Artwork Interpretation work-in-progress presentation draft

**Page 1**

Hi everyone, I am Zixun Wu. My supervisor and co-supervisor are Doctor Jeyhan Lau and Doctor Kris Ehinger. Today I will give the presentation regarding my research topic: *Artwork Interpretation*.

**Page 2**

The purpose of this research is to generate high-quality descriptions of artwork images by adapting image captioning models. Image captioning is the progress of producing the description of the content of an image.

The right picture is an example, the green box are descriptions of the above image. **(不翻页)**

**Page 3**

In recent years, A large body of literature has investigated on image captioning. most of them are designed for natural images **(click)** natural images are like the photographs from real life, limited studies have been made in the art domain. There are three major challenges of adapting image captioning models for natural images to artwork images, and make it a worth-research topic.

**(click)** the first challenge. We knowThe objects play an important role in the description of an image, so recognizing the objects is a crucial part.

But objects recognition is harder for artwork images than natural images.

**(click** to show the first two pictures) In these two natural images, We can see the objects are obvious, thus easy to recognize.

**(click** to show 3rd picture) But how about this artwork image? these objects shows strong **individualities** and are difficult to represent through **generalizable** patterns.

**Page 4**

The second challenge is different artworks express various aesthetics and various styles.

As we all know, there are different artistic schools, such as realism, impression. Different schools have their own styles of expression.

For example, both of them depict a mountain, but the appearance of them differ a lot. Therefore, the model should have the ability of encoding different styles of artworks. This is the second challenge.

(2:30)

**Page 5**

The third challenge is descriptions of artworks need to contain high-level semantics information.

(1st click to show the image)Let’s look at an example for clarity

(2nd click to show two boxes) This is an artwork image

Here are two types of descriptions. The yellow box describes what **depicts** in the image, while the green box provides some high-level understanding beyond the content of image.

For natural image, the first type is enough, but the second type is also expected for artwork image.

**Page 6**

Most of image captioning models are based on encoder-decoder architecture. The structure of encoder-decoder architecture is shown in the picture. It consists of two parts: encoder and decoder. Encoder inputs the image and outputs the image representation, this step is called visual encoding Then feed the image representation to the decoder and decoder generates the descriptions.

As the first step of image captioning, visual encoding plays a significant influence.

**Page 9**

There are three types of visual encoding: global representation, grid-based representation and detector-based representation.

**(click to show 1st image)**global representation: represent an image with a single vector, so it suffers from **excessive** compression and lack of **granularity.**

To solve the shortcoming, grid-based representation is proposed.**(click to show 2nd image)**It splits an image into equally-sized grid, each grid is represented by a feature vector. In this way, the decoder can only focus on relevant grids when generating each word of description. Which increases the granularity. Also grid-based representation has compactness since all grids cover the whole image.

**(click to show 3rd image)** The last one, detector-based representation. Firstly, An object detection model is used to extract several **salient** regions, then represent the image by the feature vectors of all salient regions.

(5:04) (5:14)

Since most salient regions contain objects, detector-based representation can provide object-level information. Additionally, the object detection model can provide the position of individual object, so it performs well on **localization** of individual objects.

Due to these advantages of detector-based representations, many recent image captioning chooses to use it. Thus The detector-based representation becomes the de-facto standard.

Although detector-based representation perform quite well for natural images, **(click)** it is more sensitive with the first two challenges of artworks than grid-based ones. which are objects in artworks are harder to recognize, and artworks have various styles.

**Page 10**

Therefore it’s essential to investigate which kind of visual encoding is suitable for artwork images in image captioning task.

And this is our first research question.

Grid-based and detector-based representation have their own advantages.

Then we will explore how to combine and utilize both kinds of visual encoding techniques?

(6:47)

**Page 11**

For the methodology, our purpose is to combine different visual encoding techniques for encoding artwork images.

Firstly, we should figure out the **feasibility** of grid-based representation and detector-based representation on artwork images.

In the experiment setup, we choose two models that use grid-based and detector-based representation respectively which are meshed-memory transformer and RSTNet.

The SemArt dataset, a dataset consists of artwork images and their descriptions is used.

In this step, We will investigate two things: 1) the performance of each visual encoding on natural images and artwork images. and 2) the performance on each school of images for each model.

The next step is to explore how to combine both of visual encoding techniques, this step should depend on result of the first step.

The study of Luo gives us a candidate direction. In their approach, alignment graph is used to incorporate two types of visual encoding for natural image.

(8:23)

**Page 12**

the artwork often comes with some attributes, such as type, school. These attributes contain rich semantic cues.

**(click**) for example ,school indicates the style of artistic expression.

**(click)** Which infers our last research question: how to incorporate attributes into the image captioning model.

**Page 13**

we have two candidate approach. The picture shows the brief structure of first one.

Without attributes, the generated words come from the output embedding. But now, the attribute embedding also contributes to the generated words.

Not every word needs the information of attributes such as the functional words (of, the, to)

so there is a controller parameter alpha which controls how much attention should be focused on the attribute when generating current word,

**Page 14**

In the The second candidate approach. The rectangle part is the normal image captioning pipeline of encoder-decoder architecture.

We establish a knowledge graph by images and their attributes. the knowledge graph can provide the contextual embedding of one image. The contextual embedding contains the information of attributes and relationship between attributes. Then we minimize the difference between contextual embedding and visual embedding. In such way, the encoder could learn to generate visual embedding involving the contextual information.

**Page 15**

This is the timeline of my research project.

Page 14

All right. That’s all of my presentation. Thanks for listening and watching.