고차 다항식 풀링을 사용한 딥 멀티모달 다선형 융합

밍 허우[1,*], **지아지아** 탕[2,1,*], **지안하이** 장[2], **완젠** 콩[2], **치빈** 자오[1,†]
[1] 일본 이화학연구소 첨단지능 프로젝트 센터, 텐서 학습 유닛
[2] 중국 항저우 디안지 대학교 컴퓨터 과학 대학
ming.hou@riken.jp, hdutangjiajia@163.com
jhzhang@hdu.edu.cn, kongwanzeng@hdu.edu.cn, qibin.zhao@riken.jp

초록

텐서 기반 다중 모드 융합 기법은 뛰어난 예측 성능을 보여주었습니다. 하지만 기존 접근 방식은 2선형 또는 3선형 풀링만 고려하기 때문에 상호 작용 순서가 제한된 다선형 융합의 완전한 표현력을 발휘하지 못한다는 한 가지 한계가 있습니다. 더 중요한 것은 단순히 특징을 한꺼번에 융합하면 복잡한 국부적 상호관계를 무시하여 예측 성능이 저하된다는 점입니다. 이 연구에서는 먼저 고차 모멘트를 고려하여 다중 모드 특징을 통합하기 위한 다항식 텐서 풀링(PTP) 블록을 제안한 다음 텐서화된 완전 연결 계층을 제안합니다. PTP를 빌딩 블록으로 취급하여 계층적 다항식 융합 네트워크(HPFN)를 구축하여 로컬 상관관계를 글로벌 상관관계로 재귀적으로 전송합니다. 여러 개의 PTP를 쌓아 올리면 계층 수에 따라 HPFN의 표현력이 기하급수적으로 증가하며, 이는 매우 심층적인 컨볼루션 연산 회로와 동등하다는 것을 보여줍니다. 다양한 실험을 통해 최첨단 성능을 달성할 수 있음을 입증했습니다.

1 소개

다중 모드 표현 학습은 인공지능과 인간 커뮤니케이션 분석 분야에서 매우 활발하게 성장하고 있는 연구 분야입니다. 감정 인식[2], 성격 특성 인식[22], 감정 분석[18]과 같은 인간의 멀티모달 작업 전반에 걸쳐 그 응용이 확산되고 있습니다. 다양한 방식(음성 언어, 시각 및 청각 신호)에서 수집된 멀티모달 신호는 일관성과 상보성의 특성을 나타냅니다[28]. 다중 양식과 그 복잡한 상호 작용을 모델링하기 위해 광범위한 연구가 진행되었습니다 [28, 14, 15, 13]. 이러한 상호 작용은 사소하지 않은 다중 모드 정렬 및 모드 간 신뢰할 수 없거나 모순되는 정보와 같은 요인으로 인해 모델링하기 어렵습니다. 멀티모달 데이터의 이질적인 속성을 탐색하여 모델의 일반화 능력을 향상시키는 것은 여전히 주요 과제로 남아 있습니다.

멀티모달 모델링의 핵심 단계는 멀티모달 융합으로, 여러 모달리티의 특징을 통합하여 보다 강력한

예측을 도출하는 것을 목표로 합니다. 일반적으로 멀티모달 특징 융합은 조기 융합, 후기 융합, 하이브리드 융합으로 분류할 수 있습니다[1]. 이 중 조기 융합은 서로 다른 소스의 연결된 신호를 모델 입력으로 활용합니다[7]. 반면 후기 융합은 각 양식을 개별적으로 모델링하여 투표 또는 평균을 통해 의사 결정 수준에서 병합하려고 시도합니다 [19, 25]. 하이브리드 융합에서 출력은 단일 모드와 초기 융합의 예측에 모두 의존합니다. 단순함에도 불구하고 앞서 언급한 기존의 융합 기법은 모두 다중 모드 특징의 연결 또는 평균화, 또는 더 일반적으로 선형 조합으로 제한됩니다. 그리고 선형 모델링은 복잡한 상호 연관성을 포착하기에 충분하지 않을 수 있습니다.

*저자들은 동등하게 기여합니다.
†해당 저자

제33회 신경 정보 처리 시스템 컨퍼런스(NeurIPS 2019), 캐나다 밴쿠버.

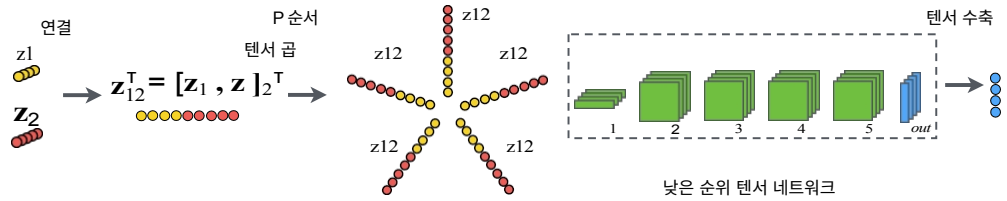


그림 1: z1과 z2를 융합하기 위한 5차 다항식 텐서 풀링(PTP) 블록의 구성도.

텐서 곱 표현을 활용하여 최근의 융합 모델[16, 27]은 2선형/삼선형 교차 모달 상호 작용을 모델링하는 데 초점을 맞추고 있으며 성능을 크게 향상시킵니다. 그럼에도 불구하고 이러한 표현은 단일 모달의 차원과 모달의 수와 관련하여 특징 차원이 기하급수적으로 증가하여 엄청난 양의 매개 변수를 생성하는 데 어려움을 겪습니다. 이 문제를 해결하기 위해 [17]에서는 낮은 순위의 텐서 인자를 학습하여 융합 파라미터를 효율적으로 줄이면서 3모달(삼선형) 상호작용을 표현할 수 있는 능력을 보존합니다.

그러나 이 모델은 상호 작용 순서를 제한함으로써 다선형 특징 상호 관계의 표현력을 충분히 발휘하지 못합니다. 즉, 상호 작용은 각 양식에 대해 선형적이며, 예를 들어 세 가지 양식에 대해 최대 3선형 상호 작용까지만 가능합니다. 더 중요한 것은 이러한 프레임워크가 최종 예측에 중요한 상호작용의 로컬 역학을 완전히 무시한 채 단순히 멀티모달 기능을 한꺼번에 융합하는 데 초점을 맞추고 있다는 점입니다. 따라서 진화하는 시간적-양식적 상관관계를 파악할 수 없으며, 특히 긴 시계열이 관련된 경우 예측이 악화될 수 있습니다.

이 작업에서는 먼저 국부적으로 혼합된 시간-양식 특징을 융합할 수 있는 다항식 텐서 풀링(PTP) 블록을 제안합니다. PTP를 사용하면 고차 모멘트를 통해 복잡한 비선형 멀티모달 상관관계를 포착할 수 있습니다. 기본 PTP 블록을 기반으로 로컬 시간-모달리티 상관관계를 재귀적으로 통합하여 글로벌 상관관계로 전송하는 계층적 아키텍처를 추가로 구축합니다. 이렇게 하면 멀티모달 시계열 데이터를 융합하는 것이 가능해집니다. 우리는 제안된 프레임워크를 계층적 다항식 융합 네트워크(HPFN)라고 부릅니다. HPFN을 사용하면 두 가지 이점이 있습니다: 1) 로컬 상호작용을 훨씬 더 세밀하게 파악할 수 있으며, 지배적인 로컬 상관관계를 글로벌 규모로 효율적으로 전송할 수 있습니다. 2) PTP를 여러 층으로 쌓아 올리면 표현력이 기하급수적으로 증가하는데, 이는 HPFN을 매우 심층적인 컨볼루션 연산 회로에 연결하면 알 수 있습니다. 두 가지 멀티모달 작업에서 HPFN의 우수한 성능을 검증합니다.

2 예선전

우리는 실수의 다방향 배열을 *텐서라고* 부릅니다[12]. P차 텐서 $\mathbf{W} \in$는 다음과 같이 나타냅니다. $\mathbf{R}^{I_1 \times \cdots \times I_P}$, *P 모드*. W의 $(i_1, \ldots, i_P)$번째 항목은 $i_p \in [I_p]$와 함께 $\mathbf{W}_{i_1,\ldots,i_P}$로 표시됩니다. 는 모든 $p \in [P]$에 대해, 여기서 $[P]$는 $\{1, 2, \ldots, P\}$ 집합을 나타냅니다. 로 표시된 *텐서 곱은* 텐서 분석의 기본 연산자입니다. 두 텐서 $\mathbf{A} \in \mathbf{R}^{I_1,\ldots,I_P}$와 $\mathbf{B} \in \mathbf{R}^{I_{P+1},\ldots,I_{P+Q}}$가 주어지면 텐서 곱은 다음과 같이 $(P + Q)$차 텐서 $\mathbf{A} \otimes \mathbf{B} \in \mathbf{R}^{I_1,\ldots,I_{P+Q}}$를 생성합니다.

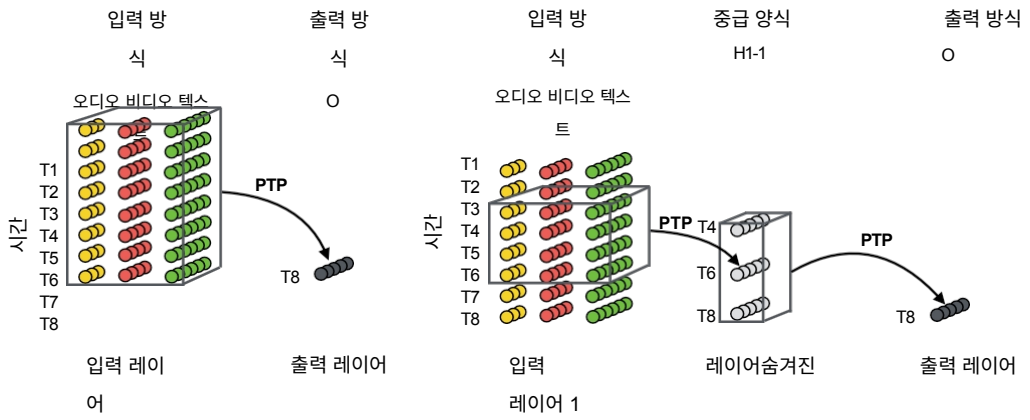
$$\mathbf{A} \otimes \mathbf{B} = \mathbf{A}_{i_1,\ldots,i_P} \cdot \mathbf{B}_{i_{P+1},\ldots,i_{P+Q}}. \tag{1}$$

텐서 곱은 벡터 입력에 대한 표준 외부 곱으로 축소됩니다. $p \in [P]$에 대한 P 벡터 $\mathbf{v}^{(p)} \in \mathbf{R}^{I_p}$

의 텐서 곱은 1순위 텐서 $\mathsf{A} = \mathbf{w}^{(1)} \otimes \cdots \otimes \mathbf{w}^{(P)}$를 산출합니다. CAN-W의 DECOMP/PARAFAC(CP) 분해[3]는 다음과 같이 1등급 텐서의 합으로 쓸 수 있습니다. $\mathsf{W} = \sum_{r=1}^{R} \mathbf{w}_r^{(1)} \otimes \cdots \otimes \mathbf{w}_r^{(P)}$, 여기서 R은 *텐서 랭크로* 정의됩니다. *텐서 네트워크*(TN) [4] 생성 고차 텐서를 드물게 상호 연결된 저차 텐서 집합으로 인수 분해하여 텐서 분해를 알라이즈합니다. TN 표현은 고차 고밀도 텐서와 관련된 차원성의 저주 효과를 크게 줄여줍니다. TN에는 CP, Tucker[26], 텐서 트레인(TT)[21], 텐서 링(TR)[32] 형식과 같은 여러 가지 특수한 경우가 포함됩니다.

3 방법론

이 섹션에서는 계층적 다항식 퓨전 프레임워크(HPFN)의 기본 구성 요소 역할을 하는 다항식 텐서 풀링(PTP)이라는 제품 풀링 전략을 소개하는 것으로 시작하겠습니다.



(a) 1레이어 HPFN (b) 2레이어 HPFN

그림 2: (a) 수신 '창' 크기가 [8 × 3]인 단일 PTP 블록이 있는 퓨전 네트워크의 예시입니다. (b) 2계층 HPFN의 예시. 입력 계층의 경우, 겹쳐진 '창'의 크기는 [4 × 3]이고 시간 차원에 따라 보폭 스텝 크기는 2입니다. 숨겨진 레이어의 경우, [3 × 1] 크기의 '창'은 이전 레이어의 모든 중간 특징을 포함합니다. H1-1은 '첫 번째' 숨겨진 레이어에 있는 특징 노드의 '첫 번째' 열 인덱스를 나타냅니다.

1) 고차 비선형 모달 내 및 모달 간 상호작용을 명시적으로 모델링하고, 2) 멀티모달 시계열의 경우 시간적 차원과 모달 차원 모두에서 스캐닝 수용 '창' 내의 로컬 상호작용을 직접 모델링할 수 있다는 점입니다.

3.1 고차 다항식 텐서 풀링(PTP)

PTP 블록의 목적은 $\{\mathbf{z}_m\}_{m=1}^{M}$ 를 하나의 조인트로 병합하는 것입니다.

고차 모멘트의 명시적 상호 작용을 활용하여 압축 표현 **Z를 생성합니다.** 그림 1은 PTP 블록의 연산 흐름도를 보여줍니다. 보다 구체적으로, M*개의* 특징 벡터 집합은 다음과 같습니다. $\{\mathbf{z}_m\}_{m=1}^{M}$ 는 먼저 긴 특징 벡터 $\mathbf{z}_{12\cdots M}$로 연결됩니다:

$$\mathbf{z}_{12\cdots M} = [1, \mathbf{z}_1^{\mathsf{T}}, \mathbf{z}_2^{\mathsf{T}}, \cdots, \mathbf{z}_M^{\mathsf{T}}]. \tag{2}$$

그런 다음 연결된 특징 벡터 $\mathbf{z}_{12\cdots M}$의 P *차텐서곱*을 사용하여 P *다항식 특징 텐서(*^{P}*) 차수를* 다음과 같이 공식화합니다.

$$\mathcal{Z}^{P} = \mathbf{z}_{12\cdots M} \otimes \mathbf{z}_{12\cdots M} \otimes \cdots \otimes \mathbf{z}_{12\cdots M}, \tag{3}$$

여기서 $\{\otimes_p\}_{p=1}^{P}$ 는 텐서 곱 연산자입니다. ZP는 가능한 모든 텐서곱을 표현할 수 있습니다. (2)에서 상수 항 '1'의 통합으로 인해 최대 P*차까지 다항식* 확장이 가능합니다. 특징 간 P*차 다항식* 상호 작용의 효과는 *풀링 가중 텐서* $\mathcal{W} = [\mathcal{W}_1, \ldots, \mathcal{W}_h, \ldots, \mathcal{W}_H]$로 완전히 측정할 수 있습니다:

$$z_h = \sum_{i_1, i_2, \cdots, i_P} w_{h i_1 i_2 \cdots i_P} \cdot z_{i_1 i_2 \cdots i_P}, \tag{4}$$

여기서 z_h *는* H *차원* 융합 벡터 **z의** h *번째* 요소를 나타내고, i_p 는 p 번째 모드에서 고차 항을 인덱싱합니다. 안타깝게도 (4)의 매개변수 $\mathcal{W}_h$ 의 매개변수 수는 (4)에서 증가합니다.

를 다항식 순서 P 로 기하급수적으로 증가시킵니다. 이 문제를 해결하기 위해 낮은 순위의 TN을 채택하여 $\mathcal{W}_h$*를* 효율적으로 근사화합니다. $\mathcal{W}_h$*가 랭크-*R CP 형식을 허용한다고 가정하면 (4)는 다음과 같이 됩니

다.

$$z_h = \sum_{i1,i2,\cdots,iP} \left(\sum_{r=1} a^h_{r} \prod_{p=1} w^{h(p)}_{r;ip} \right) \left(\prod_{p=1} z12\cdots M;i_p \right) = \sum_{r=1} a^h_{r} \prod_{p=1} \sum_{ip} w^{h(p)}_{r;ip} z12\cdots M;i_p . \quad (5)$$

명시적으로 구성된 피처 텐서는 매우 대칭적이므로 다음과 같이 가정하는 것이 합리적입니다.

$w^h_r = \mathbf{w}^{h(p)}_r$ 모든 $p \in [P]$에 대해. 따라서 $\{\{a^h, w^h_r\}_{r=1}^R\}_{h=1}^H$ 에 대한 융합 매개변수의 집합은 추정합니다. Wh가 TR 형식을 허용하는 경우 (4)에서 다음과 같은 공식을 도출할 수 있습니다.

$$z_h = \sum_{i1,i2,\cdots,iP} \left(\sum_{r1,r2,\cdots,rP} \prod_{p=1} G^{h(P)}_{rP;iP;rP+1} \right) \left(\prod_{p=1} z12\cdots M;i_p \right)$$

$$= \sum_{\substack{R_1,R_2,\cdots,R_P \\ I_p}} \prod_{P=1} G^{h(P)}_{rP;iP;rP+1} z12\cdots M;i_p = \sum_{r1,r2,\cdots,rP} \prod_{p=1} G^{h(p)}_{rp;rp+1} = \mathrm{Trace}\left(\prod_{p=1} \tilde{G}^{h(p)} \right), \quad (6)$$

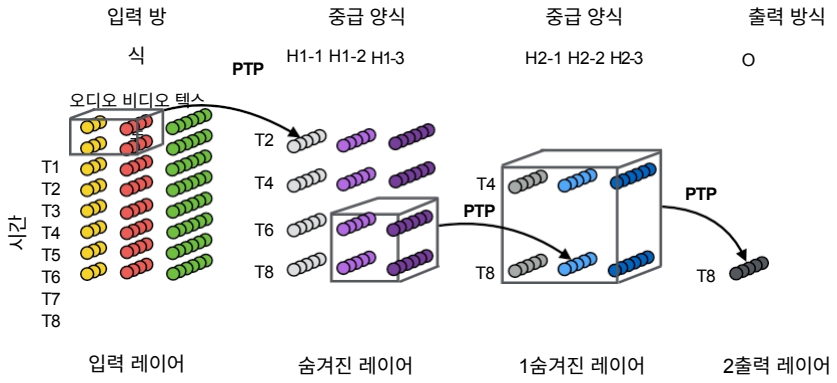


그림 3: 3계층 HPFN의 예시.

여기서 3차 코어 텐서 $\{\{G_h(p)\}_{P}$ $_{p=1}\}_{h=1}^{H}$ 는 융합 매개변수입니다. $\{r_p\}_{P}$ $_{p=1}$ 가 정의됩니다. 를 $r_{P+1} = r_1$ 의 TR-랭크로 가정하는 것도 합리적입니다. 또한 모든 p [P]에 대해 공유 $^{h} = h(p)$ 를 가정하는 것도 합리적입니다. 이렇게 하면 각 차원을 따라 융합 계산을 암시적으로 효율적으로 수행할 수 있으므로 특징 텐서와 가중 텐서 모두에서 차원성의 저주를 피할 수 있습니다.

3.2 계층적 다항식 퓨전 네트워크(HPFN)

기본적인 풀링 블록을 소개한 다음, 이제 멀티모달 데이터를 융합하기 위한 일반적인 프레임워크를 소개하겠습니다. 일반적으로 멀티모달 시계열을 '2D 피처 맵'으로 재배열하면, 상관관계 패턴이 두 차원에 걸쳐 시간적-양식적 특징이 국지적으로 혼합된 수용적인 '창'에 나타날 수 있습니다. 그런 다음 단일 PTP 블록을 해당 로컬 '창'에 연결하여 상호 작용을 측정할 수 있습니다. 계층적 아키텍처를 사용하면 여러 계층에 PTP를 스태킹하여 로컬의 시간적-양식적 상관관계 패턴을 재귀적으로 통합할 수 있습니다. 마지막으로 중요한 상관관계가 식별되어 글로벌 규모로 전송됩니다.

그림 2 (a)는 8개의 시간 단계와 3개의 모달리티에 걸친 특징을 모두 포괄하는 하나의 수신 '창'에서 단일 PTP가 작동하는 간단한 1계층 융합 네트워크를 보여줍니다. 이러한 방식으로 PTP는 '창' 내에 있는 총 24개의 혼합된 특징들 사이의 고차 비선형 상호작용을 포착할 수 있게 해줍니다. PTP가 작은 수용 '창'에 연결되면 자연스럽게 국소적 상관관계를 특징짓는 것을 관찰할 수 있습니다. 그리고 '2D 피처 맵'의 서로 다른 위치에 있는 혼합 피처의 로컬 '창'에 여러 개의 PTP 블록을 배치할 수 있습니다. 그런 다음 각 레이어의 작은 '창'에 PTP 블록을 부착하여 융합 프로세스를 여러 레이어로 분산하는 것이 간단합니다. 실제로 상위 레이어의 융합 노드는 하위 레이어에서 더 큰 효과를 수용하는 피처의 '창'에 해당합니다. 그 결과, 보다 표현력이 풍부한 로컬 및 글로벌 상관관계를 매우 유연하게 효율적으로 모델링할 수 있습니다. 제안된 프레임워크를 계층적 다항식 융합 네트워크(HPFN)라고 합니다.

그림 3은 3계층 HPFN의 인스턴스를 보여줍니다. 첫 번째 숨겨진 레이어에서 각 PTP는 2개의 시간 단계와 2개의 모달리티로 구성된 '창'에서 로컬 상호 작용을 모델링하려고 시도합니다. 예를 들어, 시간 T1과 T2에 걸친 오디오 및 비디오 특징은 시간 T2의 히든 노드 H1-1로 병합되고, 마찬가지로 시간 T2의 히든 노드 H1-3은 T1과 T2의 오디오 및 텍스트 특징을 융합하여 출력됩니다. 두 번째

숨겨진 레이어에는 이전 레이어의 중간 특징이 공급됩니다. 출력 레이어에서는 T4와 T8 시점의 두 번째 숨겨진 레이어에 있는 3개 모달리티의 중간 특징에 PTP를 적용하여 최종 특징을 얻습니다.

HPFN의 유연성 덕분에 아키텍처 설계를 위한 다양한 선택이 가능합니다. 원칙적으로 중간 레이어를 더 추가하면 훨씬 더 큰 유효 수용 '창' 내에서 더 복잡하고 고차원적인 상호 작용이 이루어집니다. '창'을 중첩하여 더 복잡한 상호 작용을 모델링할 수도 있습니다. 그림 2 (b)는 [4 3]의 융합 '윈도우'가 시간 차원을 따라 보폭 2의 크기로 겹쳐진 2계층 HPFN의 아키텍처를 보여줍니다. PTP를 컨볼루션 필터에 비유하면 더 많은 변형을 구현할 수 있습니다. CNN과 마찬가지로 PTP 연산자는 '융합 필터'로 볼 수 있습니다. 이러한 방식으로 HPFN은 일반 CNN의 아키텍처에서 몇 가지 유사한 이점을 차용할 수도 있습니다. 보다 정확하게는, 스캐닝 '창'이 시간 차원을 따라 미끄러질 때 각 계층에서 PTP '융합 필터'를 공유하여 시계열에서 반복되는 중요한 상관관계 패턴을 포착할 수 있습니다. 또한, 여러 개의 PTP '융합 필터'를 하나의 '창'에 동시에 연결하면 해당 '창'에 존재하는 여러 상관관계 패턴을 포착할 수 있습니다.

고밀도로 연결된 네트워크(DenseNets)[11]의 경험적 성공은 HPFN 아키텍처를 확장하는 데 또 다른 영감을 줍니다. 고밀도 연결의 통합은 표현 능력을 향상시킵니다.

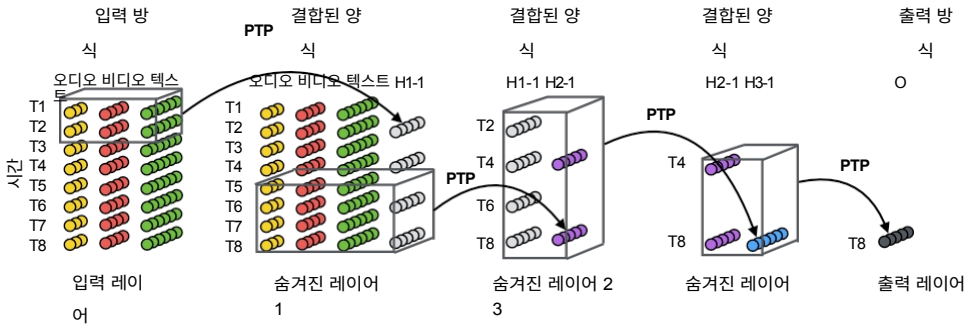


그림 4: 성장률 $k = 1$로 조밀하게 연결된 4계층 HPFN의 예입니다.

의 융합 모델입니다. 조밀한 상호 연결을 추가하면 순차적 신호를 처리하는 데 유용할 수 있습니다. 특히, 이전 레이어의 특징을 현재 레이어에 직접 포함시킴으로써 고밀도 연결을 실현할 수 있습니다. 연결에 관련된 이전 레이어의 수 k N은 성장률로 정의됩니다. 그림 4는 성장률 $k = 1$인 고밀도 HPFN의 인스턴스를 보여줍니다.

3.3 컨볼루션 산술 회로에 대한 연결

방정식 (5)는 PTP가 실제로 컨볼루션, 풀링 및 선형 변환의 결합된 연산을 수행한다는 것을 시사한다는 점이 흥미롭습니다. 이는 CNN의 특수한 변형으로 볼 수 있는 컨볼루션 연산 회로(ConvAC)[5]와 매우 유사합니다. 정류기 활성화 및 평균/최대 풀링 대신, ConAC에는 선형 활성화 및 제품 풀링 레이어가 장착되어 있습니다. 5]의 저자들은 계층적 터커 분해(HTD)[9]와의 동등성을 도출하여 심층 ConAC의 표현력을 분석합니다. 심층 ConvAC는 일반 정류기 기반 CNN보다 표현력이 뛰어나다는 것이 입증되었습니다[5]. 실제로 단일 PTP 블록은 CP 형식을 사용하는 경우 얕은 ConvAC에 해당하며, 풀링 가중 텐서에 HTD를 채택하는 경우 깊은 ConAC에 해당합니다. ConAC와 PTP의 가장 큰 차이점은 표준 ConAC의 제품 풀링은 피처의 위치에 대해 수행되는 반면, PTP의 제품 풀링은 연결된 피처의 다항식 순서에 대해 수행된다는 점입니다. PTP 블록을 여러 계층으로 쌓는 것은 본질적으로 여러 개의 HTD를 재귀적인 방식으로 사용하는 것과 같으며, 그 결과 HPFN이 훨씬 더 심층적인 ConAC에 대응하게 됩니다. 그 결과, 보다 유연한 고차적인 로컬 및 글로벌 상호 연관성을 명시적, 암시적으로 포착할 수 있으며, 이는 매우 심층적인 ConAC에 HPFN을 연결함으로써 뛰어난 표현력을 내포할 수 있습니다.

3.4 모델 복잡성

이 섹션에서는 HPFN의 모델 복잡도를 다른 두 가지 텐서 기반 모델과 비교합니다: TFN [27] 및 LMF [17]. 특징 텐서의 대칭 속성을 활용하는 PTP의 경우, 가중 텐서의 파라미터 수는 차수 P 와 무관하며 '윈도우'의 연결된 혼합 특징에 따라 선형적으로 확장됩니다. L *계층* HPFN의 경우, 파라미터의 양은 다음과 선형적으로 관련됩니다.
총 PTP '윈도우' 수 $\sum_{l=1}^{L} N_l$, 여기서 N_l 은 레이어 $l \in [L]$에 있는 '윈도우'의 수입니다. In
실제로 N_l 은 일반적으로 작고 계층을 따라 감소합니다(예: $N_1 > N_2 > \cdots > N_L$). 시간 차원에 따른 공유 전략을 채택하면 N_l 은 더욱 작아집니다. 원칙적으로 표 1에서 볼 수 있듯이 HPFN의 매개변수는 LMF보다 크거나 비슷하지만 TFN보다는 훨씬 작습니다.

표 1: TFN, LMF 및 HPFN의 모델 복잡성 비교. I_y는 출력 피처 길이입니다. M은 모달리티의 수입니다. R은 텐서 랭크입니다. PTP와 HPFN의 경우, $[T, S]$는 로컬 '창' 크기이며 $S \leq M$ 입니다. $I_{t,m}$은 시간 t에서 모달리티 m의 특징 차원입니다.

모델	TFN [비시간적]	LMF [비시간적]	PTP [시간적]	HPFN (L 레이어) [시간적]
Param.	$O(I_y \prod_{m=1}^{M} I_m)$	$O(I_y R(\sum_{m=1}^{M} I_m))$	$O(I_y R(\sum_{t=1}^{T} \sum_{m=1}^{S} I_{t,m}))$	$O(I_y R(\sum_{l=1}^{L} N_l)(\sum_{t=1}^{T} \sum_{m=1}^{S} I_{t,m}))$

4 관련 작업

멀티모달 융합 연구에는 크게 두 가지 라인이 존재합니다. **비시간적 모델은** 시간적 차원을 따라 특징을 평균화하여 각 단일모달의 관측치를 요약합니다. 이러한 모델은 멀티모달 감성 분석의 초기 연구에서 그 유용성을 발견했습니다[18, 31]. 최근에는 텐서 융합 네트워크(TFN)[27]가 텐서 곱을 활용하여 비시간적 유니모달, 바이모달, 멀티모달을 모델링합니다.

표 2: 비시간적 버전의 HPFN을 위한 네트워크 아키텍처 사양. [-]는 구성을 나타냅니다.
를 나타냅니다. PTP^k 는 레이어 'k'에서 'm'번째 융합 피처 노드를 나타냅니다.

모델계층별 구성에 대한 설명	
HPFN	$[PTP^o_1(a, v, l)]$
HPFN-L2	$[PTP^{h1}_1(a, v), PTP^{h1}_2(v, l), PTP^{h1}_3(a, l)] - [PTP^o_1(PTP^{h1}_1, PTP^{h1}_2, PTP^{h1}_3)]]$
HPFN-L2-S1	$[PTP^{h1}_1(a, v, l)] - [PTP^o_1(PTP^{h1}_1, a, v, l)]$
HPFN-L2-S2	$[PTP^{h1}_1(a, v), PTP^{h1}_2(v, l), PTP^{h1}_3(a, l)] - [PTP^o_1(PTP^{h1}_1, PTP^{h1}_2, PTP^{h1}_3, a, v, l)]$
HPFN-L3	$[PTP^{h1}_1(a, v), PTP^{h1}_2(v, l), PTP^{h1}_3(a, l)] --$ $[PTPh2_1(PTPh1_1, PTPh1_2)[h1_1(PTP1, PTP3), h1_2(PTP2, PTP3)] - [PTP1(PTP1, PTP2), o_{h2}(PTP3)]]$
HPFN-L4	$[PTP^{h1}_1(a, v), PTP^{h1}_2(v, l), PTP^{h1}_3(a, l)] --$ $[PTP^{h2}_1(PTP^{h1}_1, PTP^{h1}_2), PTP^{h2}_2(PTP^{h1}_1, PTP^{h1}_3), PTP^{h2}_3(PTP^{h1}_2, PTP^{h1}_3)] --$ $[PTP^{h3}_1(PTP^{h2}_1, PTP^{h2}_2), PTP^{h3}_2(PTP^{h2h2}_1, PTP^{h3}_3), PTP^{h3}_3(PTP^{h2}_2, PTP^{h2}_3)] - [PTP^o_1(PTP^{h3}_1, PTP^{h3}_2, PTP^{h3}_3)]$

모달리티 간의 트라이모달 상호 작용. 차원성의 저주 문제를 처리하기 위해 저순위 다중 모드 융합 네트워크(LMF)[17]는 모달리티별 저순위 인자를 사용하여 비시간적 융합의 확장성을 더욱 향상시킵니다. 이러한 모든 접근 방식은 특징의 평균 통계를 사용하여 시간적 정보를 사용하지 않고 상관관계를 한 번에 식별하려고 시도합니다. 단순하지만 시간 순서에 따라 진화하는 모달 내 및 모달 간 역학을 학습할 수 없으므로 예측 정확도가 떨어집니다.

반면에 멀티모달 시간 모델은 시간 차원을 따라 훨씬 더 세밀한 단위로 멀티모달 상호 작용을 처리합니다. 순차적 멀티모달 설정에는 장단기 메모리(LSTM)[10]가 광범위하게 사용되어 왔습니다. 그 중 멀티뷰 LSTM(MV-LSTM)[24]은 특정 모달리티에 해당하는 메모리 셀을 분할하여 뷰별 및 뷰 간 상호작용을 모두 포착하고, 양방향 컨텍스트 LSTM(BC-LSTM)[23]은 멀티모달 시계열로 맥락 의존적 감성 분석 및 감정 인식을 수행하기 위해 제안되었습니다. 메모리 융합 네트워크(MFN)[28]는 다중 뷰 게이트 메모리를 사용하여 시간 영역에 따라 모달 간 및 모달 내 상호 작용을 저장하고, 다중 주의 반복 네트워크(MARN)[29]는 다중 주의 블록을 사용하여 주의 계수를 사용하여 모달 간 역학을 발견합니다. 최근에는 반복적 다단계 융합 네트워크(RMFN)[14]가 융합을 여러 단계로 분해하고, 각 단계는 융합 결과가 이전 단계의 중간 표현을 기반으로 구축되는 신호의 하위 집합에 초점을 맞춥니다. 그러나 텐서 기반 다중 모드 융합에 비해 위의 모든 접근 방식은 선형 상호 작용만 모델링하도록 제한되어 있어 복잡한 다중 모드 상관 관계를 식별할 수 없습니다.

5 실험

5.1 실험 설정

데이터 세트. CMU-MOSI 데이터 세트 [30]는 YouTube 영화 리뷰의 2,199개 의견 비디오 클립으로 구성됩니다. 각 클립에는 높은 부정에서 높은 긍정까지 [3, 3] 범위의 지정된 감정이 할당되어 있습니다. 훈련 세트에는 1, 284개의 세그먼트가, 검증 세트에는 229개의 세그먼트가, 테스트 세트에는 686개의 세그먼트가 있습니다. IEMOCAP 데이터 세트 [2]에는 총 302개의 비디오가 포함되어 있습니다. 비디오의 세그먼트에는 이산 감정(중립, 공포, 행복, 화, 실망, 슬픔, 좌절, 흥분, 놀람)과 우세, 가치, 각성 등의 주석이 달렸습니다. 훈련, 검증 및 테스트 세트의 분할은 각각 6, 373, 1, 775 및 1, 807입니다. 두 데이터 세트의 분할은 화자 독립적이므로 지정된 화자는 세 세트 중 하나에만 속할 수 있습니다.

특징. IEMOCAP의 경우, 시간 차원을 평균화하여 음향 및 시각적 특징을 얻는 LMF [17]의 작업에 따라 사전 처리된 비시간적 입력을 채택합니다. CMU-MOSI의 경우: 시간적 특징은 텍스트 모달리티에 따라 세 가지 모달리티의 추출된 특징이 단어 수준에서 동기화되는 MFN [28]과 동일한 방식으로 활용됩니다.

비교. 메모리 융합 네트워크(MFN) [28], 다중 주의 반복 네트워크(MARN) [29], 텐서 융합 네트워크(TFN) [27], 저순위 다중 모드 융합 네트워크(LMF) [17]와 같은 최신 텐서 및 비텐서 기반 모델을 HPFN과의 비교에 포함시켰으며, 기타 기준선도 일부 포함시켰습니다. 평균 절대 오차(MAE)와 피어슨 상관관계, 정확도 및 F1 측정값을 보고합니다. HPFN의 경우, 최적의 설정을 위해 평가를 5회 반복합니다.

모델 아키텍처. 실험에 채택된 HPFN 아키텍처는 표 2에 설명되어 있으며, 여기에는 고밀도로 연결된 2계층 `HPFN-L2-S1` 및 `HPFN-L2-S2` 변형이 포함되어 있습니다.

표 3: 표 3: CMU-MOSI의 감정 분석 결과와 IEMOCAP의 감정 인식 결과.

모델	CMU-MOSI					IEMOCAP			
	MAE	Corr	Acc-2	F1	Acc-7	F1-Happy	F1-Sad	F1-Angry	F1-중립
SVM [6]	1.864	0.057	50.2	50.1	17.5	81.5	78.8	82.4	64.9
DF [20]	1.143	0.518	72.3	72.1	26.8	81.0	81.2	65.4	44.0
BC-LSTM [23]	1.079	0.581	73.9	73.9	28.7	81.7	81.7	84.2	64.1
MV-LSTM [24]	1.019	0.601	73.9	74.0	33.2	81.3	74.0	84.3	66.7
MARN [29]	0.968	0.625	77.1	77.0	34.7	83.6	81.2	84.2	65.9
MFN [28]	0.965	0.632	77.4	77.3	34.1	84.0	82.1	83.7	69.2
TFN [27]	0.970	0.633	73.9	73.4	32.1	83.6	82.8	84.2	65.4
LMF [17]	**0.912**	0.668	76.4	75.7	32.8	85.8	85.9	**89.0**	71.7
HPFN, P=[4] (오디오)	1.404	0.223	57.3	57.4	19.0	79.4	81.8	84.9	63.6
HPFN, P=[4] (비디오)	1.409	0.221	57.0	57.1	20.6	83.2	73.2	72.3	58.5
HPFN, P=[4] (텍스트)	0.975	0.634	76.4	76.4	35.1	85.3	83.0	85.6	70.8
HPFN, P=[4]	0.965	0.650	**77.5**	**77.4**	36.0	85.7	86.4	88.3	72.1
HPFN, P=[8]	0.968	0.648	77.2	77.2	**36.9**	85.7	86.5	87.9	71.8
HPFN-L2, P=[2, 2]	0.945	**0.672**	**77.5**	**77.4**	36.7	**86.2**	**86.6**	88.8	**72.5**

구현 세부 사항. LMF [17]에 따라, 저희는 PTP의 가중치 압축 실험에서 CP 형식을 '주력' 저순위 TN으로 사용합니다. 후보 CP 등급은 1, 4, 8, 16 입니다. 다른 TN 변형은 향후 작업에서 조사할 예정입니다. HPFN은 요소별 곱셈을 계산할 때 고차 모멘트를 포함하기 때문에 중간 특징의 값이 크게 달라져 예측이 불안정해질 수 있습니다. 모델을 수치적으로 더 안정적으로 만들기 위해 [8]과 유사하게 파워 정규화(요소별 부호화된 제곱근) 또는 l_2 정규화를 선택적으로 적용할 수 있습니다.

5.2 실험 결과

최첨단 모델과의 성능 비교. 먼저 감정 분석 및 감정 인식 작업에서 기준선과 최첨단 모델을 비교했습니다. 표 3의 맨 아래 줄에는 우리 모델의 성능이 기록되어 있습니다. 멀티모달 데이터에 대한 우리의 모델이 대부분의 지표에서 경쟁사보다 우수한 성능을 보였습니다. 특히 감정 과제에서 8차수의 HPFN은 'Acc-7'의 이전 최고 MARN을 2.2% 차이로 능가합니다. 전체적으로 가장 우수한 결과는 HPFN-L2가 달성했는데, 이는 계층적 융합 구조의 표현력과 효율성이 우수하다는 것을 의미합니다. 또한 유니모달 입력(텍스트)이 주어졌을 때에도 4차 HPFN이 거의 모든 다른 방법보다 훨씬 우수한 'Acc-7'(35.1) 및 'F1-Neutral'(70.8) 정밀도를 얻어 고차 상호작용 모델링이 가져다주는 이점을 보여준다는 점도 흥미로웠습니다.

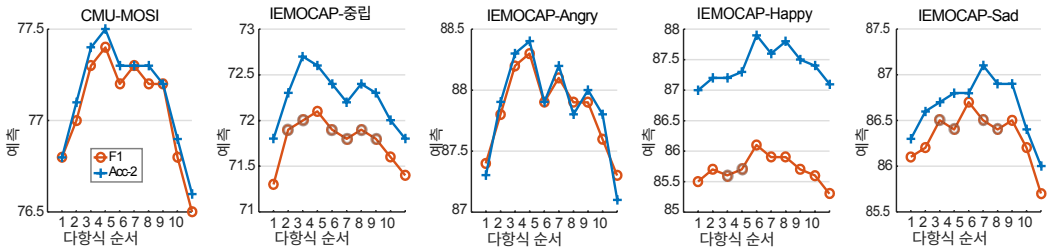


그림 5: 다항식 상호작용의 순서가 IEMOCAP 및 CMU-MOSI에 미치는 영향의 결과.

다항식 융합 순서의 효과. 퓨전 전략에서 고차 모멘트가 중요한 역할을 하므로, 우리는 서로 다른 차수가 예측 성능에 어떤 영향을 미치는지 살펴보고자 합니다. 간단하게 하기 위해 시간 차원을 평

균화하여 비시간적 멀티모달 특징에 전력 정규화가 포함된 HPFN을 직접 적용합니다. 차수 P는 1에서 10까지 다양합니다. 그림 5에서 HPFN은 테스트된 주문에 대해 상당히 우수한 정확도를 달성할 수 있습니다. 특히 CMU-MOSI의 경우 4차수에서 HPFN이 예측을 극대화하는 것을 볼 수 있습니다. IEMOCAP의 경우, '중립' 및 '화난' 감정에서 3번과 4번 순서에서 상대적으로 높은 성능 피크를 관찰할 수 있습니다. 나머지 감정의 경우, 바람직한 순서는 5에서 8까지입니다. 이러한 관찰은 멀티모달 기능을 융합할 때 고차 상호작용을 탐색하는 것이 필요하고 효과적이라는 것을 의미합니다
.

깊이와 고밀도 연결의 효과. 이 부분에서는 다양한 아키텍처 설계, 즉 깊이와 고밀도 연결이 예측 성능에 미치는 영향을 조사합니다. 깊이의 변화에 초점을 맞추기 위해 비시간적 멀티모달 특징에 아키텍처를 적용합니다. 심도

표 4: 비시간적 멀티모달 피처의 깊이 및 밀도 연결에 따른 HPFN 결과.

모델	IEMOCAP				CMU-MOSI				
	F1-Happy	F1-Sad	F1-Angry	F1-중립	MAE	Corr	Acc-2	F1	Acc-7
HPFN, P=[2]	85.7	86.2	87.8	71.9	0.973	0.635	77.1	77.0	35.9
HPFN-L2, P=[2, 2]	86.2	86.6	88.8	72.5	0.958	0.652	77.1	77.1	36.3
HPFN-L2-S1, P=[2, 2]	86.2	86.7	88.9	72.6	0.959	0.654	**77.3**	77.2	**36.5**
HPFN-L2-S2, P=[2, 2]	**86.2**	86.7	**89.0**	**72.7**	**0.957**	**0.656**	**77.3**	**77.3**	**36.5**
HPFN-L3, P=[2, 2, 1]	86.1	**86.8**	88.3	**72.7**	0.960	0.651	76.8	76.8	36.0
HPFN-L4, P=[2, 2, 2, 1]	85.8	86.4	88.1	72.5	0.992	0.634	76.6	76.5	34.6

표 5: 국지적으로 혼합된 시간-양식 특징의 모델링 결과.

모델	CMU-MOSI				
	MAE	Corr	Acc-2	F1	Acc-7
HPFN-L2, P=[2, 2](비시간적)	0.958	0.652	77.1	77.1	36.3
HPFN-L2, P=[2, 2](템포럴 오버랩, 오디오)	1.407	0.229	57.4	56.2	20.1
HPFN-L2, P=[2, 2](템포럴 오버랩, 비디오)	1.358	0.183	61.2	61.3	20.3
HPFN-L2, P=[2, 2] (템포럴 오버랩, 텍스트)	**0.933**	0.677	76.7	76.6	35.4
HPFN-L2, P=[2, 2] (템포럴 오버랩)	0.944	**0.678**	**77.5**	**77.4**	**36.7**
HPFN-L2, P=[2, 2](무게 공유)	0.955	0.667	77.0	76.9	35.7

변종에 대해 HPFN, HPFN-L2, HPFN-L3 및 HPFN-L4에서 검증합니다. 또한 밀접하게 연결된 두 가지 변종과도 비교합니다: HPFN-L2-S1과 HPFN-L2-S2입니다. 표 4에서는 2계층 및 3계층 기반 아키텍처가 1계층 및 4계층 아키텍처보다 전반적으로 더 나은 결과를 달성하는 것을 확인할 수 있습니다. 특히 HPFN-L2-S2는 두 데이터 세트 모두에서 최고의 정밀도에 도달합니다. HPFN은 너무 단순하여 복잡한 상호작용을 학습하기 어렵고, 중간 노드가 너무 많은 HPFN-L4는 이 특정 아키텍처 설계에 과도하게 적합할 가능성이 높습니다. 건너뛰기 연결을 허용하면 중간 계층을 추가하지 않고도 보다 변별력 있는 유니모달 신호의 지침을 통합할 수 있기 때문에 HPFN-L2의 성능이 더욱 향상됩니다.

혼합 시간-양식 특징 모델링의 효과. 시간적-양식적 특징의 국소적 혼합을 처리할 수 있다는 것은 우리 모델의 바람직한 특성 중 하나입니다. 이 테스트에서는 시간적 영역과 양식 영역을 모두 고려하여 모델이 어떻게 작동하는지 살펴봅니다. 입력 레이어의 '창' 크기를 [4 2]로 설정하고 시간 차원에 따라 보폭 단계를 2로 설정하여 HPFN-L2를 시간적 컨텍스트에 맞게 조정합니다. 비시간적 HPFN-L2는 시간 차원을 평균화하여 모달리티 영역만 고려합니다. 표 5는 비시간적 HPFN-L2에 비해 시간적 HPFN-L2의 우월성을 나타냅니다. 또한 시간적 방향을 따라 '창'을 스캔하여 PTP를 공유하려고 시도합니다. 여러 창에 대해 단일 PTP 유닛을 공유해도 이 설정에서 추가적인 성능 이득을 얻지 못하는 것으로 나타났습니다. 그림 6은 시간 영역의 '창' 크기에 따른 예측을 표시합니다. 가중치를 공유하지 않는 경우 중간 정도의 '창'(크기 5 및 10)이 최고 성능에 도달합니다. 반면, 가중치 공유 모달은 가장 큰 창 크기에서 상대적으로 높은 성능을 보입니다.
(20). 이는 다시 한 번 단일 PTP와의 공유가 로컬에서 진화하는 상호작용의 역학을 포착하지 못할 수 있음을 의미합니다.

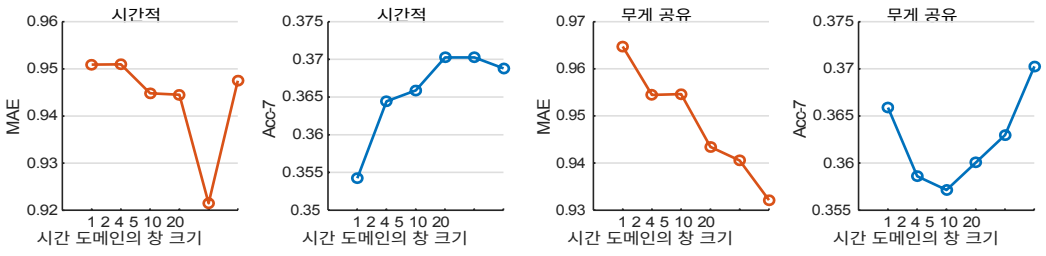


그림 6: 시간 도메인의 '창' 크기에 따른 예측 결과. 왼쪽 두 그림: 가중치 미공유 모델. 오른쪽 두 그림: 가중치 공유 모델.

6 결론

본 논문에서는 멀티모달 특징 융합을 위한 고차 다항식 다선형 풀링 블록을 제안했습니다. 이를 기반으로 다음과 같은 계층적 다항식 융합 네트워크(HPFN)를 구축했습니다.

시간 및 양식 영역 모두에서 혼합된 특징을 유연하게 융합합니다. 제안된 모델은 로컬 규모에서 글로벌 규모에 이르기까지 매우 복잡한 시간-양식 간 상관관계를 포착하는 데 효과적입니다. 실제 멀티모달 융합 작업에 대한 다양한 실험을 통해 제안된 모델의 우수한 성능을 검증했습니다. 향후 연구에서는 아키텍처 설계가 예측 성능에 어떤 영향을 미치는지 더 자세히 살펴보고자 합니다. 예를 들어, 하나의 '창'에 여러 개의 PTP 블록을 연결하고, 시간 차원에 따라 여러 개의 PTP '융합 필터'를 공유하여 보다 복잡한 상관관계 패턴을 모델링하는 것입니다.

감사

이 연구는 일본 과학기술연구개발기구의 국가핵심연구개발사업 정부 간 국제과학기술혁신협력사업(MOST-RIKEN)(보조금 번호 17K00326) 및 중국 국가자연과학재단(보조금 번호 61633010)의 일부 지원을 받아 수행되었습니다.

참조

[1] 타다스 발트루사이티스, 차이타냐 아후자, 루이-필립 모렌시. 멀티모달 머신 러닝: 설문조사 및 분류. *IEEE 트랜잭션 패턴 분석 및 기계 지능*, 41(2):423-443, 2019.

[2] 카를로스 부소, 무르타자 불루트, 이치춘, 아베 카젬자데, 에밀리 모어, 사무엘 김, 지넷 창, 이성복, 슈리칸스 나라야난. Iemocap: 대화형 정서적 조증 모션 캡처 데이터베이스. *언어 자원 및 평가*, 42(4):335, 2008.

[3] J 더글러스 캐롤과 장지지에. "에카르트-영" 분해의 n-방향 일반화를 통한 다차원 스케일링의 개인차 분석. *사이코메트리카*, 35(3):283-319, 1970.

[4] 안제이 시초키, 이남길, 이반 오셀레데츠, 안후이 판, 치빈 자오, 다닐로 만딕 등. 차원 축소 및 대규모 최적화를 위한 텐서 네트워크: 1부 저순위 텐서 분해. *기계 학습의 기초와 동향*, 9(4-5):249-429, 2016.

[5] 나다브 코헨, 오르 샤리르, 암논 샤슈아. 딥러닝의 표현력에 대해: 텐서 분석. *학습 이론 컨퍼런스*, 698-728 페이지, 2016.

[6] 코리나 코르테스와 블라디미르 바프닉. 서포트 벡터 네트워크. *기계 학습*, 20(3):273-297, 1995.

[7] 시드니 K 디멜로와 재클린 코리. 다중 모드 영향 감지 시스템에 대한 검토 및 메타 분석. *ACM 컴퓨팅 설문조사(CSUR)*, 47(3):43, 2015.

[8] 아키라 후쿠이, 박동혁, 데이렌 양, 안나 로르바흐, 트레버 대럴, 마커스 로르바흐. 시각적 질문 답변 및 시각적 접지를 위한 멀티모달 콤팩트 바이리니어 풀링. *자연어 처리의 경험적 방법에 관한 컨퍼런스 논문집*, 457-468페이지, 2016.

[9] 볼프강 핵부쉬와 스테판 쿤. 텐서 표현을 위한 새로운 체계. *푸리에 분석 및 응용 저널*, 15(5):706-722, 2009.

[10] 셉 호크라이터와 위르겐 슈미트후버. 장단기 기억. *신경 계산*, 9(8):1735-1780, 1997.

[11] 가오 황, 좡 리우, 로렌스 반 더 마텐, 킬리안 큐 와인버거. 조밀하게 연결된 컨볼루션 네트워크. *컴퓨터 비전 및 패턴 인식에 관한 IEEE 컨퍼런스 논문집*, 4700-4708페이지, 2017.

[12] 타마라 G 콜다와 브렛 W 배더. 텐서 분해와 응용. *SIAM 리뷰*, 51(3):455-500, 2009.

[13] 폴 푸 리앙, 야오 총 림, 야오-흥 허버트 차이, 루슬란 살라쿠트디노프, 루이-필립 모렌시. 다중 모달 발화 임베딩을 위한 강력하고 간단한 기준선. *컴퓨터 언어학 협회 북미 지부 컨퍼런스 논문집*, 2599-2609페이지, 2019.

[14] 폴 푸 리앙, 지인 리우, 아미르 알리 바거 자데, 루이-필립 모렌시. 반복적 다단계 융합을 통한 다중 모드 언어 분석. *자연어 처리의 경험적 방법에 관한 컨퍼런스 논문집*, 150-161페이지, 2018.

[15] 폴 푸 리앙, 아미르 자데, 루이-필립 모렌시. 감정 인식을 위한 멀티모달 로컬-글로벌 랭킹 융합. *다중 모드 상호 작용에 관한 국제 컨퍼런스 논문집*, 472-476페이지. ACM, 2018.

[16] Tsung-Yu Lin, Aruni RoyChowdhury, 수브란수 마지. 세분화된 시각 인식을 위한 이중선형 CNN 모델. *IEEE 컴퓨터 비전 국제 컨퍼런스 논문집*, 1449-1457페이지, 2015.

[17] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh 및 Louis-Philippe Morency. 모달리티별 요인을 사용한 효율적인 저순위 멀티모달 융합. *전산 언어학 협회 연례 회의 논문집*, 2247-2256페이지, 2018.

[18] 루이 필립 모렌시, 라다 미할시아, 페이랄 도시. 멀티모달 감성 분석을 향하여: 웹에서 의견 수집하기. *다중 모드 인터페이스에 관한 국제 컨퍼런스 논문집*, 169-176쪽. ACM, 2011.

[19] 에밀리 모방, 아모리 하브라드, 스테판 아야체. 후기 융합을 위한 다양한 분류기의 다수결 투표. *패턴 인식(SPR) 및 구조 및 구문 패턴 인식(SSPR)의 통계적 기법에 관한 IAPR 국제 공동 워크숍*, 153-162페이지. Springer, 2014.

[20] 베나즈 노자바나가리, 디팍 고피나스, 자얀트 쿠식, 타다스 발트루사이티스, 루이-필립 모렌시. 설득력 예측을 위한 심층 멀티모달 융합. *ACM 국제 멀티모달 상호 작용 컨퍼런스 논문집*, 284-288페이지. ACM, 2016.

[21] 이반 V 오셀레데스. 텐서-트레인 분해. *SIAM 과학 컴퓨팅 저널*, 33(5):2295-2317, 2011.

[22] 박성현, 심한석, 모이트레야 차터지, 켄지 사가에, 루이-필립 모렌시. 소셜 멀티미디어의 설득력에 대한 컴퍼지셔널 분석: 새로운 데이터 세트와 멀티모달 예측 접근법. *멀티모달 상호작용 국제 컨퍼런스 논문집*, 50-57쪽. ACM, 2014.

[23] 수잔야 포리아, 에릭 캠브리아, 데바마니유 하자리카, 나보닐 마줌더, 아미르 자데, 루이-필립 모렌시. 사용자 제작 동영상의 컨텍스트 의존적 감정 분석. *컴퓨터 언어학 협회 연례 회의 논문집*, 873-883페이지, 2017.

[24] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, Roland Goecke. 멀티뷰 구조화 학습을 위한 장단기 메모리 확장. *유럽 컴퓨터 비전 컨퍼런스*, 338-353쪽. Springer, 2016.

[25] 예카테리나 슈토바, 두베 키엘라, 장 마이야르. 블랙홀과 흰 토끼: 시각적 특징을 이용한 은유 식별. *컴퓨터 언어학 협회 북미 지부 컨퍼런스 논문집: 인간 언어 기술*, 160-170페이지, 2016.

[26] 레디야드 R 터커. 3 모드 요인 분석에 대한 몇 가지 수학적 메모. *사이코메트리카*, 31(3):279-311, 1966.

[27] 아미르 자데, 밍하이 첸, 수잔야 포리아, 에릭 캠브리아, 루이-필립 모렌시. 다중 모드 감정 분석을 위한 텐서 융합 네트워크. *자연어 처리의 경험적 방법에 관한 컨퍼런스 논문집*, 1103-1114페이지, 2017.

[28] 아미르 자데, 폴 푸 리앙, 나보닐 마줌더, 수잔야 포리아, 에릭 캄브리아, 루이-필립 모렌시. 멀티뷰 순차 학습을 위한 메모리 융합 네트워크. *인공지능에 관한 AAAI 컨퍼런스*, 2018.

[29] 아미르 자데, 폴 푸 리앙, 수잔야 포리아, 프라텍 비지, 에릭 캄브리아, 루이 필립 모렌시. 인간의 의사소통 이해를 위한 다중 주의 반복 네트워크. 2018 *인공 지능에 관한 AAAI 컨퍼런스에서*.

[30] 아미르 자데, 로완 젤러스, 엘리 핀커스, 루이-필립 모렌시. Mosi: 온라인 오피니언 비디오의 감정 강도 및 주관성 분석에 대한 다중 모드 말뭉치. *arXiv 사전 인쇄물 arXiv:1606.06259*, 2016.

[31] 아미르 자데, 로완 젤러스, 엘리 핀커스, 루이-필립 모렌시. 비디오의 멀티모달 감정 강도 분석: 얼굴 제스처와 언어 메시지. *IEEE 지능형 시스템*, 31(6):82-88, 2016.

[32] 치빈 자오, 마사시 스기야마, 롱하오 위안, 안제이 치초키. 링 구조 네트워크를 사용한 효율적인 텐서 표현 학습. *IEEE 국제 음향, 음성 및 신호 컨퍼런스 처리*, 8608-8612페이지. IEEE, 2019.