Abstract

- هاد أكبر **مجموعة بيانات مفتوحة** للطبخ، فيها **أكتر من مليون وصفة** و**13 مليون صورة أكل**. **Recipe1M+ Dataset:**
- **تعليم متعدد الوسائط:** المدربين بيقدروا يدرّبوا **نماذج ذكاء اصطناعي قوية** على بيانات متوافقة بين الصور والوصفات.
- **نموذج التضمين المشترك:** شبكة عصبية بتتعلم **تمثيل مشترك للوصفات والصور**، وبتحقق **نتائج قوية بمهام البحث عن الوصفات بالصور**.
- **تحسين الأداء:** إضافة هدف تصنيفي عالي بتحسّن البحث لدرجة **قريبة من أداء البشر**، وبتساعد على **إجراء حسابات معنوية بين المتجهات**.
- **إمكانيات مستقبلية:** هاي البيانات والنماذج بتمهّد لمجال واسع من الأبحاث بمجال **الذكاء الاصطناعي للأكل والتعلم المتعدد الوسائط**.
- **متاح للجميع: الكود، البيانات، والنماذج** مفتوحة المصدر لأي حد بده يشتغل عليها.

introduction

:

الاكل جزء أساسي من حياة الإنسان، وله علاقة كبيرة بالشاعر والشاعر. حتى لما الناس يهاجروا لبلد جديد، بيفضلوا يحافظوا على أكلهم التقليدي أكثر من لغتهم الأم. بالإضافة إلى أهميته الثقافية، الاكل بيقدم تحديات جديدة في مجال الذكاء الاصطناعي، مثل كيف تآزر العامل حتى لو اتبعت أو أتحورت (مثل ما بتحصل لما نقطّع أو نطبخ العاملين). بسبب عدم وجود عدد كبير من وصفات وصفات الأكل على الإنترنت، بقدر نعلّم ويفهم طريقة تحضير الأكليمن من خلال تحليل قائمة المكونات وطريقة الطبخ وبعض. هالتقنية مش حكيم في الطبخ، لكن ممكن تساعدنا في فهم تأثير الأكل في الصحة العامة والتراث الثقافي. لذلك نطور هالأدوات، يجب أن يكون عندنا مجموعات بيانات كبيرة ومنظمة.

التطورات الحديثة في الذكاء الاصطناعي، اللي اعتمدت على مجموعات بيانات ضخمة وتقنيات التعلم العميق، حسنت بشكل كبير قدرة الآلات على تمييز الأشياء والمشاهد. هالتقنيات ساعدت كمان في مجالات جديدة، مثل تقسيم الصور بدقة والتعرف على تفاصيلها. لو نعمل قاعدة بيانات كبيرة عن الأكل، ممكن نطور المجال أكتر، خصوصاً لأن الأكل عنده تحديات خاصة—مثل إنه نعرف إذا المكون مقطع، مشوي، مسلوق، أو نيء، وهالشي صعب على القواعد البيانات الحالية.

بالإضافة، الأكل مش ثابت مثل تصنيف الصور العادي، فهو يحتاج لطريقة مرنة عشان ناخد بالاعتبار الاختلافات في الوصفات وطرق التحضير

الشكل 1 يوضح إنه لو جمعنا صور الأكل مع وصفاتها (المكونات وطريقة الطبخ)، رح يساعد الذكاء الاصطناعي يفهم عملية الطبخ بشكل أعمق. الأبحاث الحالية تعتمد على قواعد بيانات صور أكل صغيرة أو متوسطة، مثل مجموعة Food-101 اللي حققت دقة 50.8%، وبعدين تحسنت إلى حوالي 80%—بس هادا يدل إنه حجم البيانات هو المشكلة الأساسية. في دراسات

وزملائه) حاولوا يحسبوا سعرات الأكل، لكنهم ما شاركوا البيانات التفصيلية (مثل تقسيم الصور) اللي Myers (مثل عمل تسمح للباحثين الآخرين يطوروا عليها.

١. مجموعة بيانات ضخمة (Recipe1M+):

- عندنا مليون وصفة طبخ منظمة
- كل وصفة معها صور الأكل التابعة لها
- أكبر مجموعة بيانات من هالنوع حتى الآن

٢. مهمة جديدة (im2recipe):

- الفكرة: تقدر تيجي على صورة أكلة وتلاقي الوصفة التابعة لها
- حتى لو ما تعرفش توصف الأكلة بالكلام!
- بستخدم النصوص (المقادير وطريقة التحضير) مع الصور مع بعض

٣. نموذج ذكاء اصطناعي متطور:

- بيدمج الصور والوصفات في "مساحة مشتركة"
- بيفهم العلاقة بين الصورة والنص
- بتحسن أداؤه عن طريق تصنيف الأكل لأنواعه (حلويات، أكل بحري، إلخ)

٤. نتائج مذهلة:

- النموذج بتاعنا ضرب كل النماذج القديمة
- قارب أداء البشر في بعض المهام!
- يعني تقدر تاخد صورة وتلاقي وصفتها بدقة عالية

٥. انفتاح آفاق جديدة:

- مش بس البحث عن وصفات

- ممكن يستخدم في:
  - تعديل الوصفات تلقائيًا
  - تحليل القيمة الغذائية
  - فهم التراث الثقافي من خلال الأكل
  - اكتشاف وصفات جديدة!

باختصار، هالدراسة فتحت باب كبير قدام الباحثين لفهم أعمق لعالم الطبخ باستخدام الذكاء الاصطناعي. وإحنا ناطرين شو رح يقدروا يعملوا بهالبيانات الجديدة

Related Work

سلايد 1: لمحة عن الدراسات السابقة

- من 2017 لليوم، في دراسات كتيرة تناولت الذكاء الصناعي بمجال الأكل والتعلم المتعدد الوسائط.
- المحاور الأساسية: استرجاع الوصفات من الصور، تحليل خصائص الأكل، وهيكلة الوصفات.

سلايد 2: نماذج متعددة الوسائط

- *Herranz et al.*: اقترحوا نموذج متطور بيعتمد على صور الأكل، الوصفات، المعلومات الغذائية، الموقع، قوائم المطاعم، وأنماط الأكل.
- *Min et al.*: طوروا نهج MATM اللي بدرس تأثير نمط المطبخ، نوع الوجبة، والنكهات، معتمدين على بيانات Yummly.

سلايد 3: مقارنة الوصفات وهيكلتها

- *Chang et al.*: حللوا اختلاف وصفات لنفس الطبخة (زي "كوكيز الشوكولاتة") باستخدام تجميع التشابه بين الوصفات.
- طريقتهم تعتمد على اختيار يدوي للميزات، بينما نموذجنا بيتعلمها تلقائيًا باستخدام الذكاء الصناعي.

Data Collection from Recipe Websites

سلايد 1: لمحة عن جمع البيانات

في هذا السلايد بنحكي عن كيفية جمع الوصفات. بدأنا بجمع الوصفات من **أكتر من عشرين موقع طبخ** مشهور ومعروف. هاي المواقع كانت تشمل مجموعة متنوعة من الوصفات من كل أنحاء العالم. بعد ما جمعنا الوصفات، قمنا باستخدام **عملية** وبعدين نزلنا الصور المرتبطة بكل وصفة. بعدها، نظمت البيانات ،**HTML معالجة** حتى نقدر نستخرج النصوص المهمة من بحيث تكون قابلة للترتيب والمتابعة، وكل وصفة وصورة صارت **محددة بشكل فريد** عشان نعرفها **JSON** في **صيغة** بسهولة.

سلايد 2: تنظيف البيانات وإزالة التكرار

بعد جمع البيانات، كانت هناك بعض الأشياء اللي لازم نتخلص منها عشان نعمل بيانات نظيفة ودقيقة. **إزالة الفراغات الزائدة** كانت خطوة ضرورية، لأن النصوص أحيانًا تحتوي على فراغات غير ضرورية اللي بتشوش على المعالجة. كمان زي الأكواد اللي تستخدمها المواقع عشان تعرض النصوص بشكل معين. وبالإضافة **HTML** قمنا بإزالة **الرموز الخاصة في** كانت تسبب مشاكل لأنها مش مدعومة في بعض الأنظمة. بعد هيك، عملنا **إزالة التكرار** من **ASCII** لهدول، **الحروف الغير** البيانات، يعني لو كان في وصفات متشابهة جدًا أو متكررة، استبعدناها. في النهاية، صار عندنا **أكثر من مليون وصفة** و**800 ألف صورة**.

سلايد 3: حجم البيانات والتوسعة

القاعدة اللي بنيناها هي **أكبر من أي قاعدة بيانات مشابهة** موجودة في نفس المجال. مثلاً، صار عنا **ضعف عدد الوصفات** من قواعد بيانات تانية كانت موجودة قبل هيك، و**ثماني مرات أكتر صور**. يعني حجم البيانات اللي جمعناها فعلاً ضخم وبيحتوي على تنوع هائل من الوصفات والصور. في المرحلة الجاية، رح نكمل توسعة قاعدة البيانات عن طريق البحث عن **صور أكل إضافية** باستخدام **محركات البحث عن الصور**، عشان نزيد حجم البيانات ونخليها أكثر تنوعًا ودقة

Dataset

تحديات فهم الأكل بالذكاء الاصطناعي .

- الأكل صعب تحلله لأنه:
- عندك ألف طريقة لنفس الطبق (مثل البيتزا بمكونات مختلفة)
- كل طباخ بيجهزه بطريقة
- الشكل النهائي ممكن يختلف حتى لنفس الوصفة!

. مشاكل قواعد البيانات الحالية٢

- أغلبها إما:

• **صور بس** (بدون وصفات)

(• **نصوص بس** (بدون صور

- اللي فيها الاتنين إما:

مش منظمة) HTML • صغيرة (مثلاً 101 ألف صورة مع وصفات

(• محدودة بمطبخ معين (مثل الأكل الصيني فقط

+Recipe1M . الحل: مجموعة٣

- الأضخم عالمياً:

• **مليون وصفة** منظمة

• **13 مليون صورة** أكل

- طريقة العمل:

(صورة Kوصفة، 800 Mجمعنا وصفات + صور من مواقع طبخ (1 ○

(بعد حذف المكرر Mزودناها بصور من محركات البحث (وصلت 13 ○

- أحجام مهولة:

• ضعف عدد الوصفات من أي مجموعة سابقة

!• 130 ضعف عدد الصور

. ليش هادا إنجاز؟٤

- رح يخلينا نطور ذكاء اصطناعي:

• يفهم العلاقة بين الصورة والوصفة

مسلوق) vs • يميز حتى طريقة التقطيع أو الطبخ (مشوي

ً• يعدل الوصفات تلقائيا

2.2 Data Extension using Image Search Engine

١. جمع البيانات من جوجل

.(مثال: بحث "أجنحة دجاج" (الصور الناتجة عالية الجودة، كما في الشكل ٢

.الهدف: جمع ٥٠ مليون صورة (٥٠ صورة لكل وصفة) باستخدام أدوات أوتوماتيكية

:التحديات

.ًحذف ٣٢ مليون صورة مكررة تماما

.(تصفية الصور المشوهة (معدلة الحجم/مقطوعة/مضافة عليها نصوص

.(إزالة الصور غير المرتبطة (مثل وجوه أشخاص أو ملصقات غذائية

(. حجم مجموعة البيانات (الجدول ٢١

:تحتوي على :+Recipe1M

.(مليون وصفة + ١٣.٧ مليون صورة (الأكبر في مجال الأكل

.مجموعة التدريب: ٧٢٠ ألف وصفة، ٩.٧ مليون صورة

.مجموعات الاختبار: ~١٥٥ ألف وصفة، ~٢ مليون صورة لكل منها

!الميزة: ١٣٠ ضعف عدد الصور في المجموعات السابقة

. تنظيف البيانات٣

.لمقارنة الصور ResNet18 الكشف عن التكرار: استخدام نموذج

.توزيع عادل: تخصيص الصور بشكل متساوٍ للوصفات المتشابهة

.لمعالجة البيانات بسرعة C++ السرعة: استخدام لغة

. الأهمية٤

.تدريب نماذج ذكاء اصطناعي أقوى لفهم العلاقة بين الصور والوصفات

.(مجموعة مفتوحة للباحثين لتطوير تطبيقات جديدة (مثل تعديل الوصفات تلقائيا

!باختصار: هالمجموعة رح تكون مرجع أساسي لأي بحث مستقبلي عن الأكل والذكاء الاصطناعي

Nutritional Information

سلايد 1: معالجة بيانات المكونات

في البداية، كانت قوائم المكونات في الوصفات موجودة كلها في **جملة واحدة** تشمل المكون، الكمية، والوحدة. عشان نسهّل عملية حساب **المعلومات الغذائية** لكل وصفة، قررنا نفصل هاي البيانات إلى **ثلاث حقول منفصلة**:

1. المكون (مثلاً: حليب).
2. الكمية (مثلاً: 2).
3. الوحدة (مثلاً: كوب).

عشان نحدد **أنماط الجمل** التي تحتوي على الترتيب المطلوب **(NLP)** بعد هيك، استخدمنا **أداة معالجة اللغة الطبيعية** (حرف جر) 'of' (وحدة) 'cups' ((الكمية - الوحدة - المكون). كمان، علمنا كل كلمة في الجملة (مثلاً: '2' (رقم (مكون) 'milk')).

سلايد 2: تنظيف البيانات وتطابق المكونات

بعد ما فصلنا المكونات، بدأنا نبحث عن **وحدات قابلة للقياس** (مثل الكوب، الملعقة) لأن بعض الوحدات ما بتكون قابلة للقياس (مثل "حفنة" أو "شريحة"). بعد هيك، بنينا **مجموعة من المكونات الفريدة** وقمنا بتطابق الأسماء مع قاعدة بيانات **وزارة الزراعة**

اللي بتحتوي على معلومات غذائية لأكثر من 8,000 نوع طعام. **(USDA) الأمريكية**
ثم، استخدمنا **الكلمات الأولى** في كل جملة مكون واكتشفنا أسماء المكونات المألوفة، مثل "عصير التفاح" أو "فلفل حار". في حالات معينة، إذا كانت الجملة مش واضحة أو كانت تحتوي على مكونات غير معروفة، أخذنا فقط الكلمة الأولى (مثل "سكر" أو ""ماء").

سلايد 3: المعلومات الغذائية والتصور
حصلنا على **50,637 وصفة** تحتوي على **معلومات غذائية كاملة**. بعدها، USDA، بعد ما جمعنا المكونات من قاعدة بيانات عشان نعرض كيف تكون **الوصفات متشابهه** أو مختلفة حسب **السمات الغذائية** (مثل **t-SNE** استخدمنا **التصور باستخدام** السكر، الدهون، والملح). عرضنا الوصفات باستخدام ألوان مختلفة عشان نحدد **الصحة** بناءً على محتويات الوصفة الغذائية، مثل (ما بنشوف في الشكل الأول (تصنيف الوصفات حسب النوع) وفي الشكل الثاني (تصنيف الوصفات حسب الصحة.

.

Data structure

. تركيب البيانات١

- **الطبقة الأولى (نص الوصفة):**
- اسم الطبق
- المقادير (مع الكميات لو موجودة)
- خطوات التحضير
- معلومات غذائية (سعرات، سكر، دهون) - مع نظام "إشارة المرور" للأكل الصحي

- **الطبقة الثانية (الصور):**
- 13 مليون صورة أكل مرتبطة بالوصفات
- بعض الوصفات مصنفة لنوع الوجبة (مقبلات، حلويات، إلخ)

. التمثيل البصري للبيانات (الشكل ٢٣)

- خريطة ذكاء اصطناعي تظهر علاقات الأطباق:
- مثلاً: أكلات الدجاج بتكون متقاربة مع بعض
- الحلويات بتكون في منطقة ثانية

. تحليل الصحة (الشكل ٣٤)

- الألوان بتعبر عن مدى صحة الوصفة:
  - أخضر = صحي
  - أحمر = مليان سكر/دهون
  - بيستخدم معايير بريطانية للتقييم

. ليش هادا مهم؟٤

- أول قاعدة بيانات تجمع:
  - الوصفة مكتوبة
  - صورتها
  - قيمتها الغذائية
- بتسمح للذكاء الاصطناعي:
  - يفهم إيش الأكلات المتشابهة
  - يعدل الوصفات عشان تصير أكثر صحية

Analysis

- فيه شوية تكرارات تقريبا 0.4%، وحوالي 20% من الوصفات عناوينها مش فريدة Dataset Recipe1M+ وتختلف بمتوسط 16 مكون.
- تقريباً نصف الوصفات ما كان فيها صور بالبداية، لكن بعد ما ضفنا بيانات جديدة، صار 2% بس من الوصفات اللي ما فيها صور.
- فيه 16 ألف مكون فريد بالوصفات، ومنهم 4 آلاف بيمثلوا 95% من التكرار.
- الوصفة المتوسطة بتتكون من 9 مكونات و 10 تعليمات، والصور بتختلف بشكل كبير، يعني بعض الوصفات الشهيرة عندها صور أكتر بكتير من غيرها.

الشريحة 2: تجربة مطابقة الصور

- بعد ما ضفنا البيانات، زاد عدد الوصفات اللي فيها صور من 333 ألف لـ1 مليون وصفة.
- Recipe1M+ فيه في المتوسط 13 صورة لكل وصفة.
- وطلبنا من العمال يختاروا أفضل صورة من اثنين (Amazon Mechanical Turk (AMT جربنا على منصة لوصفة معينة. اختاروا الصورة الأصلية 28.1%، والصورة المستخلصة 23.8%، و45.8% اختاروا الصورتين مع بعض.
- النتائج بتقول إن المجموعة الموسعة للبيانات فيها ضوضاء قليلة جداً مقارنة بالمجموعة الأصلية .

3 LEARNING EMBEDDINGS

الفكرة الأساسية .

نبي نعلم الذكاء الاصطناعي يفهم العلاقة بين:

- وصفة مكتوبة (المقادير + خطوات الطبخ)
- صورة الأكلة

عشان يقدر يربط بينهم، زي ما الإنسان يعرف إن صورة "باستا" تنتمي لوصفة معينة!

٢. معالجة الوصفات

المقادير:

- تفهم ترتيب المقادير حتى لو مكتوبة عشوائي (LSTM) بتتعالج بـ شبكة عصبية خاصة.
- عندنا نظام يستخرج اسم المكون الأساسي من الجملة (مثلاً: "كوب طحين" ← "طحين") بدقة ٩٩.٥٪.

طريقة الطبخ:

- كل خطوة بتتحول لـ "كود" يفهمها الآلة، وبتتدرس مع جيرانها (مثلاً: الخطوة اللي قبلها وبعدها).

٣. معالجة الصور

تتعرف على مكونات الصورة (ألوان، أشكال، ترتيب المكونات) بنستخدم شبكات جاهزة (مثل ResNet).

٤. الدمج بين الصور والنصوص

."النموذج بيجمع المعلومات من الجزئين (النص والصورة) ويحطهم في "خريطة مشتركة

:بيتم تدريبه على

- .يقرّب النقاط اللي بينتموا لنفس الأكلة
- .يبعد النقاط اللي ما بينتموش لبعض

.تحسين إضافي: تصنيف الأكل لأنواعه (حلويات، أكل بحري، إلخ) عشان يفهم العلاقات بشكل أعمق

. طريقة التدريب٥

:التدريب بيكون على مرحلتين

.. نثبت جزء الصور وندرب جزء النصوص١

.. نثبت النصوص وندرب جزء الصور٢

!الوقت المستغرق: ٣ أيام على ٤ كروت شاشة قوية

. الاختبار على البشر٦

!النموذج قدّ أداء البشر في إيجاد الوصفة المناسبة للصورة

.(بس لسه يعاني مع التفاصيل الدقيقة (مثلاً: التمييز بين أنواع السوشي المختلفة

Representation of Recipes

الشريحة 1: تمثيل الوصفات

1. **المكونات:**

a. كل وصفة فيها لستة مكونات.
b. "زيت الزيتون" ← "2 ملعقة كبيرة من زيت الزيتون" (مثال: "من النص، بنستخرج أسماء المكونات").
c. بنستخدم تقنية **Word2Vec** عشان نعمل تمثيل لكل مكون.
d. استخراج المكونات بيتم باستخدام **LSTM** بدقة 99.5% **ثنائي الاتجاه**.

2. **تعليمات الطبخ:**
a. التعليمات بتكون طويلة عادة (المتوسط 208 كلمات).
b. بنستخدم نموذج **LSTM** لمعالجة التعليمات **ذو مرحلتين:**
i. أول شي، بنمثل كل تعليم باستخدام متجه **"skip-instructions"**.
ii. بعدين بندرّب على تسلسل هاي المتجهات لتمثيل كل التعليمات **LSTM**.

الشريحة 2: "Skip-Instructions" والدمج

1. **"Skip-Instructions":**
a. بيعتمد على تقنية **Skip-Thoughts**، اللي بترمز الجمل وبتتوقع الجمل اللي قبلها أو بعدها.
b. التعديلات بتشمل إضافة رموز بداية ونهاية للوصفة واستخدام **LSTM** بدلاً من **GRU**.
c. كل تعليم بنحوله لـ **متجه ثابت الطول**.

2. **النموذج النهائي:**
a. بندخل الترميز النهائي للتعليمات في **النموذج المشترك للدمج** (joint embedding model) لتمريره لمرحلة تالية.

3.2 Representation of Food Images

١. طريقة تمثيل الصور

- بنستخدم نوعين من شبكات الذكاء الاصطناعي لتحليل الصور:
  - **VGG-16**: شبكة قديمة لكنها قوية، بتستخدم فلاتر صغيرة لتحليل الصور.
  - **ResNet-50**: شبكة أحدث، فيها "قفزات ذكية" (residual connections) تسمح بيها تتدرب بشكل أعمق وأدق.

٢. ليش ResNet أحسن؟

- بتتجنب مشكلة **ضعف الإشارة** في الشبكات العميقة (اللي بتخلي الأداء يضعف كل ما زاد عدد الطبقات).
- مجربة ومضمونة: حققت نتائج قياسية في مسابقات مثل ImageNet.

. كيف بتندمج مع باقي النموذج؟٣

- .(بنشيل آخر طبقة (المصنفة) عشان نستخرج **خصائص عامة** من الصورة (مثل القوام والألوان
- .الخصائص دي بتتحول لـ "أرقام" (متجهات) تدخل للنموذج المشترك اللي يفهم العلاقة بين الصورة والوصفة

. ليه هادا مهم؟ ٤

- .(بقدر يفرق حتى بين التفاصيل الدقيقة (مثل طريقة التقطيع أو الشوي ResNet **دقة أعلى**:
- .**توفير وقت**: لأنها متدربة مسبقًا على ملايين الصور العامة، فبتتعلم بسرعة لما نطبقها على الأكل

JOINT NEURAL EMBEDDING

الشريحة 1: الهدف من الدمج العصبي المشترك

إحنا هنا بنحاول نعمل نموذج يجمع بين الوصفة والصورة ويقدر يربط بينهم. يعني لو عندك صورة أكلة، النموذج هيقدر يعرف شو هي الوصفة الخاصة بها. كيف؟ من خلال:

وهو نوع من الشبكات العصبية) علشان نتعامل مع المكونات. كل مكون مثل "زيت) LSTM تمثيل المكونات: بنستخدم زيتون" يتم معالجته بطريقة معينة علشان نعرفه بشكل دقيق.

تاني LSTM تمثيل التعليمات: بعدين بنحتاج نفهم كيفية تحضير الأكلة، فبنعمل تمثيل خاص للتعليمات من خلال استخدام علشان نعرف كيف يتم الطهي بالضبط.

الشريحة 2: تمثيل الوصفة والصورة

المكونات والتعليمات: الوصفة عبارة عن مكونات (مثل "2 كوب دقيق") وتعليمات (مثل "اخلط المكونات"). بنحول هاي المكونات والتعليمات لتمثيلات رياضية باستخدام تقنيات مثل word2vec يعرفها النموذج علشان.

فضاء مشترك بين الوصفة والصورة: بعد ما نحول المكونات والتعليمات لتمثيلات، بنحتاج نحطهم في فضاء مشترك مع الصورة. يعني لو الصورة لطبق معين، نحاول نعمل تمثيل مشترك للطبق مع الوصفة والصورة بحيث يرتبطوا ببعض.

الشريحة 3: التدريب على النموذج

هدف التدريب: إحنا بدنا نعلم النموذج كيف يميز بين الوصفات الصحيحة والصور المناسبة لها. بنستخدم التشابه الكوني علشان نقيس إذا الصورة متطابقة مع الوصفة أو لا.

دالة الخسارة: النموذج يتعلم عن طريق تقليل الفرق بين تمثيل الوصفة والصورة المطابقة لها (بزيادة التشابه بيناتهم) وفي نفس الوقت التقليل من التشابه بين الوصفات والصور غير المطابقة.

ببساطة، النموذج بيمرّ بتدريب علشان يتعلم يربط بين الصور والوصفات الصحيحة5

5_ SEMANTIC REGULARIZATION

. الفكرة الأساسية١

- عشان نجعل النموذج يفهم العلاقة بين الصور والوصفات بشكل أعمق، خلقنا **تصنيف مشترك** للأكل.

النموذج إجباري يتعلم إنه:

- صورة "بيتزا" ووصفة "بيتزا" لازم يعطوا نفس التصنيف.
- التصنيفات المشتركة بتخلق "لغة موحدة" بين النص والصورة.

. أنواع التصنيفات٢

- استخدمنا مصدرين:

غطت ١٣٪ من البيانات. - (Food-101 ١. **قوائم جاهزة** (مثل

."). ٢. **كلمات متكررة** في عناوين الوصفات (مثل "سلطة دجاج"، "خضار مشوي

- ."بعد التنظيف: عندنا ١,٠٤٧ نوع أكل يغطي ٥٠٪ من الوصفات (الباقي مكتوب "خلفية

. كيف اشتغلت؟٣

- .أضفنا **طبقة تصنيف واحدة** للنصوص والصور معًا (في الشكل ٦

- النتيجة:

• تحسن دقة البحث (مثلاً: إيجاد الوصفة من الصورة).

- .(!• صار ممكن نعمل "عمليات حسابية" على الأكل (مثل: بيتزا - جبنة + فواكه = فطيرة فواكه

. التحديات٤

- .بعض الوصفات ما عندها تصنيف واضح

- (بعض العناوين ممكن تنتمي لأكثر من نوع (بنختار الأكثر شيوعًا

Classification & Semantic Regularization

الفكرة بشكل عام هي تحسين طريقة التمثيل الرياضي للوصفات والصور المرتبطة فيها. الهدف هو إنه نخلي الوصفات والصور اللي بتتعلق مع بعض بشكل دلالي (يعني تكون مرتبطة من حيث المعنى) تكون قريبة من بعض .في الفضاء الرياضي اللي بنمثل فيه البيانات

:الطريقة

تمثيل الوصفة: بنستخدم نوع من الشبكات العصبية لتمثيل الوصفة بناءً على مكونين رئيسيين: المكونات (المقادير) والتعليمات (طريقة التحضير).

تمثيل الصورة: كمان بنمثل الصورة باستخدام شبكة عصبية تانية.

دمج التمثيلات: بعد ما نكون مثلنا الوصفة والصورة كل واحدة على حدة، بندمجهم مع بعض في "فضاء مشترك" بحيث تكون الصور اللي مرتبطة بوصفات معينة قريبة من بعضها.

التدريب: الهدف هو إنه نخلي النموذج يتعلم كيف يربط بين الصورة والوصفة، بحيث الصور المرتبطة بوصفات معينة تكون قريبة في الفضاء الرياضي.

التصنيف: بنضيف خطوة أخيرة للتأكد من أنه النموذج يستطيع تصنيف الوصفات والصور بشكل صحيح بناءً على التشابه الدلالي بيناتهم.

الهدف النهائي هو تحسين قدرة النموذج على فهم العلاقة بين الوصفات والصور بشكل أفضل.

Optimization

- **التدريب على مرحلتين:**
- أولاً، ندرب شبكة الوصفات مع تثبيت شبكة الصور (اللي تم تدريبها مسبقًا على مهمة تصنيف الصور في ImageNet).
- بعدها، ندرب شبكة الصور مع تثبيت شبكة الوصفات.
- **السبب:** هاي الطريقة بتمنع حدوث تذبذب أو عدم استقرار في التدريب المشترك.
- **التدريب النهائي:** بعد التمرين الأولي، نعيد تدريب الشبكتين مع بعض، مع تغييرات قليلة.
- **التحقق من الأداء (MedR):** نستخدم **ترتيب الميديان** عشان نضمن استقرار الأداء.

6.1Implementation Details

طريقة التقييم .

- المهمة: نعطي النموذج **صورة أكلة** ويجب يلاقي **وصفتها** من بين ألف وصفة (أو العكس).
- معايير النجاح:
- **MedR**: متوسط ترتيب الوصفة الصحيحة (كلما قل أفضل).
- **R@K**: نسبة الصور اللي الوصفة الصحيحة تظهر ضمن أول K نتيجة.

. مقارنة مع الطرق القديمة٢

- كانت ضعيفة: (CCA النماذج التقليدية (مثل

يعني الوصفة الصحيحة في المتوسط تكون رقم 25 في القائمة!).) MedR = 25 • أسوأ نموذج:

.MedR = 15.7 • أحسن نموذج قديم:

- .المشكلة: الطرق دي بتلخص الوصفة في "متوسط كلمات" فتفقد التفاصيل

. أداء نموذجنا٣

- .(الوصفة الصحيحة في المركز الخامس في المتوسط) MedR = 5.2 **نتيجة مذهلة**:
- .**دقة 65%** في إيجاد الوصفة ضمن أول 10 نتائج
- .من 7.2 لـ5.2 MedR ساعد يقلل (semantic regularization)**التصنيف الداعم**

. نجاحات وإخفاقات٤

- .(**ينجح مع**: أكلات شكلها مميز (مثل الكيك أو السلطات
- **يخفق مع**:

• المكونات المخفية (مثل اللحمة تحت صوص اللازانيا).

• المكونات المتشابهة (مثل الجمبري والسلمون).

. مقارنة مع أحدث الأبحاث٥

- .(!الوصفة الصحيحة دايماً أول نتيجة) MedR = 1 النماذج اللي بعدنا (مثل [19]) وصلت لـ

(Ablation Study) . اختبار قطع الغيار١

- VGG-16: **تفوز** على **ResNet-50 شبكة**
  - .(مقابل 15.3 MedR 7.9)حققت **خطأ أقل بمرتين** ResNet
  - !"ليش؟ لأنها بتقدر تفرق حتى بين "دجاج مشوي" و"دجاج مقلي

- **التصنيف الداعم (Semantic Regularization)**:
  - خفض الخطأ من 7.2 لـ5.2 - يعني الوصفة الصحيحة بتكون في **المتوسط رقم 5** بدل رقم 7!

٢. الإنسان ضد الآلة .

- **الذكاء الاصطناعي vs البشر**:
  - في المهام السهلة (مثل تمييز "بيتزا" من "سلطة")، الآلة **تغلب البشر** (84.8% vs 81.6%).
  - لكن في التفاصيل الدقيقة (مثل أنواع السوشي أو العصائر)، لسة **البشر أحسن**.
- **مشاكل النموذج**:
  - ما بقدر يشوف المكونات المخبية (مثل اللحمة تحت الجبنة في اللازانيا).
  - بيفوت بين المكونات المتشابهة (مثل الجمبري والسلمون).

٣. الفرق بين Recipe1M و Recipe1M+

- **Recipe1M+ أحسن بكتير**:
  - بتحسن الدقة **بـ13%** (R@10: 76.3% من 67.5%).
  - قللت الخطأ **بـ45%**، Food-101 حتى في مجموعة!
- **ليش؟** (لأنها بتحتوي على **13 مليون صورة** بدل 800 ألف في النسخة القديمة).

٤. ليش هادا مهم؟ .

- **لأصحاب المطاعم!**: تقدر تاخذ صورة من الزبون وتلاقي الوصفة تلقائيًا
- **للمهتمين بالصحة.**: تقدر تبحث عن وصفات بناءً على قيمتها الغذائية
- **للمطورين.**: أول مجموعة بيانات كبيرة تجمع بين الصور، الوصفات، والمعلومات الغذائية

.

Analysis of the Learned Embedding

بكل بساطة، في هذا الجزء من البحث، الناس كانوا يحاولوا يتعلموا كيف الربط بين الصور والوصفات الغذائية يعمل بشكل دلالي، يعني كيف يمكن للكمبيوتر فهم العلاقة بين صورة لطبق وطريقة تحضيره.

الشرح بالتفصيل:

1. **تنشيط الخلايا العصبية**:
   في الشبكة العصبية، كل خلية (أو "وحدة") بتستجيب بطريقة معينة حسب المعلومات المدخلة. هنا، الباحثين كانوا بيشوفوا إذا كان في خلايا عصبية "تنشط" أو تفاعل مع مفاهيم معينة زي المكونات أو تعليمات الطبخ. لو لاحظوا أن الخلايا العصبية تتفاعل مع صورة أو وصفة معينة بطريقة "دلالية" (يعني تتفاعل مع المفاهيم مثل "دجاج" أو "سلطة")، هذا بيكون دليل على أن النموذج فاهم هذه المفاهيم.
2. **العمليات الحسابية على المتجهات**:
   المتجهات هنا بتكون تمثيلات رقمية لوصفات وصور. الباحثين استخدموا عمليات حسابية على هذه المتجهات زي الجمع والطرح علشان يتأكدوا إذا النموذج فاهم العلاقة بين المكونات. مثلاً، لو عندك "بيتزا دجاج" و "بيتزا"، المفروض النتيجة تكون "دجاج سلطة" لو استخدموا العملية الحسابية بشكل صحيح.
3. **العمليات الحسابية الجزئية**:
   هذه فكرة أنه بدل ما يكون عندك "طبق" واحد كامل، ممكن تدمج جزئين مع بعض وتعمل عملية حسابية على المتجهات مثل مثلاً إذا كان عندك "معكرونة" و "سلطة"، وتلعب مع المعاملات (مثل تغيير النسبة بين المعكرونة والسلطة)، فالنموذج لازم يقدر يظهر لك طبق مختلف بناءً على كيف قمت بتغيير النسبة بين المفاهيم.

باختصار:

النموذج الي كانوا بيشتغلوا عليه هو عبارة عن شبكة عصبية تعلمت كيف تربط بين الصور والوصفات الدلالية (اللي هي المفاهيم زي مكونات الأكل) بطريقة رياضية، ويقدر يعمل عمليات حسابية على هذه المفاهيم ليفهمها بشكل أفضل.

الخاتمة

الحساب الكسري في مساحة التضمين (الشكل ١٢ .)

- **الفكرة:**
  - !تخيل إنك تقدر تخلط وصفتين مع بعض بنسب مختلفة، زي ما تخلط عصير برتقال وفراولة
  - :هنا النموذج بيعمل **"خلط رقمي"** بين وصفات (مثل سلطة وباستا) ويطلع نتائج منطقية

- لو $x=0$ $x=0$: **باستا** ١٠٠٪ النتيجة بتكون.
- لو $x=0.5$ $x=0.5$: باستا بالسلطة ("خليط متساوي (مثل").
- لو $x=1$ $x=1$: **سلطة** ١٠٠٪.

- **الفرق بين Recipe1M و Recipe1M+**:
  - النموذج المدرب على **Recipe1M+** بيعطي نتائج أوضح، لأنه اتعلم على بيانات أكثر وأدق.
  - مثال: لو بدك تحول "تاكو" لـ "سلطة" عن طريق استبدال "التورتيلا" بـ "الخس"، النموذج الجديد بيفهمها أحسن!

الخلاصة (القسم ٢٧).

- **الإنجازات**:

١. مجموعة **Recipe1M+** هي الأضخم في العالم (مليون وصفة + ١٣ مليون صورة).

٢. النموذج حقق **دقة قريبة من البشر** في إيجاد الوصفة من الصورة.

٣. البيانات الجديدة **ما فيها ضوضاء كتير** (رغم أنها من الإنترنت).

- **تطبيقات مستقبلية**:
  - **تعديل الوصفات**: مثلاً، تخلي الطبق أخف أو أسرع طبخًا.
  - **توليد وصفات جديدة** من الصور!
  - استخدام التقنية في مجالات تانية زي:
    - تعليمات تركيب الأثاث.
    - برامج تعليمية.

لماذا هذا مثير؟

- النموذج بقدر **يبدع في المطبخ رقمياً**، مثل ما يعمل "خلط" بين الوصفات!
- ممكن يوفر وقتك لما تبحث عن وصفة بتعتمد على الصور اللي عندك.
- (ليس فقط الطبخ) **يفهم أي إجراءات متسلسلة** يفتح الباب لذكاء اصطناعي.

"!مستقبل الطهي رح يكون بيد الآلة، وإحنا بس بنقولها إيش بدنا نأكل"

(فكرة "الخلط الرقمي" بين الوصفات (الشكل ١٢

- **إزاي بشتغل؟**
  - النموذج بقدر **يمزج بين وصفتين** بنسب مختلفة، زي ما تخلط حليب وفراولة عشان تسوي ميلك شيك!
  - مثال:
    - لو قلنا "٧٠٪ باستا + ٣٠٪ سلطة"، النموذج بيرجع وصفة **باستا باردة** فيها خضار.
    - لو غيرنا النسب، النتيجة بتتغير كأنك بتعدل على "عيار الخلاط"!
- **الفرق بين النسختين:**
  - النموذج اللي اتدرب على **Recipe1M+**) النسخة المطورة) بيعطي **نتائج أدق،** لأنه شاف صور ووصفات أكتر.
  - مثلاً: لو بدك تحول "برجر" لـ "سلطة"، النسخة الجديدة بتفهم إنك بدك تستبدل الخبز بورق الخس!

. التطبيقات المستقبلية٢

- **تعديل الوصفات:**
  - تقدر تطلب من النموذج: "بدي وصفة شوربة بس بأقل سعرات حرارية".
- **إبداع في الطبخ:**
  - النموذج يقترح عليك وصفات جديدة من مزج أفكارك (مثلاً: "كيكة الجبن + كنافة!").
- **استخدامات أبعد من المطبخ:**
  - ممكن تطبق نفس الفكرة على:
    - تعليمات تركيب الأثاث (تعديل الخطوات حسب الأدوات المتوفرة).
    - صناعة الروبوتات (توقع كيف رح تطلع الحركة قبل ما تعملها).

ليش هادا مثير للاهتمام؟

- النموذج مش بس **يفهم** الوصفات، يقدر **يبدع فيها** !

- ممكن يوفر وقتك ويخليك تكتشف أكلات جديدة بلمسة زر.
- يفتح أبواب لذكاء اصطناعي **"مبدع"** في مجالات كثيرة غير الطبخ.

"