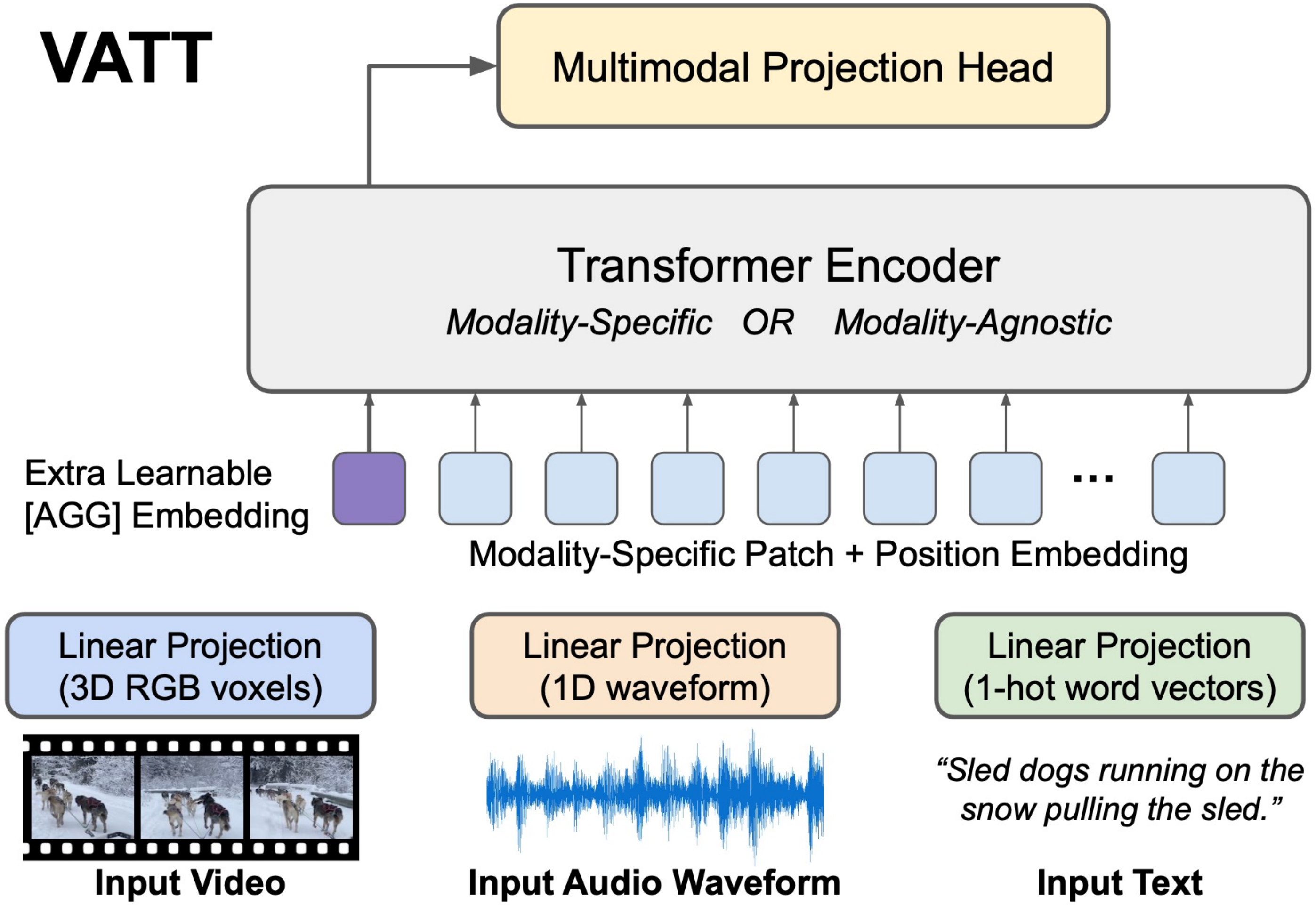
Previous study: MMV: Multi-Modal Versatile Networks

- Video, Audio, Text의 input에 각각 feature extractor가 있다.
- 각 모달리티에 적합한 아키텍쳐를 갖추고 있다.

problem . Data integrated와 computing cost 에 대한 문제를 가짐



VATT

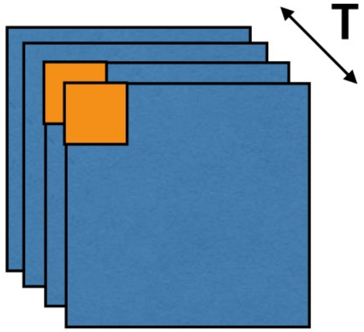


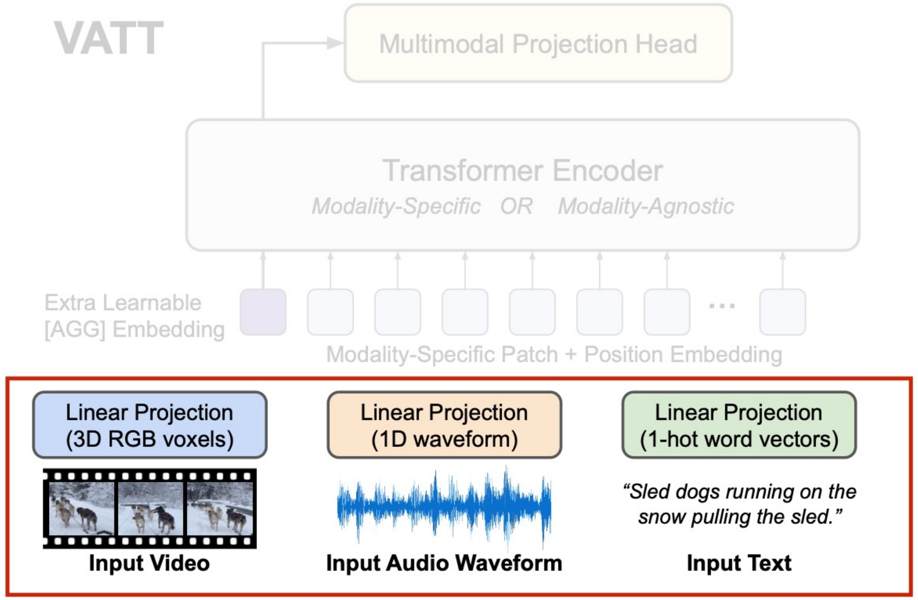
- Tokenization
- Positional Encoding
- DropToken
- TR Backbone
- Common Space Projection
- Multi-modal Contrastive Learning

VATT

Architecture: Tokenization + Positional Encoding (Video, Audio)

Video: we partition an entire video clip of size $T \times H \times W$ to a sequence of $\lceil T/t \rceil \cdot \lceil H/h \rceil \cdot \lceil W/w \rceil$ patches, where each patch contains $t \times h \times w \times 3$ voxels. We apply a linear projection on the entire voxels in each patch to get a d-dimensional vector representation. This projection is performed by a learnable weight $\boldsymbol{W}_{vp} \in \mathbb{R}^{t \cdot h \cdot w \cdot 3 \times d}$.



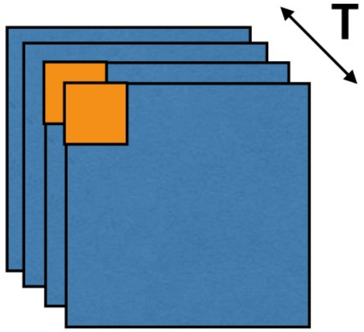


Audio: the raw audio waveform is a 1D input with length T', and we partition it to $\lceil T'/t' \rceil$ segments each containing t' waveform amplitudes. Similar to video, we apply a linear projection with a learnable weight $\boldsymbol{W}_{ap} \in \mathbb{R}^{t' \times d}$ to all elements in a patch to get a d-dimensional vector representation. We use $\lceil T'/t' \rceil$ learnable embeddings to encode the position of each waveform segment.

VATT

Architecture: Tokenization + Positional Encoding (Video, Audio)

Video: we partition an entire video clip of size $T \times H \times W$ to a sequence of $\lceil T/t \rceil \cdot \lceil H/h \rceil \cdot \lceil W/w \rceil$ patches, where each patch contains $t \times h \times w \times 3$ voxels. We apply a linear projection on the entire voxels in each patch to get a d-dimensional vector representation. This projection is performed by a learnable weight $\boldsymbol{W}_{vp} \in \mathbb{R}^{t \cdot h \cdot w \cdot 3 \times d}$.



$$\boldsymbol{E}_{\text{Temporal}} \in \mathbb{R}^{\lceil T/t \rceil \times d}$$

$$\boldsymbol{E}_{\text{Horizontal}} \in \mathbb{R}^{\lceil H/h \rceil \times d}$$

$$\boldsymbol{E}_{\text{Vertical}} \in \mathbb{R}^{\lceil W/w \rceil \times d}$$

$$\boldsymbol{e}_{i,j,k} = \boldsymbol{e}_{\text{Temporal}_i} + \boldsymbol{e}_{\text{Horizontal}_j} + \boldsymbol{e}_{\text{Vertical}_k}$$

Audio: the raw audio waveform is a 1D input with length T', and we partition it to $\lceil T'/t' \rceil$ segments each containing t' waveform amplitudes. Similar to video, we apply a linear projection with a learnable weight $\boldsymbol{W}_{ap} \in \mathbb{R}^{t' \times d}$ to all elements in a patch to get a d-dimensional vector representation. We use $\lceil T'/t' \rceil$ learnable embeddings to encode the position of each waveform segment.

use $\lceil T'/t' \rceil$ learnable embeddings

VATT

Architecture: Positional Encoding (text)

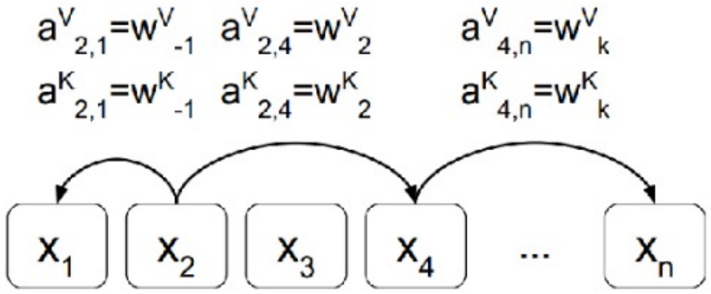
- 기존 position encoding 대신에 relative positional encoding 사용

→ attention score를 구하는 과정에서 key, query의 상대적인 거리를 더해주는 방법

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V)$$

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}} \qquad \downarrow \qquad e_{ij} = \frac{x_i W^Q (x_j W^K + \boxed{a_{ij}^K})^T}{\sqrt{d_z}}$$

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V + \boxed{a_{ij}^V})$$



$$a_{ij}^K = w^K_{\mathrm{clip}(j-i,k)}$$

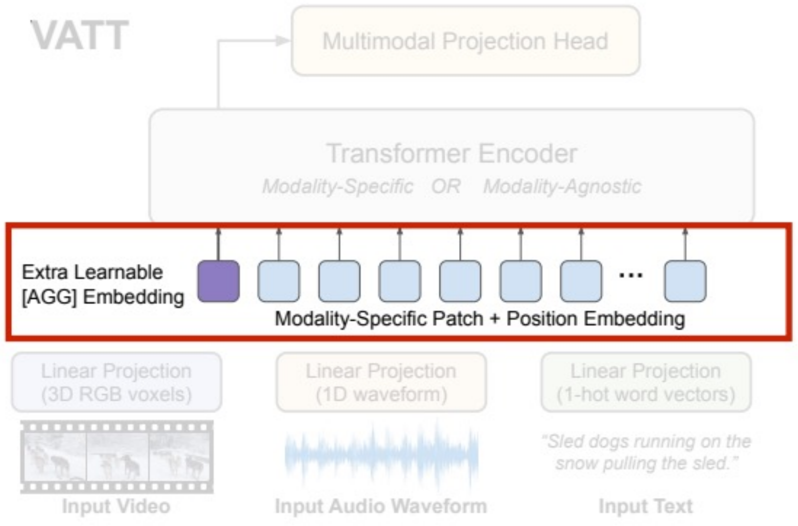
$$a_{ij}^V = w^V_{\mathrm{clip}(j-i,k)}$$

$$\mathrm{clip}(x,k) = \max(-k, \min(k,x))$$

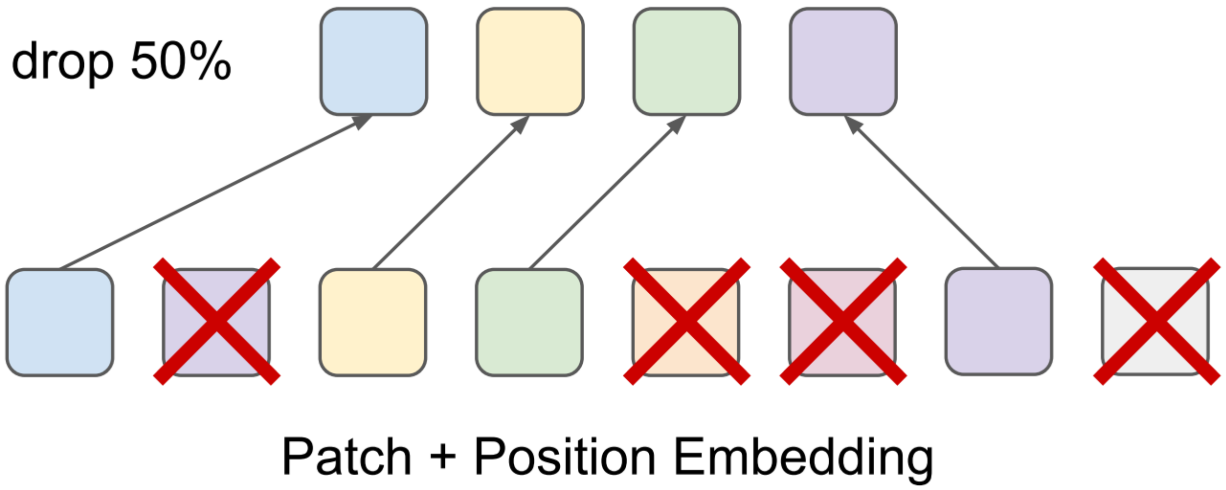
VATT

Architecture: DropToken

- Training 과정에서 연산량을 효과적으로 줄이는 간단한 방법
- Video & Audio input에 적용



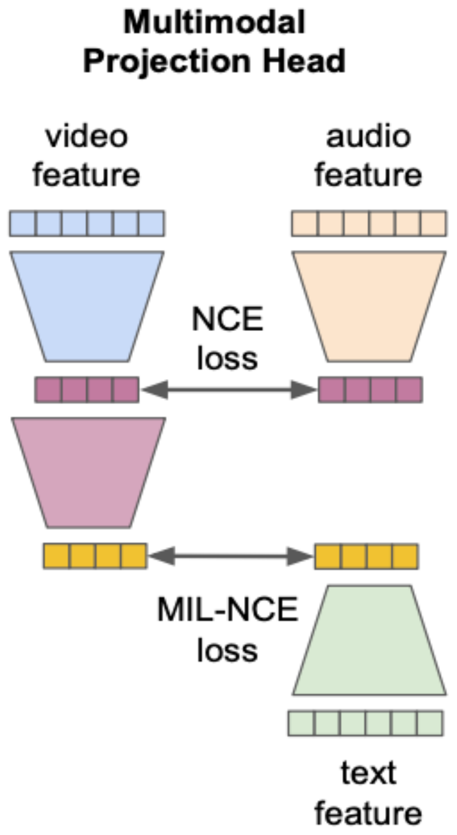
<DropToken>



VATT

Architecture: Common Space Projection

- 서로 다른 modality를 비교하기 위해 같은 feature space에 투영하여 비교
- Vision & Audio : same-d space
- Vision & Text : low-d space



- Video : 2-layer (d-512)
- Audio : 1 layer (d-512)
- Text : 1 layer (d-256)

VATT

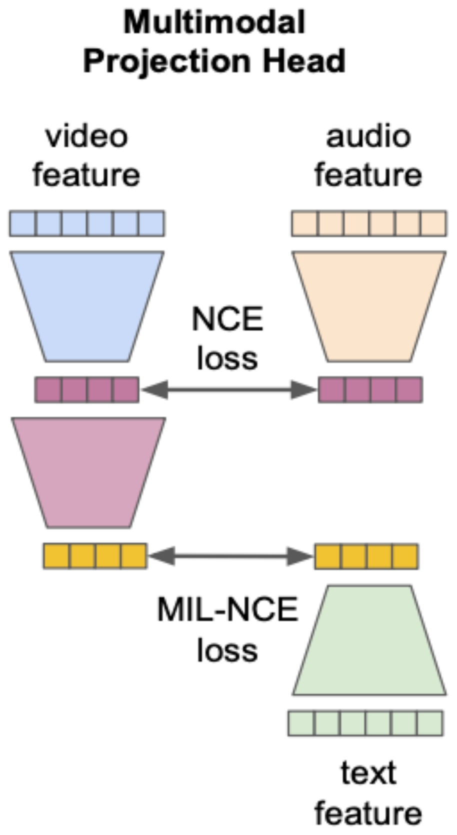
Training: Multi-modal Contrastive Learning

- Self-supervised learning에 Contrastive Learning을 사용
- Vision & Audio : Noise-Contrastive Estimation (NCE) loss
- Vision & Text : Multiple-Instance-Learning-NCE (MIL-NCE) loss

$$\mathcal{L} = \text{NCE}(\boldsymbol{z}_{v,va}, \boldsymbol{z}_{a,va}) + \lambda \text{MIL-NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\})$$

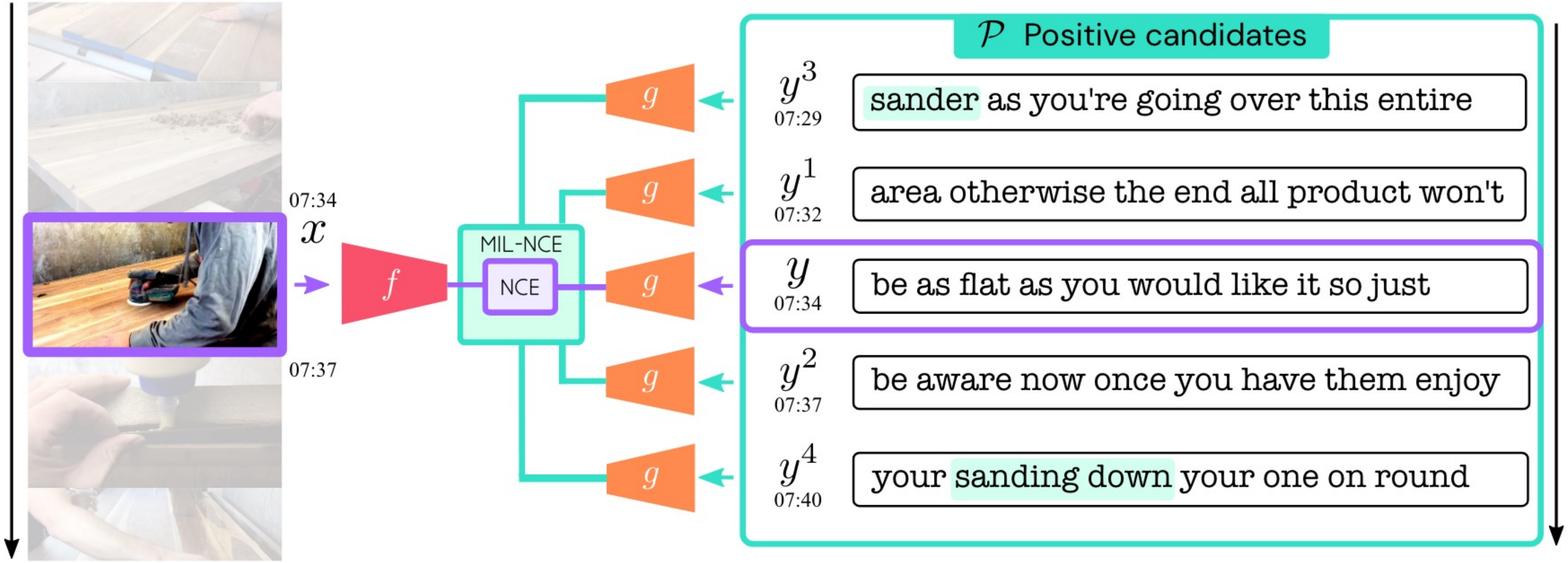
$$\text{NCE}(\boldsymbol{z}_{v,va}, \boldsymbol{z}_{a,va}) = -\log\left(\frac{\exp(\boldsymbol{z}_{v,va}^{\top}\boldsymbol{z}_{a,va}/\tau)}{\sum_{i=1}^{B}\exp({\boldsymbol{z}^{i}}_{v,va}^{\top}{\boldsymbol{z}^{i}}_{a,va}/\tau)}\right),$$

$$\text{MIL-NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\}) = -\log\left(\frac{\sum_{\boldsymbol{z}_{t,vt}\in\mathcal{P}(\boldsymbol{z}_{v,vt})}\exp(\boldsymbol{z}_{v,vt}^{\top}\boldsymbol{z}_{t,vt}/\tau)}{\sum_{\boldsymbol{z}_{t,vt}\in\mathcal{P}(\boldsymbol{z}_{v,vt})\cup\mathcal{N}(\boldsymbol{z}_{v,vt})}\exp(\boldsymbol{z}_{v,vt}^{\top}\boldsymbol{z}_{t,vt}/\tau)}\right)$$



VATT

- Vision & Text : Multiple-Instance-Learning-NCE (MIL-NCE) loss



$$\text{MIL-NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\}) = -\log\left(\frac{\sum_{\boldsymbol{z}_{t,vt}\in\mathcal{P}(\boldsymbol{z}_{v,vt})}\exp(\boldsymbol{z}_{v,vt}^{\top}\boldsymbol{z}_{t,vt}/\tau)}{\sum_{\boldsymbol{z}_{t,vt}\in\mathcal{P}(\boldsymbol{z}_{v,vt})\cup\mathcal{N}(\boldsymbol{z}_{v,vt})}\exp(\boldsymbol{z}_{v,vt}^{\top}\boldsymbol{z}_{t,vt}/\tau)}\right)$$

VATT

Conclusion

- Transformer기반 self-supervised multimodal representation learning 제안
- Modality-specific & Modality-agnostic
- multi-modal self-supervised pre-training으로 기존 large-scale labeled data에 의존성을 줄임
- DropToken: 간단하고 효과적으로 complexity를 줄여주는 방법