How to write the introduction

for a research paper

in three easy steps

Kate Saenko

Introduction

This talk will show how to write an introduction to a research paper in 3 easy steps

- STEP1: make a plan
- STEP2: write!
- STEP3: edit edit edit

Step 1: make a plan

Step 1: make a plan

Write down **in bullets:**

- your contribution
- what problem is it solving?
- what is the evaluation that proves it?

Step 1: make a plan

example

Contribution:

- Propose dense caption task
- Design new end-to-end model

Problem:

- Current methods fail...
- No one has tried this...

Evaluation:

- Compare to baseline X on dataset Y
- Tease a promising result

Kate Saenko

DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Justin Johnson* Andrej Karpathy* Li Fei-Fei
Department of Computer Science, Stanford University
{jcjohns,karpathy,feifeili}@cs.stanford.edu

Abstract

We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The dense captioning task generalizes object detection when the descriptions consist of a single word, and Image Captioning when one predicted region covers the full image. To address the localization and description task jointly we propose a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization. The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences. We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.

1. Introduction

Our ability to effortlessly point out and describe all aspects of an image relies on a strong semantic understanding of a visual scene and all of its elements. However, despite numerous potential applications, this ability remains a challenge for our state of the art visual recognition systems. In the last few years there has been significant progress in image classification [39, 26, 53, 45], where the task is to assign one label to an image. Further work has pushed these advances along two orthogonal directions: First, rapid progress in object detection [40, 14, 46] has identified models that efficiently identify and label multiple salient regions of an image. Second, recent advances in image captioning [3, 32, 21, 49, 51, 8, 4] have expanded the complexity of the label space from a fixed set of categories to sequence of words able to express significantly richer concepts.

However, despite encouraging progress along the label density and label complexity axes, these two directions have remained separate. In this work we take a step towards unifying these two inter-connected tasks into one joint framework. First, we introduce the dense captioning task (see Figure 1), which requires a model to predict a set of descriptions across regions of an image. Object detection is hence recovered as a special case when the target labels consist of one word, and image captioning is recovered when all images consist of one region that spans the full image.

Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

Additionally, we develop a Fully Convolutional Localization Network (FCLN) for the dense captioning task. Our model is inspired by recent work in image captioning [49, 21, 32, 8, 4] in that it is composed of a Convolutional Neural Network and a Recurrent Neural Network language model. However, drawing on work in object detection [38], our second core contribution is to introduce a new dense localization layer. This layer is fully differentiable and can be inserted into any neural network that processes images to enable region-level training and predictions. Internally, the localization layer predicts a set of regions of interest in the image and then uses bilinear interpolation [19, 16] to smoothly crop the activations in each region.

We evaluate the model on the large-scale Visual Genome dataset, which contains 94,000 images and 4,100,000 region captions. Our results show both performance and speed improvements over approaches based on previous state of the art. We make our code and data publicly available to support further progress on the dense captioning task.

* Both authors contributed equally to this work.

1

https://cs.stanford.edu/people/karpathy/densecap.pdf

One idea is better than many!

Good: reader has an “aha!” moment

Bad: too many things at once

Step 1: expand plan

Re-write bullets into following structure*:

- task
- state-of-the-art
- flaw in state-of-the-art
- your idea/solution
- proof it works

*this is usually easier to do for beginners than starting with this structure

Step 1: expand plan Example

title

abstract

figure 1

task

SotA

flaw

idea

proof

DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Justin Johnson* Andrej Karpathy* Li Fei-Fei
Department of Computer Science, Stanford University
{jcjohns,karpathy,feifeili}@cs.stanford.edu

Abstract

We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient [illegible]anguage. The dense caption-[illegible]tection when the descriptions [illegible] Image Captioning when one [illegible]ll image. To address the local-[illegible]ointly we propose a Fully Con-[illegible]ork (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization. The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences. We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.

1. Introduction

[illegible]essly point out and describe all aspects [illegible]n a strong semantic understanding of a [illegible] of its elements. However, despite nu-[illegible]plications, this ability remains a chal-[illegible]of the art visual recognition systems. [illegible]In the last few years there has been significant progress in image classification [39, 26, 53, 45], where the task is [illegible]to an image. Further work has pushed [illegible]g two orthogonal directions: First, rapid [illegible]tection [40, 14, 46] has identified mod-[illegible]entify and label multiple salient regions [illegible]d, recent advances in image captioning [3, 32, 21, 49, 51, 8, 4] have expanded the complexity of the label space from a fixed set of categories to sequence of [illegible]s significantly richer concepts.

[illegible] encouraging progress along the label [illegible]mplexity axes, these two directions have

[illegible]uted equally to this work.

Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

remained separate. In this work we take a step to [illegible]fying these two inter-connected tasks into one jo[illegible] work. First, we introduce the dense captioning [illegible] Figure 1), which requires a model to predict a set [illegible]tions across regions of an image. Object detectio[illegible] recovered as a special case when the target labels consist of one word, and image captioning is recovered when all images consist of one region that spans the full image.

Additionally, we develop a Fully Convolutional Localization Network (FCLN) for the dense captioning task. Our model is inspired by recent work in image captioning [49, 21, 32, 8, 4] in that it is composed of a Convolutional Neural Network and a Recurrent Neural Network language model. However, drawing on work in object detection [38], our second core contribution is to introduce a new dense localization layer. This layer is fully differentiable and can be inserted into any neural network that processes images to enable region-level training and predictions. Internally, the localization layer predicts a set of regions of interest in the image and then uses bilinear interpolation [19, 16] to smoothly crop the activations in each region.

We evaluate the model on the large-scale Visual Genome dataset, which contains 94,000 images and [illegible] captions. Our results show both performan[illegible] provements over approaches based on prev[illegible] art. We make our code and data publicly [illegible] port further progress on the dense captioni[illegible]

1

Kate Saenko

Step 1: expand plan

task

What can it do?

- State the task your model solves
- Usually only need one sentence
- Can mention applications

Why do we need this?

- Avoid reader frustration!
- (reads to page 7): oh, they're doing detection, not classification!? $#@#$%!!

Kate Saenko

DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Justin Johnson* Andrej Karpathy* Li Fei-Fei
Department of Computer Science, Stanford University
{jcjohns,karpathy,feifeili}@cs.stanford.edu

Abstract

We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The dense captioning task generalizes object detection when the descriptions consist of a single word, and Image Captioning when one predicted region covers the full image. To address the localization and description task jointly we propose a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization. The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences. We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.

1. Introduction

task

…essly point out and describe all aspects …n a strong semantic understanding of a …of its elements. However, despite nu… …plications, this ability remains a chal… …of the art visual recognition systems. In the last few years there has been significant progress in image classification [39, 26, 53, 45], where the task is to assign one label to an image. Further work has pushed these advances along two orthogonal directions: First, rapid progress in object detection [40, 14, 46] has identified models that efficiently identify and label multiple salient regions of an image. Second, recent advances in image captioning [3, 32, 21, 49, 51, 8, 4] have expanded the complexity of the label space from a fixed set of categories to sequence of words able to express significantly richer concepts.

However, despite encouraging progress along the label density and label complexity axes, these two directions have remained separate. In this work we take a step towards unifying these two inter-connected tasks into one joint framework. First, we introduce the dense captioning task (see Figure 1), which requires a model to predict a set of descriptions across regions of an image. Object detection is hence recovered as a special case when the target labels consist of one word, and image captioning is recovered when all images consist of one region that spans the full image.

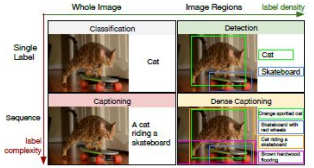


Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

Additionally, we develop a Fully Convolutional Localization Network (FCLN) for the dense captioning task. Our model is inspired by recent work in image captioning [49, 21, 32, 8, 4] in that it is composed of a Convolutional Neural Network and a Recurrent Neural Network language model. However, drawing on work in object detection [38], our second core contribution is to introduce a new dense localization layer. This layer is fully differentiable and can be inserted into any neural network that processes images to enable region-level training and predictions. Internally, the localization layer predicts a set of regions of interest in the image and then uses bilinear interpolation [19, 16] to smoothly crop the activations in each region.

We evaluate the model on the large-scale Visual Genome dataset, which contains 94,000 images and 4,100,000 region captions. Our results show both performance and speed improvements over approaches based on previous state of the art. We make our code and data publicly available to support further progress on the dense captioning task.

* Both authors contributed equally to this work.

1

Step 1: expand plan

State-of-the-art

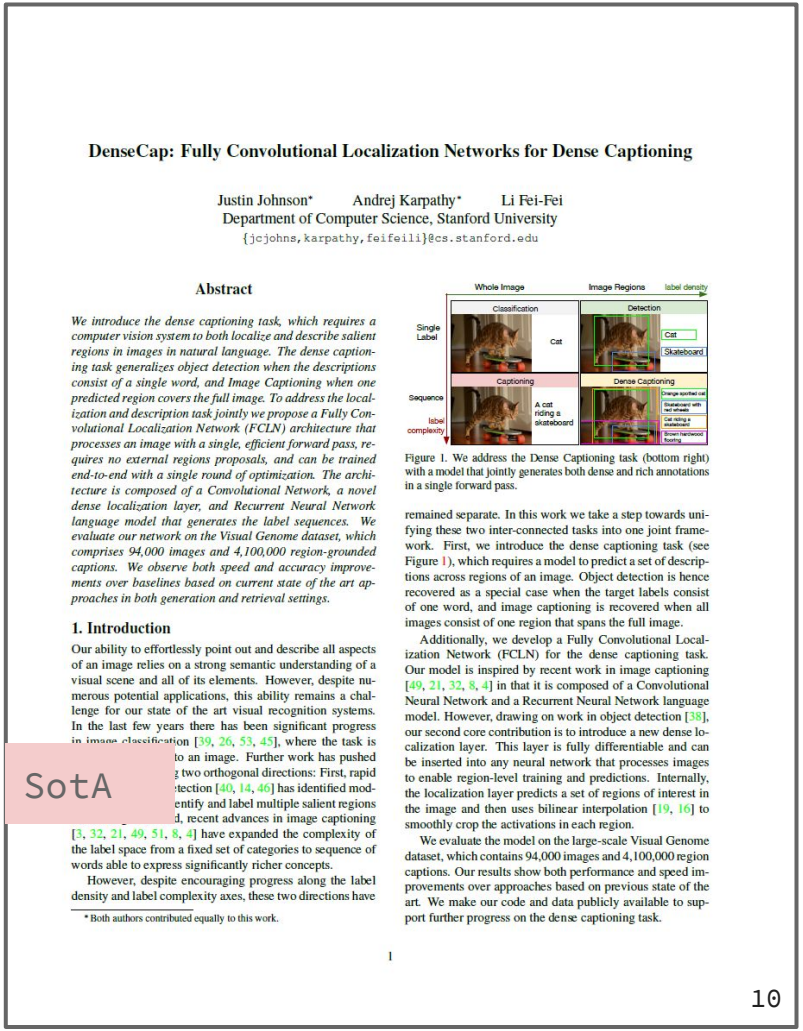
What is the best we can do now?

- Briefly cite recent work (direct competition)
- Get the reader excited about area
- “Significant progress”, “exciting results”

Common mistakes:

- Writing more than one paragraph (that is what related work is for!)
- Diminishing prior work. You’re standing on the shoulders of giants. Praise is better.

Kate Saenko



DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Justin Johnson* Andrej Karpathy* Li Fei-Fei
Department of Computer Science, Stanford University
{jcjohns,karpathy,feifeili}@cs.stanford.edu

Abstract

We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The dense captioning task generalizes object detection when the descriptions consist of a single word, and Image Captioning when one predicted region covers the full image. To address the localization and description task jointly we propose a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization. The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences. We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.

1. Introduction

Our ability to effortlessly point out and describe all aspects of an image relies on a strong semantic understanding of a visual scene and all of its elements. However, despite numerous potential applications, this ability remains a challenge for our state of the art visual recognition systems. In the last few years there has been significant progress in image classification [39, 26, 53, 45], where the task is to an image. Further work has pushed two orthogonal directions: First, rapid tection [40, 14, 46] has identified mod- entify and label multiple salient regions d, recent advances in image captioning [3, 32, 21, 49, 51, 8, 4] have expanded the complexity of the label space from a fixed set of categories to sequence of words able to express significantly richer concepts.

However, despite encouraging progress along the label density and label complexity axes, these two directions have remained separate. In this work we take a step towards unifying these two inter-connected tasks into one joint framework. First, we introduce the dense captioning task (see Figure 1), which requires a model to predict a set of descriptions across regions of an image. Object detection is hence recovered as a special case when the target labels consist of one word, and image captioning is recovered when all images consist of one region that spans the full image.

Additionally, we develop a Fully Convolutional Localization Network (FCLN) for the dense captioning task. Our model is inspired by recent work in image captioning [49, 21, 32, 8, 4] in that it is composed of a Convolutional Neural Network and a Recurrent Neural Network language model. However, drawing on work in object detection [38], our second core contribution is to introduce a new dense localization layer. This layer is fully differentiable and can be inserted into any neural network that processes images to enable region-level training and predictions. Internally, the localization layer predicts a set of regions of interest in the image and then uses bilinear interpolation [19, 16] to smoothly crop the activations in each region.

We evaluate the model on the large-scale Visual Genome dataset, which contains 94,000 images and 4,100,000 region captions. Our results show both performance and speed improvements over approaches based on previous state of the art. We make our code and data publicly available to support further progress on the dense captioning task.

Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

* Both authors contributed equally to this work.

1

SotA

Step 1: expand plan

flaw

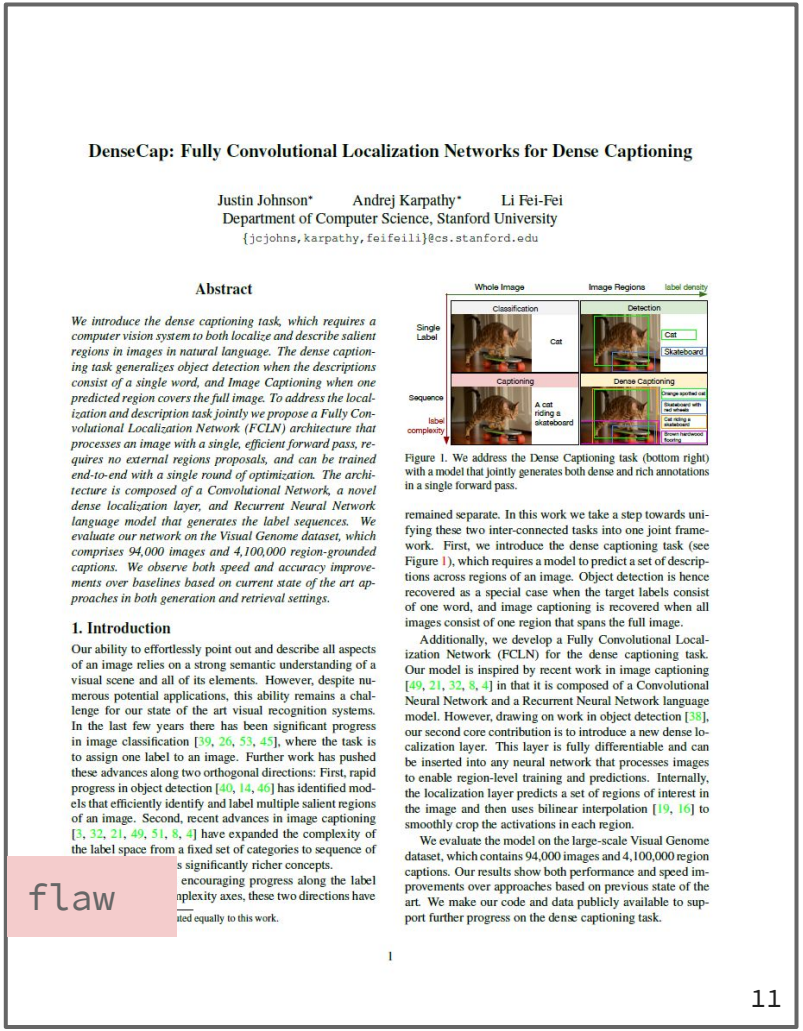
What can't we do yet?

- What is the critical flaw in SotA? (you'll be providing a solution to this later...)
- Why should the reader care? No, really?
- Why can't X, Y or Z be used to solve this?

Engage the reader

- Use an example: text, or visual (see Figure 1)
- At the end, the reader should be thinking, wow, this is really a important problem! I really wish someone would try to solve it!!

Kate Saenko



DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Justin Johnson* Andrej Karpathy* Li Fei-Fei
Department of Computer Science, Stanford University
{jcjohns,karpathy,feifeili}@cs.stanford.edu

Abstract

We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The dense captioning task generalizes object detection when the descriptions consist of a single word, and Image Captioning when one predicted region covers the full image. To address the localization and description task jointly we propose a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization. The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences. We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.

1. Introduction

Our ability to effortlessly point out and describe all aspects of an image relies on a strong semantic understanding of a visual scene and all of its elements. However, despite numerous potential applications, this ability remains a challenge for our state of the art visual recognition systems. In the last few years there has been significant progress in image classification [39, 26, 53, 45], where the task is to assign one label to an image. Further work has pushed these advances along two orthogonal directions: First, rapid progress in object detection [40, 14, 46] has identified models that efficiently identify and label multiple salient regions of an image. Second, recent advances in image captioning [3, 32, 21, 49, 51, 8, 4] have expanded the complexity of the label space from a fixed set of categories to sequence of s significantly richer concepts.

encouraging progress along the label nplexity axes, these two directions have

uted equally to this work.

Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

remained separate. In this work we take a step towards unifying these two inter-connected tasks into one joint framework. First, we introduce the dense captioning task (see Figure 1), which requires a model to predict a set of descriptions across regions of an image. Object detection is hence recovered as a special case when the target labels consist of one word, and image captioning is recovered when all images consist of one region that spans the full image.

Additionally, we develop a Fully Convolutional Localization Network (FCLN) for the dense captioning task. Our model is inspired by recent work in image captioning [49, 21, 32, 8, 4] in that it is composed of a Convolutional Neural Network and a Recurrent Neural Network language model. However, drawing on work in object detection [38], our second core contribution is to introduce a new dense localization layer. This layer is fully differentiable and can be inserted into any neural network that processes images to enable region-level training and predictions. Internally, the localization layer predicts a set of regions of interest in the image and then uses bilinear interpolation [19, 16] to smoothly crop the activations in each region.

We evaluate the model on the large-scale Visual Genome dataset, which contains 94,000 images and 4,100,000 region captions. Our results show both performance and speed improvements over approaches based on previous state of the art. We make our code and data publicly available to support further progress on the dense captioning task.

1

flaw

Step 1: expand plan

idea

What you propose to fix the flaw

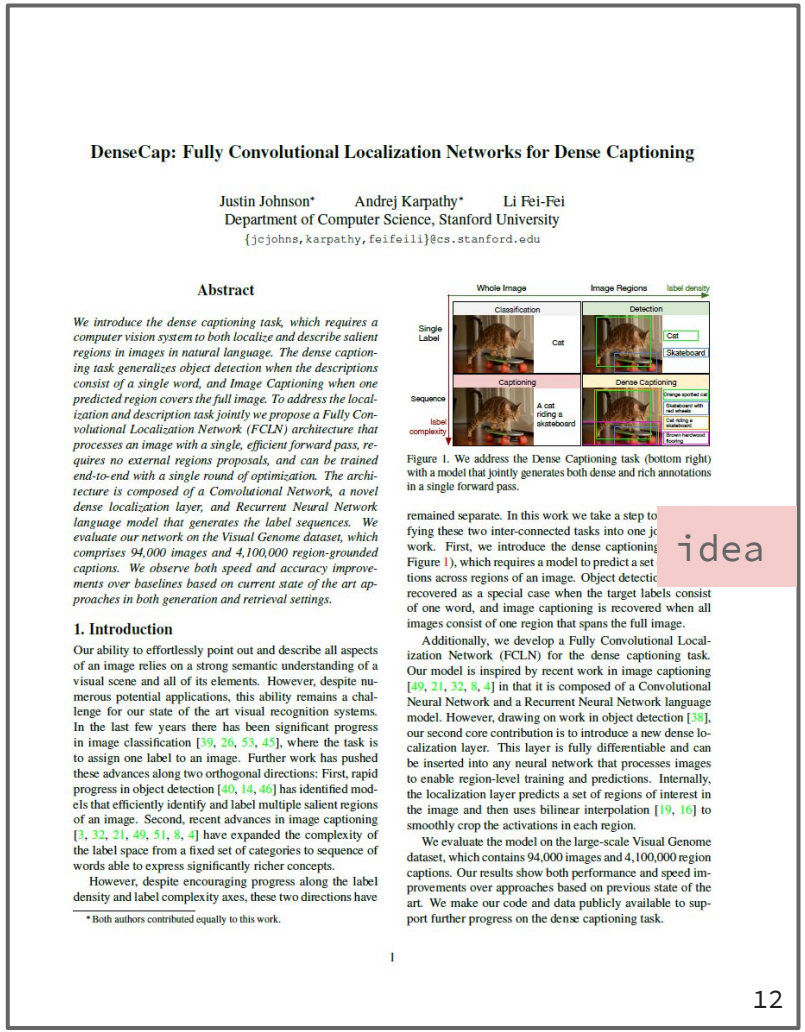
- The “hero”
- Comes to the rescue

Keep it brief

- Give the main intuition
- “We solve this by adding a new DNN layer”
- Leave the gory details to the method section!

Tip: if you have 2 ideas, describe each ‘flaw’ before you introduce each idea

Kate Saenko



DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Justin Johnson* Andrej Karpathy* Li Fei-Fei
Department of Computer Science, Stanford University
{jcjohns,karpathy,feifeili}@cs.stanford.edu

Abstract

We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The dense captioning task generalizes object detection when the descriptions consist of a single word, and Image Captioning when one predicted region covers the full image. To address the localization and description task jointly we propose a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization. The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences. We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.

1. Introduction

Our ability to effortlessly point out and describe all aspects of an image relies on a strong semantic understanding of a visual scene and all of its elements. However, despite numerous potential applications, this ability remains a challenge for our state of the art visual recognition systems. In the last few years there has been significant progress in image classification [39, 26, 53, 45], where the task is to assign one label to an image. Further work has pushed these advances along two orthogonal directions: First, rapid progress in object detection [40, 14, 46] has identified models that efficiently identify and label multiple salient regions of an image. Second, recent advances in image captioning [3, 32, 21, 49, 51, 8, 4] have expanded the complexity of the label space from a fixed set of categories to sequence of words able to express significantly richer concepts.

However, despite encouraging progress along the label density and label complexity axes, these two directions have remained separate. In this work we take a step to fying these two inter-connected tasks into one jo work. First, we introduce the dense captioning Figure 1), which requires a model to predict a set tions across regions of an image. Object detectio recovered as a special case when the target labels consist of one word, and image captioning is recovered when all images consist of one region that spans the full image.

Additionally, we develop a Fully Convolutional Localization Network (FCLN) for the dense captioning task. Our model is inspired by recent work in image captioning [49, 21, 32, 8, 4] in that it is composed of a Convolutional Neural Network and a Recurrent Neural Network language model. However, drawing on work in object detection [38], our second core contribution is to introduce a new dense localization layer. This layer is fully differentiable and can be inserted into any neural network that processes images to enable region-level training and predictions. Internally, the localization layer predicts a set of regions of interest in the image and then uses bilinear interpolation [19, 16] to smoothly crop the activations in each region.

We evaluate the model on the large-scale Visual Genome dataset, which contains 94,000 images and 4,100,000 region captions. Our results show both performance and speed improvements over approaches based on previous state of the art. We make our code and data publicly available to support further progress on the dense captioning task.

* Both authors contributed equally to this work.

1

Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

idea

Step 1: expand plan

figure 1

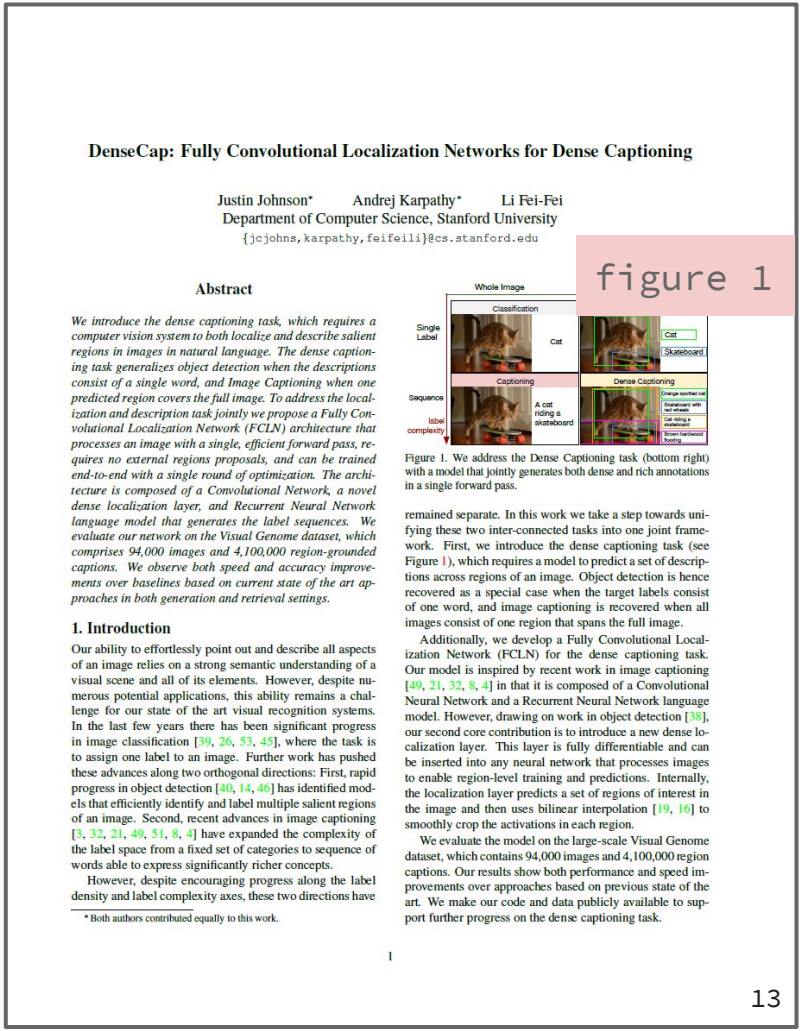
"Must have" for a vision paper

- Convey the flaw and idea visually
- Like an "ad" for your paper

Many people ONLY look at this

- If it looks interesting, they will also check the main result table
- Caption should be self-explanatory, since they may not read the main text

Kate Saenko



DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Justin Johnson* Andrej Karpathy* Li Fei-Fei
Department of Computer Science, Stanford University
{jcjohns,karpathy,feifeili}@cs.stanford.edu

figure 1

Abstract

We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The dense captioning task generalizes object detection when the descriptions consist of a single word, and Image Captioning when one predicted region covers the full image. To address the localization and description task jointly we propose a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization. The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences. We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.

1. Introduction

Our ability to effortlessly point out and describe all aspects of an image relies on a strong semantic understanding of a visual scene and all of its elements. However, despite numerous potential applications, this ability remains a challenge for our state of the art visual recognition systems. In the last few years there has been significant progress in image classification [39, 26, 53, 45], where the task is to assign one label to an image. Further work has pushed these advances along two orthogonal directions: First, rapid progress in object detection [40, 14, 46] has identified models that efficiently identify and label multiple salient regions of an image. Second, recent advances in image captioning [3, 32, 21, 49, 51, 8, 4] have expanded the complexity of the label space from a fixed set of categories to sequence of words able to express significantly richer concepts.

However, despite encouraging progress along the label density and label complexity axes, these two directions have remained separate. In this work we take a step towards unifying these two inter-connected tasks into one joint framework. First, we introduce the dense captioning task (see Figure 1), which requires a model to predict a set of descriptions across regions of an image. Object detection is hence recovered as a special case when the target labels consist of one word, and image captioning is recovered when all images consist of one region that spans the full image.

Additionally, we develop a Fully Convolutional Localization Network (FCLN) for the dense captioning task. Our model is inspired by recent work in image captioning [49, 21, 32, 8, 4] in that it is composed of a Convolutional Neural Network and a Recurrent Neural Network language model. However, drawing on work in object detection [38], our second core contribution is to introduce a new dense localization layer. This layer is fully differentiable and can be inserted into any neural network that processes images to enable region-level training and predictions. Internally, the localization layer predicts a set of regions of interest in the image and then uses bilinear interpolation [19, 16] to smoothly crop the activations in each region.

We evaluate the model on the large-scale Visual Genome dataset, which contains 94,000 images and 4,100,000 region captions. Our results show both performance and speed improvements over approaches based on previous state of the art. We make our code and data publicly available to support further progress on the dense captioning task.

Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

* Both authors contributed equally to this work.

1

Step 1: expand plan

figure 1

OK figure :| illustrates the task well

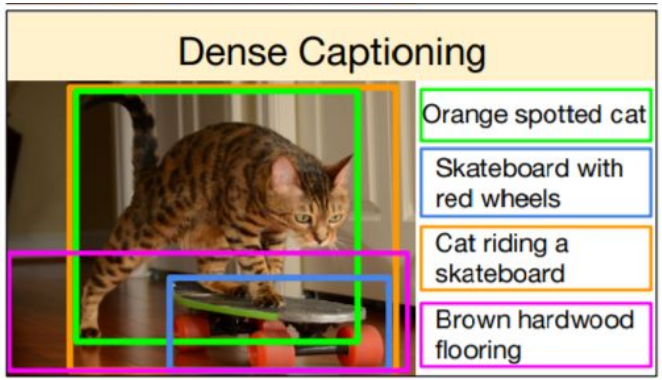


Figure 1. We address the Dense Captioning task with a model that jointly generates both dense and rich annotations in a single forward pass.

Better :) shows flaw in prior work

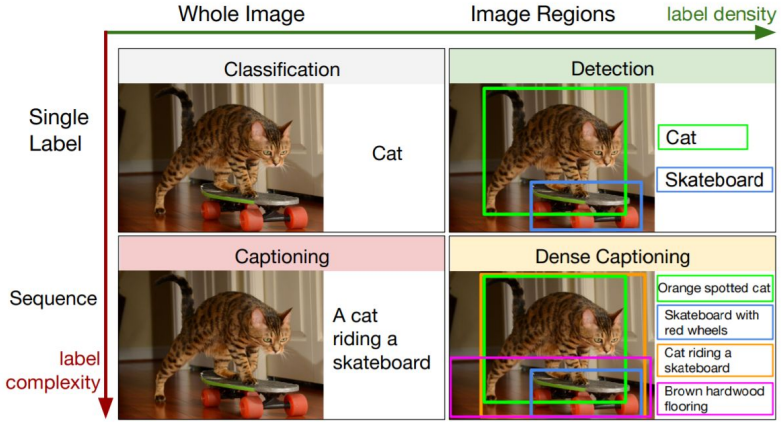


Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

Step 1: expand plan

Proof

Does it work?

- Say a bit about evaluation and baselines
- Don't give all away, but tease
- E.g. "our method lowers reconstruction error by 20% compared to previous methods"

DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Justin Johnson* Andrej Karpathy* Li Fei-Fei
Department of Computer Science, Stanford University
{jcjohns,karpathy,feifeili}@cs.stanford.edu

Abstract

We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The dense captioning task generalizes object detection when the descriptions consist of a single word, and Image Captioning when one predicted region covers the full image. To address the localization and description task jointly we propose a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization. The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences. We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.

1. Introduction

Our ability to effortlessly point out and describe all aspects of an image relies on a strong semantic understanding of a visual scene and all of its elements. However, despite numerous potential applications, this ability remains a challenge for our state of the art visual recognition systems. In the last few years there has been significant progress in image classification [39, 26, 53, 45], where the task is to assign one label to an image. Further work has pushed these advances along two orthogonal directions: First, rapid progress in object detection [40, 14, 46] has identified models that efficiently identify and label multiple salient regions of an image. Second, recent advances in image captioning [3, 32, 21, 49, 51, 8, 4] have expanded the complexity of the label space from a fixed set of categories to sequence of words able to express significantly richer concepts.

However, despite encouraging progress along the label density and label complexity axes, these two directions have

* Both authors contributed equally to this work.

Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

remained separate. In this work we take a step towards unifying these two inter-connected tasks into one joint framework. First, we introduce the dense captioning task (see Figure 1), which requires a model to predict a set of descriptions across regions of an image. Object detection is hence recovered as a special case when the target labels consist of one word, and image captioning is recovered when all images consist of one region that spans the full image.

Additionally, we develop a Fully Convolutional Localization Network (FCLN) for the dense captioning task. Our model is inspired by recent work in image captioning [49, 21, 32, 8, 4] in that it is composed of a Convolutional Neural Network and a Recurrent Neural Network language model. However, drawing on work in object detection [38], our second core contribution is to introduce a new dense localization layer. This layer is fully differentiable and can be inserted into any neural network that processes images to enable region-level training and predictions. Internally, the localization layer predicts a set of regions of interest in the image and then uses bilinear interpolation [19, 16] to smoothly crop the activations in each region.

We evaluate the model on the large-scale Visual Genome dataset, which contains 94,000 images and 4,100,000 region captions. Our results show both performance improvements over approaches based on previous art. We make our code and data publicly available to support further progress on the dense captioning

proof

Kate Saenko

Step 1: expand plan

title

Specific:

- DenseCap: Fully Convolutional Localization Networks for Dense Captioning :)
- A Network for Describing Images :(
- A Deep Learning Approach Based on Convolutional Neural Network… :(

Memorable:

- If using acronym/abbreviation make sure it is easy to spell and pronounce, e.g. FCLN :(vs DenseCap :)
- Best if conveys central idea, "dense captioning"
- "It's a bird! It's a plane! …" and other catchphrases: can be corny, proceed with caution !!!

Kate Saenko

title

DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Justin Johnson* Andrej Karpathy* Li Fei-Fei
Department of Computer Science, Stanford University
{jcjohns,karpathy,feifeili}@cs.stanford.edu

Abstract

We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The dense captioning task generalizes object detection when the descriptions consist of a single word, and Image Captioning when one predicted region covers the full image. To address the localization and description task jointly we propose a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization. The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences. We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.

1. Introduction

Our ability to effortlessly point out and describe all aspects of an image relies on a strong semantic understanding of a visual scene and all of its elements. However, despite numerous potential applications, this ability remains a challenge for our state of the art visual recognition systems. In the last few years there has been significant progress in image classification [39, 26, 53, 45], where the task is to assign one label to an image. Further work has pushed these advances along two orthogonal directions: First, rapid progress in object detection [40, 14, 46] has identified models that efficiently identify and label multiple salient regions of an image. Second, recent advances in image captioning [3, 32, 21, 49, 51, 8, 4] have expanded the complexity of the label space from a fixed set of categories to sequence of words able to express significantly richer concepts.

However, despite encouraging progress along the label density and label complexity axes, these two directions have remained separate. In this work we take a step towards unifying these two inter-connected tasks into one joint framework. First, we introduce the dense captioning task (see Figure 1), which requires a model to predict a set of descriptions across regions of an image. Object detection is hence recovered as a special case when the target labels consist of one word, and image captioning is recovered when all images consist of one region that spans the full image.

Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

Additionally, we develop a Fully Convolutional Localization Network (FCLN) for the dense captioning task. Our model is inspired by recent work in image captioning [49, 21, 32, 8, 4] in that it is composed of a Convolutional Neural Network and a Recurrent Neural Network language model. However, drawing on work in object detection [38], our second core contribution is to introduce a new dense localization layer. This layer is fully differentiable and can be inserted into any neural network that processes images to enable region-level training and predictions. Internally, the localization layer predicts a set of regions of interest in the image and then uses bilinear interpolation [19, 16] to smoothly crop the activations in each region.

We evaluate the model on the large-scale Visual Genome dataset, which contains 94,000 images and 4,100,000 region captions. Our results show both performance and speed improvements over approaches based on previous state of the art. We make our code and data publicly available to support further progress on the dense captioning task.

* Both authors contributed equally to this work.

1

Step 1: expand plan

Abstract

Concise:

- Summary of task and contributions
- Write this last! Copy-paste, rephrase, done!

Audience:

- Meant for general audience
- “Grandmother” should get the gist

Kate Saenko

DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Justin Johnson* Andrej Karpathy* Li Fei-Fei
Department of Computer Science, Stanford University
{jcjohns,karpathy,feifeili}@cs.stanford.edu

Abstract

We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient [illegible]language. The dense captioning [illegible]tection when the descriptions [illegible] Image Captioning when one [illegible]l image. To address the localization [illegible]intly we propose a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization. The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences. We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.

abstract

Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

1. Introduction

Our ability to effortlessly point out and describe all aspects of an image relies on a strong semantic understanding of a visual scene and all of its elements. However, despite numerous potential applications, this ability remains a challenge for our state of the art visual recognition systems. In the last few years there has been significant progress in image classification [39, 26, 53, 45], where the task is to assign one label to an image. Further work has pushed these advances along two orthogonal directions: First, rapid progress in object detection [40, 14, 46] has identified models that efficiently identify and label multiple salient regions of an image. Second, recent advances in image captioning [3, 32, 21, 49, 51, 8, 4] have expanded the complexity of the label space from a fixed set of categories to sequence of words able to express significantly richer concepts.

However, despite encouraging progress along the label density and label complexity axes, these two directions have remained separate. In this work we take a step towards unifying these two inter-connected tasks into one joint framework. First, we introduce the dense captioning task (see Figure 1), which requires a model to predict a set of descriptions across regions of an image. Object detection is hence recovered as a special case when the target labels consist of one word, and image captioning is recovered when all images consist of one region that spans the full image.

Additionally, we develop a Fully Convolutional Localization Network (FCLN) for the dense captioning task. Our model is inspired by recent work in image captioning [49, 21, 32, 8, 4] in that it is composed of a Convolutional Neural Network and a Recurrent Neural Network language model. However, drawing on work in object detection [38], our second core contribution is to introduce a new dense localization layer. This layer is fully differentiable and can be inserted into any neural network that processes images to enable region-level training and predictions. Internally, the localization layer predicts a set of regions of interest in the image and then uses bilinear interpolation [19, 16] to smoothly crop the activations in each region.

We evaluate the model on the large-scale Visual Genome dataset, which contains 94,000 images and 4,100,000 region captions. Our results show both performance and speed improvements over approaches based on previous state of the art. We make our code and data publicly available to support further progress on the dense captioning task.

* Both authors contributed equally to this work.

1


GOOD MORNING, CECILIA!
DRAFT
I TOOK THE LIBERTY OF WRITING MYSELF WHILE YOU SLEPT!
WHAT?
DRAFT
YOUR DRAFT IS FINISHED!
THAT'S AMAZING!!!
YOU DON'T HAVE TO DO ANY MORE WRITING!
DRAFT
WOW!!
I'M TOTALLY DREAMING, AREN'T I?
THE UNICORN GAVE IT AWAY.
DRAFT
JORGE CHAM © 2016
WWW.PHDCOMICS.COM

Step 2: write!

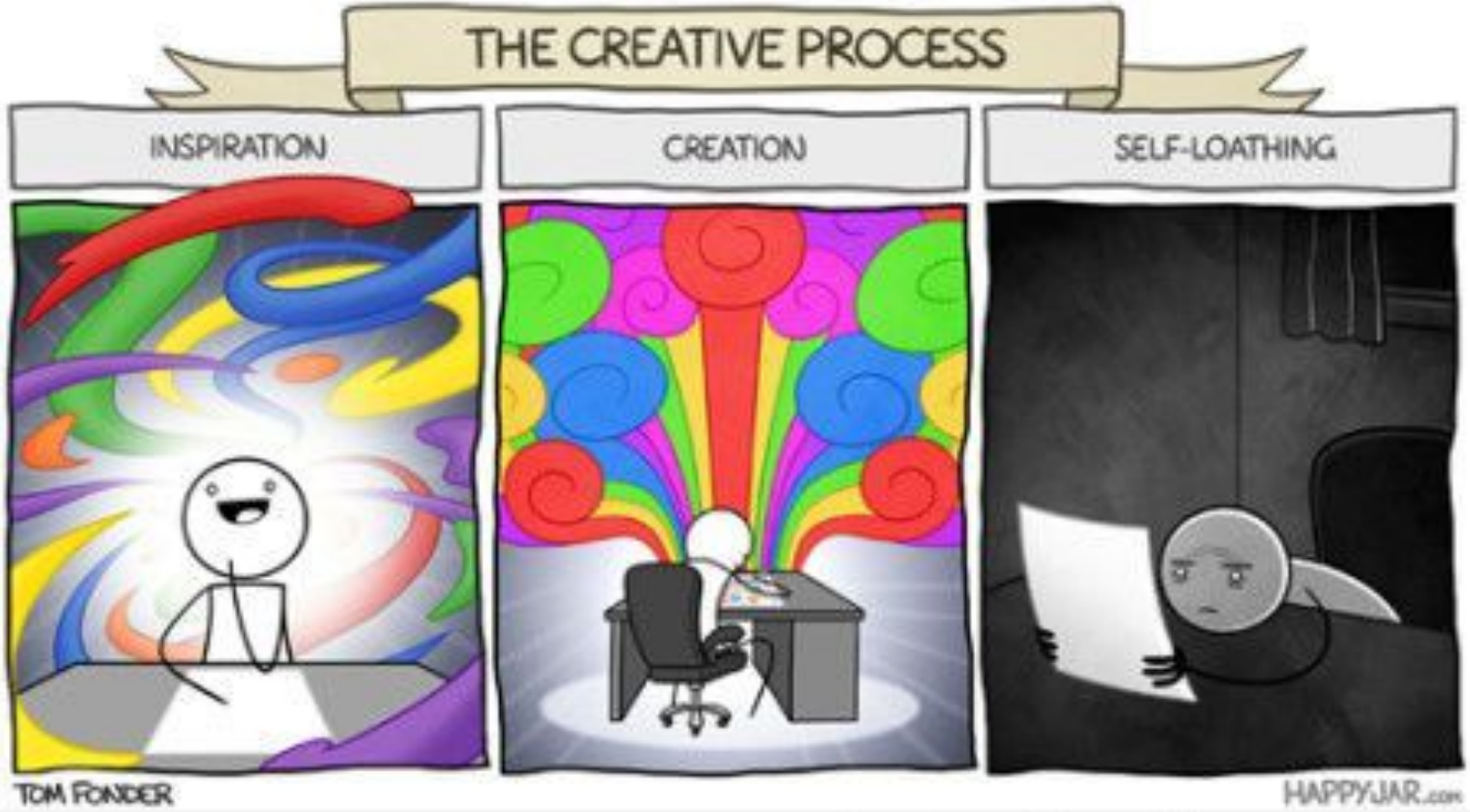
Step 2: write!

By now, you should have your introduction outlined in several dozen bullets, so the rest should be much easier

This would be a good time to run it by others

Expand each bullet into sentence/paragraph

Fill in details along the way



Step 3: edit edit edit

Step 3: edit edit edit

Some lucky people can write near-perfect text in one go

Most of us are not so lucky

I re-write my drafts many times, sometimes dozens

Important to get feedback from your advisor/labmates

That's it!

Now you just have the rest of the paper to write :)