RITSUMEIKAN

TOP GLOBAL UNIVERSITY JAPAN

Discovering Fashion Trends through Multimodal Image Captioning and Topic Modeling

Beyond Borders

ICT for Human Enhancement

Agenda

- Background
- Related Work
- Proposed Approach
- Experiments
- Results
- Conclusions

Background

- Present Scenario in Fashion Trends:
  - Fashion trends evolve rapidly, influenced by social media and cultural shifts
- Current Approaches:
  - Time series forecasting
  - CNNs for image classification
- Issues in the Current Approaches:
  - Traditional methods do not differentiate between new and current fashion trends

Research Goal

- Forecast future fashion trends based on present visual data
- Following technology needs to be applied
  - **Detailed Image Captioning:** Generate captions capturing visual and textual attributes
  - **Theme Discovery:** Use topic modeling to discover themes in fashion trends
  - **Trend Insights:** Provide trend modeling into the evolution and influencers of fashion trends

Related Work: Image Captioning

- Xuewen et al., (2022) Fashion captioning:
  - Used a novel LSTM-based encoder-decoder framework for expressive fashion captioning
- Chen et al., (2022) Attribute conditioned fashion image captioning:
  - Proposed a multi-modal method using semantic attributes for caption generation, tested on the FACAD170k dataset

RITSUMEIKAN
TOP GLOBAL UNIVERSITY JAPAN

Related Work: Image Topic Modeling

- Kyeong et al., (2024) See, caption, cluster: Large-scale image analysis using caption and topic modeling:
  - Proposed a new approach to image topic modeling via image captioning
  - Improved topic modeling analysis with t-SNE visualization techniques

Beyond Borders

ICT for Human Enhancement

Proposed Approach

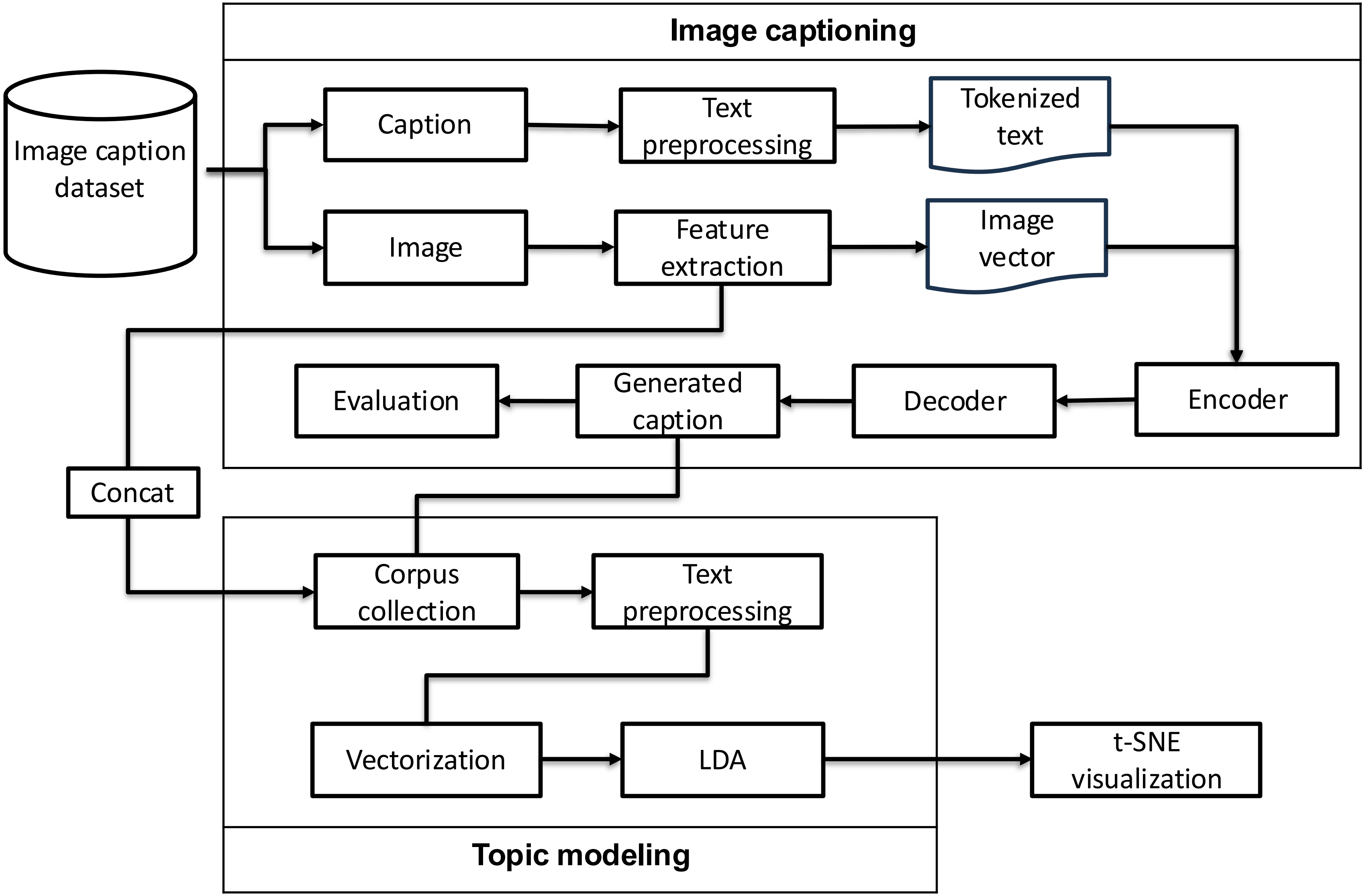
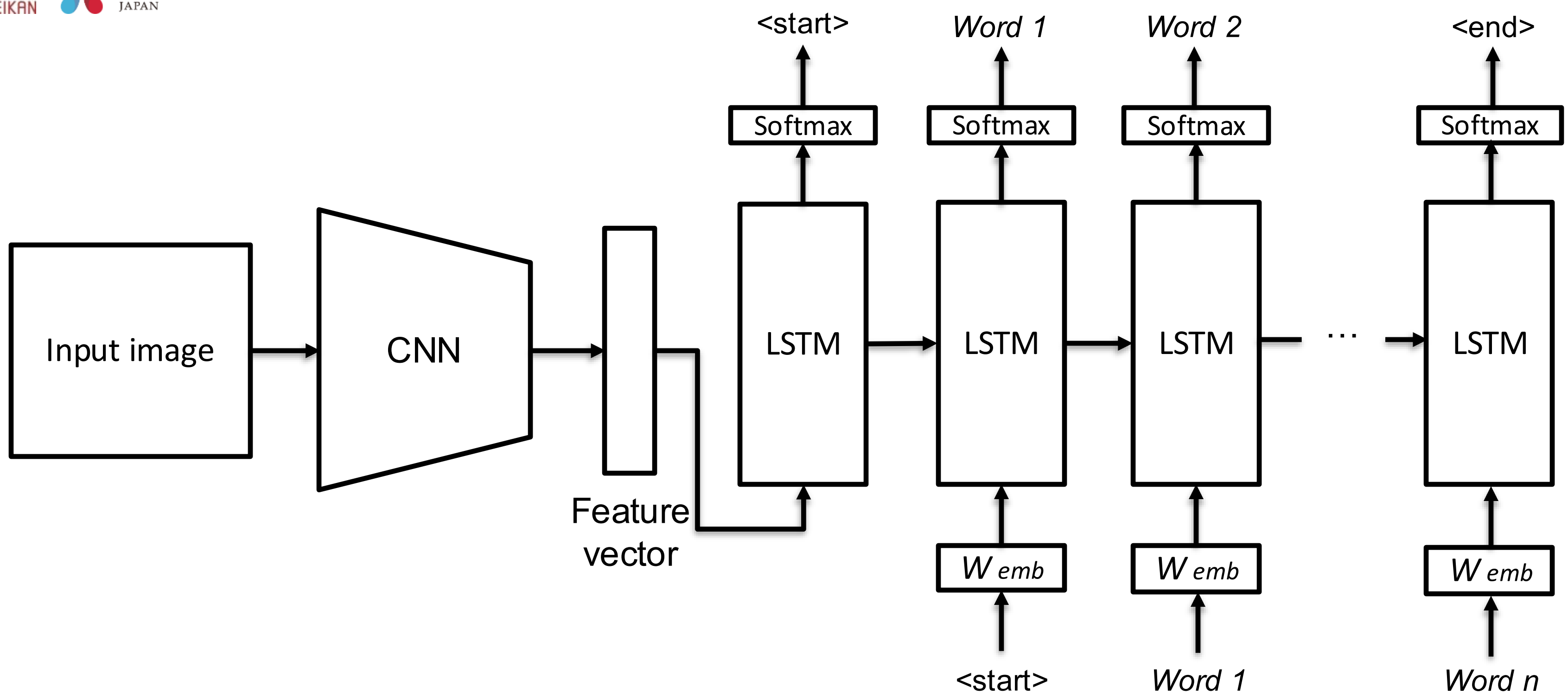


Image Captioning Architecture



- Encoder
  - Pre-train the DenseNet201 model
- Decoder
  - The LSTM (Long-Short Term Memory) network

Experiments

1. LSTM encoder-decoder framework
   - Processes input into a fixed-size context vector for output generation

2. Attention-based LSTM encoder-decoder framework
   - Uses attention to focus on specific input parts, improving long-sequence handling

Experiments: Dataset

- DeepFashion-MultiModal:
  - A dataset with 44,096 high-resolution human images, including 12,701 full-body images
  - Textual descriptions accompany each image, with data split into 80% training and 20% testing
  - 5-fold cross-validation is applied
  - Source: https://github.com/yumingj/DeepFashion-MultiModal

Evaluation Metric

- BiLingual Evaluation Understudy (BLEU) metric:
  - Cumulative N-gram scores
  - Formula:

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} w_n \log P_n\right) \quad BP = \begin{cases} 1 & if\ c > r \\ e^{\left(1-\frac{r}{c}\right)} & if\ c \leq r \end{cases}$$

    - BP : Brevity Penalty
    - w_n : Weight for the n-gram precision
    - P_n : Precision for n-grams
    - c : The length of the candidate translation
    - r : The length of the reference translation

Results: Model Loss and SE

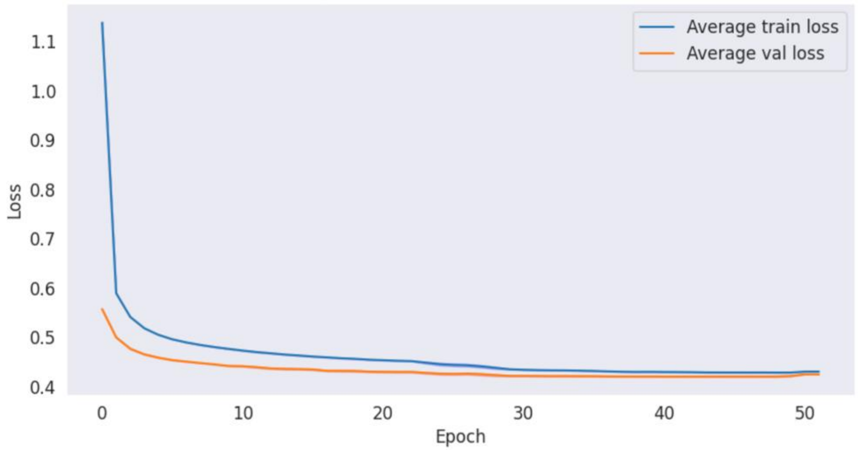


Fig 1. Experiment 1 Average model loss with standard error

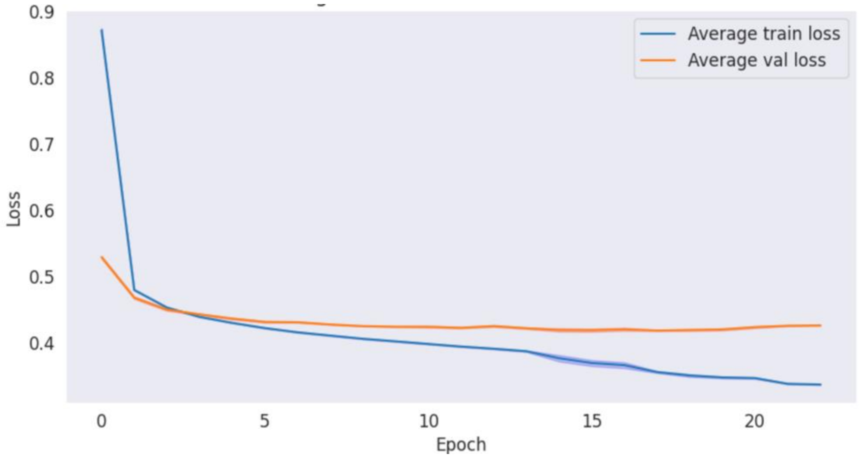


Fig 2. Experiment 2 Average model loss with standard error

	Experiment 1		Experiment 2	
	Standard Error for Train Loss	Standard Error for Validation Loss	Standard Error for Train Loss	Standard Error for Validation Loss
Fold 1	0.0134152087778	0.0032698838603	0.0214774532411	0.0047631773175
Fold 2	0.0147427840058	0.0034458647783	0.0237168522338	0.0056338693359
Fold 3	0.0165403906054	0.0038279949171	0.0214543406312	0.0049878685657
Fold 4	0.0141054721086	0.0034386553075	0.0246567067385	0.0057161718295
Fold 5	0.0142577491675	0.0034117771355	0.0242290208128	0.0053907884398

Results: BLEU Evaluation

	Aggregate 1-gram BLEU score	Aggregate 2-gram BLEU score	Aggregate 3-gram BLEU score	Aggregate 4-gram BLEU score
Experiment 1	0.565223204	0.417449557	0.337129166	0.290864990
Experiment 2	0.574543203	0.422793542	0.341547735	0.294696234

- Results indicate high performance of the model
- Experiment 2 shows a slight improvement over Experiment 1 across all BLEU metrics

Results: Example 1

Actual Caption	The upper clothing has **long sleeves**, **cotton fabric** and **graphic patterns**. The neckline of it is **crew**. The lower clothing is of **three-point length**. The fabric is **cotton** and it has **graphic patterns**. There is an **accessory** on her wrist. This person wears a **ring**.
Experiment 1 Predicted caption (LSTM encoder-decoder model)	startseq the upper clothing has **long sleeves cotton fabric** and **graphic patterns** it has **round** neckline the lower clothing is of **threepoint length** the fabric is **cotton** and it has **graphic patterns** there is an **accessory** on her wrist there is **ring** on her finger endseq
Experiment 2 predicted caption (Attention-based LSTM encoder-decoder model)	startseq the **shirt** this person wears has **long sleeves** and it is with **cotton fabric** and **graphic patterns** the neckline of the shirt is **crew** this person wears **threepoint** shorts with **cotton fabric** and **pure color patterns** there is an **accessory** on her wrist there is **ring** on her finger endseq

Beyond Borders

ICT for Human Enhancement

Results: Example 2

Actual Caption	His **sweater** has **long sleeves**, **cotton fabric** and **solid color patterns**. The neckline of it is **round**. This gentleman wears a **long trousers**. The trousers are with **denim fabric** and **lattice patterns**. The **outer clothing** this man wears is with **cotton fabric** and **pure color patterns**.
Experiment 1 Predicted caption	startseq the upper clothing has **long sleeves cotton fabric** and **pure color patterns** it has **round** neckline the lower clothing is of long length the fabric is **cotton** and it has **pure color patterns** the **outer clothing** is with **cotton fabric** and **solid color patterns** endseq
Experiment 2 predicted caption	startseq the **sweater** this man wears has **long sleeves** and it is with **cotton fabric** and **solid color patterns** the neckline of the sweater is **crew** this man wears **long trousers** with **denim fabric** and **solid color patterns** the **outer clothing** this man wears is with **cotton fabric** and **pure color patterns** endseq

Conclusions

- The proposed approach combining image captioning and topic modeling demonstrates significant potential in identifying fashion trends
- The experiments resulted in high-model performance
- **Limitation:** The model did not generate caption address some detail fashion attributes
- **Future work:** Enhancing image topic modeling by incorporating both images and captions will be focused

Key Parameters and Settings

- Optimizer: **Adam optimizer**
- Loss Function: **Categorical cross-entropy**
- Batch Size: **32**
- Callbacks: **ModelCheckpoint, EarlyStopping, ReduceLROnPlateau**
- Vocabulary size: **106**

Formulae

- Standard Error Calculation

$$SE = \frac{\sigma}{\sqrt{n}}$$

  - σ : The standard deviation of the model loss
  - n : The number of observations (folds)

- Standard Deviation Calculation

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  - n : The number of observations
  - x_i : Each individual observation
  - $\bar{x}$: Mean of the observations

Example 1: BLEU Score

	Experiment 1	Experiment 2
Aggregate 1-gram BLEU score	0.5869565217391	0.4313725490196
Aggregate 2-gram BLEU score	0.4978213403988	0.3217598666159
Aggregate 3-gram BLEU score	0.4324654180905	0.2227749892306
Aggregate 4-gram BLEU score	0.3584256720161	0.1602984865564

Example 2: BLEU Score

	Experiment 1	Experiment 2
Aggregate 1-gram BLEU score	0.4999999999999	0.5849056603773
Aggregate 2-gram BLEU score	0.3651483716701	0.4860163590932
Aggregate 3-gram BLEU score	0.2930286576063	0.4147644670645
Aggregate 4-gram BLEU score	0.2411652393904	0.3593424504128