# :Summary of paper 1

1. Introduction
2. Applications
3. Multimodal Representations
4. Translation
5. Alignment
6. Fusion
7. Co-learning
8. Conclusion

**1-introduction:**

**- اي هو ال modality و بشرح ليه مهم ال Multimodal Machine Learning**

**- التصنيفات القديمة كانت بتعتمد علي ال (early fusion and late fusion) ف علشان كدا قالوا ان في 5 تحديات اساسيه اعمق من الطريقه دي**

**- Representation**

**- Translation**

**- Alignment**

**- Fusion**

**-Co-learning**

Table 1: A summary of applications enabled by multimodal machine learning. For each application area we identify the core technical challenges that need to be addressed in order to tackle it.

| | Challenges | | | | |
|---|---|---|---|---|---|
| Applications | Representation | Translation | Alignment | Fusion | Co-learning |
| **Speech recognition** | | | | | |
| Audio-visual speech recognition | ✓ | | ✓ | ✓ | ✓ |
| **Event detection** | | | | | |
| Action classification | ✓ | | | ✓ | ✓ |
| Multimedia event detection | ✓ | | | ✓ | ✓ |
| **Emotion and affect** | | | | | |
| Recognition | ✓ | | ✓ | ✓ | ✓ |
| Synthesis | ✓ | ✓ | | | |
| **Media description** | | | | | |
| Image description | ✓ | ✓ | ✓ | | ✓ |
| Video description | ✓ | ✓ | ✓ | ✓ | ✓ |
| Visual question-answering | ✓ | | ✓ | ✓ | ✓ |
| Media summarization | ✓ | ✓ | | ✓ | |
| **Multimedia retrieval** | | | | | |
| Cross modal retrieval | ✓ | ✓ | ✓ | | ✓ |
| Cross modal hashing | ✓ | | | | ✓ |
| **Multimedia generation** | | | | | |
| (Visual) speech and sound synthesis | ✓ | ✓ | | | |
| Image and scene generation | ✓ | ✓ | | | |

**-Applications:2**

**- Multimedia Data Analysis (Automatic Captioning )**

**- Human-Computer Interaction – HCI (Affective Computing / Emotion Recognition &)**

**( Audio-Visual Speech Recognition – AVSR**

**Healthcare -**

*** The Historical Perspective**

First Stage:

Second Stage:

1-Hidden Markov Models (HMM)

2-Dynamic Time Warping (DTW)

Third Stage:

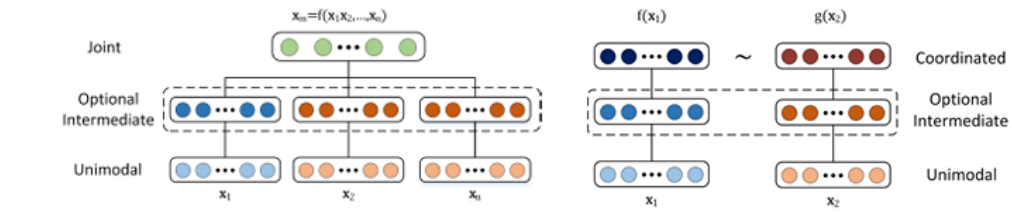
Fourth Stage:

1-Deep Neural Networks

3-**Multimodal Representations**

*** Joint Representations** $xm = f(x1, \ldots, xn)$,

***Coordinated Representations** $f(x1) \sim g(x2)$,




(a) Joint representation (b) Coordinated representations

**<u>Joint Representations:</u>**

**Equation =** $xm = f(x1, \ldots, xn)$,

- الفكرة الأساسية :نأخذ الإشارات القادمة من الأنماط المختلفة (نص، صورة، صوت) ونقوم بإسقاطها (Project) معاً في فضاء رياضي مشترك واحد.

-حيث Xn التمثيلات الفردية لكل نمط، والدالة f (والتي غالباً ما تكون شبكة عصبية عميقة) تدمجهم لتخرج بتمثيل موحد Xm

- متى نستخدمها؟ تُستخدم بشكل أساسي في المهام التي تكون فيها جميع الأنماط متوفرة أثناء مرحلة التدريب (Training) وأيضاً أثناء التوقع .(Inference) أبسط مثال لها هو ربط المتجهات بجانب بعضها.(Concatenation)

**الورقة تقسم الطرق المتقدمة لبناء التمثيلات المشتركة إلى 3أنواع:**

- **Neural Networks**

كيف تعمل؟ يبدأ كل نمط بتمثيل نفسه عبر طبقات عصبية خاصة به(Unimodal Layers) ، ثم تلتقي هذه المسارات في طبقة مخفية مشتركة (Hidden Layer) تقوم بإسقاطهم في الفضاء الموحد.

المميزات :يمكن تدريبها من البداية للنهاية(End-to-End) ، كما يسهل عمل تدريب مسبق (Pre-training) لها باستخدام بيانات غير معلمة.(Unlabeled Data)

العيوب :لا تتعامل بشكل مرن وطبيعي مع البيانات المفقودة(Missing Data) ؛ فلو اختفى نمط أثناء التوقع، يواجه النموذج مشكلة.

- **Graphical Models**

كيف تعمل؟ تعتمد على المتغيرات العشوائية الكامنة (Latent Random Variables) باستخدام هياكل مثل "آلات بولتزمان العميقة.(Deep Boltzmann Machines - DBM) "

المميزات :طبيعتها توليدية(Generative) ، وهي ممتازة جداً في التعامل مع البيانات المفقودة .حتى لو اختفى نمط كامل، يستطيع النموذج توليده أو تعويضه من النمط الآخر.

العيوب :صعبة جداً في التدريب، وتستهلك قدرة حوسبية عالية.(High computational cost)

- **Sequential Representations**

كيف تعمل؟ لتمثيل البيانات متغير الطول عبر الزمن (مثل الفيديو أو الصوت أو الجمل المكتوبة)، نستخدم الشبكات التكرارية مثل (RNN) و .(LSTM) الحالة المخفية (Hidden State) للشبكة عند اللحظة T تعبر عن تلخيص لكل ما مر به النموذج عبر الزمن.

**<u>Coordinated Representations:</u>**

**Equation =** $f(x1) \sim g(x2)$

الفكرة الأساسية :هنا لا نخلط الأنماط في فضاء واحد؛ كل نمط يحتفظ بفضائه الرياضي المستقل وطبقاته المنفصلة .ولكن، نقوم بفرض قيود رياضية (Constraints) لضمان وجود تنسيق وتشابه بين الفضاءات

حيث f و g دالتان مستقلتان تماماً لكل نمط، والعلامة (~) تعني وجود علاقة تنسيقية (مثل تقليل المسافة بينهما).

**الورقة تقسم التمثيلات المنسقة إلى نوعين:**

**-Similarity Models**

**كيف تعمل؟** نتدرب على تصغير المسافة (Distance) بين النمطين إذا كانا يعبران عن نفس الشيء.

نموذج **DeViSE**ونموذج **WSABIE**

**-Structured Models**

تذهب لأبعد من مجرد "التشابه" وتفرض شروطاً هيكلية محددة تناسب التطبيق

**(Cross-modal Hashing):**ضغط البيانات عالية الأبعاد إلى أكواد ثنائية (Binary Codes) لتسهيل البحث السريع والملائم بين الأنماط

**(Order Embeddings):**فرض قيد "ترتيبي أو هرمي .(Partial Order) "مثلاً: متجه صورة "امرأة تمشي مع كلبها" يجب أن يكون جزءاً تفرعياً تحت متجه النص العام "امرأة تمشي"، مما يخلق تسلسلاً مفهومياً داخل الفضاء المنسق.

**Joint Representation** : مناسب لما تكون البيانات متوفرة دائماً معاً في كل الأوقات، وهي بمثابة "مزيج كامل

**Coordinated Representations:** ناسب جداً لعمليات البحث واسترجاع البيانات(Information Retrieval) ؛ لأنك لو تملك نصاً فقط، يمكنك البحث به في فضاء الصور المستقل بفضل التنسيق المسبق بين الفضائين.

**الـ Joint Representations:** بتدمج كله في فضاء واحد .شرطها الأساسي :لازم كل الأنماط تكون موجودة وقت التوقع .(Inference / Test time) ممتازة في: التعرف على الكلام(AVSR) ، وتحليل المشاعر، وتفسير الإيماءات. وتتميز بأنها ممكن تدمج أكثر من نمطين عادي (صوت + صورة + نص).

**Coordinated Representations**: بتسيب كل نمط في فضائه بس منسقين .ميزتها الكبرى :مش شرط الأنماط كلها تكون موجودة وقت التست؛ ممكن يكون معاك نمط واحد وتستدعي التاني (زي تبحث بنص يرجعلك صورة). ممتازة في: الـRetrieval ، الـTranslation ، الـGrounding ، والـ Zero-shot learning. لكن عيبها الحالي (وقت نشر الورقة) : إنها محصورة غالباً بين نمطين اثنين فقط.

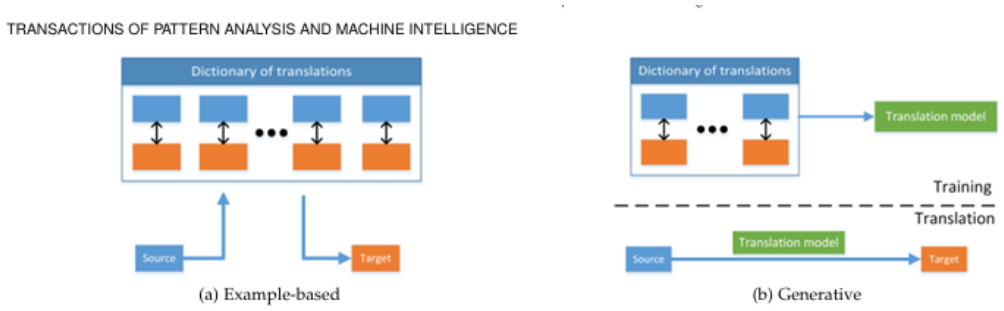
| REPRESENTATION | MODALITIES | REFERENCE |
|---|---|---|
| **Joint** | | |
| Neural networks | Images + Audio<br>Images + Text | [150], [157], [235]<br>[191] |
| Graphical models | Images + Text<br>Images + Audio | [206]<br>[108] |
| Sequential | Audio + Video<br>Images + Text | [100], [158]<br>[173] |
| **Coordinated** | | |
| Similarity | Images + Text<br>Video + Text | [64], [110]<br>[166], [239] |
| Structured | Images + Text<br>Audio + Articulatory | [33], [220], [256]<br>[228] |

## 4-TRANSLATION

في هذا القسم، الباحثون بيناقشوا مشكلة في غاية الأهمية :كيف نأخذ بيانات من نمط معين (مثلاً: صورة) ونقوم بتوليد أو ترجمة بيانات متوافقة معها تماماً في نمط آخر (مثلاً: وصف نصي لهذه الصورة)؟

الورقة وضعت تصنيفاً دقيقاً جداً (Taxonomy) لتلخيص كل أبحاث الترجمة دي، وقسمتها بناءً على **عاملين أساسيين**:

1. طبيعة النموذج :(Model Texture) هل النموذج يعتمد على القواعد واسترجاع البيانات(Example-based)، أم هو نموذج توليدي ذكي(Generative) ؟
2. اتجاة الترجمة :(Directionality) هل الترجمة بتشتغل في اتجاه واحد فقط(Unidirectional) ، أم في الاتجاهين تبادلياً (Bidirectional




**الفئة الأولى: النماذج القائمة على الأمثلة(Example-based Models)**

النماذج دي هي الأقدم والأبسط فكرة، ومبتخترعش كلام أو صور جديدة من عندها؛ بل بتعتمد على قاموس أو قاعدة بيانات ضخمة (Dictionary/Database) تم تدريبها عليها سابقاً.

الورقة بتقسمها لنوعين:

**1.النماذج القائمة على الاسترجاع:(Retrieval-based Models)**

- كيف تعمل؟ لو عطيته صورة جديدة وعايزه يترجمها لنص، النموذج مبيكتبش جملة جديدة. هو بيروح يدور في قاعدة البيانات اللي عنده على أقرب صورة تشبه الصورة دي في الفضاء المنسق(Coordinated Space) ، وياخد النص المكتوب تحت الصورة القديمة ويعطيه لك كترجمة!
- المميزات :الجمل الناتجة دايماً بتكون صحيحة لغوياً ومكتوبة بشكل بشري ممتاز (لأنها جمل حقيقية مسجلة مسبقاً).
- العيوب :النموذج مقيد باللي شافه في قاعدة البيانات؛ يعني لو الصورة فيها تفاصيل فريدة أو تركيب جديد، مش هيعرف يوصفه بدقة، لأنه هياخد أقرب جملة جاهزة وخلاص.

**2.النماذج القائمة على التعديل:(Out-of-context / Modification-based)**

- كيف تعمل؟ دي خطوة أذكى شوية من الاسترجاع المحض. النموذج بياخد الجملة المسترجعة من قاعدة البيانات، ويبدأ يعدل في بعض كلماتها عشان تناسب الصورة الجديدة.
- مثال :لو الجملة الجاهزة هي *"كلب أسود يجري في العشب"*، والصورة الجديدة فيها كلب بني، النموذج بيستبدل كلمة "أسود" بـ "بني" بناءً على خوارزميات معينة

**3.النماذج القائمة على الدمج أو التركيب(Combination-based Models)**

- الفكرة :هذه الطريقة أعمق وأعقد من مجرد تعديل كلمة في جملة واحدة. هنا، النموذج بيقوم بـ استرجاع جمل كثيرة جداً ومختلفة من قاعدة البيانات (كل جملة منها بتوصف جزء أو عنصر من الصورة الجديدة)، وبعد كده بيعمل لهم دمج وتركيب (Combine) مع بعض عشان يطلع جملة واحدة هجينة وجديدة!
- مثال دقيق :تخيل صورتك الجديدة فيها (شجرة، وكلب، وكرة).
  - النموذج يروح لقاعدة البيانات ويسترجع جملة لشجرة" *:شجرة خضراء كبيرة في الحديقة."*
  - ويسترجع جملة لكلب" *:كلب صغير يجري ويلهث."*
  - ويسترجع جملة لكرة" *:كرة حمراء تدحرج على الأرض."*
- الآن يأتي دور الـ :Combination يقوم النموذج بدمج وتركيب الأجزاء المهمة من الـ 3 جمل دول، ويصيغ جملة مرغوبة مثل" *:كلب صغير يجري بجانب شجرة خضراء كبيرة وبقربه كرة."*

**الفئة الثانية: النماذج التوليدية(Generative Models)**

هنا بقى قمة الـDeep Learning ؛ النماذج دي مابتسرقش جمل أو صور جاهزة، بل بتفهم المعنى وبتولد (Generate)حاجة جديدة تماماً من الصفر حرفياً كلمة كلمة أو بكسل بكسل.

الورقة بتقسم النماذج التوليدية لـ 3 أبنية رياضية شهيرة جداً:

**1.النماذج القائمة على القواعد والنحو:(Grammar-based Models)**

- كيف تعمل؟ النماذج دي بتجبر الكمبيوتر إنه يولد النص بناءً على قالب أو قواعد نحوية صارمة.(Templates)

- مثال :النموذج بيحدد الأول: (الفاعل، الفعل، المفعول، المكان). الكمبيوتر يشوف الصورة ويقول: الفاعل = ولد، الفعل = يلعب، المفعول = كرة، المكان = حديقة. وبعدين يركبهم في قالب جاهز" *:الولد يلعب بالكرة في الحديقة."*

- المميزات :بتضمن إن الجملة مش هتطلع مكسرة أو برة سياق القواعد.

- العيوب :الجمل بتطلع ناشفة، مكررة، وبتفتقد للمرونة الإبداعية البصرية.

**.2نماذج التشفير وفك التشفير:(Encoder-Decoder Models)**

- الفكرة العبقرية :ده الهيكل الأكثر شهرة واستخداماً اليوم) زي أنظمة الـ Image Captioning الحديثة.(

- كيف تعمل؟ .1الـ) Encoder المشفر :(بياخذ النمط الأول (مثلاً الصورة) ويضغط كل تفاصيلها البصرية في متجه رقمي .(Embedding Vector) غالباً بنستخدم هنا شبكة .2 .(CNN) الـ) Decoder مفك الشفرة :( بياخذ المتجه الرقمي ده، ويبدأ يفك تشفيره لتوليد النمط التاني (النص) كلمة ورا كلمة بالتتابع. غالباً بنستخدم هنا شبكة تكرارية RNN) أو.(LSTM

**.3النماذج المستمرة أو المتواصلة:(Continuous Generation Models)**

- كيف تعمل؟ دي مخصصة لما نكون بنترجم بين أنماط كلها مستمرة عبر الزمن (Sequence-to-Sequence) وبتتحرك مع بعض، زي دمج تتابع الفيديو لتوليد تتابع صوتي، أو العكس.

- **التحدي الأكبر هنا :**الحفاظ على التزامن والمحاذاة الزمنية (Temporal consistency) بين الأنماط المستمرة بحيث تفضل الترجمة ماشية مع الحركة لحظة بلحظة.

## 5-ALIGNMENT

| ALIGNMENT | MODALITIES | REFERENCE |
|---|---|---|
| **Explicit** | | |
| Unsupervised | Video + Text<br>Video + Audio | [136], [210], [211]<br>[160], [215], [259] |
| Supervised | Video + Text<br>Image + Text | [24], [260]<br>[113], [138], [168] |
| **Implicit** | | |
| Graphical models | Audio/Text + Text | [194], [224] |
| Neural networks | Image + Text<br>Video + Text | [102], [236], [238]<br>[244], [249] |

**ما هي الـ Alignment ببساطة؟**

تخيل عندك صورة مكتوب تحتها وصف" *.ولد يرتدي قميصاً أحمر ويمسك كرة ."*مهمة الـ Alignment هي أن الكمبيوتر يقدر يحدد بدقة) يعمل:**Link(**

- أن كلمة "ولد "تقابل هذه المنطقة البكسلية المحددة في الصورة
- أن كلمة "كرة "تقابل الدائرة الصغيرة التي في يده.

### (Explicit Alignment) المحاذاة الصريحة

هنا، الـ Alignment هو الهدف الأساسي والنهائي للنموذج .(Main Objective)إحنا داخلين الكود عشان نخلي الكمبيوتر يوفق المكونات دي ببعضها وبس.

**المحاذاة غير الخاضعة للإشراف:(Unsupervised Explicit Alignment)**

- **الفكرة :**لا نملك أي بيانات مساعدة (Labels) تخبرنا مسبقاً من يطابق من .النموذج بيعتمد على قيود رياضية زي الترتيب الزمني .

- **الخوارزمية الأشهر .(DTW (Dynamic Time Wrapping:** وهي خوارزمية برمجة ديناميكية ممتازة لمطابقة تتابعين زمنيين مختلفين في السرعة (مثلاً: مطابقة ملف صوتي لخطاب شخص مع فيديو لحركة وجهه، أو مطابقة خطوات وصفة طبخ مكتوبة مع فيديو الطباخ وهو بيعملها) .

- تم تطويرها لاحقاً باستخدام الـ CCA والـ Deep CCA لمطابقة العلاقات المعقدة وغير الخطية بين الأنماط (Deep Canonical Time Warping).

- النماذج الرسومية :(Graphical Models) مثل الـ Hidden Markov Models (HMM) استخدمت هنا بقوة لمطابقة لقطات الأفلام مع النصوص المكتوبة لها.

**ب. المحاذاة الخاضعة للإشراف:(Supervised Explicit Alignment)**

- الفكرة :هنا نملك بيانات تم تدريب النموذج عليها تحتوي على إشارات مرجعية واضحة (مثال: مربعات بصرية تحدد بدقة الكلمة المقابلة لها في النص).

- تستخدم في مهام مثل الـ Image co-reference لمطابقة الكلمات مع كائنات الصورة بصورة مباشرة ومدعومة بداتا سيت قوية.

**المحاذاة الضمنية(Implicit Alignment)**

هنا بقى الذكاء؛ الـ Alignment ليس هو الهدف النهائي للنموذج، بل هو خطوة وسيطة خفية (Intermediate / Latent step) تحدث داخل الشبكة لمساعدتها على حل مشكلة أخرى تماماً .

- المثال الأشهر عالمياً :الـ ( Attention Mechanism آلية الانتباه (داخل الشبكات العصبية (Neural Networks) والـ Transformers.

- كيف تعمل كـImplicit Alignment ؟ عندما تقوم شبكة بترجمة صورة إلى نص(Image Captioning) ، وهي تكتب كلمة *"دبدوب"*، تقوم طبقة الـ Attention تلقائياً بعمل "محاذاة وضبط بؤرة" على بكسلات الدبدوب داخل الصورة .نحن لم نطلب منها رسمياً استخراج مكان الدبدوب(Explicit) ، ولكنها اضطرت لعمل هذه المحاذاة داخلياً لكي تستطيع توليد الكلمة الصحيحة في النمط النصي .

- يتم استخدام هذا النوع بكثرة في أنظمة استرجاع الصور بناءً على النص، والترجمة الآلية المتقدمة، ونماذج الـ Neural Networks والـ Graphical Models الحديثة كما يظهر في الجدول

**التحديات الخاصة بالـAlignment (Challenges)**

الورقة بتذكر أن مهام المحاذاة تواجه 3 صعوبات تاريخية:

1. **الاعتمادية طويلة المدى :(Long-range dependencies)** مطابقة عنصر في أول الفيديو مع كلمة في آخر النص .
2. **الغموض :(Ambiguity)** كلمة واحدة قد تشير لأكثر من كائن في الصورة، أو كائن ب Visual مريب يحتمل أكثر من معنى نصي .
3. **صعوبة التقييم البشري والآلي :**جمع داتا سيت معمول لها لبلة (Labeling) على مستوى الأجزاء الصغيرة جداً هو أمر مرهق ومكلف للغاية .

## 6-Fusion

في هذا القسم، الباحثون بيناقشوا المشكلة الكلاسيكية والأكثر شهرة في المجال :كيف نأخذ معلومات مستخرجة من أنماط مختلفة (مثل الصوت والصورة والنص) ونقوم بدمجها معاً للقيام بتوقع أو تصنيف معين / Prediction) (Classification)

في هذا القسم، الباحثون بيناقشوا المشكلة الكلاسيكية والأكثر شهرة في المجال :**كيف نأخذ معلومات مستخرجة من أنماط مختلفة (مثل الصوت والصورة والنص) ونقوم بدمجها معاً للقيام بتوقع أو تصنيف معين / Prediction) Classification)**

الورقة وضعت تصنيفاً (Taxonomy) عبقرياً ومنظماً جداً للـ Fusion وقسمته إلى فئتين رئيسيتين )بدلاً من مجرد التصنيف القديم والسطحي المبكر والمتأخر:(

1. **الدمج القائم على النموذج :(Model-Agnostic Fusion)** وهي طرق لا تعتمد على نوع الشبكة العصبية نفسها، بل على "مكان" وكيفية دمج البيانات.
2. **الدمج المعتمد على النموذج :(Model-Based Fusion)** وهي طرق يتم فيها تصميم الـ Fusion داخل البنية الرياضية للنموذج نفسه.

**الفئة الأولى: الدمج القائم على النموذج(Model-Agnostic Fusion)**

الأساليب دي ميزتها إنك تقدر تستخدم معاها أي تصنيف (Classifier) أنت عايزه) زي SVM أو Random Forest أو حتى Neural Network بسيطة .(وهي تنقسم إلى 3 أنواع تاريخية شهيرة:

**أ. الدمج المبكر:(Early Fusion / Feature-level)**

- **كيف يعمل؟** نأخذ الميزات الخام (Features) المستخرجة من كل نمط فوراً، ونقوم بربطها بجانب بعضها البعض (Concatenation) في مصفوفة أو متجه واحد كبير، وبعد ذلك نرسل هذا المتجه الموحد إلى نموذج التصنيف.
- **المميزات :**بيسمح للنموذج إنه يلمح العلاقات والروابط المتبادلة بين الأنماط في مرحلة مبكرة جداً (Early correlation).
- **العيوب :**بيواجه مشكلة "تزامن الوقت (Time synchronization) "لو البيانات سريعة، كما أنه يؤدي إلى "لعنة الأبعاد العالية (Curse of dimensionality) "لأن المتجه الناتج بيبقى ضخم جداً.

**ب. الدمج المتأخر:(Late Fusion / Decision-level)**

- **كيف يعمل؟** هنا كل نمط بيشتغل في مسار مستقل تماماً؛ النمط الأول (الصورة) بيدخل على نموذج ويطلع "قرار أو توقع"، والنموذج الثاني (الصوت) بيطلع "قرار أو توقع". في النهاية تماماً، نأخذ هذه القرارات النهائية وندمجها معاً باستخدام معادلات بسيطة مثل) المتوسط الحسابي، أو التصويت بالأغلبية.(Majority voting
- **المميزات :**مرن جداً؛ لو حصل مشكلة أو اختفاء لنمط معين، بقية الأنماط بتشتغل عادي وتطلع قرارها دون أن ينهار النظام. كما أنه يتعامل بسهولة مع البيانات ذات الأبعاد المختلفة.
- **العيوب :**بيفقد تماماً القدرة على فهم العلاقات المتبادلة بين الأنماط (يقفل الباب أمام أي تفاعل مبكر بين الصوت والصورة مثلاً).

**ج. الدمج الهجين:(Hybrid Fusion)**

- **كيف يعمل؟** يجمع بين الطريقتين؛ يأخذ ميزات من بعض الأنماط ويدمجها مبكراً، وفي نفس الوقت يترك أنماطاً أخرى تطلع قرارات مستقلة ليتم دمجها في النهاية.

**الفئة الثانية: الدمج المعتمد على النموذج(Model-Based Fusion)**

هنا بقى الـ Fusion مبني جوه المعادلات الرياضية والـ Architecture بتاعة الـ Deep Learning نفسه، والورقة بتقسمه لـ 3 أبنية رئيسية:

**أ. النماذج القائمة على الشبكات العصبية:(Neural Network Models)**

- أصبحت هي المعيار الحديث، حيث يتم استخدام طبقات مخفية (Hidden Layers) لخلط ودمج المعلومات بعد معالجتها غير الخطية. ومن أشهرها استخدام الـ **Attention-based Fusion**، حيث تتعلم الشبكة ديناميكياً "أي نمط يجب التركيز عليه أكثر" في هذه اللحظة (مثلاً: التركيز على حركة الشفايف لو الصوت فيه ضوضاء عالية).

**ب. النماذج الرسومية الاحتمالية:(Graphical Models)**

- مثل Generative Adversarial Networks) أو آلات بولتزمان .(DBM ميزتها الكبرى هنا إنها بتقدر تدمج البيانات مع الحفاظ على التوزيع الاحتمالي لكل نمط، وهي ممتازة لو الداتا فيها قيم مفقودة (Missing values).

**ج. نماذج النواة المتعددة:(Multiple Kernel Learning - MKL)**

- دي امتداد لـ خوارزميات الـSVM ؛ حيث يتم تخصيص "نواة (Kernel) "معينة لكل نمط من الأنماط (تتناسب مع طبيعته الرياضية)، وبعد ذلك يتم دمج هذه النوايا عبر جمع خطي (Linear combination) أثناء تدريب النموذج.

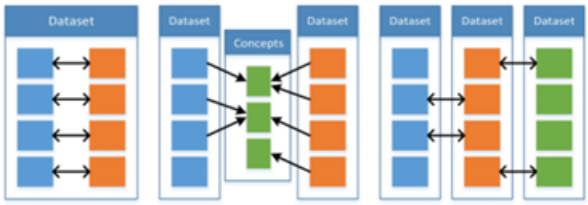
الورقة بتقولك إن الـ Fusion هو أقدم تحدي في الـMultimodal ، والكل كان فاكر إن الموضوع محصور في "أدمج في الأول ولا في الآخر؟(Early vs Late) "، لكن مع الـ Deep Learning اتضح إن الـ ImplicitوالـAttention based fusionهو المستقبل، لأنه بيسمح للكمبيوتر إنه يقرر بنفسه وبشكل مرن إزاي ومتى يدمج الإشارات دي بناءً على الموقف.

12

Table 5: A summary of our taxonomy of multimodal fusion approaches. OUT — output type (class — classification or reg — regression), TEMP — is temporal modeling possible.

| FUSION TYPE | OUT | TEMP | TASK | REFERENCE |
|---|---|---|---|---|
| **Model-agnostic** | | | | |
| Early | class | no | Emotion rec. | [35] |
| Late | reg | yes | Emotion rec. | [175] |
| Hybrid | class | no | Multimedia event detection | [122] |
| **Model-based** | | | | |
| Kernel-based | class<br>class | no<br>no | Object class.<br>Emotion rec. | [32], [69]<br>[94], [189] |
| Graphical models | class<br>reg<br>class | yes<br>yes<br>no | AVSR<br>Emotion rec.<br>Media class. | [78]<br>[14]<br>[97] |
| Neural networks | class<br>class<br>reg | yes<br>no<br>yes | Emotion rec.<br>AVSR<br>Emotion rec. | [100], [232]<br>[157]<br>[39] |

7-Co-learning




(a) Parallel (b) Non-parallel (c) Hybrid

في التحديات السابقة) مثل الـ Fusion أو الـTranslation) ، كنا نحتاج دائماً إلى وجود الأنماط معاً أثناء عملية التوقع أو التوليد. لكن في الـ Co-learning، الفلسفة مختلفة تماماً وعملية جداً؛ السؤال هنا هو :كيف نستخدم المعرفة المتاحة بكثرة في نمط معين (مستفيض البيانات) لمساعدة وتدريب نموذج يعمل على نمط آخر تماماً (شحيح البيانات أو يصعب جمعها)؟

الباحثون في الورقة يوضحون أن هذا التحدي يكون مفيداً جداً عندما نريد بناء نموذج يعمل في وقت التست (Test time) على نمط واحد فقط(Unimodal) ، ولكننا نريد استغلال البيانات متعددة الأنماط المتوفرة لدينا أثناء مرحلة التدريب (Training)فقط لجعل هذا النموذج أقوى وأكثر ذكاءً.

الورقة تقسم أبحاث الـ Co-learning إلى ثلاث فئات رئيسية بناءً على طريقة نقل المعرفة:

**لفئة الأولى: النقل الموازي(Parallel Co-learning)**

- الفكرة الأساسية :نستخدم هذه الطريقة عندما نملك مجموعة بيانات تحتوي على النمطين معاً لنفس العينات (أي أن البيانات موازية لبعضها: لكل صورة هناك نص يشرحها، أو لكل ملف صوتي هناك فيديو لحركة الشفايف).
- كيف يعمل؟ بدلاً من دمجهم، نقوم بجعل كل نمط يتدرب في مسار مستقل، ولكننا نفرض قيوداً تجعل طبقاتهم المخفية (Hidden layers) تتبادل المعرفة أو تتشابه في التمثيلات.
- الخوارزمية الأشهر :التعلم التعاوني .(Co-training) حيث يرى النموذج الأول نمطاً واحداً (مثل النص) والنموذج الثاني يرى النمط الآخر (مثل الصورة)، ويقوم كل نموذج بمساعدة الآخر في تسمية البيانات غير المعلمة (Unlabeled data) بناءً على مدى ثقته في التوقع، مما يرفع كفاءة الاثنين معاً.

**الفئة الثانية: النقل غير الموازي(Non-parallel Co-learning)**

- الفكرة الأساسية :هنا المعضلة أكبر؛ نحن لا نملك بيانات متطابقة أو موازية تجمع النمطين لنفس العينات (مثلاً: لدينا داتا سيت ضخمة جداً لنصوص طبية، وداتا سيت أخرى منفصلة تماماً لأشعة طبية لمرضى آخرين).
- كيف يعمل؟ نستخدم طرقاً لنقل المعرفة عبر المفاهيم المشتركة (Concepts) أو التصنيفات المشتركة بين النمطين.
- أبرز التطبيقات :الـ ) Transfer Learningالتعلم بنقل الأثر (والـ .Zero-Shot Learning
  - على سبيل المثال: نأخذ فضاء لغوياً نظيفاً ومدرَّباً على ملايين النصوص) مثل(Word2Vec ، ونستخدم الهيكلية والمعاني التي تعلمها النص لكي نُهيكل ونُرتب فضاء الرؤية الحاسوبية (Computer Vision)حتى لو كانت صور الحيوانات فيه شحيحة، فيبدأ نموذج الصور في فهم الفروق بين الكائنات بناءً على العلاقات المفهومية التي استعارها من فضاء النصوص.

**الفئة الثالثة: النقل الهجين أو الاستعاري(Hybrid / Transfer Co-learning)**

- الفكرة الأساسية :نستخدم نمطاً مساعداً كـ "جسر" أو مرجع لنقل المعرفة لنمط آخر تماماً، ويتم ذلك عبر طريقتين شهيرتين:
  - أ. النماذج المتعددة الأنماط الزائفة :(Multimodal Pseudo-inputs) إذا كان لدينا نقص في أحد الأنماط، نستخدم نموذج ترجمة (Translation) لتوليد بيانات زائفة أو تقريبية للنمط المفقود، ثم نستخدم هذا المزيج لتدريب النموذج النهائي.
  - ب. تقطير المعرفة :(Knowledge Distillation) تدريب شبكة ضخمة جداً ومتعددة الأنماط (صوت + صورة + نص)، ثم نقوم بـ "تقطير" وتركيز معرفتها داخل شبكة صغيرة جداً وبسيطة تعمل على نمط واحد فقط (مثل الصوت فقط)، لتصبح هذه الشبكة الصغيرة قادرة على رؤية أبعاد في الصوت لم تكن لتراها لولا توجيه الشبكة الكبيرة متعددة الأنماط أثناء التدريب.

| DATA PARALLELISM | TASK | REFERENCE |
|---|---|---|
| **Parallel** | | |
| Co-training | Mixture | [22], [115] |
| Transfer learning | AVSR<br>Lip reading | [157]<br>[148] |
| **Non-parallel** | | |
| Transfer learning | Visual classification<br>Action recognition | [64]<br>[134] |
| Concept grounding | Metaphor class.<br>Word similarity | [188]<br>[107] |
| Zero shot learning | Image class.<br>Thought class. | [64], [198]<br>[165] |
| **Hybrid data** | | |
| Bridging | MT and image ret.<br>Transliteration | [174]<br>[154] |

## ملخص مختصر للورقة(Abstract Summary)

الورقة هي دراسة مرجعية وتصنيفية (Survey & Taxonomy) شاملة لمجال التعلم الآلي متعدد الأنماط (Multimodal Machine Learning).تغطي الورقة الأبحاث التي تدمج وتتعامل مع أنماط بيانات مختلفة (النصوص، الصور، الفيديو، إشارات الصوت، البيانات الحركية). تهدف الورقة إلى توحيد المصطلحات وتوفير هيكل رياضي وهندسي واضح يجمع شتات الأبحاث السابقة بدل الاعتماد على التصنيفات السطحية القديمة.

## الفكرة الرئيسية والمساهمة(Main Contribution)

المساهمة الأساسية :(The Core Contribution) صياغة تصنيف هيكلي جديد (New Taxonomy) يقسم تحديات المجال إلى 5تحديات كبرى وثابتة-(Representation, Translation, Alignment, Fusion, Co learning) :.

إعادة تعريف وتفصيل المفاهيم الرياضية لكل تحدي، وربط الأبحاث التاريخية والحديثة (حتى تاريخ نشر الورقة) بهذه التحديات الخمسة لتسهيل مقارنتها.

مناقشة أزمات التقييم (Evaluation) والتوجهات المستقبلية للمجال.

## الـ Architecture أو Framework المقترح

بما أن الورقة Survey Paperوليست ورقة تجريبية(Empirical Paper) ، فهي لا تقترح الـ Architecture خاصاً بها لتدريبه. المساهمة هنا هي Conceptual Framework(إطار مفاهيمي تنظيمي) يوضح كيف يجب تصميم أي Architecture حديث في هذا المجال بناءً على التحديات الخمسة.

## الـ Datasets المستخدمة في الأبحاث المرجعية

الورقة استعرضت عشرات الداتا سيتس الشهيرة عبر جداولها بناءً على المهام:

للمهام الصوتية-البصرية :(AVSR) داتا سيتس مثل *Grid*و *CUAVE*.

لوصف الصور والفيديو :(Media Description & Captioning) داتا سيتس مثل *MS COCO*, *Flickr30K*, و *MSR-VTT*.

لتحليل المشاعر والتفاعل البشري :(Affect Recognition) داتا سيتس مثل *IEMOCAP*و *CMU-MOSEI*.

للإجابة على الأسئلة البصرية :داتا سيتس *VQA*.

## نقاط القوة والضعف(Pros & Cons)

نقاط القوة:

تنظيم وتأصيل عبقري للمجال أخرجه من التصنيف الضيق (Early/Late fusion) إلى أفق أوسع (التحديات الخمسة).

ربط التقنيات الإحصائية الكلاسيكية القديمة بطفرات الـ Deep Learning بشكل سلس وتاريخي متسلسل.

الدقة الشديدة في نقاش مشكلة التقييم.(Evaluation)

نقاط الضعف:

التركيز الأكبر كان محصوراً في ثلاثة أنماط رئيسية (نص، صوت، صورة/فيديو)، بينما تم إعطاء مساحة أقل للأنماط الصناعية أو الطبية أو إشارات المستشعرات الحيوية.(Sensors)

النماذج المذكورة في الـ Coordinated غلب عليها حصر العلاقة بين نمطين اثنين فقط، ولم تتوسع في هندسة التنسيق متعدد الأنماط الفائق.(Multi-modality > 2)

**الـ Research Gap الأشياء التي لم تعالجها الورقة**

التعامل مع البيانات الواقعية المشوشة :(In-the-wild Data) أغلب النماذج تفترض أن الأنماط متوفرة دائماً ومزامنتها الزمنية مثالية، والورقة تركت فجوة بحثية في كيفية التعامل مع الأنماط المفقودة ديناميكياً (Missing modalities)أو المشوشة جداً أثناء الـInference.

الشبكات الديناميكية بالكامل :(Dynamic Architectures) غياب النماذج التي تقود فيها طبيعة البيانات في نمط معين تغيير هيكل (Architecture) الشبكة الأخرى بشكل لحظي.