

BiProjection Multimodal Transformer for Multimodal Classification Problems

Diego Aarón Moreno Galván

Asesor: Dr. Adrián Pastor López Monroy, (CIMAT)
Coasesor: Dr. Luis Carlos González Gurrola, (UACH)

Centro de Investigación en Matemáticas
Departamento de Ciencias de la Computación

xx de noviembre de 2022



1 Introducción

2 BiProjection Multimodal Transformer (BPMuLT)

3 Resultados

4 Conclusiones

5 Referencias

1 Introducción

Panorama General

Conjunto de Datos y SotA

Estudios de Problemáticas

2 BiProjection Multimodal Transformer (BPMuLT)

3 Resultados

4 Conclusiones

5 Referencias

1 Introducción

Panorama General

Conjunto de Datos y SotA

Estudios de Problemáticas

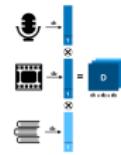
② BiProjection Multimodal Transformer (BPMuLT)

3 Resultados

4 Conclusiones

5 Referencias

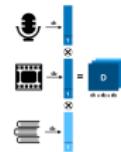
Introducción



- Información multimodal: texto, imágenes y audio



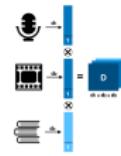
Introducción



- Información multimodal: texto, imágenes y audio
 - Ejemplos de tareas: Detección de sentimientos, clasificación de películas por rating y por género



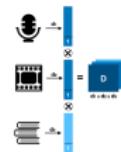
Introducción



- Información multimodal: texto, imágenes y audio
 - Ejemplos de tareas: Detección de sentimientos, clasificación de películas por rating y por **género**
 - Importancia: Recomendaciones, contenido apropiado.



Introducción



- Información multimodal: texto, imágenes y audio
 - Ejemplos de tareas: Detección de sentimientos, clasificación de películas por rating y por género
 - Importancia: Recomendaciones, contenido apropiado.
 - Objetivo: Diseñar una estrategia para mejorar la representación entre las modalidades, que ayude en problemas de clasificación supervisada y que sea competitiva o superior al estado del arte.



1 Introducción

Panorama General

Conjunto de Datos y SotA

Estudios de Problemáticas

② BiProjection Multimodal Transformer (BPMuLT)

3 Resultados

4 Conclusiones

5 Referencias

Dataset principal: Moviescope

- Detección de género de películas

Two clownfish, Marlin and his wife Coral are admiring their new home in the New Caledonia Barrier Reef and their clutch of eggs that are due to hatch in a few days. Suddenly, a barracuda attacks them, leaving Marlin unconscious before eating Coral and all but one of their eggs. Marlin names this egg Nemo, a name that Coral liked. The (...)

T: family: 0.81 | fantasy: 0.80 | sci-fi: 0.16



Figura 1: Ejemplo de datos moviescope.

Dataset principal: Moviescope

- Detección de género de películas
 - 13 géneros: action, animation, biography, comedy, crime, drama, family, fantasy, horror, mystery, romance, sci-fi y thriller

Two clownfish, Marlin and his wife Coral are admiring their new home in the New Caledonia Barrier Reef and their clutch of eggs that are due to hatch in a few days. Suddenly, a barracuda attacks them, leaving Marlin unconscious before eating Coral and all but one of their eggs. Marlin names this egg Nemo, a name that Coral liked. The (...)

T: family: 0.81 | fantasy: 0.80 | sci-fi: 0.16



Figura 1: Ejemplo de datos moviescope.

Dataset principal: Moviescope

- Detección de género de películas
- 13 géneros: action, animation, biography, comedy, crime, drama, family, fantasy, horror, mystery, romance, sci-fi y thriller
- Alrededor de 5000 películas

Two clownfish, Marlin and his wife Coral are admiring their new home in the New Caledonia Barrier Reef and their clutch of eggs that are due to hatch in a few days. Suddenly, a barracuda attacks them, leaving Marlin unconscious before eating Coral and all but one of their eggs. Marlin names this egg Nemo, a name that Coral liked. The (...)

T: family: 0.81 | fantasy: 0.80 | sci-fi: 0.16



Figura 1: Ejemplo de datos moviescope

Dataset principal: Moviescope

- Detección de género de películas
- 13 géneros: action, animation, biography, comedy, crime, drama, family, fantasy, horror, mystery, romance, sci-fi y thriller
- Alrededor de 5000 películas
- Usa de tráiler (audio, texto y video), póster y metadatos (año, presupuesto, actores, etc.)

Two clownfish, Marlin and his wife Coral are admiring their new home in the New Caledonia Barrier Reef and their clutch of eggs that are due to hatch in a few days. Suddenly, a barracuda attacks them, leaving Marlin unconscious before eating Coral and all but one of their eggs. Marlin names this egg Nemo, a name that Coral liked. The (...)

T: family: 0.81 | fantasy: 0.80 | sci-fi: 0.16



Figura 1: Ejemplo de datos moviescope.

Dataset principal: Moviescope

- Detección de género de películas
- 13 géneros: action, animation, biography, comedy, crime, drama, family, fantasy, horror, mystery, romance, sci-fi y thriller
- Alrededor de 5000 películas
- Usa de tráiler (audio, texto y video), póster y metadatos (año, presupuesto, actores, etc.)
- Video: 200 vectores correspondiente 200 frames (1 cada 10), preprocesados por una VGG16 (al igual que póster)

Two clownfish, Marlin and his wife Coral are admiring their new home in the New Caledonia Barrier Reef and their clutch of eggs that are due to hatch in a few days. Suddenly, a barracuda attacks them, leaving Marlin unconscious before eating Coral and all but one of their eggs. Marlin names this egg Nemo, a name that Coral liked. The (...)

T: family: 0.81 | fantasy: 0.80 | sci-fi: 0.16



Figura 1: Ejemplo de datos moviescope.



Dataset principal: Moviescope

- Detección de género de películas
- 13 géneros: action, animation, biography, comedy, crime, drama, family, fantasy, horror, mystery, romance, sci-fi y thriller
- Alrededor de 5000 películas
- Usa de tráiler (audio, texto y video), póster y metadatos (año, presupuesto, actores, etc.)
- Video: 200 vectores correspondiente 200 frames (1 cada 10), preprocesados por una VGG16 (al igual que póster)
- Audio: 200 vectores parte del spectrograma log-mel de potencia

Two downfish, Marlin and his wife Coral are admiring their new home in the New Caledonia Barrier Reef and their clutch of eggs that are due to hatch in a few days. Suddenly, a barracuda attacks them, leaving Marlin unconscious before eating Coral and all but one of their eggs. Marlin names this egg Nemo, a name that Coral liked. The (...)

T: family: 0.81 | fantasy: 0.80 | sci-fi: 0.16



Figura 1: Ejemplo de datos moviescope.



Modelo MuLT-GMU Late Fusion

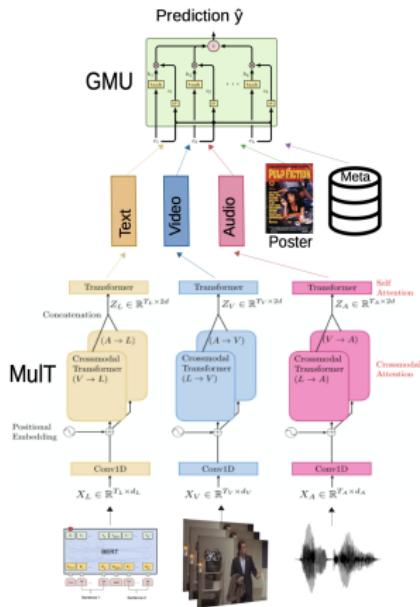


Figura 2: Modelo **MuLT-GMU** Late Fusion de Rodríguez I., et. al. (2021), para datos de detección de géneros de películas. Se usa GMU de Arévalo J. et. al. (2017).

Dataset experimental: MM-IMDb

- Detección de género de películas

MM-IMDb

Genre	Train	Develop	Test	Genre	Train	Develop	Test
Drama	8424	1401	4142	Family	978	172	518
Comedy	5108	873	2611	Biography	788	144	411
Romance	3226	548	1590	War	806	128	401
Thriller	3113	512	1567	History	680	118	345
Crime	2293	382	1163	Music	634	100	311
Action	2155	351	1044	Animation	586	105	306
Adventure	1611	278	821	Musical	503	85	253
Horror	1603	275	825	Western	423	72	210
Documentary	1234	219	629	Sport	379	64	191
Mystery	1231	209	617	Short	281	48	142
Sci-Fi	1212	193	586	Film-Noir	202	34	102
Fantasy	1162	186	585				

Dataset experimental: MM-IMDb

- Detección de género de películas
- 25,959 películas con sus plots, pósters, géneros y metadatos

MM-IMDb

Genre	Train	Develop	Test	Genre	Train	Develop	Test
Drama	8424	1401	4142	Family	978	172	518
Comedy	5108	873	2611	Biography	788	144	411
Romance	3226	548	1590	War	806	128	401
Thriller	3113	512	1567	History	680	118	345
Crime	2293	382	1163	Music	634	100	311
Action	2155	351	1044	Animation	586	105	306
Adventure	1611	278	821	Musical	503	85	253
Horror	1603	275	825	Western	423	72	210
Documentary	1234	219	629	Sport	379	64	191
Mystery	1231	209	617	Short	281	48	142
Sci-Fi	1212	193	586	Film-Noir	202	34	102
Fantasy	1162	186	585				

Dataset experimental: MM-IMDb

- Detección de género de películas
- 25,959 películas con sus plots, pósters, géneros y metadatos
- 23 géneros

MM-IMDb

Genre	Train	Develop	Test	Genre	Train	Develop	Test
Drama	8424	1401	4142	Family	978	172	518
Comedy	5108	873	2611	Biography	788	144	411
Romance	3226	548	1590	War	806	128	401
Thriller	3113	512	1567	History	680	118	345
Crime	2293	382	1163	Music	634	100	311
Action	2155	351	1044	Animation	586	105	306
Adventure	1611	278	821	Musical	503	85	253
Horror	1603	275	825	Western	423	72	210
Documentary	1234	219	629	Sport	379	64	191
Mystery	1231	209	617	Short	281	48	142
Sci-Fi	1212	193	586	Film-Noir	202	34	102
Fantasy	1162	186	585				

Dataset experimental: MM-IMDb

- Detección de género de películas
- 25,959 películas con sus plots, pósters, géneros y metadatos
- 23 géneros
- 4096 poster features preprocesados por una ResNet

MM-IMDb

Genre	Train	Develop	Test	Genre	Train	Develop	Test
Drama	8424	1401	4142	Family	978	172	518
Comedy	5108	873	2611	Biography	788	144	411
Romance	3226	548	1590	War	806	128	401
Thriller	3113	512	1567	History	680	118	345
Crime	2293	382	1163	Music	634	100	311
Action	2155	351	1044	Animation	586	105	306
Adventure	1611	278	821	Musical	503	85	253
Horror	1603	275	825	Western	423	72	210
Documentary	1234	219	629	Sport	379	64	191
Mystery	1231	209	617	Short	281	48	142
Sci-Fi	1212	193	586	Film-Noir	202	34	102
Fantasy	1162	186	585				

Dataset experimental: IEMOCAP

- Detección de emociones

IEMOCAP

Emotion	Train	Develop	Test
Neutral	954	358	383
Happy	338	116	135
Sad	690	188	193
Angry	735	136	227

Dataset experimental: IEMOCAP

- Detección de emociones
- Emociones (4): felicidad, enojo, tristeza y neutral

IEMOCAP

Emotion	Train	Develop	Test
Neutral	954	358	383
Happy	338	116	135
Sad	690	188	193
Angry	735	136	227

Dataset experimental: IEMOCAP

- Detección de emociones
- Emociones (4): felicidad, enojo, tristeza y neutral
- 12 horas de vídeo

IEMOCAP

Emotion	Train	Develop	Test
Neutral	954	358	383
Happy	338	116	135
Sad	690	188	193
Angry	735	136	227

Dataset experimental: IEMOCAP

- Detección de emociones
- Emociones (4): felicidad, enojo, tristeza y neutral
- 12 horas de vídeo
- Monólogos leídos y dichos por actores en la universidad de CMU

IEMOCAP

Emotion	Train	Develop	Test
Neutral	954	358	383
Happy	338	116	135
Sad	690	188	193
Angry	735	136	227

Dataset experimental: IEMOCAP

- Detección de emociones
- Emociones (4): felicidad, enojo, tristeza y neutral
- 12 horas de vídeo
- Monólogos leídos y dichos por actores en la universidad de CMU
- Se capturan grabaciones, transcripciones y características de los gestos faciales

IEMOCAP

Emotion	Train	Develop	Test
Neutral	954	358	383
Happy	338	116	135
Sad	690	188	193
Angry	735	136	227

Dataset experimental: CMU-MOSEI

- Detección de emociones e intensidad de ellas

CMU-MOSEI Unaligned

Emotion	Train	Develop	Test
Anger	3267	318	1015
Disgust	2738	273	744
Fear	1263	169	371
Happiness	7587	945	2220
Sadness	4026	509	1066
Surprise	1465	197	393

Dataset experimental: CMU-MOSEI

- Detección de emociones e intensidad de ellas
- Emociones (6): felicidad, enojo, tristeza, disgusto, miedo y sorpresa

CMU-MOSEI Unaligned

Emotion	Train	Develop	Test
Anger	3267	318	1015
Disgust	2738	273	744
Fear	1263	169	371
Happiness	7587	945	2220
Sadness	4026	509	1066
Surprise	1465	197	393

Dataset experimental: CMU-MOSEI

- Detección de emociones e intensidad de ellas
- Emociones (6): felicidad, enojo, tristeza, disgusto, miedo y sorpresa
- 23,457 monólogos extraídos de videos de YouTube

CMU-MOSEI Unaligned

Emotion	Train	Develop	Test
Anger	3267	318	1015
Disgust	2738	273	744
Fear	1263	169	371
Happiness	7587	945	2220
Sadness	4026	509	1066
Surprise	1465	197	393

Dataset experimental: CMU-MOSEI

- Detección de emociones e intensidad de ellas
- Emociones (6): felicidad, enojo, tristeza, disgusto, miedo y sorpresa
- 23,457 monólogos extraídos de videos de YouTube
- Monólogos escogidos de manera aleatoria y distintos temas con balance en género

CMU-MOSEI Unaligned

Emotion	Train	Develop	Test
Anger	3267	318	1015
Disgust	2738	273	744
Fear	1263	169	371
Happiness	7587	945	2220
Sadness	4026	509	1066
Surprise	1465	197	393

Dataset experimental: CMU-MOSEI

- Detección de emociones e intensidad de ellas
- Emociones (6): felicidad, enojo, tristeza, disgusto, miedo y sorpresa
- 23,457 monólogos extraídos de videos de YouTube
- Monólogos escogidos de manera aleatoria y distintos temas con balance en género
- Se toman los features preprocesados de audio y video

CMU-MOSEI Unaligned

Emotion	Train	Develop	Test
Anger	3267	318	1015
Disgust	2738	273	744
Fear	1263	169	371
Happiness	7587	945	2220
Sadness	4026	509	1066
Surprise	1465	197	393

1 Introducción

Panorama General

Conjunto de Datos y SotA

Estudios de Problemáticas

2 BiProjection Multimodal Transformer (BPMuLT)

3 Resultados

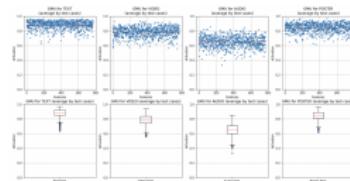
4 Conclusiones

5 Referencias

Estudios de Problemáticas

Adaptación 1: Enriquecer modalidades

- ¿Cómo se podría activar mejor cada modalidad?

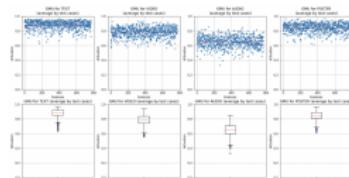


Adaptación 2: Mejorar la fusión

Adaptación 3: Mejorar el aprendizaje

Estudios de Problemáticas

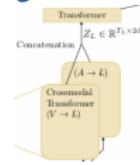
Adaptación 1: Enriquecer modalidades



- ¿Cómo se podría activar mejor cada modalidad?

Adaptación 2: Mejorar la fusión

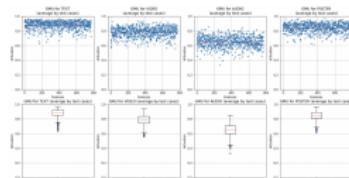
- ¿Cómo optimizar la fusión de información sin perjudicar desempeño?



Adaptación 3: Mejorar el aprendizaje

Estudios de Problemáticas

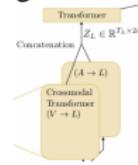
Adaptación 1: Enriquecer modalidades



- ¿Cómo se podría activar mejor cada modalidad?

Adaptación 2: Mejorar la fusión

- ¿Cómo optimizar la fusión de información sin perjudicar desempeño?



Adaptación 3: Mejorar el aprendizaje

- ¿Cómo reducir el desv. grad. colocando estratégicamente conexiones residuales?

1 Introducción

2 BiProjection Multimodal Transformer (BPMuLT)

3 Resultados

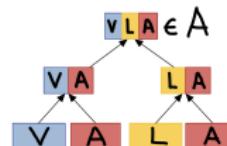
4 Conclusiones

5 Referencias

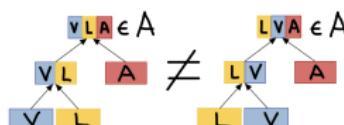
Soluciones a problemáticas

Adaptación 1: Enriquecer modalidades

- Proyectar en conjunto 2 modalidades distintas a una tercera en lugar de proyectar cada una por separado



(a) Ejemplo: Proyección por separado



(b) Ejemplo: Biproyección

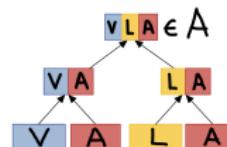
Adaptación 2: Mejorar la fusión

Adaptación 3: Mejorar el aprendizaje

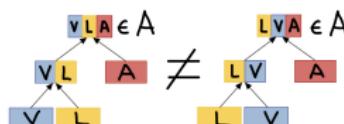
Soluciones a problemáticas

Adaptación 1: Enriquecer modalidades

- Proyectar en conjunto 2 modalidades distintas a una tercera en lugar de proyectar cada una por separado



(a) Ejemplo: Proyección por separado



(b) Ejemplo: Biproyección

Adaptación 2: Mejorar la fusión

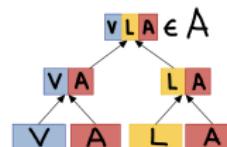
- Se realiza una combinación pesada con un módulo nuevo basado en GMU basado en Arévalo J, et. al. (2017)

Adaptación 3: Mejorar el aprendizaje

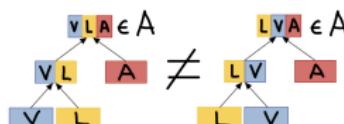
Soluciones a problemáticas

Adaptación 1: Enriquecer modalidades

- Proyectar en conjunto 2 modalidades distintas a una tercera en lugar de proyectar cada una por separado



(a) Ejemplo: Proyección por separado



(b) Ejemplo: Biproyección

Adaptación 2: Mejorar la fusión

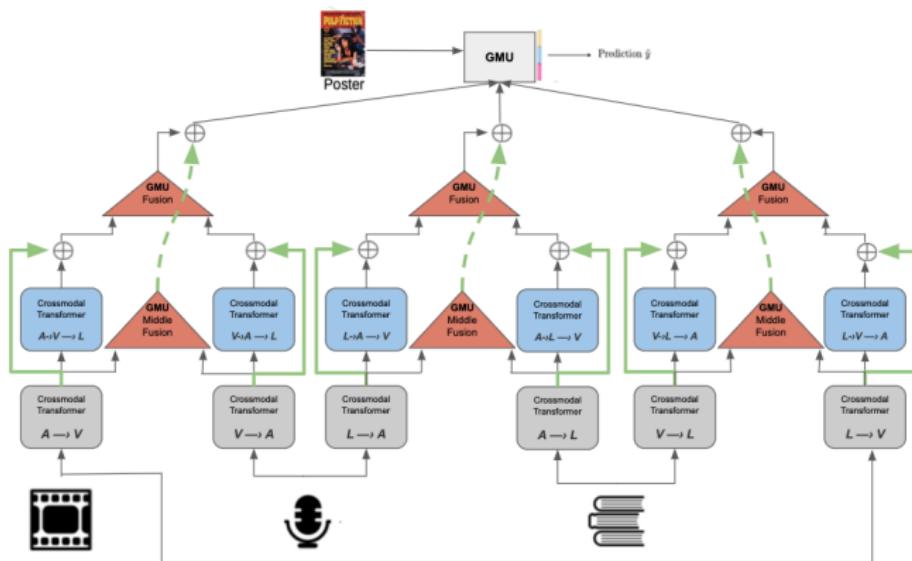
- Se realiza una combinación pesada con un módulo nuevo basado en GMU basado en Arévalo J, et. al. (2017)

Adaptación 3: Mejorar el aprendizaje

- Se colocan de manera estratégica conexiones res. dentro de la arquitectura

Detalles de la Arquitectura

BiProjection Multimodal Transformer



1 Introducción

2 BiProjection Multimodal Transformer (BPMuLT)

3 Resultados

4 Conclusiones

5 Referencias

Resultados en Moviescope

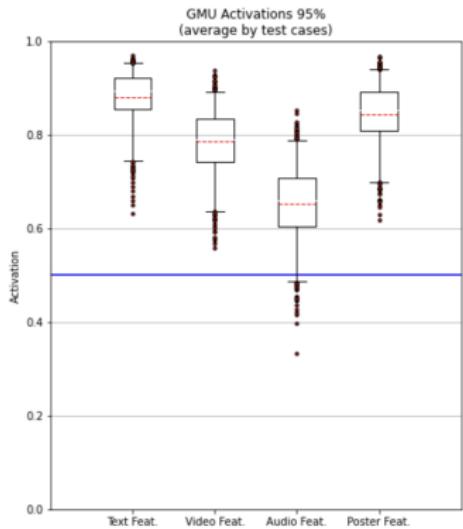
Comparación del BPMuLT vs. Otros

Model	μ AP	mAP	sAP
MMBT	77.4 ± 0.7	74 ± 0.8	85.1 ± 0.7
Fast-MA	74.9	67.5	82.3
MuLT	78.9 ± 0.3	75.7 ± 0.5	85.6 ± 0.3
MuLT-GMU	79.8 ± 0.4	76.0 ± 0.9	86.1 ± 0.4
BPMuLT (ours)	81.4 ± 0.3	78.0 ± 0.5	88.1 ± 0.3

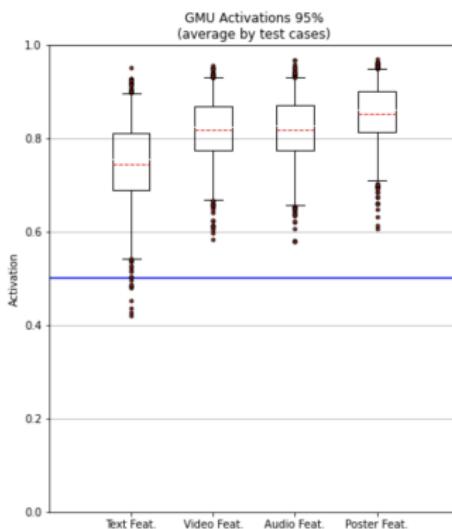
Tabla 1: Metrics reported correspond to average precision, micro (μ AP), macro (mAP) and sample (sAP) averaged.

- Mejora por más de un punto en ambos scores
- Se logra cumplir objetivos sin perjudicar clasificación

Comparación de Activación del GMU final para MuLT-GMU y BPMuLT



(a) MuLT-GMU



(b) BPMuLT

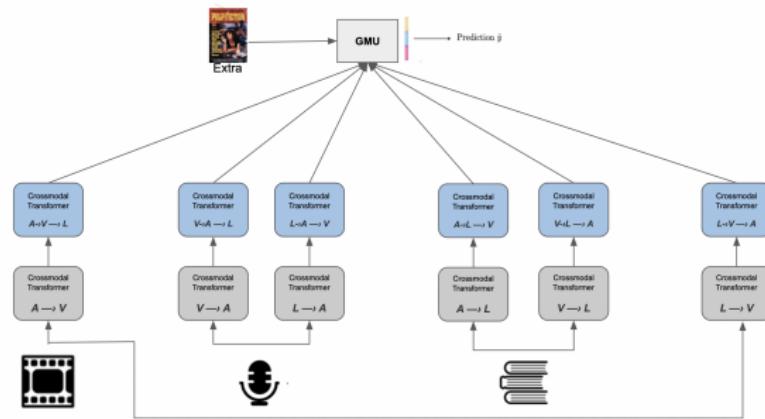
Estudios de Ablación Experimental

BPMuLT sin Módulos FGMU



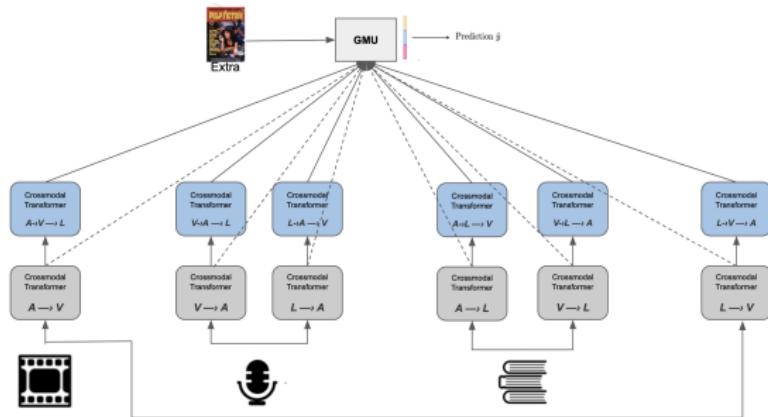
Estudios de Ablación Experimental

BPMuLT sin Módulos FGCU ni Conexiones Residuales



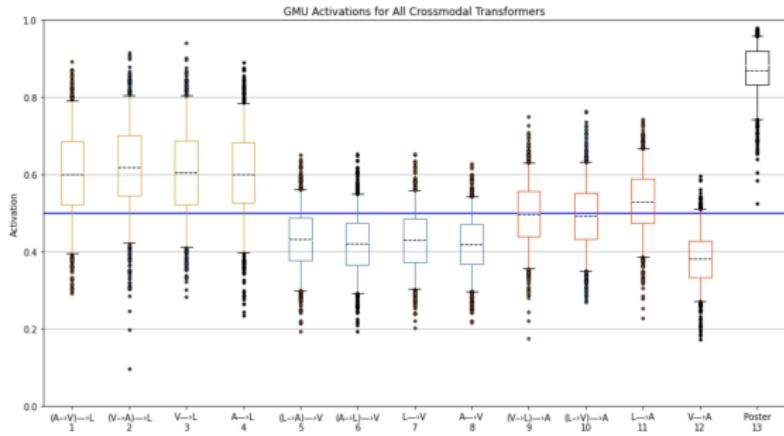
Estudios de Ablación Experimental

BPMuLT con Todos los Cross. Transformers al GMU



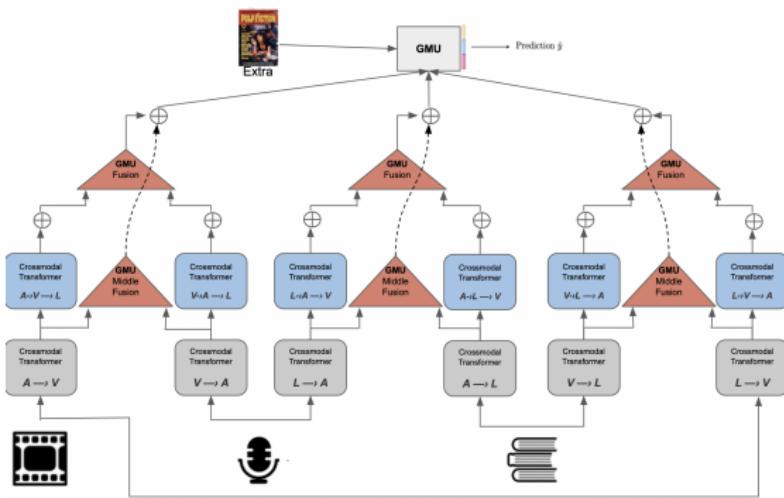
Estudios de Ablación Experimental

BPMuLT con los Cross. Transformers finales al GMU



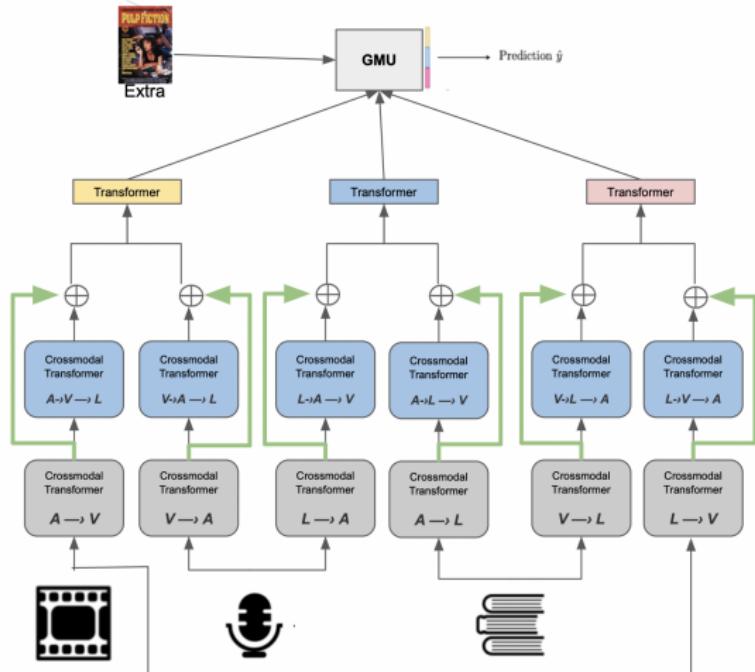
Estudios de Ablación Experimental

BPMuLT sin Conexiones Residuales



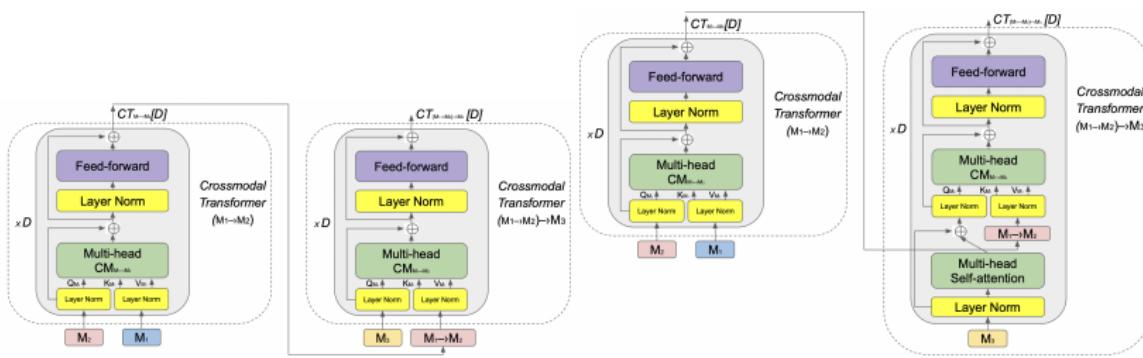
Estudios de Ablación Experimental

BPMuLT con Transformer de Fusión



Estudios de Ablación Experimental

BPMuLT con Translating Attention



Estudios de Ablación Experimental

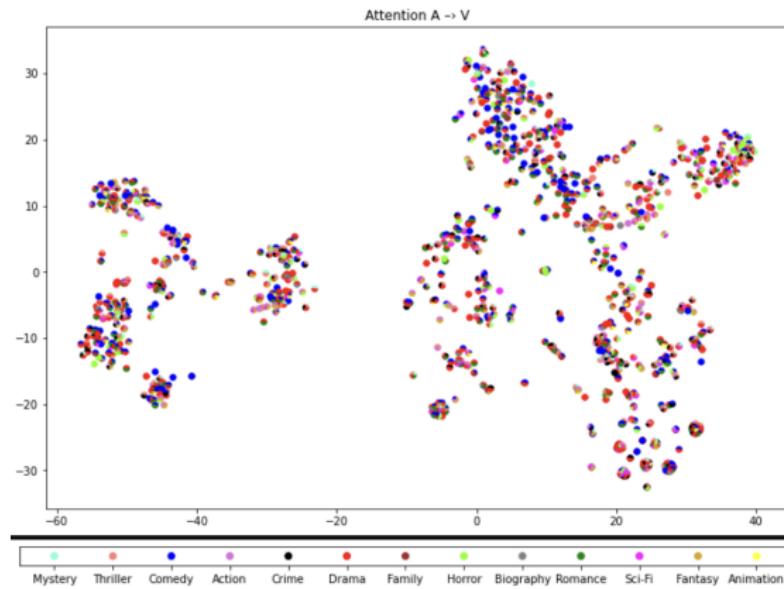
Tabla de resultados

Modalities	Model	μ AP	mAP	sAP
TVAP	BPMuLT-translating (ours)	81.2 ± 0.3	77.8 ± 0.5	87.0 ± 0.3
	BPMuLT-with-transformer (ours)	79.97 ± 0.1	76.6 ± 0.3	86.2 ± 0.3
	BPMuLT-no-FGMU-Middle (ours)	80.95 ± 0.3	77.6 ± 0.5	86.9 ± 0.4
	BPMuLT-no-RC (ours)	81.2 ± 0.3	77.7 ± 0.5	87.1 ± 0.3
	BPMuLT-no-FGMU (ours)	79.9 ± 0.25	76.54 ± 0.3	86.2 ± 0.4
	BPMuLT-no-FGMU-nor-RC (ours)	80.5 ± 0.5	76.9 ± 0.5	86.4 ± 0.3
	BPMuLT-all-transformers (ours)	79.8 ± 0.4	76.4 ± 0.3	86.1 ± 0.5
	BPMuLT-no-parallel (ours)	81.4 ± 0.3	78.0 ± 0.5	87.2 ± 0.4
	BPMuLT (ours)	81.7 ± 0.2	78.1 ± 0.1	92.4 ± 0.1

Tabla 2: Metrics reported corresponding to average precision, micro (μ AP), macro (mAP), and sample (sAP) averaged.

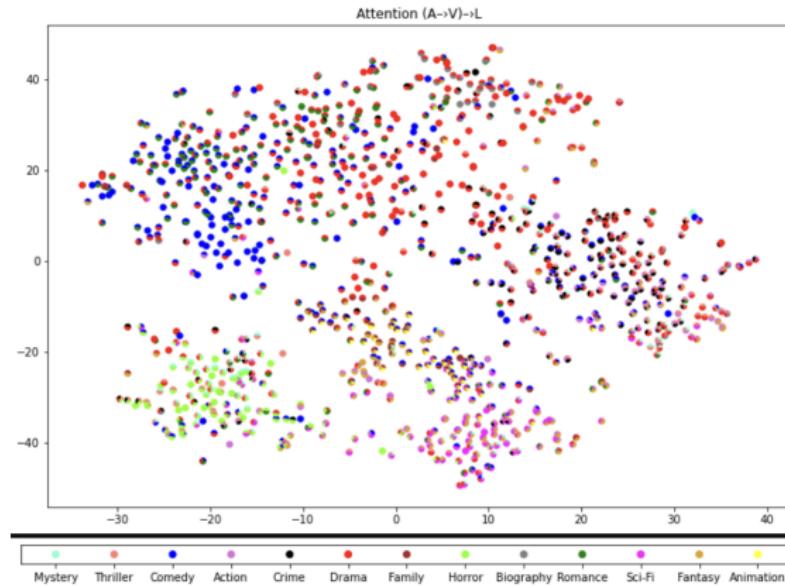
Estudio TSNE activación dada FGMUs

Atención A → V



Estudio TSNE activación dada FGMUs

Atención (A→V)→L



Resultados en Otros Conjuntos

MM-IMDb

MM-IMDB

Model	μ -F1	m -F1	W -F1	s -F1
GMU	63.0	51.4	61.7	63
MuLT-GMU	$66.3 \pm .6$	$61.1 \pm .6$	-	-
ConcatBERT	$65.9 \pm .2$	$60.5 \pm .3$	-	-
MMBT	$66.8 \pm .1$	$61.6 \pm .2$	-	-
BPMuLT-no-parallel (ours)	$68.9 \pm .1$	$58.7 \pm .3$	$68.8 \pm .1$	$69.4 \pm .2$
BPMuLT (ours)	$68.9 \pm .1$	$58.7 \pm .3$	$68.8 \pm .1$	$69.4 \pm .2$

Resultados en Otros Conjuntos

IEMOCAP Alineado

IEMOCAP Aligned

Model	Neutral		Happy		Sad		Angry		Average	
	Acc	AUC								
MuLT	71.0	77.2	83.5	71.2	85.0	89.3	85.5	92.4	81.3	82.5
ModTrans-MMEmoRe	71.1	76.7	85.0	74.2	86.6	88.4	88.1	93.2	82.7	83.1
BPMuLT-no-p (ours)	71.9	78.4	87.0	73.0	85.8	89.0	85.5	92.3	82.6	83.2
BPMuLT (ours)	65.0	70.3	85.5	73.5	78.8	76.4	81.1	85.6	77.6	76.4

Resultados en Otros Conjuntos

IEMOCAP NO Alineado

IEMOCAP Unaligned

Model	Neutral		Happy		Sad		Angry		Average	
	Acc	F1								
MuLT	62.5	59.7	84.8	81.9	77.7	74.1	73.9	70.2	74.7	71.5
BPMuLT-no-p (ours)	59.6	45.7	85.6	79.0	79.4	70.3	75.8	65.4	75.1	65.1
BPMuLT (ours)	59.2	44.0	85.6	79.0	79.4	70.3	75.6	66.2	74.9	64.9

Resultados en Otros Conjuntos

CMU-MOSEI

CMU-MOSEI Unaligned

Model	Angry W-Acc/F1	Disgust W-Acc/F1	Fear W-Acc/F1	Happy W-Acc/F1	Sad W-Acc/F1	Surprised W-Acc/F1	Average W-Acc/F1
MuLT	64.9/47.5	71.6/49.3	62.9/25.3	67.2/ 75.4	64.0/48.3	61.4/25.6	65.4/45.2
FE2E	67.0/ 49.6	77.7 /57.1	63.8/26.8	65.4/72.6	65.2/ 49.0	66.7 / 29.1	67.6/ 47.4
MESM (0.5)	66.8/49.3	75.6/56.4	65.8/28.9	64.1/72.3	63.0/46.6	65.7/27.2	66.8/46.8
BPMuLT-no-p (ours)	66.7/26.5	69.0/ 75.6	66.2 / 48.7	74.7 /50.8	67.3 /48.5	62.9/26.5	68.3±.3 / 46.1±.4
BPMuLT (ours)	69.3/26.4	68.0/ 74.9	66.0 / 48.5	74.5 /50.9	67.1 /48.3	63.4/25.1	68.0±.6 / 45.7±.6

1 Introducción

2 BiProjection Multimodal Transformer (BPMuLT)

3 Resultados

4 Conclusiones

5 Referencias

Conclusiones

- Se logra subir la activación/atención de las modalidades que se consideraban «malas» sin tener un perjuicio.

Conclusiones

- Se logra subir la activación/atención de las modalidades que se consideraban «malas» sin tener un perjuicio.
- Las conexiones residuales se colocan de tal forma que el score mejora y se reduce el desvanecimiento del gradiente.

Conclusiones

- Se logra subir la activación/atención de las modalidades que se consideraban «malas» sin tener un perjuicio.
- Las conexiones residuales se colocan de tal forma que el score mejora y se reduce el desvanecimiento del gradiente.
- Se obtiene un modelo más interpretable y menos costoso gracias a la fusión GMU.

Conclusiones

- Se logra subir la activación/atención de las modalidades que se consideraban «malas» sin tener un perjuicio.
- Las conexiones residuales se colocan de tal forma que el score mejora y se reduce el desvanecimiento del gradiente.
- Se obtiene un modelo más interpretable y menos costoso gracias a la fusión GMU.
- Se obtiene un modelo capaz de manejar secuencias multimodales de distintas tareas e incluso interpretar una modalidad con distintos canales.

1 Introducción

2 BiProjection Multimodal Transformer (BPMuLT)

3 Resultados

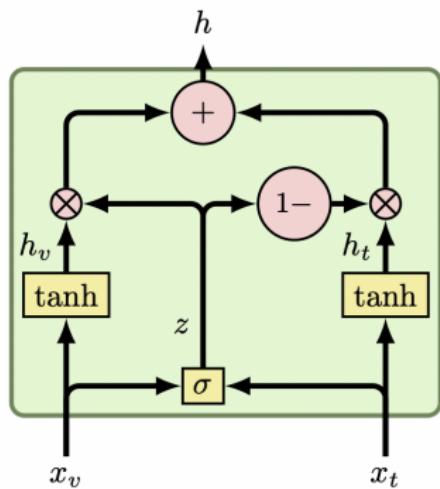
4 Conclusiones

5 Referencias

- Rodriguez-Bribiesca, I., Lopez-Monroy, A. P., and Montes-y-Gomez, M. Multimodal weighted fusion of transformers for movie genre classification. Proceedings of the Third Workshop on Multimodal Artificial Intelligence, Association for Computational Linguistics, 2021, 1–5.
- Arevalo, John, Solorio, Thamar, Montes-y-Gómez, Manuel, González, Fabio A. GATED MULTIMODAL UNITS FOR INFORMATION FUSION, Workshop track - ICLR, 2017, arXiv:1702.01992v1.
- Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, Ruslan Salakhutdinov, Multimodal Transformer for Unaligned Multimodal Language Sequences, Association for Computational Linguistics, 2019, pages 6558–6569.
- Cascante-Bonilla, P., Sitaraman, K., Luo, M., and Ordóñez, V. MovieScope: Large-scale analysis of movies using multiple modalities. arXiv preprint arXiv:1908.03180, 2019, 0–5.
- Kiela, D., Bhooshan, S., Firooz, H., Perez, E., and Testuggine, D. Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950, 2019, 0–5.
- Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Provost, E. M., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. Iemocap: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 0–5.
- Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. ACL, 2018, 0–5

¡Gracias por su atención!

Gated Multimodal Units (GMU)



(a) Arquitectura GMU

$$\begin{aligned}
 h_v &= \tanh(W_v \cdot x_v) \\
 h_t &= \tanh(W_t \cdot x_t) \\
 z &= \sigma(W_z \cdot [x_v, x_t]) \\
 h &= z * h_v + (1 - z) * h_t \\
 \Theta &= \{W_v, W_t, W_z\}
 \end{aligned}$$

(b) Ecuaciones GMU

Figura 6: GMU de Arévalo, et. al. (2017), bimodal.

Gated Multimodal Units (GMU)

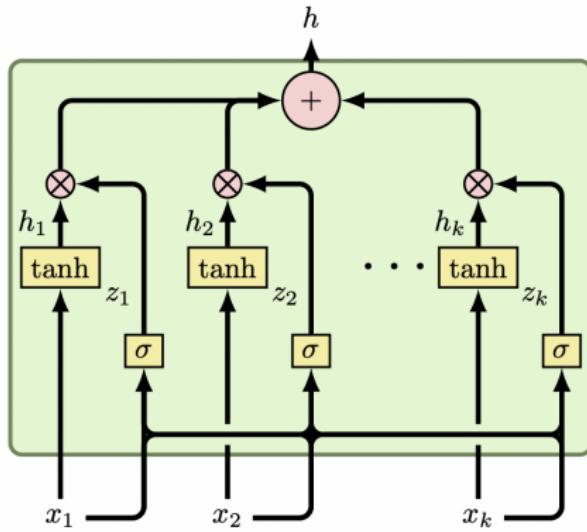


Figura 7: GMU de Arévalo, et. al. (2017), para varias clases.